Predicting Customer Churn with R - Logistic Regression & Random Forest

Lucas Daniel Zarzeczny

Monday, June 15, 2020

```
library(plyr)
library(corrplot)
library(ggplot2)
library(gridExtra)
library(ggthemes)
library(caret)
library(MASS)
library(randomForest)
library(party)
library(readxl)
library(plotly)
library(dplyr)
library(wesanderson)
library(viridis)
library(MASS)
library(rpart)
library(rpart.plot)
library(randomForest)
library(caret)
```

Step 1: Load the Data

```
# Working Directory
getwd()
```

```
## [1] "/Users/lucasdanielzarzeczny/Desktop"
```

```r
setwd("/Users/lucasdanielzarzeczny/Desktop")
getwd()
```

```
## [1] "/Users/lucasdanielzarzeczny/Desktop"
```

```r
# Download the dataset
churn_data <-
  read_excel('/Users/lucasdanielzarzeczny/Desktop/WA_Fn-UseC_-Telco-Customer-Churn.xlsx')
churn_data <- data.frame(churn_data )

# Structure of the Dataset
str(churn_data)
```

```
## 'data.frame':    7043 obs. of  21 variables:
##  $ customerID      : chr  "7590-VHVEG" "5575-GNVDE" "3668-QPYBK" "7795-CFOCW" ...
##  $ gender          : chr  "Female" "Male" "Male" "Male" ...
##  $ SeniorCitizen   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Partner         : chr  "Yes" "No" "No" "No" ...
##  $ Dependents      : chr  "No" "No" "No" "No" ...
##  $ tenure          : num  1 34 2 45 2 8 22 10 28 62 ...
##  $ PhoneService    : chr  "No" "Yes" "Yes" "No" ...
##  $ MultipleLines   : chr  "No phone service" "No" "No" "No phone service" ...
##  $ InternetService : chr  "DSL" "DSL" "DSL" "DSL" ...
##  $ OnlineSecurity  : chr  "No" "Yes" "Yes" "Yes" ...
##  $ OnlineBackup    : chr  "Yes" "No" "Yes" "No" ...
##  $ DeviceProtection: chr  "No" "Yes" "No" "Yes" ...
##  $ TechSupport     : chr  "No" "No" "No" "Yes" ...
##  $ StreamingTV     : chr  "No" "No" "No" "No" ...
##  $ StreamingMovies : chr  "No" "No" "No" "No" ...
##  $ Contract        : chr  "Month-to-month" "One year" "Month-to-month" "One year" ...
##  $ PaperlessBilling: chr  "Yes" "No" "Yes" "No" ...
##  $ PaymentMethod   : chr  "Electronic check" "Mailed check" "Mailed check" "Bank transfer (automatic)" ...
##  $ MonthlyCharges  : num  29.9 57 53.9 42.3 70.7 ...
##  $ TotalCharges    : num  29.9 1889.5 108.2 1840.8 151.7 ...
##  $ Churn           : chr  "No" "No" "Yes" "No" ...
```

Step 2: Data Wrangling

```r
# Churn is already in a binary format ("Yes", "No")
unique(churn_data$Churn)
```

```
## [1] "No"  "Yes"
```

```r
# Check the number of missing values in each column
sapply(churn_data, function(x) sum(is.na(x)))
```

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

```r
# There are 11 missing values in TotalCharges

# Remove all the rows with the missing values
churn_data <- churn_data[complete.cases(churn_data), ]
sapply(churn_data, function(x) sum(is.na(x)))
```

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

```r
churn_data <- churn_data[complete.cases(churn_data), ]
# Confirmed - no missing values in the data set now

# Data Wrangling
# I simply want rows to say yes or not for columns that require a binary response
unique(churn_data$OnlineSecurity) # change no internet service to "No"
```

```
## [1] "No"                  "Yes"                 "No internet service"
```

```r
unique(churn_data$OnlineBackup)  # change no internet service to "No"
```

```
## [1] "Yes"                 "No"                  "No internet service"
```

```r
unique(churn_data$DeviceProtection) # change no internet service to "No"
```

```
## [1] "No"                  "Yes"                 "No internet service"
```

```r
unique(churn_data$TechSupport) # change no internet service to "No"
```

```
## [1] "No"                  "Yes"                 "No internet service"
```

```r
unique(churn_data$StreamingTV) # change no internet service to "No"
```

```
## [1] "No"                  "Yes"                 "No internet service"
```

```r
unique(churn_data$StreamingMovies) # change no internet service to "No"
```

```
## [1] "No"                  "Yes"                 "No internet service"
```

```r
# Function to Change the "no internet service" to "No" (Mutate)
require(plyr)
require(dplyr)

churn_data <- mutate(churn_data, OnlineSecurity = replace(OnlineSecurity,
                                                          OnlineSecurity == "No internet service", "No"))
churn_data <- mutate(churn_data, OnlineBackup = replace(OnlineBackup,
                                                        OnlineBackup == "No internet service", "No"))
churn_data <- mutate(churn_data, DeviceProtection = replace(DeviceProtection, DeviceProtection == "No internet service", "No"))
churn_data <- mutate(churn_data, TechSupport = replace(TechSupport,
                                                       TechSupport == "No internet service", "No"))
churn_data <- mutate(churn_data, StreamingTV = replace(StreamingTV, StreamingTV == "No internet service", "No"))
churn_data <- mutate(churn_data, StreamingMovies = replace(StreamingMovies,
                                                           StreamingMovies == "No internet service", "No"))
```

```r
# Validate the results (Should all be "Yes", "No")
unique(churn_data$OnlineSecurity)
```

```
## [1] "No"  "Yes"
```

```r
unique(churn_data$OnlineBackup)
```

```
## [1] "Yes" "No"
```

```r
unique(churn_data$DeviceProtection)
```

```
## [1] "No"  "Yes"
```

```r
unique(churn_data$TechSupport)
```

```
## [1] "No"  "Yes"
```

```r
unique(churn_data$StreamingTV)
```

```
## [1] "No"  "Yes"
```

```r
unique(churn_data$StreamingMovies)
```

```
## [1] "No"  "Yes"
```

```r
# Double check the other columns
unique(churn_data$gender)
```

```
## [1] "Female" "Male"
```

```r
unique(churn_data$Partner)
```

```
## [1] "Yes" "No"
```

```r
unique(churn_data$Dependents)
```

```
## [1] "No"  "Yes"
```

```r
unique(churn_data$PhoneService)
```

```
## [1] "No"  "Yes"
```

```r
unique(churn_data$MultipleLines)
```

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

```r
unique(churn_data$Contract)
```

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

```r
unique(churn_data$PaperlessBilling)
```

```
## [1] "Yes" "No"
```

```r
unique(churn_data$PaymentMethod)
```

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

```r
unique(churn_data$Churn)
```

```
## [1] "No"  "Yes"
```

```r
unique(churn_data$SeniorCitizen)
```

```
## [1] 0 1
```

```r
# Multiple lines has "No phone service" as well
churn_data <- mutate(churn_data, MultipleLines = replace(MultipleLines,
                                                         MultipleLines == "No phone service", "No"))

# Double check Multiple Lines
unique(churn_data$MultipleLines)
```

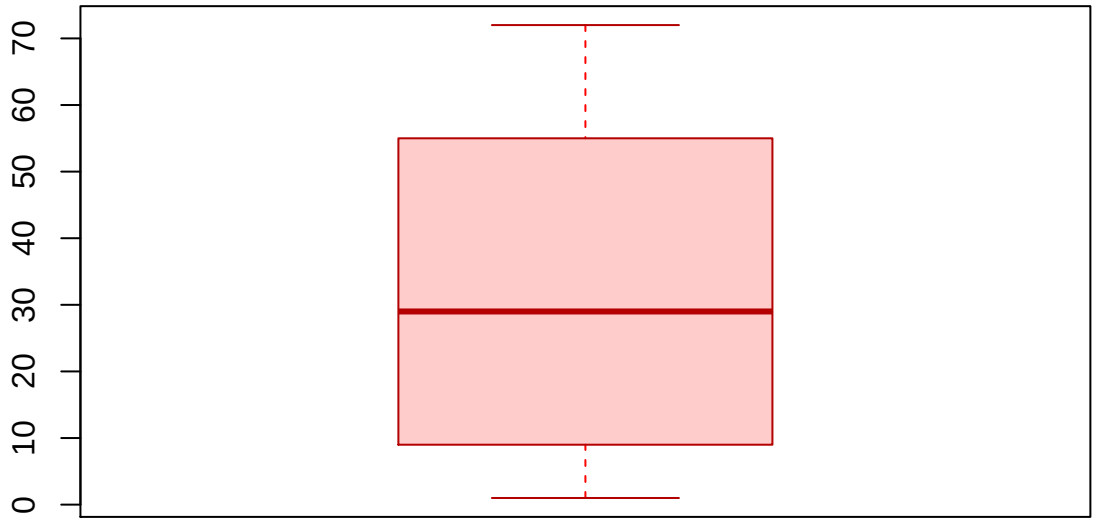
```
## [1] "No"  "Yes"
```

```r
# Change Senior Citizen column from 0 & 1 to Yes and No for consistency
churn_data$SeniorCitizen <- as.factor(mapvalues(churn_data$SeniorCitizen,
                                          from=c("0","1"),
                                          to=c("No", "Yes")))
# Double check Senior Citizen column
unique(churn_data$SeniorCitizen)
```

```
## [1] No  Yes
## Levels: No Yes
```

```r
# Check Tenure Column Distribution
c1 <- rainbow(10)
c2 <- rainbow(10, alpha=0.2)
c3 <- rainbow(10, v=0.7)
```

```r
boxplot(churn_data$tenure , col=c2, medcol=c3, whiskcol=c1,
        staplecol=c3, boxcol=c3, outcol=c3, pch=23, cex=2)
```




```r
# Tenure between 0 and 70+ months

# Group tenure into groups  "0-12 Month", "12-24 Month", "24-48 Months", "48-60 Month", "> 60 Month"
min(churn_data$tenure)
```

```
## [1] 1
```

```r
max(churn_data$tenure)
```

```
## [1] 72
```

```r
# Create a function to create the churn groups

group_tenure <- function(tenure){
  if (tenure >= 0 & tenure <= 12){
    return('0-12 Month')
  }else if(tenure > 12 & tenure <= 24){
    return('12-24 Month')
  }else if (tenure > 24 & tenure <= 48){
    return('24-48 Month')
```

```
  }else if (tenure > 48 & tenure <=60){
    return('48-60 Month')
  }else if (tenure > 60){
    return('> 60 Month')
  }
}
churn_data$tenure_group <- sapply(churn_data$tenure,group_tenure)
churn_data$tenure_group <- as.factor(churn_data$tenure_group)

# View churn_data$tenure_group
unique(churn_data$tenure_group)
```

```
## [1] 0-12 Month  24-48 Month 12-24 Month > 60 Month  48-60 Month
## Levels: > 60 Month 0-12 Month 12-24 Month 24-48 Month 48-60 Month
```

```
# Drop the tenure column then
churn_data$tenure <- NULL

# Check the structure of the data
str(churn_data)
```

```
## 'data.frame':    7032 obs. of  21 variables:
##  $ customerID      : chr  "7590-VHVEG" "5575-GNVDE" "3668-QPYBK" "7795-CFOCW" ...
##  $ gender          : chr  "Female" "Male" "Male" "Male" ...
##  $ SeniorCitizen   : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Partner         : chr  "Yes" "No" "No" "No" ...
##  $ Dependents      : chr  "No" "No" "No" "No" ...
##  $ PhoneService    : chr  "No" "Yes" "Yes" "No" ...
##  $ MultipleLines   : chr  "No" "No" "No" "No" ...
##  $ InternetService : chr  "DSL" "DSL" "DSL" "DSL" ...
##  $ OnlineSecurity  : chr  "No" "Yes" "Yes" "Yes" ...
##  $ OnlineBackup    : chr  "Yes" "No" "Yes" "No" ...
##  $ DeviceProtection: chr  "No" "Yes" "No" "Yes" ...
##  $ TechSupport     : chr  "No" "No" "No" "Yes" ...
##  $ StreamingTV     : chr  "No" "No" "No" "No" ...
##  $ StreamingMovies : chr  "No" "No" "No" "No" ...
##  $ Contract        : chr  "Month-to-month" "One year" "Month-to-month" "One year" ...
##  $ PaperlessBilling: chr  "Yes" "No" "Yes" "No" ...
##  $ PaymentMethod   : chr  "Electronic check" "Mailed check" "Mailed check" "Bank transfer (automatic)" ...
```

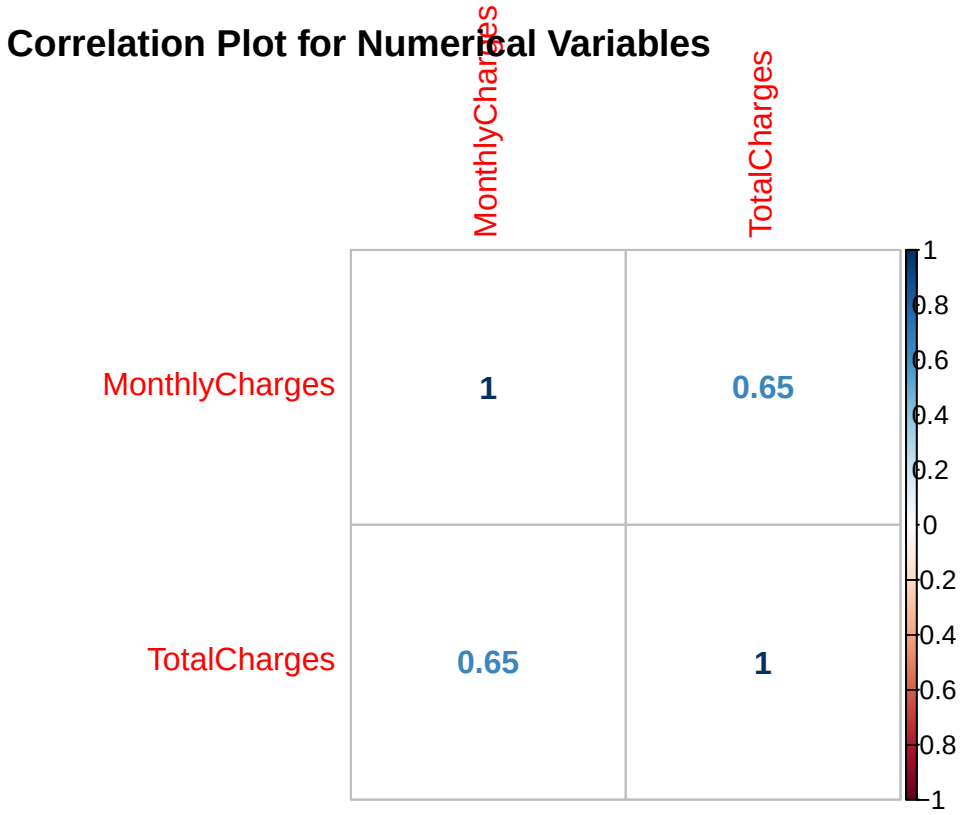
```
##  $ MonthlyCharges  : num  29.9 57 53.9 42.3 70.7 ...
##  $ TotalCharges    : num  29.9 1889.5 108.2 1840.8 151.7 ...
##  $ Churn           : chr  "No" "No" "Yes" "No" ...
##  $ tenure_group    : Factor w/ 5 levels "> 60 Month","0-12 Month",..: 2 4 2 4 2 2 3 2 4 1 ...
```

```
# Customer ID does not add value to the data either
churn_data$customerID <- NULL

# Check the structure of the data
str(churn_data)
```

```
## 'data.frame':    7032 obs. of  20 variables:
##  $ gender          : chr  "Female" "Male" "Male" "Male" ...
##  $ SeniorCitizen   : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Partner         : chr  "Yes" "No" "No" "No" ...
##  $ Dependents      : chr  "No" "No" "No" "No" ...
##  $ PhoneService    : chr  "No" "Yes" "Yes" "No" ...
##  $ MultipleLines   : chr  "No" "No" "No" "No" ...
##  $ InternetService : chr  "DSL" "DSL" "DSL" "DSL" ...
##  $ OnlineSecurity  : chr  "No" "Yes" "Yes" "Yes" ...
##  $ OnlineBackup    : chr  "Yes" "No" "Yes" "No" ...
##  $ DeviceProtection: chr  "No" "Yes" "No" "Yes" ...
##  $ TechSupport     : chr  "No" "No" "No" "Yes" ...
##  $ StreamingTV     : chr  "No" "No" "No" "No" ...
##  $ StreamingMovies : chr  "No" "No" "No" "No" ...
##  $ Contract        : chr  "Month-to-month" "One year" "Month-to-month" "One year" ...
##  $ PaperlessBilling: chr  "Yes" "No" "Yes" "No" ...
##  $ PaymentMethod   : chr  "Electronic check" "Mailed check" "Mailed check" "Bank transfer (automatic)" ...
##  $ MonthlyCharges  : num  29.9 57 53.9 42.3 70.7 ...
##  $ TotalCharges    : num  29.9 1889.5 108.2 1840.8 151.7 ...
##  $ Churn           : chr  "No" "No" "Yes" "No" ...
##  $ tenure_group    : Factor w/ 5 levels "> 60 Month","0-12 Month",..: 2 4 2 4 2 2 3 2 4 1 ...
```

Step 3: Independent Variable Correlation

```
# Which independent variables contribute the most to the dependent variable
numeric.var <- sapply(churn_data, is.numeric)
corr.matrix <- cor(churn_data[,numeric.var])
corrplot(corr.matrix, main="\n\nCorrelation Plot for Numerical Variables", method="number")
```




```
# Based on this matrix, MonthlyCharges and TotalCharges are correlated thus removing one of the columns.
# Will remove TotalCharges since most analysis is done on a monthly basis
```

Step 4: Data Exploration

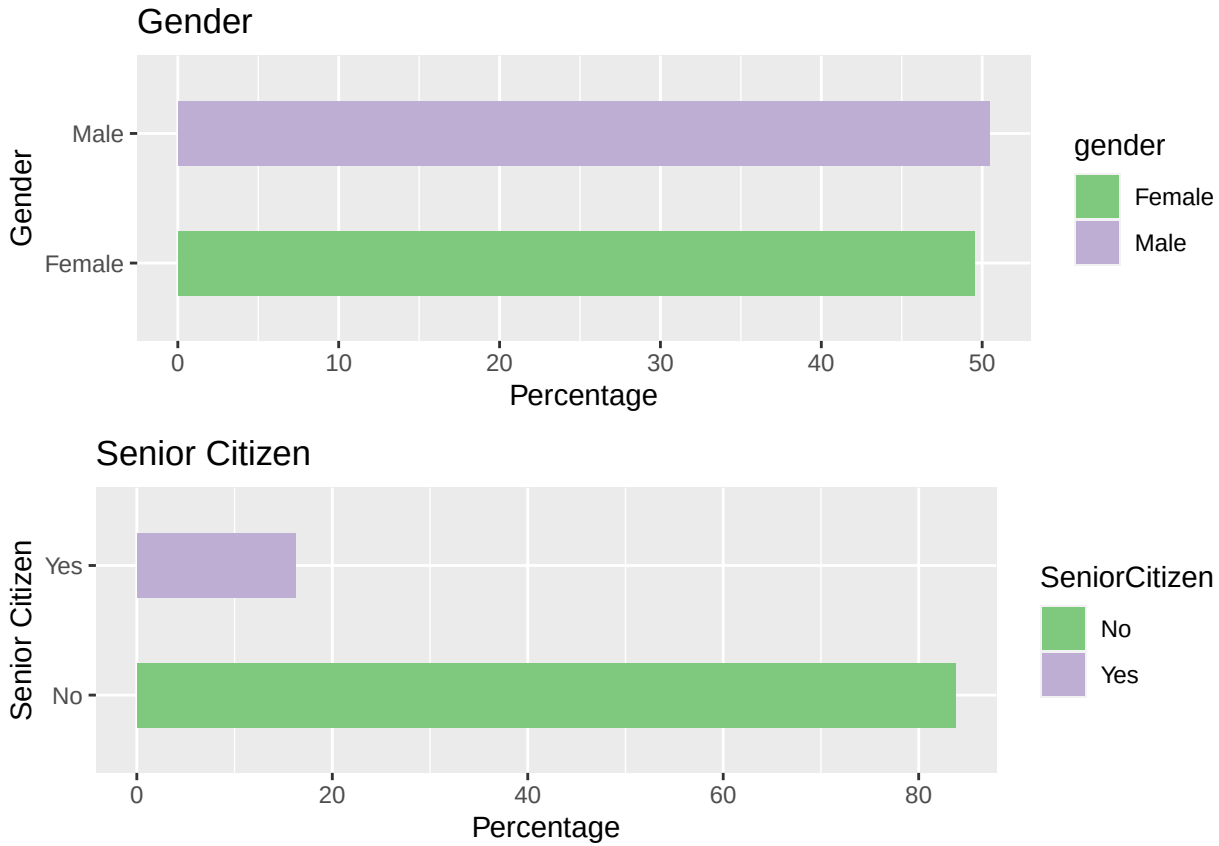
```
churn_data$TotalCharges <- NULL

# Check the structure of the data
str(churn_data)
```

```
## 'data.frame':    7032 obs. of  19 variables:
##  $ gender          : chr  "Female" "Male" "Male" "Male" ...
##  $ SeniorCitizen   : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
```

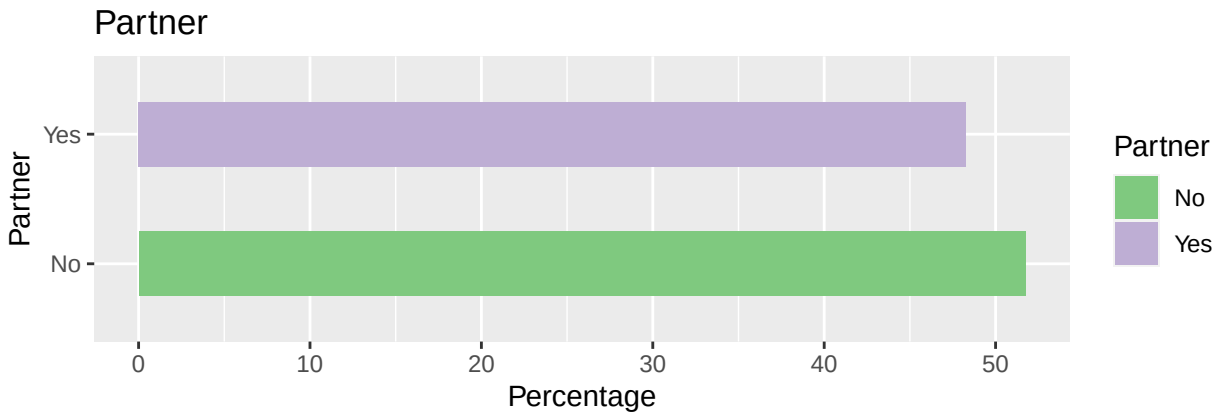
```
##  $ Partner         : chr  "Yes" "No" "No" "No" ...
##  $ Dependents      : chr  "No" "No" "No" "No" ...
##  $ PhoneService    : chr  "No" "Yes" "Yes" "No" ...
##  $ MultipleLines   : chr  "No" "No" "No" "No" ...
##  $ InternetService : chr  "DSL" "DSL" "DSL" "DSL" ...
##  $ OnlineSecurity  : chr  "No" "Yes" "Yes" "Yes" ...
##  $ OnlineBackup    : chr  "Yes" "No" "Yes" "No" ...
##  $ DeviceProtection: chr  "No" "Yes" "No" "Yes" ...
##  $ TechSupport     : chr  "No" "No" "No" "Yes" ...
##  $ StreamingTV     : chr  "No" "No" "No" "No" ...
##  $ StreamingMovies : chr  "No" "No" "No" "No" ...
##  $ Contract        : chr  "Month-to-month" "One year" "Month-to-month" "One year" ...
##  $ PaperlessBilling: chr  "Yes" "No" "Yes" "No" ...
##  $ PaymentMethod   : chr  "Electronic check" "Mailed check" "Mailed check" "Bank transfer (automatic)" ...
##  $ MonthlyCharges  : num  29.9 57 53.9 42.3 70.7 ...
##  $ Churn           : chr  "No" "No" "Yes" "No" ...
##  $ tenure_group    : Factor w/ 5 levels "> 60 Month","0-12 Month",..: 2 4 2 4 2 2 3 2 4 1 ...
```

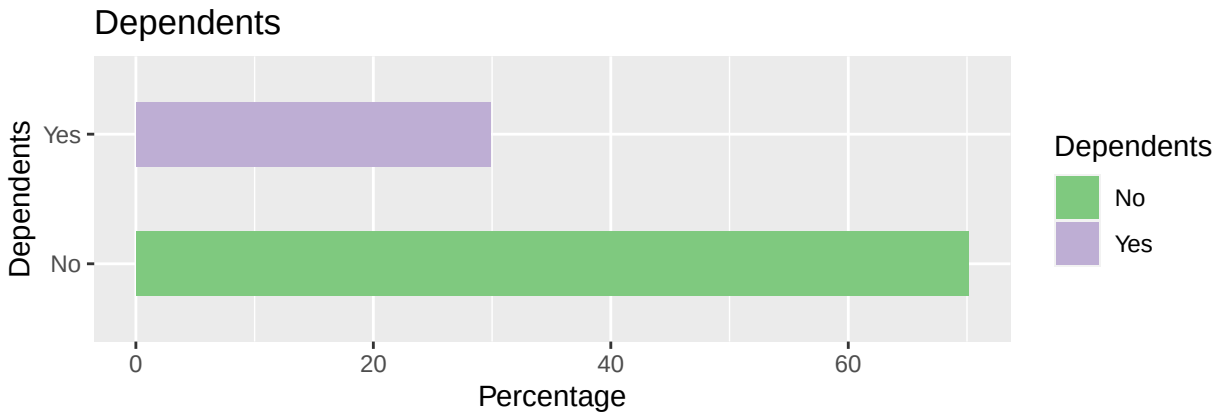
```r
# Bar plots for all categorical variables

bar_gender <- ggplot(churn_data, aes(x=gender, fill=gender)) +
  ggtitle("Gender") + xlab("Gender") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) +
  ylab("Percentage") + coord_flip() + scale_fill_brewer(palette="Accent")

bar_senior <- ggplot(churn_data, aes(x=SeniorCitizen, fill=SeniorCitizen)) +
  ggtitle("Senior Citizen") +
  xlab("Senior Citizen") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) +
  ylab("Percentage") + coord_flip() + scale_fill_brewer(palette="Accent")

bar_partner <- ggplot(churn_data, aes(x=Partner, fill=Partner)) +
  ggtitle("Partner") + xlab("Partner") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) +
  ylab("Percentage") + coord_flip() + scale_fill_brewer(palette="Accent")

bar_dependents <- ggplot(churn_data, aes(x=Dependents, fill=Dependents)) +
  ggtitle("Dependents") + xlab("Dependents") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) +
```

```
  ylab("Percentage") + coord_flip() + scale_fill_brewer(palette="Accent")

grid.arrange(bar_gender, bar_senior, ncol=1)
```




```
grid.arrange(bar_partner, bar_dependents, ncol=1)
```

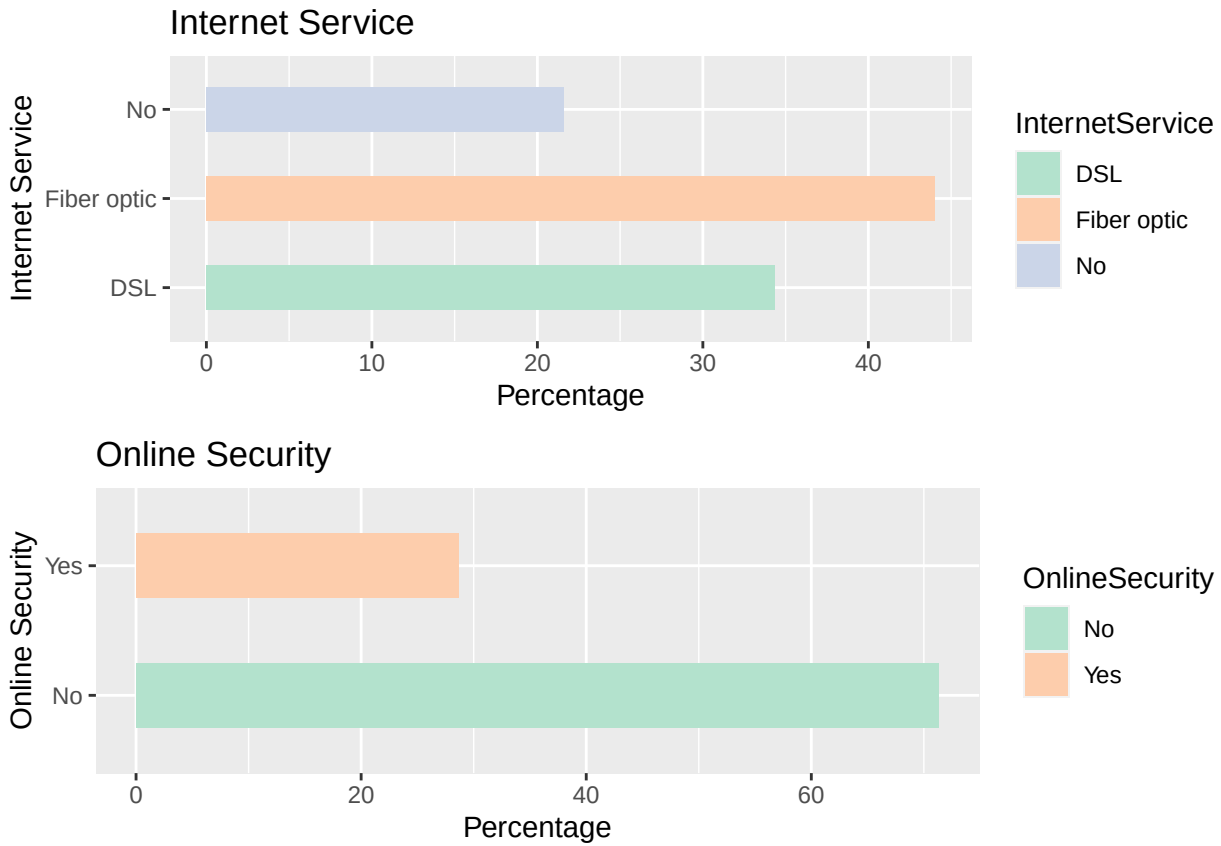





```
bar_phoneservice <- ggplot(churn_data, aes(x=PhoneService, fill=PhoneService)) +
  ggtitle("Phone Service") + xlab("Phone Service") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) +
  ylab("Percentage") + coord_flip() + scale_fill_brewer(palette="Pastel2")

bar_multiplelines <- ggplot(churn_data, aes(x=MultipleLines, fill=MultipleLines)) +
  ggtitle("Multiple Lines") + xlab("Multiple Lines") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) +
  ylab("Percentage") + coord_flip() + scale_fill_brewer(palette="Pastel2")

bar_internetservice <- ggplot(churn_data, aes(x=InternetService, fill=InternetService)) +
  ggtitle("Internet Service") + xlab("Internet Service") +
```

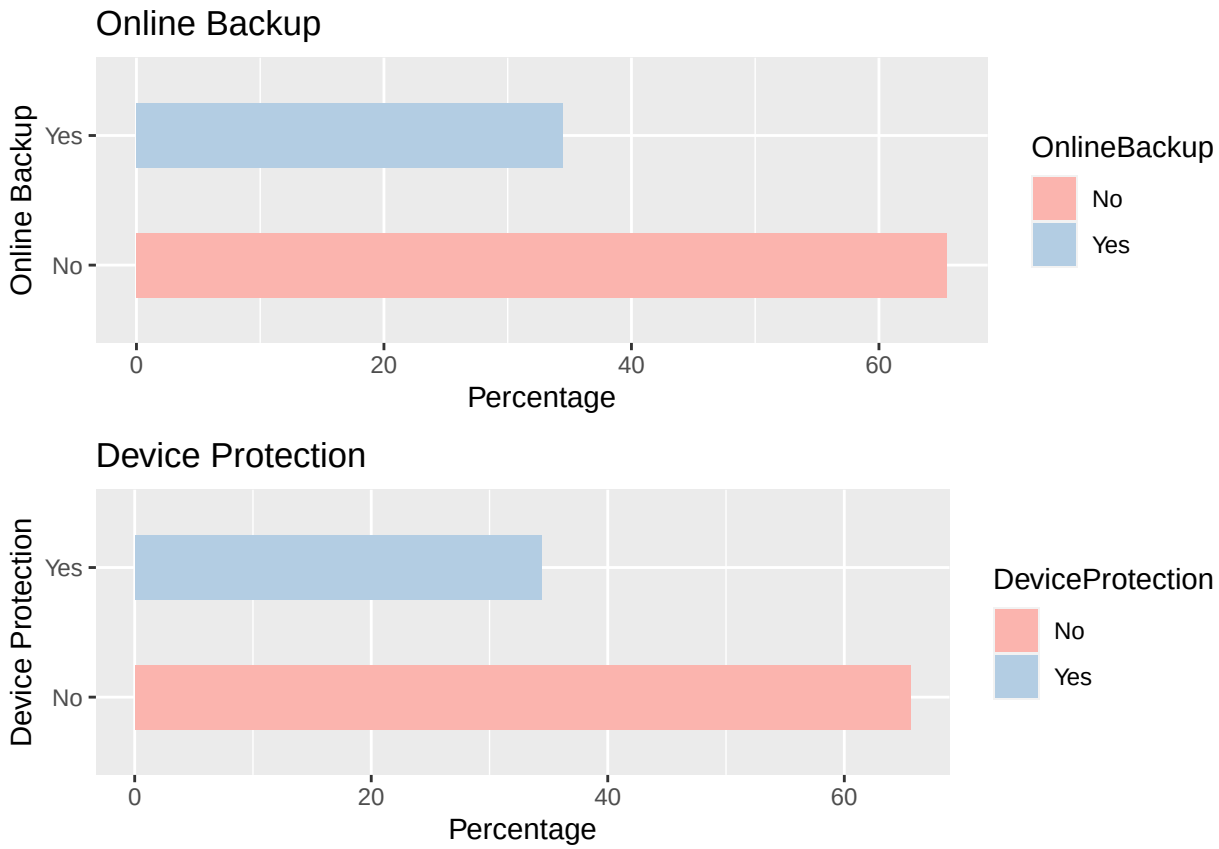
```
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) +
  ylab("Percentage") + coord_flip() + scale_fill_brewer(palette="Pastel2")

bar_onlinesecurity <- ggplot(churn_data, aes(x=OnlineSecurity, fill=OnlineSecurity)) +
  ggtitle("Online Security") + xlab("Online Security") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) +
  ylab("Percentage") + coord_flip() + scale_fill_brewer(palette="Pastel2")

grid.arrange(bar_phoneservice, bar_multiplelines, ncol=1)
```

```
grid.arrange(bar_internetservice, bar_onlinesecurity, ncol=1)
```

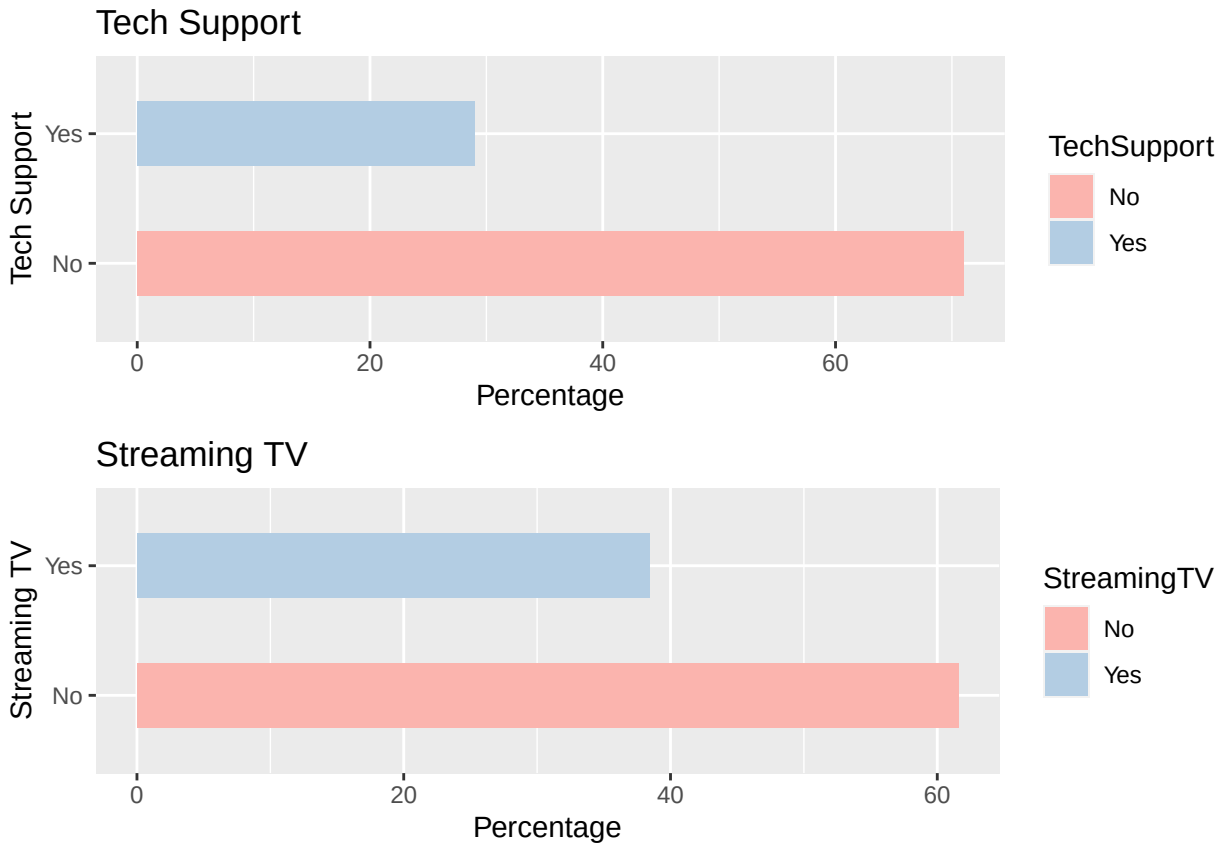



```
bar_onlinebackup <- ggplot(churn_data, aes(x=OnlineBackup, fill=OnlineBackup)) +
  ggtitle("Online Backup") + xlab("Online Backup") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) +
  ylab("Percentage") + coord_flip() + scale_fill_brewer(palette="Pastel1")

bar_deviceprotection <- ggplot(churn_data, aes(x=DeviceProtection, fill=DeviceProtection)) +
  ggtitle("Device Protection") + xlab("Device Protection") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") +
  coord_flip() + scale_fill_brewer(palette="Pastel1")
```

```
bar_techsupport <- ggplot(churn_data, aes(x=TechSupport, fill=TechSupport)) +
  ggtitle("Tech Support") + xlab("Tech Support") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") +
  coord_flip() + scale_fill_brewer(palette="Pastel1")

bar_steamingtv <- ggplot(churn_data, aes(x=StreamingTV, fill=StreamingTV)) +
  ggtitle("Streaming TV") + xlab("Streaming TV") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") +
  coord_flip() + scale_fill_brewer(palette="Pastel1")

grid.arrange(bar_onlinebackup, bar_deviceprotection, ncol=1)
```

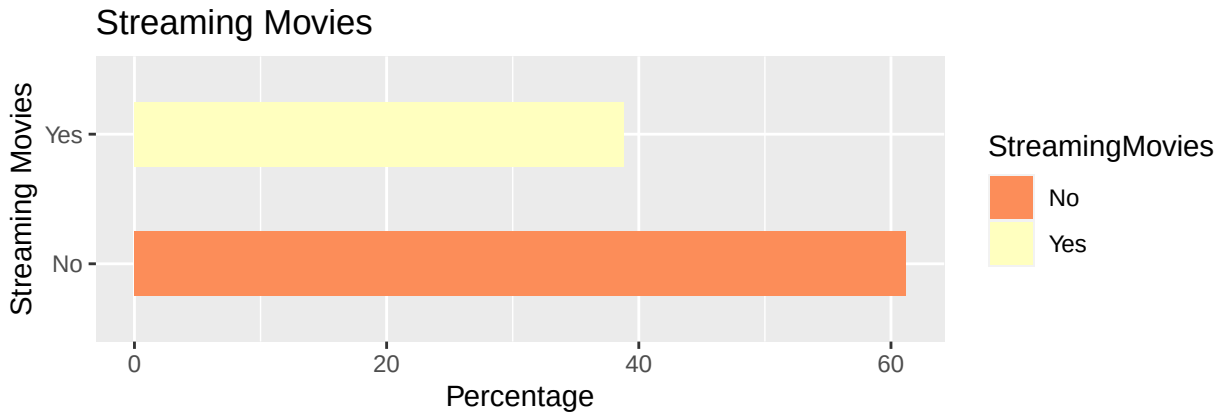


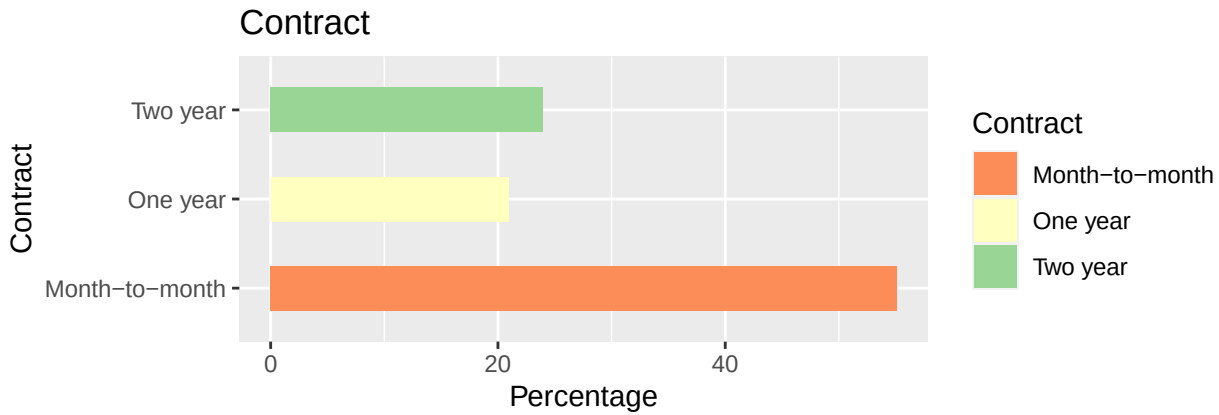
```
grid.arrange(bar_techsupport, bar_steamingtv, ncol=1)
```



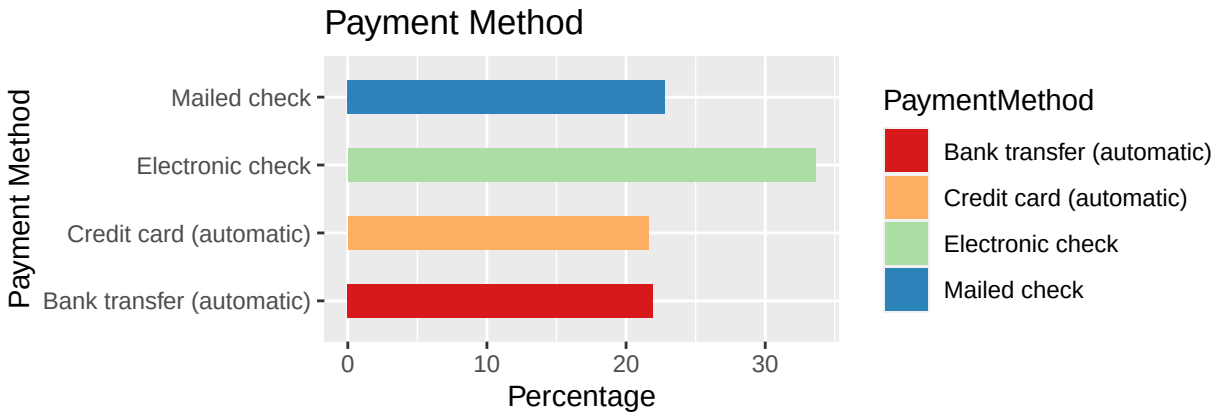

```
bar_steamingmovies <- ggplot(churn_data, aes(x=StreamingMovies, fill=StreamingMovies)) +
  ggtitle("Streaming Movies") + xlab("Streaming Movies") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") +
  coord_flip() + scale_fill_brewer(palette="Spectral")

bar_contract <- ggplot(churn_data, aes(x=Contract, fill=Contract)) + ggtitle("Contract") +
  xlab("Contract") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") +
  coord_flip()  + scale_fill_brewer(palette="Spectral")
```

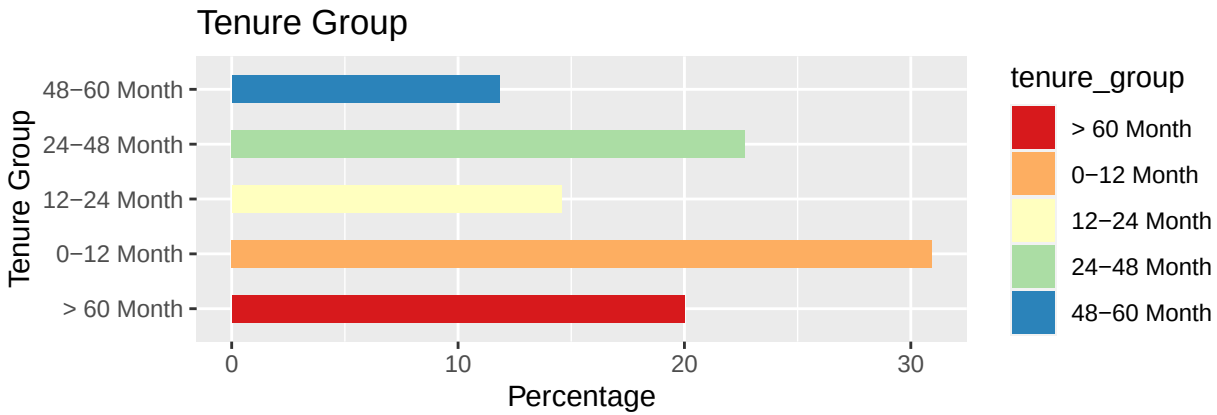
```
bar_paperlessbilling <- ggplot(churn_data, aes(x=PaperlessBilling, fill=PaperlessBilling)) +
  ggtitle("Paperless Billing") + xlab("Paperless Billing") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") +
  coord_flip()  + scale_fill_brewer(palette="Spectral")

bar_paymentmethod <- ggplot(churn_data, aes(x=PaymentMethod, fill=PaymentMethod)) +
  ggtitle("Payment Method") + xlab("Payment Method") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") +
  coord_flip()  + scale_fill_brewer(palette="Spectral")

bar_tenuregroup <- ggplot(churn_data, aes(x=tenure_group, fill=tenure_group)) +
  ggtitle("Tenure Group") + xlab("Tenure Group") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") +
  coord_flip()  + scale_fill_brewer(palette="Spectral")

grid.arrange(bar_steamingmovies, bar_contract, ncol=1)
```



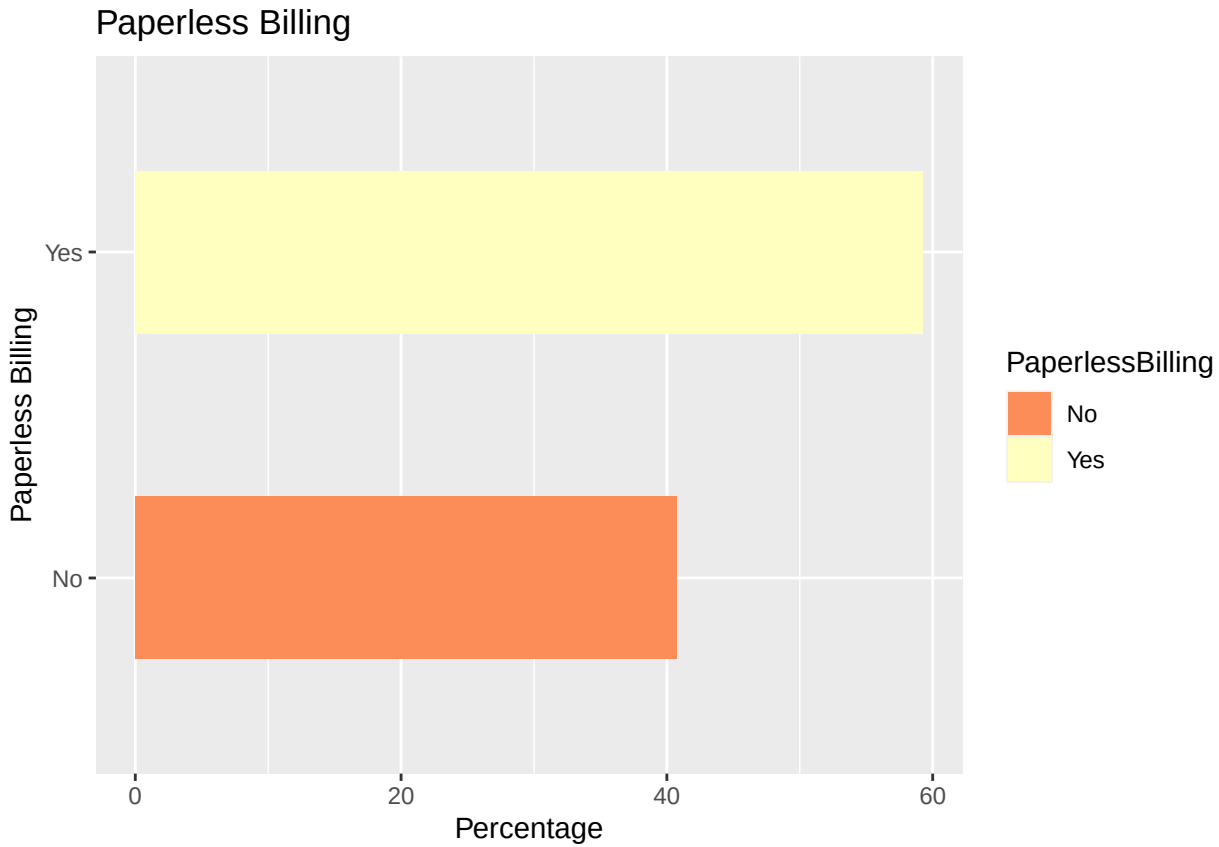



```
grid.arrange(bar_paymentmethod, bar_tenuregroup, ncol=1)
```






```
grid.arrange(bar_paperlessbilling, ncol=1)
```




```
# Distribution Analysis
 # Gender - 50/50
 # Senior Citizen - Majority are NO
 # Partner - about 50/50
 # Dependents - most do not have a dependent
 # Phone Service - most have it
 # Multiple Lines - about 60% have mulitple lines
 # Internet Service - majority is Fiber Optic and DSL
 # Online Security - must do not have online security
 # Online Backup - over 60% do not have online backup
 # Device Protection - over 60% do not have device protection
 # Tech Support - over 60% do not have tech support
```

```
# Steaming TV - over 60% do not have steaming tv
# Steaming Movies - over 60% do not steam movies
# Contract - majority have month to month contract
# Paperless Billing - more than 60% do have paperless billing
# Payment Method - evenly distributed
# Tenure Group - majority is 0-12 months followed by 24-48 months
```

Step 5: Machine Learning

```
# Machine Learning
# Logistic Regression

#Changing all character variables into factors
str(churn_data)
```

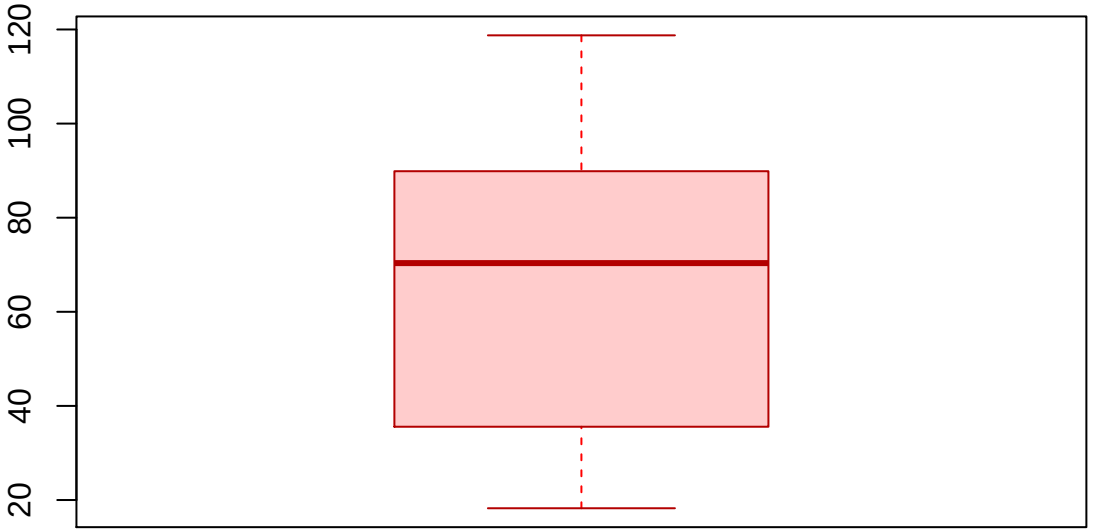
```
## 'data.frame':    7032 obs. of  19 variables:
##  $ gender          : chr  "Female" "Male" "Male" "Male" ...
##  $ SeniorCitizen   : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Partner         : chr  "Yes" "No" "No" "No" ...
##  $ Dependents      : chr  "No" "No" "No" "No" ...
##  $ PhoneService    : chr  "No" "Yes" "Yes" "No" ...
##  $ MultipleLines   : chr  "No" "No" "No" "No" ...
##  $ InternetService : chr  "DSL" "DSL" "DSL" "DSL" ...
##  $ OnlineSecurity  : chr  "No" "Yes" "Yes" "Yes" ...
##  $ OnlineBackup    : chr  "Yes" "No" "Yes" "No" ...
##  $ DeviceProtection: chr  "No" "Yes" "No" "Yes" ...
##  $ TechSupport     : chr  "No" "No" "No" "Yes" ...
##  $ StreamingTV     : chr  "No" "No" "No" "No" ...
##  $ StreamingMovies : chr  "No" "No" "No" "No" ...
##  $ Contract        : chr  "Month-to-month" "One year" "Month-to-month" "One year" ...
##  $ PaperlessBilling: chr  "Yes" "No" "Yes" "No" ...
##  $ PaymentMethod   : chr  "Electronic check" "Mailed check" "Mailed check" "Bank transfer (automatic)" ...
##  $ MonthlyCharges  : num  29.9 57 53.9 42.3 70.7 ...
##  $ Churn           : chr  "No" "No" "Yes" "No" ...
##  $ tenure_group    : Factor w/ 5 levels "> 60 Month","0-12 Month",..: 2 4 2 4 2 2 3 2 4 1 ...
```

```
churn_data$gender <- as.factor(churn_data$gender)
churn_data$Partner <- as.factor(churn_data$Partner)
churn_data$Dependents <- as.factor(churn_data$Dependents)
```

```r
churn_data$PhoneService <- as.factor(churn_data$PhoneService)
churn_data$MultipleLines <- as.factor(churn_data$MultipleLines)
churn_data$InternetService <- as.factor(churn_data$InternetService)
churn_data$OnlineSecurity <- as.factor(churn_data$OnlineSecurity)
churn_data$OnlineBackup <- as.factor(churn_data$OnlineBackup)
churn_data$DeviceProtection <- as.factor(churn_data$DeviceProtection)
churn_data$TechSupport <- as.factor(churn_data$TechSupport)
churn_data$StreamingTV <- as.factor(churn_data$StreamingTV)
churn_data$StreamingMovies <- as.factor(churn_data$StreamingMovies)
churn_data$Contract <- as.factor(churn_data$Contract)
churn_data$PaperlessBilling <- as.factor(churn_data$PaperlessBilling)
churn_data$PaymentMethod <- as.factor(churn_data$PaymentMethod)
churn_data $Churn<- as.factor(churn_data$Churn)
str(churn_data)
```

```
## 'data.frame':    7032 obs. of  19 variables:
##  $ gender          : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 1 2 1 1 2 ...
##  $ SeniorCitizen   : Factor w/ 2 levels "No","Yes": 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 ...
##  $ PhoneService    : Factor w/ 2 levels "No","Yes": 1 2 2 1 2 2 2 1 2 2 ...
##  $ MultipleLines   : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 2 1 2 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 1 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 2 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 4 4 1 3 3 2 4 3 1 ...
##  $ MonthlyCharges  : num  29.9 57 53.9 42.3 70.7 ...
##  $ Churn           : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1 1 2 1 ...
##  $ tenure_group    : Factor w/ 5 levels "> 60 Month","0-12 Month",..: 2 4 2 4 2 2 3 2 4 1 ...
```

```r
# Distribution of Monthly Charges
boxplot(churn_data$MonthlyCharges , col=c2, medcol=c3, whiskcol=c1, staplecol=c3, boxcol=c3, outcol=c3, pch=23, cex=2)
```




```
# Group Monthly Charges into
min(churn_data$MonthlyCharges)
```

```
## [1] 18.25
```

```
max(churn_data$MonthlyCharges)
```

```
## [1] 118.75
```

```
# Creating a factor for Montly Charges
group_monthlycharges <- function(monthlycharges){
  if (monthlycharges >= 0 & monthlycharges <= 40){
    return('$0-40 Monthly Charges')
  }else if(monthlycharges > 40 & monthlycharges <= 60){
    return('$41-60 Monthly Charges')
  }else if (monthlycharges > 60 & monthlycharges <= 80){
    return('$61-80 Monthly Charges')
  }else if (monthlycharges > 80 & monthlycharges <= 100){
    return('$81-100 Monthly Charges')
  }else if (monthlycharges > 100){
    return('> $100 Monthly Charges')
  }
}
```

```
churn_data$group_monthlycharges <- sapply(churn_data$MonthlyCharges,group_monthlycharges)
churn_data$group_monthlycharges <- as.factor(churn_data$group_monthlycharges)
unique(churn_data$group_monthlycharges)
```

```
## [1] $0-40 Monthly Charges   $41-60 Monthly Charges  $61-80 Monthly Charges
## [4] $81-100 Monthly Charges > $100 Monthly Charges
## 5 Levels: > $100 Monthly Charges ... $81-100 Monthly Charges
```

```
# Drop the old monthly charges column
churn_data$MonthlyCharges <- NULL

# Double check the structure of the dataset
str(churn_data)
```

```
## 'data.frame':    7032 obs. of  19 variables:
##  $ gender              : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 1 2 1 1 2 ...
##  $ SeniorCitizen       : Factor w/ 2 levels "No","Yes": 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 ...
##  $ PhoneService        : Factor w/ 2 levels "No","Yes": 1 2 2 1 2 2 2 1 2 2 ...
##  $ MultipleLines       : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 2 1 2 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 1 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 2 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 4 4 1 3 3 2 4 3 1 ...
##  $ Churn               : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1 1 2 1 ...
##  $ tenure_group        : Factor w/ 5 levels "> 60 Month","0-12 Month",..: 2 4 2 4 2 2 3 2 4 1 ...
##  $ group_monthlycharges: Factor w/ 5 levels "> $100 Monthly Charges",..: 2 3 3 3 4 5 5 2 1 3 ...
```

```
# Confirmed, all factors

# Split the data into testing and training sets
intrain<- createDataPartition(churn_data$Churn,p=0.7,list=FALSE)
set.seed(2017)
```

```r
training_set <- churn_data[intrain,]
testing_set <- churn_data[-intrain,]

# Confirm that I have two sets and the split is accurate
dim(training_set)
```

```
## [1] 4924   19
```

```r
dim(testing_set)
```

```
## [1] 2108   19
```

```r
# Fit the logistic regression model to the training set
LogisticModel <- glm(Churn ~ .,family=binomial(link="logit"),data=training_set)
print(summary(LogisticModel))
```

```
##
## Call:
## glm(formula = Churn ~ ., family = binomial(link = "logit"), data = training_set)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9375  -0.6735  -0.2903   0.6910   3.1365
##
## Coefficients:
##                                         Estimate Std. Error z value
## (Intercept)                            -2.1735573  0.6720903  -3.234
## genderMale                             -0.0002869  0.0775454  -0.004
## SeniorCitizenYes                        0.1933685  0.1009237   1.916
## PartnerYes                             -0.0704779  0.0933190  -0.755
## DependentsYes                          -0.1796586  0.1090659  -1.647
## PhoneServiceYes                        -0.4353858  0.2521212  -1.727
## MultipleLinesYes                        0.2465774  0.0998762   2.469
## InternetServiceFiber optic              1.1581873  0.2326613   4.978
## InternetServiceNo                      -1.0264697  0.3086463  -3.326
## OnlineSecurityYes                      -0.3740466  0.1052052  -3.555
## OnlineBackupYes                        -0.1331457  0.0972790  -1.369
## DeviceProtectionYes                     0.0254733  0.1004243   0.254
## TechSupportYes                         -0.3374536  0.1085169  -3.110
## StreamingTVYes                          0.2097973  0.1190648   1.762
```

```
## StreamingMoviesYes                             0.2841585  0.1224020   2.322
## ContractOne year                              -0.6996166  0.1285468  -5.443
## ContractTwo year                              -1.5795879  0.2193649  -7.201
## PaperlessBillingYes                            0.3743348  0.0890850   4.202
## PaymentMethodCredit card (automatic)          -0.1603473  0.1369763  -1.171
## PaymentMethodElectronic check                  0.2923704  0.1116994   2.617
## PaymentMethodMailed check                     -0.0798604  0.1355068  -0.589
## tenure_group0-12 Month                         1.7442193  0.2056344   8.482
## tenure_group12-24 Month                        0.8675760  0.2031100   4.271
## tenure_group24-48 Month                        0.5894408  0.1845738   3.194
## tenure_group48-60 Month                        0.1823309  0.2040229   0.894
## group_monthlycharges$0-40 Monthly Charges      0.3094602  0.6163700   0.502
## group_monthlycharges$41-60 Monthly Charges     0.2891101  0.4454565   0.649
## group_monthlycharges$61-80 Monthly Charges    -0.0526724  0.2835763  -0.186
## group_monthlycharges$81-100 Monthly Charges   -0.1829547  0.1752146  -1.044
##                                               Pr(>|z|)
## (Intercept)                                   0.001221 **
## genderMale                                    0.997048
## SeniorCitizenYes                              0.055367 .
## PartnerYes                                    0.450107
## DependentsYes                                 0.099507 .
## PhoneServiceYes                               0.084187 .
## MultipleLinesYes                              0.013556 *
## InternetServiceFiber optic                    6.42e-07 ***
## InternetServiceNo                             0.000882 ***
## OnlineSecurityYes                             0.000377 ***
## OnlineBackupYes                               0.171094
## DeviceProtectionYes                           0.799761
## TechSupportYes                                0.001873 **
## StreamingTVYes                                0.078062 .
## StreamingMoviesYes                            0.020259 *
## ContractOne year                              5.25e-08 ***
## ContractTwo year                              5.99e-13 ***
## PaperlessBillingYes                           2.65e-05 ***
## PaymentMethodCredit card (automatic)          0.241751
## PaymentMethodElectronic check                 0.008858 **
## PaymentMethodMailed check                     0.555629
## tenure_group0-12 Month                         < 2e-16 ***
## tenure_group12-24 Month                       1.94e-05 ***
```

```
## tenure_group24-48 Month                     0.001405 **
## tenure_group48-60 Month                     0.371494
## group_monthlycharges$0-40 Monthly Charges   0.615619
## group_monthlycharges$41-60 Monthly Charges  0.516326
## group_monthlycharges$61-80 Monthly Charges  0.852646
## group_monthlycharges$81-100 Monthly Charges 0.296404
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5702.8  on 4923  degrees of freedom
## Residual deviance: 4092.8  on 4895  degrees of freedom
## AIC: 4150.8
##
## Number of Fisher Scoring iterations: 6
```

```r
# Feature Selection importance with Chi-Squared Distribution
anova(LogisticModel, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Churn
##
## Terms added sequentially (first to last)
##
##
##                     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                 4923     5702.8
## gender               1     0.16      4922     5702.6   0.68870
## SeniorCitizen        1   106.09      4921     5596.5 < 2.2e-16 ***
## Partner              1   137.37      4920     5459.1 < 2.2e-16 ***
## Dependents           1    45.28      4919     5413.9 1.704e-11 ***
## PhoneService         1     0.05      4918     5413.8   0.82351
## MultipleLines        1     6.36      4917     5407.4   0.01165 *
## InternetService      2   481.90      4915     4925.5 < 2.2e-16 ***
## OnlineSecurity       1   179.65      4914     4745.9 < 2.2e-16 ***
## OnlineBackup         1    68.14      4913     4677.8 < 2.2e-16 ***
```

```
## DeviceProtection      1    40.37      4912     4637.4 2.100e-10 ***
## TechSupport           1    80.26      4911     4557.1 < 2.2e-16 ***
## StreamingTV           1     0.05      4910     4557.1   0.81663
## StreamingMovies       1     0.53      4909     4556.6   0.46691
## Contract              2   258.68      4907     4297.9 < 2.2e-16 ***
## PaperlessBilling      1    15.66      4906     4282.2 7.599e-05 ***
## PaymentMethod         3    35.86      4903     4246.3 7.997e-08 ***
## tenure_group          4   147.62      4899     4098.7 < 2.2e-16 ***
## group_monthlycharges  4     5.97      4895     4092.8   0.20116
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Some of the top features are
 # tenure_group < 2e-16 *** p-value
 # contract 5.72e-15 *** p-value

# Based on the chi-square test the top features are
 # tenure_group  < 2.2e-16 ***
 # Contract < 2.2e-16 ***
 # InternetService < 2.2e-16 ***
 # OnlineSecurity < 2.2e-16 ***
 # OnlineBackup  < 2.2e-16 ***
 # PaperlessBilling 2.021e-05 ***

# Based on the deviance column,
 #  Adding InternetService, Contract and tenure_group significantly reduces the residual deviance.
 # The other variables such as PaymentMethod and Dependents seem to improve the model
  # less even though they all have low p-values.

# Assessing the predictive ability of the Logistic Regression Model

testing_set$Churn <- as.character(testing_set$Churn)
testing_set$Churn[testing_set$Churn=="No"] <- "0"
testing_set$Churn[testing_set$Churn=="Yes"] <- "1"
fitted.results <- predict(LogisticModel,newdata=testing_set,type='response')
fitted.results <- ifelse(fitted.results > 0.5,1,0)
misClasificError <- mean(fitted.results != testing_set$Churn)
print(paste('Logistic Regression Accuracy',1-misClasificError))
```

```
## [1] "Logistic Regression Accuracy 0.799335863377609"
```

```r
# Logistic Confusion Matrix

print("Confusion Matrix for Logistic Regression")
```

```
## [1] "Confusion Matrix for Logistic Regression"
```

```r
table(testing_set$Churn, fitted.results > 0.5)
```

```
##
##     FALSE TRUE
##   0  1402  146
##   1   277  283
```

```r
# Odds Ratio Analysis
# Odds ratio is what the odds of an event happening.
exp(cbind(OR=coef(LogisticModel), confint(LogisticModel)))
```

```
## Waiting for profiling to be done...
```

```
##                                               OR      2.5 %    97.5 %
## (Intercept)                            0.1137722 0.03033946 0.4231613
## genderMale                             0.9997132 0.85875564 1.1638832
## SeniorCitizenYes                       1.2133298 0.99533275 1.4785210
## PartnerYes                             0.9319483 0.77617934 1.1190999
## DependentsYes                          0.8355555 0.67417831 1.0339995
## PhoneServiceYes                        0.6470150 0.39603235 1.0647481
## MultipleLinesYes                       1.2796382 1.05230458 1.5567379
## InternetServiceFiber optic             3.1841561 2.02636573 5.0471843
## InternetServiceNo                      0.3582695 0.19509873 0.6545943
## OnlineSecurityYes                      0.6879448 0.55929021 0.8449148
## OnlineBackupYes                        0.8753376 0.72332487 1.0592279
## DeviceProtectionYes                    1.0258005 0.84258838 1.2491918
## TechSupportYes                         0.7135851 0.57641495 0.8821600
## StreamingTVYes                         1.2334280 0.97689611 1.5581647
## StreamingMoviesYes                     1.3286435 1.04562958 1.6897743
## ContractOne year                       0.4967757 0.38510404 0.6376118
## ContractTwo year                       0.2060600 0.13218219 0.3129274
## PaperlessBillingYes                    1.4540239 1.22151261 1.7322027
## PaymentMethodCredit card (automatic)   0.8518479 0.65083084 1.1137299
## PaymentMethodElectronic check          1.3395991 1.07691715 1.6688571
## PaymentMethodMailed check              0.9232452 0.70806127 1.2046094
```

```
## tenure_group0-12 Month                       5.7214331 3.83864098 8.5997854
## tenure_group12-24 Month                      2.3811320 1.60422441 3.5588281
## tenure_group24-48 Month                      1.8029799 1.26006148 2.5997219
## tenure_group48-60 Month                      1.2000113 0.80434818 1.7914347
## group_monthlycharges$0-40 Monthly Charges    1.3626893 0.40766952 4.5704534
## group_monthlycharges$41-60 Monthly Charges   1.3352387 0.55814065 3.2013262
## group_monthlycharges$61-80 Monthly Charges   0.9486908 0.54397370 1.6537256
## group_monthlycharges$81-100 Monthly Charges 0.8328059 0.59053484 1.1739089
```

```
# The OR is a way to present the strength of association between risk factors/exposures and outcomes.
# If the OR is <1, odds are decreased for an outcome
# OR >1 means the odds are increased for a given outcome.

# Likely events
 # Customers choosing InternetServiceFiber optic (6.4368732)
 # tenure_group0-12 Month (8.3184940)
 # group_monthlycharges$0-40 Monthly Charges (7.1006580)
 # group_monthlycharges$41-60 Monthly Charges (5.0153739)

# Least Likely events
 # genderMale (0.1959533)
 # ContractTwo year (0.2812260)


# Random Forest Prediction and Confusion Matrix
training_set$Churn <- as.character(training_set$Churn)
training_set$Churn[training_set$Churn=="No"] <- "0"
training_set$Churn[training_set$Churn=="Yes"] <- "1"
training_set$Churn <- as.factor(training_set$Churn)
testing_set$Churn <- as.factor(testing_set$Churn)

#Check unique values
unique(training_set$Churn)
```

```
## [1] 0 1
## Levels: 0 1
```

```
unique(testing_set$Churn)
```

```
## [1] 0 1
```

```
## Levels: 0 1
```

```
#Check Structure
str(training_set)
```

```
## 'data.frame':    4924 obs. of  19 variables:
##  $ gender              : Factor w/ 2 levels "Female","Male": 1 2 2 1 1 1 2 2 2 2 ...
##  $ SeniorCitizen       : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Partner             : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 2 2 1 2 1 ...
##  $ Dependents          : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 2 1 1 1 ...
##  $ PhoneService        : Factor w/ 2 levels "No","Yes": 1 2 1 2 2 2 2 2 2 2 ...
##  $ MultipleLines       : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 2 1 1 2 2 ...
##  $ InternetService     : Factor w/ 3 levels "DSL","Fiber optic",..: 1 1 1 2 2 2 1 3 2 2 ...
##  $ OnlineSecurity      : Factor w/ 2 levels "No","Yes": 1 2 2 1 1 1 2 1 1 1 ...
##  $ OnlineBackup        : Factor w/ 2 levels "No","Yes": 2 2 1 1 1 1 1 1 1 2 ...
##  $ DeviceProtection    : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1 1 2 2 ...
##  $ TechSupport         : Factor w/ 2 levels "No","Yes": 1 1 2 1 1 2 1 1 1 1 ...
##  $ StreamingTV         : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 2 1 1 2 2 ...
##  $ StreamingMovies     : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 2 1 1 2 2 ...
##  $ Contract            : Factor w/ 3 levels "Month-to-month",..: 1 1 2 1 1 1 1 3 2 1 ...
##  $ PaperlessBilling    : Factor w/ 2 levels "No","Yes": 2 2 1 2 2 2 2 1 1 2 ...
##  $ PaymentMethod       : Factor w/ 4 levels "Bank transfer (automatic)",..: 3 4 1 3 3 3 4 2 2 1 ...
##  $ Churn               : Factor w/ 2 levels "0","1": 1 2 1 2 2 2 1 1 1 2 ...
##  $ tenure_group        : Factor w/ 5 levels "> 60 Month","0-12 Month",..: 2 2 4 2 2 4 3 3 5 5 ...
##  $ group_monthlycharges: Factor w/ 5 levels "> $100 Monthly Charges",..: 2 3 3 4 5 1 3 2 1 1 ...
```

```
str(testing_set)
```

```
## 'data.frame':    2108 obs. of  19 variables:
##  $ gender              : Factor w/ 2 levels "Female","Male": 2 2 1 2 1 2 1 2 2 1 ...
##  $ SeniorCitizen       : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Partner             : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 1 2 2 2 1 ...
##  $ Dependents          : Factor w/ 2 levels "No","Yes": 1 2 1 2 2 1 1 2 2 2 ...
##  $ PhoneService        : Factor w/ 2 levels "No","Yes": 2 2 1 2 2 2 2 2 2 2 ...
##  $ MultipleLines       : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 1 2 1 2 1 ...
##  $ InternetService     : Factor w/ 3 levels "DSL","Fiber optic",..: 1 2 1 1 1 3 1 1 2 1 ...
##  $ OnlineSecurity      : Factor w/ 2 levels "No","Yes": 2 1 2 2 1 1 1 2 1 1 ...
##  $ OnlineBackup        : Factor w/ 2 levels "No","Yes": 1 2 1 2 1 1 2 2 2 1 ...
##  $ DeviceProtection    : Factor w/ 2 levels "No","Yes": 2 1 1 1 2 1 1 1 1 1 ...
##  $ TechSupport         : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 1 2 2 1 1 ...
```

```
##  $ StreamingTV         : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 1 1 1 2 2 ...
##  $ StreamingMovies     : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 2 2 ...
##  $ Contract            : Factor w/ 3 levels "Month-to-month",..: 2 1 1 2 1 1 3 1 1 1 ...
##  $ PaperlessBilling    : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 1 2 1 2 2 ...
##  $ PaymentMethod       : Factor w/ 4 levels "Bank transfer (automatic)",..: 4 2 4 1 2 4 2 2 3 4 ...
##  $ Churn               : Factor w/ 2 levels "0","1": 1 1 1 1 2 2 1 1 2 2 ...
##  $ tenure_group        : Factor w/ 5 levels "> 60 Month","0-12 Month",..: 4 3 2 1 2 2 5 5 4 3 ...
##  $ group_monthlycharges: Factor w/ 5 levels "> $100 Monthly Charges",..: 3 5 2 3 3 2 3 3 5 4 ...
```

```
# Machine Learning Model 2:  Random Forest
rfModel <- randomForest(Churn ~., data = training_set)
print(rfModel)
```

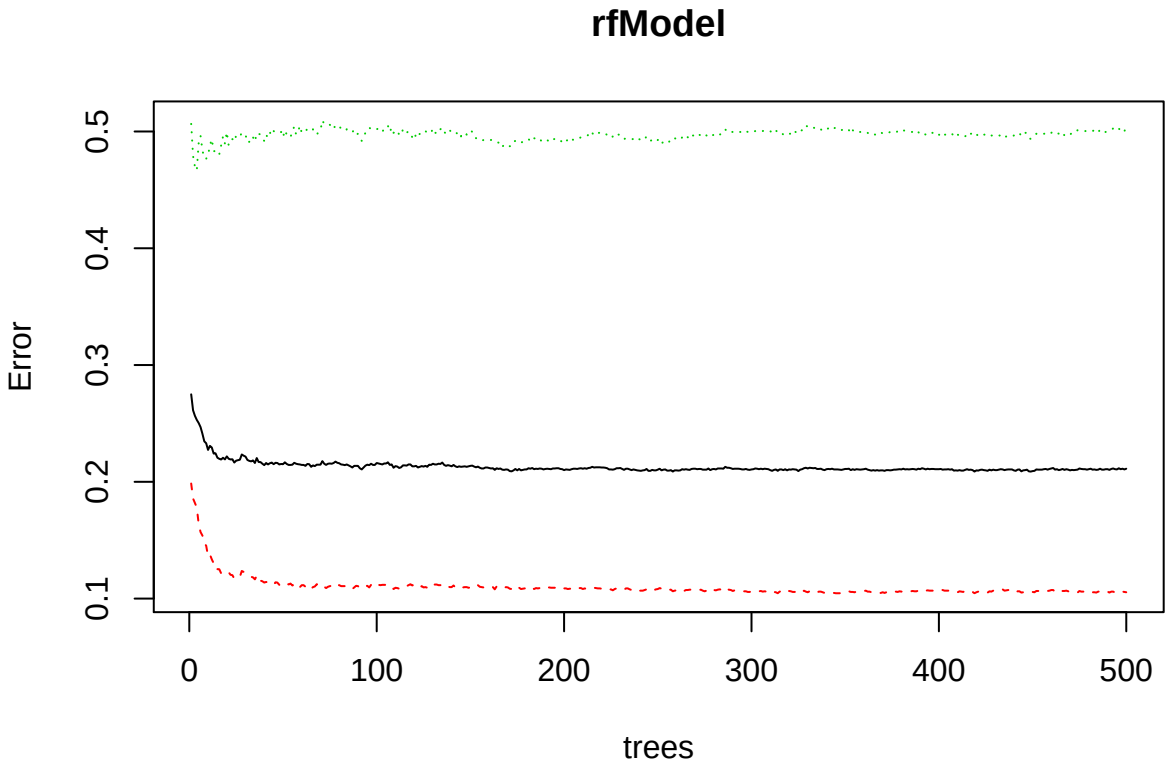
```
##
## Call:
##  randomForest(formula = Churn ~ ., data = training_set)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 4
##
##         OOB estimate of  error rate: 21.12%
## Confusion matrix:
##      0   1 class.error
## 0 3234 381   0.1053942
## 1  659 650   0.5034377
```

```
# Logistic Regression Model still performs better
# Error rate is 50% when predicting Yes!

# Random Forest Prediction and Confusion Matrix
pred_rf <- predict(rfModel, testing_set)
caret::confusionMatrix(pred_rf, testing_set$Churn)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1384  292
##          1  164  268
##
```

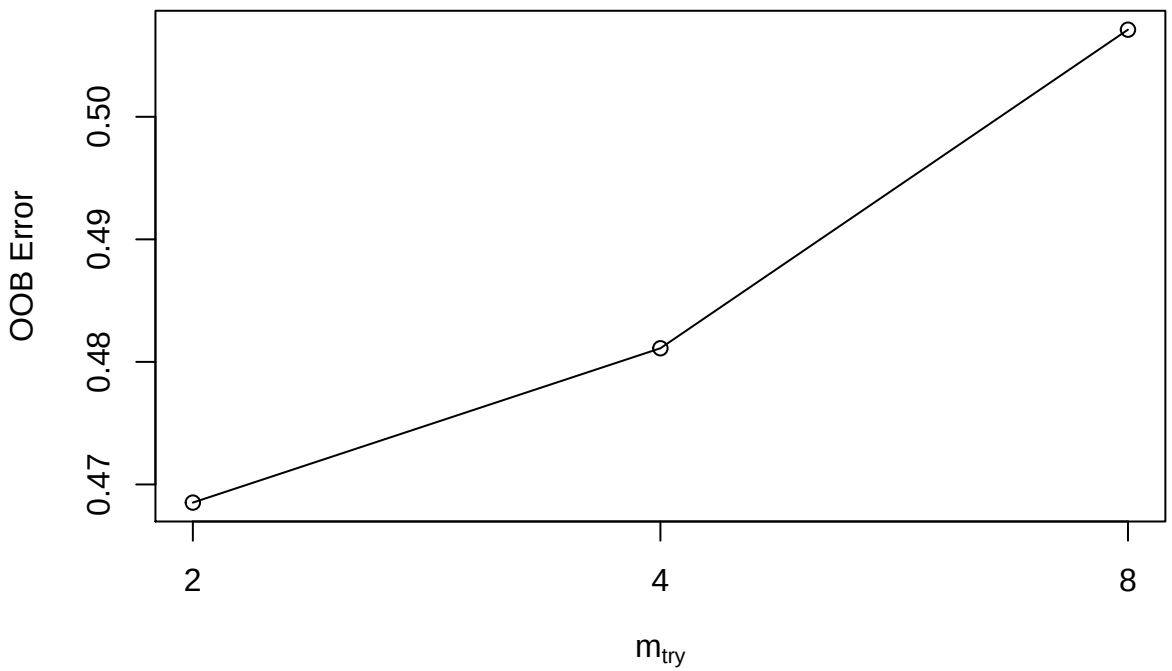
```
##                Accuracy : 0.7837
##                  95% CI : (0.7655, 0.8011)
##     No Information Rate : 0.7343
##     P-Value [Acc > NIR] : 9.376e-08
##
##                   Kappa : 0.4019
##
##  Mcnemar's Test P-Value : 2.726e-09
##
##             Sensitivity : 0.8941
##             Specificity : 0.4786
##          Pos Pred Value : 0.8258
##          Neg Pred Value : 0.6204
##              Prevalence : 0.7343
##          Detection Rate : 0.6565
##    Detection Prevalence : 0.7951
##       Balanced Accuracy : 0.6863
##
##        'Positive' Class : 0
##
```

```r
# Plot Model
plot(rfModel)
```




```
# Use this plot to help us determine the number of trees
# As the number of trees increases, the OOB error rate decreases, and then becomes almost constant
# Not able to decrease the OOB error rate after about 100 to 200 trees.

# Tune Random Forest Model (specifiying trees)
tune_model <- tuneRF(training_set[, -18], training_set[, 18], stepFactor = 0.5, plot = TRUE, ntreeTry = 200, trace = TRUE, improve = 0.05)
```

```
## mtry = 4  OOB error = 48.11%
## Searching left ...
## mtry = 8     OOB error = 50.71%
## -0.05403124 0.05
## Searching right ...
## mtry = 2     OOB error = 46.85%
## 0.02617138 0.05
```




```
# Use this plot to give us some ideas on the number of mtry to choose.
# OOB error rate is at the lowest when mtry is 2.
# Number of variables available for splitting at each tree node
# For regression models, it is the number of predictor variables divided by 3 (rounded down)
# Using mtry=2

# Fit the Random Forest Model to Training Set after Tuning
randomForestModel_New <- randomForest(Churn ~., data = training_set, ntree = 200, mtry = 2, importance = TRUE, proximity = TRUE)
print(randomForestModel_New)
```

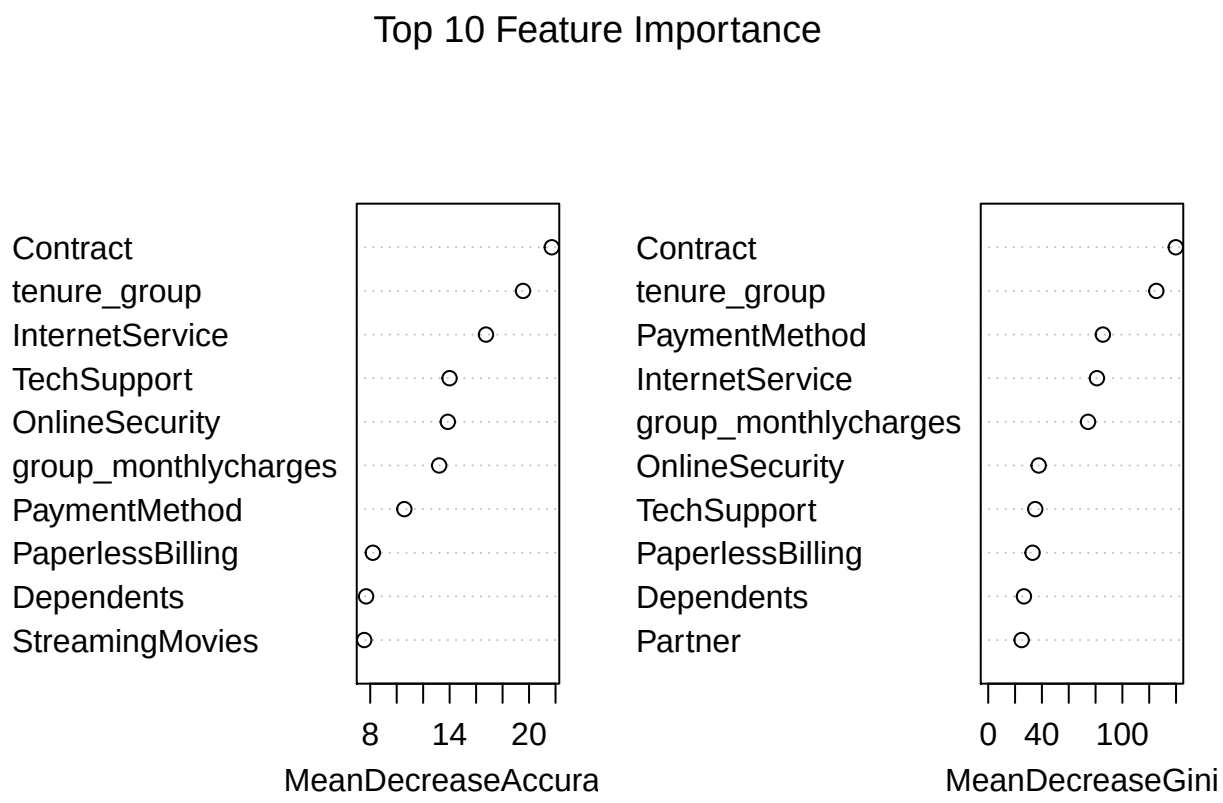
```
##
## Call:
##  randomForest(formula = Churn ~ ., data = training_set, ntree = 200,      mtry = 2, importance = TRUE, proximity = TRUE)
##                Type of random forest: classification
##                      Number of trees: 200
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 20.76%
```

```
## Confusion matrix:
##      0   1 class.error
## 0 3303 312  0.08630705
## 1  710 599  0.54239878
```

```
# OBB estimate of error rate decreased slightly to 20.98% from 21.77%

# Random Forest Predictions and Confusion Matrix After Tuning
pred_rf_new <- predict(randomForestModel_New, testing_set)
caret::confusionMatrix(pred_rf_new, testing_set$Churn)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1418  322
##          1  130  238
##
##                Accuracy : 0.7856
##                  95% CI : (0.7674, 0.8029)
##     No Information Rate : 0.7343
##     P-Value [Acc > NIR] : 3.016e-08
##
##                   Kappa : 0.3829
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9160
##             Specificity : 0.4250
##          Pos Pred Value : 0.8149
##          Neg Pred Value : 0.6467
##              Prevalence : 0.7343
##          Detection Rate : 0.6727
##    Detection Prevalence : 0.8254
##       Balanced Accuracy : 0.6705
##
##        'Positive' Class : 0
##
```

```
# Accuracy : 0.8017  vs.  0.7932
# Sensitivity : 0.9244  vs.  0.8979
# Mcnemar's Test P-Value: < 2.2e-16 vs. 1.205e-08 (nice improvement)
# Tuning did boost the performance of the Random Forest Model

# Random Forest Feature Importance
varImpPlot(randomForestModel_New, sort=T, n.var = 10, main = 'Top 10 Feature Importance')
```




```
# Contract and Tenure Group are at the top as shown before.
# Logistic Model Accuracy: .805028462998102
# Random Forest Model Accuracy: 0.8017
# Logistic Regression Model performs slightly better
```