Practice

Oscar Monroy

1/28/2021

```
library(readr)
```

```
## Warning: package 'readr' was built under R version 3.6.3
```

```
data <- read_csv("myMTdata.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   Administrative = col_double(),
##   Administrative_Duration = col_double(),
##   Informational = col_double(),
##   Informational_Duration = col_double(),
##   ProductRelated = col_double(),
##   ProductRelated_Duration = col_double(),
##   BounceRates = col_double(),
##   ExitRates = col_double(),
##   PageValues = col_double(),
##   SpecialDay = col_double(),
##   Month = col_character(),
##   OperatingSystems = col_double(),
##   Browser = col_double(),
##   Region = col_double(),
##   TrafficType = col_double(),
##   VisitorType = col_character(),
##   Weekend = col_logical(),
##   Revenue = col_logical()
## )
```

```
apply(apply(data, 2, is.na), 2, sum) # Check for any missing value
```

```
##          Administrative Administrative_Duration           Informational
##                       0                       0                       0
##  Informational_Duration          ProductRelated ProductRelated_Duration
##                       0                       0                       0
##             BounceRates               ExitRates              PageValues
##                       0                       0                       0
##             SpecialDay                   Month        OperatingSystems
##                       0                       0                       0
```

```
##                Browser                 Region            TrafficType
##                      0                      0                      0
##            VisitorType                Weekend                Revenue
##                      0                      0                      0
```

```
# Looks like we're free of any missing values
summary(data) # Now we check the summary to see if anything appears strange
```

```
##  Administrative   Administrative_Duration Informational
##  Min.   : 0.000   Min.   :   0.00         Min.   : 0.0000
##  1st Qu.: 0.000   1st Qu.:   0.00         1st Qu.: 0.0000
##  Median : 1.000   Median :   5.00         Median : 0.0000
##  Mean   : 2.315   Mean   :  83.68         Mean   : 0.4274
##  3rd Qu.: 4.000   3rd Qu.:  96.00         3rd Qu.: 0.0000
##  Max.   :23.000   Max.   :2720.50         Max.   :10.0000
##  Informational_Duration ProductRelated   ProductRelated_Duration
##  Min.   :   0.00        Min.   :  0.00   Min.   :    0.0
##  1st Qu.:   0.00        1st Qu.:  7.00   1st Qu.:  169.8
##  Median :   0.00        Median : 17.00   Median :  576.0
##  Mean   :  28.73        Mean   : 29.63   Mean   : 1120.8
##  3rd Qu.:   0.00        3rd Qu.: 35.00   3rd Qu.: 1380.2
##  Max.   :1652.00        Max.   :501.00   Max.   :23888.8
##   BounceRates         ExitRates         PageValues        SpecialDay
##  Min.   :0.000000   Min.   :0.00000   Min.   :  0.000   Min.   :0.00000
##  1st Qu.:0.000000   1st Qu.:0.01432   1st Qu.:  0.000   1st Qu.:0.00000
##  Median :0.003333   Median :0.02636   Median :  0.000   Median :0.00000
##  Mean   :0.024559   Mean   :0.04562   Mean   :  5.682   Mean   :0.05574
##  3rd Qu.:0.020000   3rd Qu.:0.05017   3rd Qu.:  0.000   3rd Qu.:0.00000
##  Max.   :0.200000   Max.   :0.20000   Max.   :361.764   Max.   :1.00000
##     Month           OperatingSystems    Browser          Region
##  Length:1184        Min.   :1.000    Min.   : 1.000   Min.   :1.000
##  Class :character   1st Qu.:2.000    1st Qu.: 2.000   1st Qu.:1.000
##  Mode  :character   Median :2.000    Median : 2.000   Median :3.000
##                     Mean   :2.102    Mean   : 2.282   Mean   :3.182
##                     3rd Qu.:3.000    3rd Qu.: 2.000   3rd Qu.:4.000
##                     Max.   :8.000    Max.   :13.000   Max.   :9.000
##   TrafficType     VisitorType         Weekend         Revenue
##  Min.   : 1.000   Length:1184        Mode :logical   Mode :logical
##  1st Qu.: 2.000   Class :character   FALSE:918       FALSE:1012
##  Median : 2.000   Mode  :character   TRUE :266       TRUE :172
##  Mean   : 4.061
##  3rd Qu.: 4.000
##  Max.   :20.000
```

```
# Given the context of this data set, there's nothing that's really farfectched here.
```

1) Please access your data from the link which follows this document. The descriptions for each field are given in data_description.txt (follows the download link). Please describe the dataset to us:

a) how many non-numeric fields are present

We'll check which of these fields are numeric or not now:

```r
c_data <- unlist(lapply(data, class)) # Grabs class type of all fields
table(c_data)
```

```
## c_data
## character   logical   numeric
##         2         2        14
```

We have a total of 4 non-numeric fields available. However, since the variables OperatingSystems, Browser, Region, and TrafficType could be considered more as factors as they're mainly used for classification, then we would have 8 non-numeric fields.

b) how many numeric fields are present, clearly identify which are discrete and which should be treated as continuous?

As seen above, we see that there are 14 numeric fields, but now to determine which are discrete. To do this, we could just count up all the non-duplicated values in each of the numeric fields:

```r
cn_data <- which(c_data == "numeric")
sort(apply(apply(data[, cn_data], 2, function(x) !(duplicated(x))), 2, sum))
```

```
##              SpecialDay        OperatingSystems                  Region
##                       6                       7                       9
##           Informational                 Browser             TrafficType
##                      11                      12                      18
##          Administrative          ProductRelated  Informational_Duration
##                      22                     140                     184
##              PageValues            BounceRates Administrative_Duration
##                     266                     332                     500
##              ExitRates ProductRelated_Duration
##                     661                    1050
```

For the purpose of time, we'll use any value above 22 as our threshold for what constitutes as continuous and everything less than or equal to 22 is discrete.

c) how many observations were in your dataset

We can simply check the number of dimensions in the data to determine how many observations there are:

```r
dim(data)[1]
```

```
## [1] 1184
```

Looks like we have 1184 observations.

a) Duration was measured in 3 different ways, please construct a total duration variable and provide an appropriate statistical summary for your duration variable (you can define a "statistical summary" for us)

Before creating the TotalDuration variable, I will define a statistical summary as the standard deviation, mean, median, min, and max.

```
td <- data[[2]] + data[[4]] + data[[6]]
data.frame("Mean" = mean(td), "SD" = sd(td), "Median" = median(td), "Min" = min(td), "Max" = max(td))
```

```
##       Mean       SD   Median Min      Max
## 1 1233.168 1923.643 640.1333   0 25390.01
```

b) The field "Browser" has numerous values, but two of the values dominate. Please recode/reconstruct "Browser" in such a way that there are only three possible values – the two dominant values and all other. Then, tell us whether there is evidence that duration differs by the value of your new "Browser" variable.

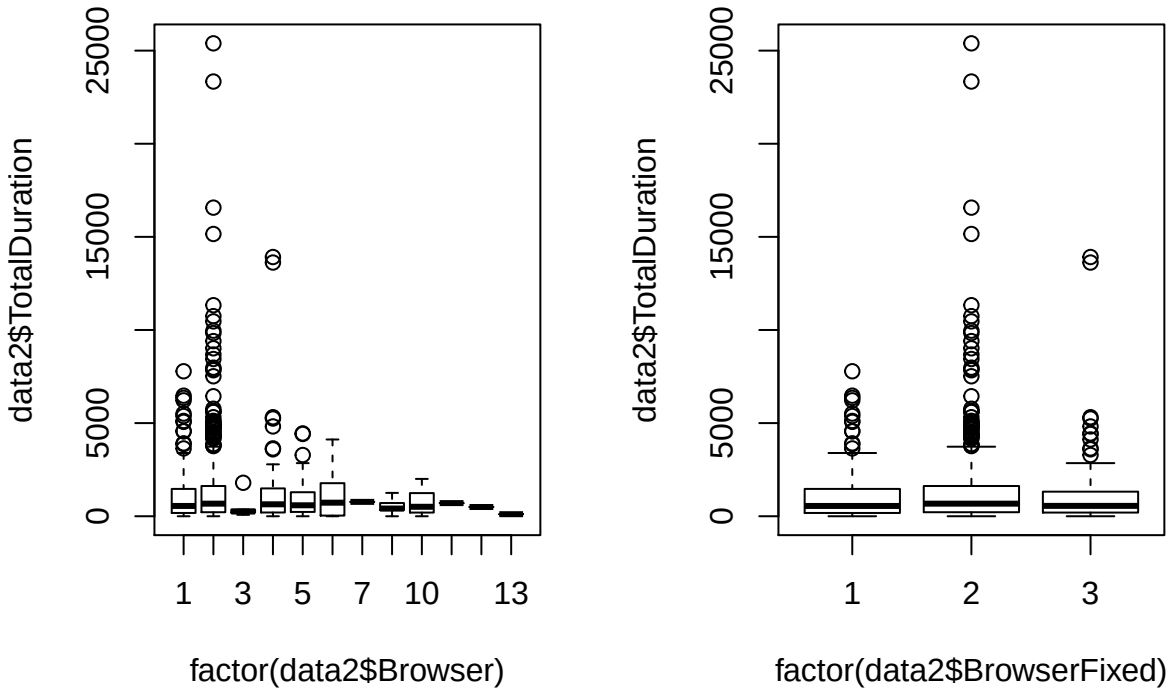
```
table(data$Browser)
```

```
##
##   1   2   3   4   5   6   7   8  10  11  12  13
## 248 759   5  81  44  18   1  12  13   1   1   1
```

```
# We can see 1 and 2 are the most dominant values here, so we'll all the other values into one.
# We'll use 3 as the blanket value.
gt2 <- which(data$Browser > 2)
bf <- data$Browser
bf[gt2] <- 3
table(bf)
```

```
## bf
##   1   2   3
## 248 759 177
```

```
data2 <- cbind(data, "TotalDuration" = td, "BrowserFixed" = bf)
par(mfrow = c(1, 2))
plot(data2$TotalDuration ~ factor(data2$Browser))
plot(data2$TotalDuration ~ factor(data2$BrowserFixed))
```



Judging by these two graphs, it looks like there's some evidence that the duration differs by the new blanket value in Browser.

c) Which Month/VisitorType (omit VisitorType = Other) combination has the highest proportion of Revenue = TRUE?

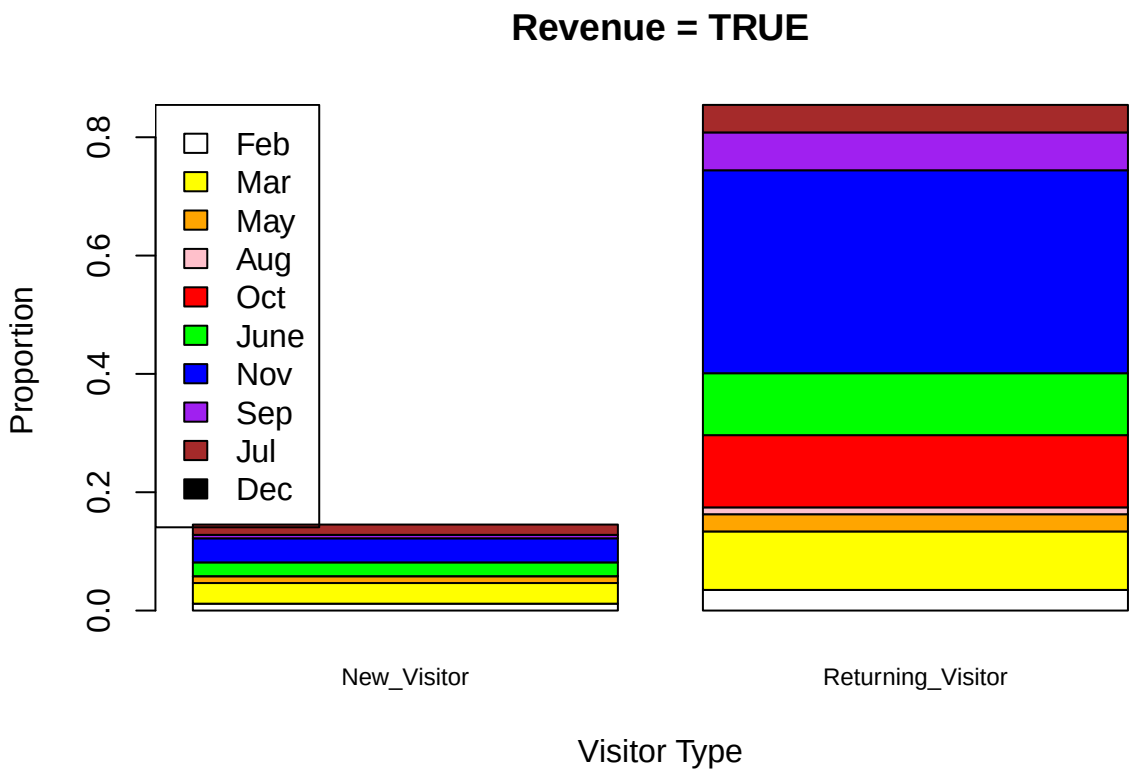
```
data_no <- data2[-which(data2$VisitorType == "Other"), ]
rev_true <- which(data2$Revenue == TRUE)
table(data_no$Month[rev_true], data_no$VisitorType[rev_true]) / 172
```

```
##
##        New_Visitor Returning_Visitor
##   Aug  0.011627907       0.034883721
##   Dec  0.034883721       0.098837209
##   Jul  0.011627907       0.029069767
##   June 0.000000000       0.011627907
##   Mar  0.000000000       0.122093023
##   May  0.023255814       0.104651163
##   Nov  0.040697674       0.343023256
##   Oct  0.005813953       0.063953488
##   Sep  0.017441860       0.046511628
```

Here we can see that combination of returning visitors in November has the highest proportion of Revenue = TRUE.

2) The client is most interested in having a better understanding of the process of revenue (Revenue) generation as it relates to the information collected for this study. Please construct a visualization (graphic) which will help your client. Your graphic should use at minimum the revenue variable and also one other variable (more than one other is acceptable and encouraged). Also, please record yourself on video explaining the graphic as if you were sharing a Zoom screen with the client. Assume the client has a college degree and above average intelligence, but has no prior training in statistics beyond an introductory course completed more than 10 years ago. Your video should be no longer than 2 minutes and should be smaller than 100MB (please see the links on video compression for assistance)

```
t1 <- table(data_no$Month[rev_true], data_no$VisitorType[rev_true])
t1_p <- prop.table(t1)
barplot(t1_p, main = "Revenue = TRUE",
        col = c("white", "yellow", "orange", "pink", "red", "green", "blue", "purple", "brown", "black"),
        ylab = "Proportion", xlab = "Visitor Type", cex.names = 0.75)
legend("topleft", legend = unique(data_no$Month),
       fill = c("white", "yellow", "orange", "pink", "red", "green", "blue", "purple", "brown", "black"))
```



3)

```
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 3.6.3
```

```
## randomForest 4.6-14

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

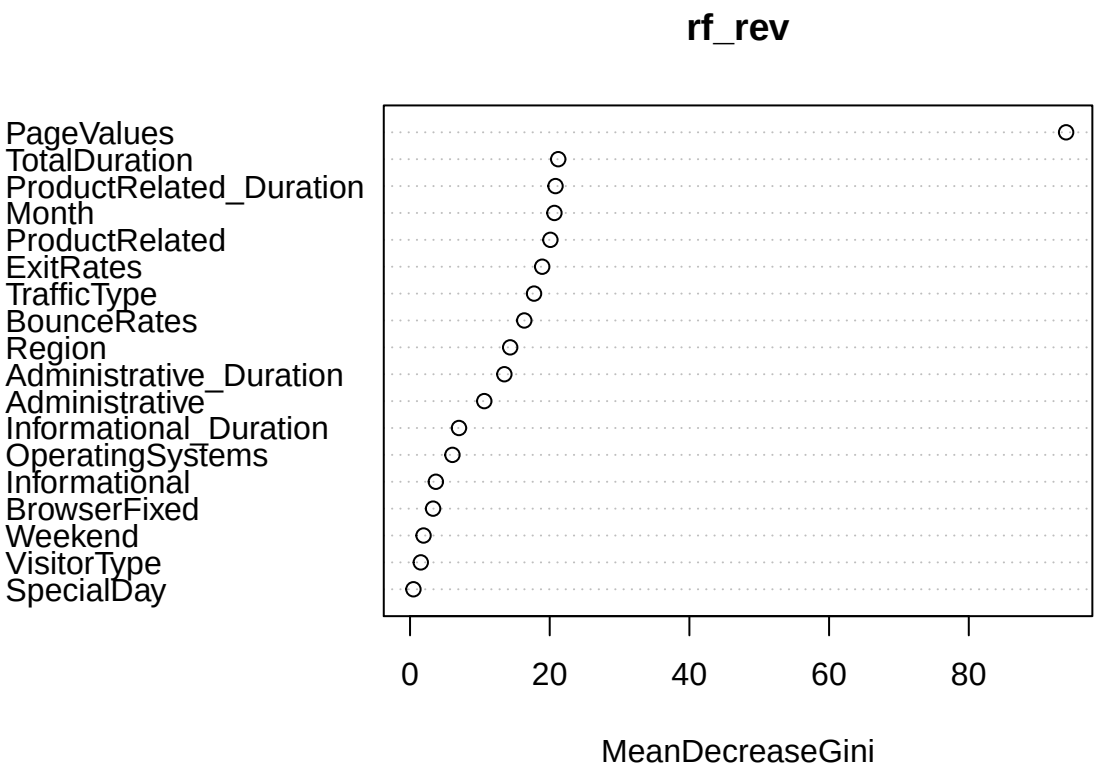
```
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.6.3

## Loading required package: lattice

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 3.6.3

##
## Attaching package: 'ggplot2'

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

```
data_no2 <- data_no[, -13]
data_no2$BrowserFixed <- factor(data_no2$BrowserFixed)
data_no2$OperatingSystems <- factor(data_no2$OperatingSystems)
data_no2$Region <- factor(data_no2$Region)
data_no2$TrafficType <- factor(data_no2$TrafficType)
data_no2$Revenue <- factor(data_no2$Revenue)
data_no2$Weekend <- factor(data_no2$Weekend)
data_no2$VisitorType <- factor(data_no2$VisitorType)
data_no2$Month <- factor(data_no2$Month)

rf_rev <- tuneRF(x = data_no2[, -17], y = data_no2$Revenue, doBest = T, plot = F)
```

```
## mtry = 4  OOB error = 9.83%
## Searching left ...
## mtry = 2     OOB error = 10.93%
## -0.112069 0.05
## Searching right ...
## mtry = 8     OOB error = 10.68%
## -0.0862069 0.05
```

```
varImpPlot(rf_rev)
```




```
varImp(rf_rev)
```

```
##                            Overall
## Administrative          10.6306147
## Administrative_Duration 13.5045291
## Informational            3.6979211
## Informational_Duration   7.0073943
## ProductRelated          20.0974170
## ProductRelated_Duration 20.8298109
## BounceRates             16.3570070
## ExitRates               18.9239292
## PageValues              93.9370009
## SpecialDay               0.4849217
## Month                   20.6715816
## OperatingSystems         6.0731032
## Region                  14.3418790
## TrafficType             17.7563407
## VisitorType              1.5415627
## Weekend                  1.9341435
## TotalDuration           21.2191155
## BrowserFixed             3.3009601
```

Here I used a RandomFOrest model as its simple and gets a fairly accurate model without doing much work. While it was a good choice for me to use, the model could still use some improvements as I had no time to properly train the model, although it should still get the job done. WIth this RandomForest model, we can

see that the most important predictor in determining whether or not a user generates revenue is PageValues, followed by ProductRelated pages, the Month, ExitRates, and also TotalDuration. These are the factors that should be looked at more closely when wanting to find the proper combination of factors to generate revenue in a webpage. On the otherhand, SpecialDay, VisitorType, and Weekend appear to be the least important factors. This model should give decent success if predicting a new set of users and whether or not they generate revenue.