randomForest.R

sigsp

2022-03-22

```
## Author: Stephen E. Porter
## Title: Random Forest (Task 2)
## Course: WGU D209: Data Mining I
## Instructor: Dr. Festus Elleh
options(warn=-1)

# Libraries
library(tidyverse)
```

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

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(caret)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

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

```
library(randomForest)
```

```
## randomForest 4.7-1
```

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

```
##
## Attaching package: 'randomForest'
```

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

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

```r
# Import CSV as data frame
df <- read.csv(file = 'C:/WGU/D209 Data Mining I/churn_clean.csv')

# Checking for nulls
sapply(df, function(x) sum(is.na(x)))
```

```
##            CaseOrder          Customer_id          Interaction
##                    0                    0                    0
##                  UID                 City                State
##                    0                    0                    0
##               County                  Zip                  Lat
##                    0                    0                    0
##                  Lng           Population                 Area
##                    0                    0                    0
##             TimeZone                  Job             Children
##                    0                    0                    0
##                  Age               Income              Marital
##                    0                    0                    0
##               Gender                Churn   Outage_sec_perweek
##                    0                    0                    0
##                Email             Contacts Yearly_equip_failure
##                    0                    0                    0
##               Techie             Contract           Port_modem
##                    0                    0                    0
##               Tablet      InternetService                Phone
##                    0                    0                    0
##             Multiple       OnlineSecurity         OnlineBackup
##                    0                    0                    0
##     DeviceProtection          TechSupport          StreamingTV
##                    0                    0                    0
##      StreamingMovies     PaperlessBilling        PaymentMethod
##                    0                    0                    0
##               Tenure        MonthlyCharge    Bandwidth_GB_Year
##                    0                    0                    0
##                Item1                Item2                Item3
##                    0                    0                    0
##                Item4                Item5                Item6
##                    0                    0                    0
##                Item7                Item8
##                    0                    0
```

```r
dim(df)
```

```
## [1] 10000    50
```

```r
str(df)
```

```
## 'data.frame':    10000 obs. of  50 variables:
##  $ CaseOrder          : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Customer_id        : chr  "K409198" "S120509" "K191035" "D90850" ...
##  $ Interaction        : chr  "aa90260b-4141-4a24-8e36-b04ce1f4f77b" "fb76459f-c047-4a9d-8af9-e0f7d4
##  $ UID                : chr  "e885b299883d4f9fb18e39c75155d990" "f2de8bef964785f41a2959829830fb8a"
##  $ City               : chr  "Point Baker" "West Branch" "Yamhill" "Del Mar" ...
##  $ State              : chr  "AK" "MI" "OR" "CA" ...
##  $ County             : chr  "Prince of Wales-Hyder" "Ogemaw" "Yamhill" "San Diego" ...
```

```
##  $ Zip                : int  99927 48661 97148 92014 77461 31030 37847 73109 34771 45237 ...
##  $ Lat                : num  56.3 44.3 45.4 33 29.4 ...
##  $ Lng                : num  -133.4 -84.2 -123.2 -117.2 -95.8 ...
##  $ Population         : int  38 10446 3735 13863 11352 17701 2535 23144 17351 20193 ...
##  $ Area               : chr  "Urban" "Urban" "Urban" "Suburban" ...
##  $ TimeZone           : chr  "America/Sitka" "America/Detroit" "America/Los_Angeles" "America/Los_A
##  $ Job                : chr  "Environmental health practitioner" "Programmer, multimedia" "Chief Fi
##  $ Children           : int  0 1 4 1 0 3 0 2 2 1 ...
##  $ Age                : int  68 27 50 48 83 83 79 30 49 86 ...
##  $ Income             : num  28562 21705 9610 18925 40074 ...
##  $ Marital            : chr  "Widowed" "Married" "Widowed" "Married" ...
##  $ Gender             : chr  "Male" "Female" "Female" "Male" ...
##  $ Churn              : chr  "No" "Yes" "No" "No" ...
##  $ Outage_sec_perweek : num  7.98 11.7 10.75 14.91 8.15 ...
##  $ Email              : int  10 12 9 15 16 15 10 16 20 18 ...
##  $ Contacts           : int  0 0 0 2 2 3 0 0 2 1 ...
##  $ Yearly_equip_failure: int  1 1 1 0 1 1 1 0 3 0 ...
##  $ Techie             : chr  "No" "Yes" "Yes" "Yes" ...
##  $ Contract           : chr  "One year" "Month-to-month" "Two Year" "Two Year" ...
##  $ Port_modem         : chr  "Yes" "No" "Yes" "No" ...
##  $ Tablet             : chr  "Yes" "Yes" "No" "No" ...
##  $ InternetService    : chr  "Fiber Optic" "Fiber Optic" "DSL" "DSL" ...
##  $ Phone              : chr  "Yes" "Yes" "Yes" "Yes" ...
##  $ Multiple           : chr  "No" "Yes" "Yes" "No" ...
##  $ OnlineSecurity     : chr  "Yes" "Yes" "No" "Yes" ...
##  $ OnlineBackup       : chr  "Yes" "No" "No" "No" ...
##  $ DeviceProtection   : chr  "No" "No" "No" "No" ...
##  $ TechSupport        : chr  "No" "No" "No" "No" ...
##  $ StreamingTV        : chr  "No" "Yes" "No" "Yes" ...
##  $ StreamingMovies    : chr  "Yes" "Yes" "Yes" "No" ...
##  $ PaperlessBilling   : chr  "Yes" "Yes" "Yes" "Yes" ...
##  $ PaymentMethod      : chr  "Credit Card (automatic)" "Bank Transfer(automatic)" "Credit Card (aut
##  $ Tenure             : num  6.8 1.16 15.75 17.09 1.67 ...
##  $ MonthlyCharge      : num  172 243 160 120 150 ...
##  $ Bandwidth_GB_Year  : num  905 801 2055 2165 271 ...
##  $ Item1              : int  5 3 4 4 4 3 6 2 5 2 ...
##  $ Item2              : int  5 4 4 4 4 3 5 2 4 2 ...
##  $ Item3              : int  5 3 2 4 4 3 6 2 4 2 ...
##  $ Item4              : int  3 3 4 2 3 2 4 5 3 2 ...
##  $ Item5              : int  4 4 4 5 4 4 1 2 4 5 ...
##  $ Item6              : int  4 3 3 4 4 3 5 3 3 2 ...
##  $ Item7              : int  3 4 3 3 4 3 5 4 4 3 ...
##  $ Item8              : int  4 4 3 3 5 3 5 5 4 3 ...
```

```r
# Renaming unclear columns named Item1 through Item8 for improved readability &
# confirming they have been renamed correctly

df <- df %>%
  rename(
    Response = Item1,
    Fix = Item2,
    Replacement = Item3,
    Reliability = Item4,
    Options = Item5,
```

```
    Respectful = Item6,
    Courteous = Item7,
    Listening = Item8
  )

colnames(df)
```

```
##  [1] "CaseOrder"           "Customer_id"         "Interaction"
##  [4] "UID"                 "City"                "State"
##  [7] "County"              "Zip"                 "Lat"
## [10] "Lng"                 "Population"          "Area"
## [13] "TimeZone"            "Job"                 "Children"
## [16] "Age"                 "Income"              "Marital"
## [19] "Gender"              "Churn"               "Outage_sec_perweek"
## [22] "Email"               "Contacts"            "Yearly_equip_failure"
## [25] "Techie"              "Contract"            "Port_modem"
## [28] "Tablet"              "InternetService"     "Phone"
## [31] "Multiple"            "OnlineSecurity"      "OnlineBackup"
## [34] "DeviceProtection"    "TechSupport"         "StreamingTV"
## [37] "StreamingMovies"     "PaperlessBilling"    "PaymentMethod"
## [40] "Tenure"              "MonthlyCharge"       "Bandwidth_GB_Year"
## [43] "Response"            "Fix"                 "Replacement"
## [46] "Reliability"         "Options"             "Respectful"
## [49] "Courteous"           "Listening"
```

```
# Several columns will not be useful in analysis and therefore will be dropped.
to_drop <- c('CaseOrder', 'Customer_id', 'Interaction', 'UID', 'City',
             'County', 'Zip', 'Lat', 'Lng', 'TimeZone', 'Job')

dfDropped = df[,!(names(df) %in% to_drop)]
str(dfDropped)
```

```
## 'data.frame':    10000 obs. of  39 variables:
##  $ State               : chr  "AK" "MI" "OR" "CA" ...
##  $ Population          : int  38 10446 3735 13863 11352 17701 2535 23144 17351 20193 ...
##  $ Area                : chr  "Urban" "Urban" "Urban" "Suburban" ...
##  $ Children            : int  0 1 4 1 0 3 0 2 2 1 ...
##  $ Age                 : int  68 27 50 48 83 83 79 30 49 86 ...
##  $ Income              : num  28562 21705 9610 18925 40074 ...
##  $ Marital             : chr  "Widowed" "Married" "Widowed" "Married" ...
##  $ Gender              : chr  "Male" "Female" "Female" "Male" ...
##  $ Churn               : chr  "No" "Yes" "No" "No" ...
##  $ Outage_sec_perweek  : num  7.98 11.7 10.75 14.91 8.15 ...
##  $ Email               : int  10 12 9 15 16 15 10 16 20 18 ...
##  $ Contacts            : int  0 0 0 2 2 3 0 0 2 1 ...
##  $ Yearly_equip_failure: int  1 1 1 0 1 1 1 0 3 0 ...
##  $ Techie              : chr  "No" "Yes" "Yes" "Yes" ...
##  $ Contract            : chr  "One year" "Month-to-month" "Two Year" "Two Year" ...
##  $ Port_modem          : chr  "Yes" "No" "Yes" "No" ...
##  $ Tablet              : chr  "Yes" "Yes" "No" "No" ...
##  $ InternetService     : chr  "Fiber Optic" "Fiber Optic" "DSL" "DSL" ...
##  $ Phone               : chr  "Yes" "Yes" "Yes" "Yes" ...
##  $ Multiple            : chr  "No" "Yes" "Yes" "No" ...
##  $ OnlineSecurity      : chr  "Yes" "Yes" "No" "Yes" ...
##  $ OnlineBackup        : chr  "Yes" "No" "No" "No" ...
```

```
##  $ DeviceProtection    : chr  "No" "No" "No" "No" ...
##  $ TechSupport         : chr  "No" "No" "No" "No" ...
##  $ StreamingTV         : chr  "No" "Yes" "No" "Yes" ...
##  $ StreamingMovies     : chr  "Yes" "Yes" "Yes" "No" ...
##  $ PaperlessBilling    : chr  "Yes" "Yes" "Yes" "Yes" ...
##  $ PaymentMethod       : chr  "Credit Card (automatic)" "Bank Transfer(automatic)" "Credit Card (aut
##  $ Tenure              : num  6.8 1.16 15.75 17.09 1.67 ...
##  $ MonthlyCharge       : num  172 243 160 120 150 ...
##  $ Bandwidth_GB_Year   : num  905 801 2055 2165 271 ...
##  $ Response            : int  5 3 4 4 4 3 6 2 5 2 ...
##  $ Fix                 : int  5 4 4 4 4 3 5 2 4 2 ...
##  $ Replacement         : int  5 3 2 4 4 3 6 2 4 2 ...
##  $ Reliability         : int  3 3 4 2 3 2 4 5 3 2 ...
##  $ Options             : int  4 4 4 5 4 4 1 2 4 5 ...
##  $ Respectful          : int  4 3 3 4 4 3 5 3 3 2 ...
##  $ Courteous           : int  3 4 3 3 4 3 5 4 4 3 ...
##  $ Listening           : int  4 4 3 3 5 3 5 5 4 3 ...
```

```
# Convert character columns to factors so they can be used in randomForest

dfDropped[sapply(dfDropped, is.character)] <- lapply(dfDropped[sapply(dfDropped, is.character)], as.fac
str(dfDropped)
```

```
## 'data.frame':    10000 obs. of  39 variables:
##  $ State               : Factor w/ 52 levels "AK","AL","AR",..: 1 23 38 5 45 11 44 37 10 36 ...
##  $ Population          : int  38 10446 3735 13863 11352 17701 2535 23144 17351 20193 ...
##  $ Area                : Factor w/ 3 levels "Rural","Suburban",..: 3 3 3 2 2 3 2 2 2 1 ...
##  $ Children            : int  0 1 4 1 0 3 0 2 2 1 ...
##  $ Age                 : int  68 27 50 48 83 83 79 30 49 86 ...
##  $ Income              : num  28562 21705 9610 18925 40074 ...
##  $ Marital             : Factor w/ 5 levels "Divorced","Married",..: 5 2 5 2 4 3 5 2 4 2 ...
##  $ Gender              : Factor w/ 3 levels "Female","Male",..: 2 1 1 2 2 1 2 1 3 1 ...
##  $ Churn               : Factor w/ 2 levels "No","Yes": 1 2 1 1 2 1 2 2 1 1 ...
##  $ Outage_sec_perweek  : num  7.98 11.7 10.75 14.91 8.15 ...
##  $ Email               : int  10 12 9 15 16 15 10 16 20 18 ...
##  $ Contacts            : int  0 0 0 2 2 3 0 0 2 1 ...
##  $ Yearly_equip_failure: int  1 1 1 0 1 1 1 0 3 0 ...
##  $ Techie              : Factor w/ 2 levels "No","Yes": 1 2 2 2 1 1 2 2 1 1 ...
##  $ Contract            : Factor w/ 3 levels "Month-to-month",..: 2 1 3 3 1 2 1 1 1 3 ...
##  $ Port_modem          : Factor w/ 2 levels "No","Yes": 2 1 2 1 2 2 1 1 2 2 ...
##  $ Tablet              : Factor w/ 2 levels "No","Yes": 2 2 1 1 1 1 1 1 1 1 ...
##  $ InternetService     : Factor w/ 3 levels "DSL","Fiber Optic",..: 2 2 1 1 2 3 1 1 1 2 ...
##  $ Phone               : Factor w/ 2 levels "No","Yes": 2 2 2 2 1 2 2 1 2 2 ...
##  $ Multiple            : Factor w/ 2 levels "No","Yes": 1 2 2 1 1 2 1 1 1 1 ...
##  $ OnlineSecurity      : Factor w/ 2 levels "No","Yes": 2 2 1 2 1 2 1 1 2 2 ...
##  $ OnlineBackup        : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 2 1 2 2 1 ...
##  $ DeviceProtection    : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 1 1 1 2 ...
##  $ TechSupport         : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 1 2 1 1 1 ...
##  $ StreamingTV         : Factor w/ 2 levels "No","Yes": 1 2 1 2 2 1 2 1 1 1 ...
##  $ StreamingMovies     : Factor w/ 2 levels "No","Yes": 2 2 2 1 1 2 2 1 1 2 ...
##  $ PaperlessBilling    : Factor w/ 2 levels "No","Yes": 2 2 2 2 1 1 1 2 2 2 ...
##  $ PaymentMethod       : Factor w/ 4 levels "Bank Transfer(automatic)",..: 2 1 2 4 4 3 3 4 1 4 ...
##  $ Tenure              : num  6.8 1.16 15.75 17.09 1.67 ...
##  $ MonthlyCharge       : num  172 243 160 120 150 ...
##  $ Bandwidth_GB_Year   : num  905 801 2055 2165 271 ...
```

```
##  $ Response            : int  5 3 4 4 4 3 6 2 5 2 ...
##  $ Fix                 : int  5 4 4 4 4 3 5 2 4 2 ...
##  $ Replacement         : int  5 3 2 4 4 3 6 2 4 2 ...
##  $ Reliability         : int  3 3 4 2 3 2 4 5 3 2 ...
##  $ Options             : int  4 4 4 5 4 4 1 2 4 5 ...
##  $ Respectful          : int  4 3 3 4 4 3 5 3 3 2 ...
##  $ Courteous           : int  3 4 3 3 4 3 5 4 4 3 ...
##  $ Listening           : int  4 4 3 3 5 3 5 5 4 3 ...
```

```r
# Split dfDropped into training and testing subsets
set.seed(22)
trainId = createDataPartition(dfDropped$Churn, times = 1, p = 0.7, list = FALSE)

dfTrain = dfDropped[trainId,]
dfTest = dfDropped[-trainId,]

# Summary Statistics
summary(dfTrain)
```

```
##      State        Population             Area         Children
##  TX     : 419   Min.   :     0.0   Rural   :2321   Min.   : 0.000
##  NY     : 404   1st Qu.:   743.8   Suburban:2335   1st Qu.: 0.000
##  CA     : 380   Median :  2991.0   Urban   :2344   Median : 1.000
##  PA     : 371   Mean   :  9822.3                   Mean   : 2.085
##  IL     : 287   3rd Qu.: 13247.0                   3rd Qu.: 3.000
##  OH     : 262   Max.   :111850.0                   Max.   :10.000
##  (Other):4877
##       Age            Income                 Marital           Gender
##  Min.   :18.00   Min.   :   368.5   Divorced     :1500   Female   :3520
##  1st Qu.:35.00   1st Qu.: 18969.3   Married      :1306   Male     :3316
##  Median :53.00   Median : 32936.7   Never Married:1347   Nonbinary: 164
##  Mean   :53.17   Mean   : 39472.8   Separated    :1406
##  3rd Qu.:71.00   3rd Qu.: 52638.8   Widowed      :1441
##  Max.   :89.00   Max.   :258900.7
##
##  Churn      Outage_sec_perweek     Email          Contacts
##  No :5145   Min.   : 0.09975   Min.   : 1.00   Min.   :0.0000
##  Yes:1855   1st Qu.: 8.05141   1st Qu.:10.00   1st Qu.:0.0000
##             Median :10.02823   Median :12.00   Median :1.0000
##             Mean   :10.02875   Mean   :12.01   Mean   :0.9977
##             3rd Qu.:11.98748   3rd Qu.:14.00   3rd Qu.:2.0000
##             Max.   :21.20723   Max.   :23.00   Max.   :7.0000
##
##  Yearly_equip_failure Techie              Contract    Port_modem Tablet
##  Min.   :0.000        No :5833   Month-to-month:3832   No :3635   No :4870
##  1st Qu.:0.000        Yes:1167   One year      :1488   Yes:3365   Yes:2130
##  Median :0.000                   Two Year      :1680
##  Mean   :0.395
##  3rd Qu.:1.000
##  Max.   :4.000
##
##     InternetService Phone      Multiple   OnlineSecurity OnlineBackup
##  DSL        :2391   No : 647   No :3796   No :4565       No :3845
##  Fiber Optic:3082   Yes:6353   Yes:3204   Yes:2435       Yes:3155
##  None       :1527
```

```
##
##
##
##
##  DeviceProtection TechSupport StreamingTV StreamingMovies PaperlessBilling
##  No :3954         No :4369    No :3574    No :3563        No :2902
##  Yes:3046         Yes:2631    Yes:3426    Yes:3437        Yes:4098
##
##
##
##
##
##                    PaymentMethod      Tenure      MonthlyCharge
##  Bank Transfer(automatic):1583   Min.   : 1.00   Min.   : 79.98
##  Credit Card (automatic) :1435   1st Qu.: 7.84   1st Qu.:139.97
##  Electronic Check        :2356   Median :28.90   Median :167.48
##  Mailed Check            :1626   Mean   :34.41   Mean   :172.36
##                                  3rd Qu.:61.49   3rd Qu.:202.44
##                                  Max.   :72.00   Max.   :290.16
##
##  Bandwidth_GB_Year    Response         Fix         Replacement
##  Min.   : 155.5    Min.   :1.00   Min.   :1.000   Min.   :1.000
##  1st Qu.:1220.4    1st Qu.:3.00   1st Qu.:3.000   1st Qu.:3.000
##  Median :3069.3    Median :3.00   Median :3.000   Median :3.000
##  Mean   :3378.6    Mean   :3.49   Mean   :3.502   Mean   :3.492
##  3rd Qu.:5580.2    3rd Qu.:4.00   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :7159.0    Max.   :7.00   Max.   :7.000   Max.   :8.000
##
##   Reliability       Options        Respectful      Courteous     Listening
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.0   Min.   :1.000
##  1st Qu.:3.000   1st Qu.:3.000   1st Qu.:3.000   1st Qu.:3.0   1st Qu.:3.000
##  Median :3.000   Median :3.000   Median :3.000   Median :3.0   Median :3.000
##  Mean   :3.492   Mean   :3.492   Mean   :3.498   Mean   :3.5   Mean   :3.479
##  3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.0   3rd Qu.:4.000
##  Max.   :7.000   Max.   :7.000   Max.   :8.000   Max.   :7.0   Max.   :8.000
##
```

```
summary(dfTest)
```

```
##      State        Population             Area         Children
##  TX     : 184   Min.   :     0.0   Rural   :1006   Min.   : 0.000
##  PA     : 179   1st Qu.:   728.8   Suburban:1011   1st Qu.: 0.000
##  NY     : 154   Median :  2721.0   Urban   : 983   Median : 1.000
##  CA     : 146   Mean   :  9603.2                   Mean   : 2.094
##  IL     : 126   3rd Qu.: 12834.0                   3rd Qu.: 3.000
##  FL     :  99   Max.   :102433.0                   Max.   :10.000
##  (Other):2112
##       Age            Income                  Marital          Gender
##  Min.   :18.00   Min.   :   348.7   Divorced     :592   Female   :1505
##  1st Qu.:35.00   1st Qu.: 19738.3   Married      :605   Male     :1428
##  Median :52.00   Median : 33623.6   Never Married:609   Nonbinary:  67
##  Mean   :52.85   Mean   : 40586.5   Separated    :608
##  3rd Qu.:71.00   3rd Qu.: 54718.6   Widowed      :586
##  Max.   :89.00   Max.   :256998.4
##
```

```
##  Churn      Outage_sec_perweek     Email          Contacts
##  No :2205   Min.   : 0.1201    Min.   : 1.00   Min.   :0.000
##  Yes: 795   1st Qu.: 7.9531    1st Qu.:10.00   1st Qu.:0.000
##             Median : 9.9921    Median :12.00   Median :1.000
##             Mean   : 9.9391    Mean   :12.04   Mean   :0.986
##             3rd Qu.:11.9067    3rd Qu.:14.00   3rd Qu.:2.000
##             Max.   :19.6571    Max.   :21.00   Max.   :6.000
##
##  Yearly_equip_failure Techie              Contract    Port_modem Tablet
##  Min.   :0.000        No :2488   Month-to-month:1624   No :1531   No :2139
##  1st Qu.:0.000        Yes: 512   One year      : 614   Yes:1469   Yes: 861
##  Median :0.000                   Two Year      : 762
##  Mean   :0.405
##  3rd Qu.:1.000
##  Max.   :6.000
##
##     InternetService Phone      Multiple   OnlineSecurity OnlineBackup
##  DSL        :1072   No : 286   No :1596   No :1859       No :1649
##  Fiber Optic:1326   Yes:2714   Yes:1404   Yes:1141       Yes:1351
##  None       : 602
##
##
##
##
##  DeviceProtection TechSupport StreamingTV StreamingMovies PaperlessBilling
##  No :1660         No :1881    No :1497    No :1547        No :1216
##  Yes:1340         Yes:1119    Yes:1503    Yes:1453        Yes:1784
##
##
##
##
##
##                    PaymentMethod      Tenure       MonthlyCharge
##  Bank Transfer(automatic): 646   Min.   : 1.005   Min.   : 79.98
##  Credit Card (automatic) : 648   1st Qu.: 8.110   1st Qu.:140.00
##  Electronic Check        :1042   Median :38.710   Median :167.48
##  Mailed Check            : 664   Mean   :34.806   Mean   :173.25
##                                  3rd Qu.:61.420   3rd Qu.:200.14
##                                  Max.   :71.994   Max.   :290.16
##
##  Bandwidth_GB_Year    Response          Fix         Replacement
##  Min.   : 169.4    Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:1274.1    1st Qu.:3.000   1st Qu.:3.000   1st Qu.:3.000
##  Median :3597.9    Median :3.000   Median :4.000   Median :3.000
##  Mean   :3424.5    Mean   :3.493   Mean   :3.512   Mean   :3.476
##  3rd Qu.:5598.1    3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :7084.8    Max.   :7.000   Max.   :7.000   Max.   :7.000
##
##   Reliability      Options        Respectful      Courteous       Listening
##  Min.   :1.00   Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:3.00   1st Qu.:3.000   1st Qu.:3.000   1st Qu.:3.000   1st Qu.:3.000
##  Median :4.00   Median :4.000   Median :3.000   Median :4.000   Median :4.000
##  Mean   :3.51   Mean   :3.495   Mean   :3.495   Mean   :3.531   Mean   :3.534
##  3rd Qu.:4.00   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.000
```

```
##  Max.   :7.00   Max.   :7.000   Max.   :7.000   Max.   :7.000   Max.   :7.000
##
```

```
# Export prepared data sets
write.csv(dfTrain, "C:\\WGU\\D209 Data Mining I\\PA Task 2\\D209_dfTrain.csv", row.names = FALSE)
write.csv(dfTest, "C:\\WGU\\D209 Data Mining I\\PA Task 2\\D209_dfTest.csv", row.names = FALSE)

# Random Forest
rf <- randomForest(Churn~., data = dfTrain, proximity=TRUE)
rf
```

```
##
## Call:
##  randomForest(formula = Churn ~ ., data = dfTrain, proximity = TRUE)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 6
##
##         OOB estimate of  error rate: 10.9%
## Confusion matrix:
##       No  Yes class.error
## No  4766  379  0.07366375
## Yes  384 1471  0.20700809
```

```
pred <- predict(rf, dfTest)
confusionMatrix(pred, dfTest$Churn)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   No  Yes
##        No  2047  192
##        Yes  158  603
##
##                Accuracy : 0.8833
##                  95% CI : (0.8713, 0.8946)
##     No Information Rate : 0.735
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.6964
##
##  Mcnemar's Test P-Value : 0.07774
##
##             Sensitivity : 0.9283
##             Specificity : 0.7585
##          Pos Pred Value : 0.9142
##          Neg Pred Value : 0.7924
##              Prevalence : 0.7350
##          Detection Rate : 0.6823
##    Detection Prevalence : 0.7463
##       Balanced Accuracy : 0.8434
##
##        'Positive' Class : No
##
```

```
str(pred)
```

```
##  Factor w/ 2 levels "No","Yes": 1 1 1 2 1 2 2 2 1 1 ...
##  - attr(*, "names")= chr [1:3000] "1" "5" "9" "12" ...
```

```
churnPred<- ifelse(as.character(pred) == "Yes", 1, 0)
churnActual <- ifelse(as.character(dfTest$Churn) == "Yes", 1, 0)

mse <- mean((churnActual - churnPred)^2)
```