MSDS 660 Week 6 Assignment

Jeremy Beard

2022-08-10

Introduction This week's assignment will focus on building different types of regression models. We will be working with a "churn" dataset which contains data describing customer churn within a corporation. It was supplied by the instructor of this MSDS 660 course. It is a mostly clean dataset with few null values. Train and Test datasets will be created from this dataset.

After creating the model, we will work to improve the model through VIF analysis for collinearity, analyzing correlation coefficients, and running StepAIC analyses.

Finally, after the model has been refined, predictions will be made from the train and test data created before, using the refined regression model.

First, we'll load the libraries necessary, set the seed, and load in the data:

```
# load libraries
library(data.table)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:data.table':
##
##     between, first, last

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(car)
```

```
## Loading required package: carData

##
## Attaching package: 'car'

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

```
library(caret)
```

```
## Loading required package: ggplot2

## Loading required package: lattice
```

```r
library(caTools)
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```r
library(MASS)
```

```
##
## Attaching package: 'MASS'

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

```r
library(ggcorrplot)

# set the seed
set.seed(1)

#load data as datatable
dt <- read.csv("C:\\Users\\jerem\\OneDrive\\Documents\\School\\_REGIS\\2022-05_Summer\\MSDS660\\Week6\\
dt <- as.data.table(dt)
```

Next, we'll compute some summaries, remove the ID column as it doesn't provide useful statistical information, and display the unique entries in each column, for later factoring:

```r
str(dt)
```

```
## Classes 'data.table' and 'data.frame':   7043 obs. of  21 variables:
##  $ customerID      : chr  "7590-VHVEG" "5575-GNVDE" "3668-QPYBK" "7795-CFOCW" ...
##  $ gender          : chr  "Female" "Male" "Male" "Male" ...
##  $ SeniorCitizen   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Partner         : chr  "Yes" "No" "No" "No" ...
##  $ Dependents      : chr  "No" "No" "No" "No" ...
##  $ tenure          : int  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" ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```r
summary(dt)
```

```
##   customerID           gender          SeniorCitizen      Partner         
##  Length:7043        Length:7043        Min.   :0.0000   Length:7043       
##  Class :character   Class :character   1st Qu.:0.0000   Class :character  
##  Mode  :character   Mode  :character   Median :0.0000   Mode  :character  
##                                        Mean   :0.1621                     
##                                        3rd Qu.:0.0000                     
##                                        Max.   :1.0000                     
##                                                                           
##   Dependents            tenure      PhoneService       MultipleLines     
##  Length:7043        Min.   : 0.00   Length:7043        Length:7043       
##  Class :character   1st Qu.: 9.00   Class :character   Class :character  
##  Mode  :character   Median :29.00   Mode  :character   Mode  :character  
##                     Mean   :32.37                                        
##                     3rd Qu.:55.00                                        
##                     Max.   :72.00                                        
##                                                                          
##  InternetService    OnlineSecurity     OnlineBackup       DeviceProtection  
##  Length:7043        Length:7043        Length:7043        Length:7043       
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##                                                                             
##  TechSupport        StreamingTV        StreamingMovies      Contract        
##  Length:7043        Length:7043        Length:7043        Length:7043       
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##                                                                             
##  PaperlessBilling   PaymentMethod      MonthlyCharges    TotalCharges   
##  Length:7043        Length:7043        Min.   : 18.25   Min.   :  18.8  
##  Class :character   Class :character   1st Qu.: 35.50   1st Qu.: 401.4  
##  Mode  :character   Mode  :character   Median : 70.35   Median :1397.5  
##                                        Mean   : 64.76   Mean   :2283.3  
##                                        3rd Qu.: 89.85   3rd Qu.:3794.7  
##                                        Max.   :118.75   Max.   :8684.8  
##                                                         NA's   :11      
##     Churn          
##  Length:7043       
##  Class :character  
##  Mode  :character  
##                    
##                    
##                    
##                    
```

It looks like there are a LOT of categorical and nominal data. I should create factors from these columns so they are more easily able to be worked with.

```r
dt <- dt[, !"customerID"]
str(dt)
```

```
## Classes 'data.table' and 'data.frame':   7043 obs. of  20 variables:
##  $ gender          : chr  "Female" "Male" "Male" "Male" ...
##  $ SeniorCitizen   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Partner         : chr  "Yes" "No" "No" "No" ...
##  $ Dependents      : chr  "No" "No" "No" "No" ...
##  $ tenure          : int  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" ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```r
summary(dt)
```

```
##     gender          SeniorCitizen      Partner           Dependents       
##  Length:7043        Min.   :0.0000   Length:7043        Length:7043       
##  Class :character   1st Qu.:0.0000   Class :character   Class :character  
##  Mode  :character   Median :0.0000   Mode  :character   Mode  :character  
##                     Mean   :0.1621                                        
##                     3rd Qu.:0.0000                                        
##                     Max.   :1.0000                                        
##                                                                           
##      tenure      PhoneService       MultipleLines      InternetService   
##  Min.   : 0.00   Length:7043        Length:7043        Length:7043       
##  1st Qu.: 9.00   Class :character   Class :character   Class :character  
##  Median :29.00   Mode  :character   Mode  :character   Mode  :character  
##  Mean   :32.37                                                           
##  3rd Qu.:55.00                                                           
##  Max.   :72.00                                                           
##                                                                          
##  OnlineSecurity     OnlineBackup       DeviceProtection   TechSupport       
##  Length:7043        Length:7043        Length:7043        Length:7043       
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##                                                                             
##  StreamingTV        StreamingMovies      Contract         PaperlessBilling  
##  Length:7043        Length:7043        Length:7043        Length:7043       
```

```
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  PaymentMethod      MonthlyCharges    TotalCharges       Churn
##  Length:7043        Min.   : 18.25   Min.   :  18.8   Length:7043
##  Class :character   1st Qu.: 35.50   1st Qu.: 401.4   Class :character
##  Mode  :character   Median : 70.35   Median :1397.5   Mode  :character
##                     Mean   : 64.76   Mean   :2283.3
##                     3rd Qu.: 89.85   3rd Qu.:3794.7
##                     Max.   :118.75   Max.   :8684.8
##                                      NA's   :11
```

```r
unique(dt$gender)
```

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

```r
unique(dt$SeniorCitizen)
```

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

```r
unique(dt$Partner)
```

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

```r
unique(dt$Dependents)
```

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

```r
#unique(dt$tenure)
unique(dt$PhoneService)
```

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

```r
unique(dt$MultipleLines)
```

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

```r
unique(dt$InternetService)
```

```
## [1] "DSL"         "Fiber optic" "No"
```

```r
unique(dt$OnlineSecurity)
```

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

```r
unique(dt$OnlineBackup)
```

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

```r
unique(dt$DeviceProtection)
```

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

```r
unique(dt$TechSupport)
```

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

```r
unique(dt$StreamingTV)
```

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

```r
unique(dt$StreamingMovies)
```

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

```r
unique(dt$Contract)
```

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

```r
unique(dt$PaperlessBilling)
```

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

```r
unique(dt$PaymentMethod)
```

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

```r
#unique(dt$MonthlyCharges)
#unique(dt$TotalCharges)
unique(dt$Churn)
```

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

Now we're ready to change all 'char' columns to be factors, based on the unique entries from each column:

```r
# run tests

#Factor class and relable as benign or malignant
dt$Churn <- factor(dt$Churn, labels = c('No', 'Yes'))
dt$gender <- factor(dt$gender, labels = c('Male', 'Female'))
dt$Partner <- factor(dt$Partner, labels = c('No', 'Yes'))
dt$Dependents <- factor(dt$Dependents, labels = c('No', 'Yes'))
dt$PhoneService <- factor(dt$PhoneService, labels = c('No', 'Yes'))
head(dt$MultipleLines)
```

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

```r
dt$MultipleLines <- factor(dt$MultipleLines, labels = c('No', 'Yes', 'No phone service'))
dt$InternetService <- factor(dt$InternetService, labels = c('No', 'DSL', 'Fiber optic'))
dt$OnlineSecurity <- factor(dt$OnlineSecurity, labels = c('No', 'Yes', 'No internet service'))
dt$OnlineBackup <- factor(dt$OnlineBackup, labels = c('No', 'Yes', 'No internet service'))
dt$DeviceProtection <- factor(dt$DeviceProtection, labels = c('No', 'Yes', 'No internet service'))
dt$TechSupport <- factor(dt$TechSupport, labels = c('No', 'Yes', 'No internet service'))
dt$StreamingTV <- factor(dt$StreamingTV, labels = c('No', 'Yes', 'No internet service'))
dt$StreamingMovies <- factor(dt$StreamingMovies, labels = c('No', 'Yes', 'No internet service'))
dt$Contract <- factor(dt$Contract, labels = c('Month-to-month', 'One year', 'Two year'))
dt$PaperlessBilling <- factor(dt$PaperlessBilling, labels = c('No', 'Yes'))
dt$PaymentMethod <- factor(dt$PaymentMethod, labels = c('Electronic check', 'Mailed check', 'Bank trans
```

Next we'll find and remove any and all null values:

```r
str(dt)
```

```
## Classes 'data.table' and 'data.frame':   7043 obs. of  20 variables:
##  $ gender          : Factor w/ 2 levels "Male","Female": 1 2 2 2 1 1 2 1 1 2 ...
##  $ SeniorCitizen   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Partner         : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 1 1 2 1 ...
##  $ Dependents      : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 2 1 1 2 ...
```

```
## $ tenure          : int  1 34 2 45 2 8 22 10 28 62 ...
## $ PhoneService    : Factor w/ 2 levels "No","Yes": 1 2 2 1 2 2 2 1 2 2 ...
## $ MultipleLines   : Factor w/ 3 levels "No","Yes","No phone service": 2 1 1 2 1 3 3 2 3 1 ...
## $ InternetService : Factor w/ 3 levels "No","DSL","Fiber optic": 1 1 1 1 2 2 2 1 2 1 ...
## $ OnlineSecurity  : Factor w/ 3 levels "No","Yes","No internet service": 1 3 3 3 1 1 1 3 1 3 ...
## $ OnlineBackup    : Factor w/ 3 levels "No","Yes","No internet service": 3 1 3 1 1 1 3 1 1 3 ...
## $ DeviceProtection: Factor w/ 3 levels "No","Yes","No internet service": 1 3 1 3 1 3 1 1 3 1 ...
## $ TechSupport     : Factor w/ 3 levels "No","Yes","No internet service": 1 1 1 3 1 1 1 1 3 1 ...
## $ StreamingTV     : Factor w/ 3 levels "No","Yes","No internet service": 1 1 1 1 1 3 3 1 3 1 ...
## $ StreamingMovies : Factor w/ 3 levels "No","Yes","No internet service": 1 1 1 1 1 3 1 1 3 1 ...
## $ Contract        : Factor w/ 3 levels "Month-to-month",..: 1 2 1 2 1 1 1 1 1 2 ...
## $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 1 2 1 2 2 2 1 2 1 ...
## $ PaymentMethod   : Factor w/ 4 levels "Electronic check",..: 3 4 4 1 3 3 2 4 3 1 ...
## $ MonthlyCharges  : num  29.9 57 53.9 42.3 70.7 ...
## $ TotalCharges    : num  29.9 1889.5 108.2 1840.8 151.7 ...
## $ Churn           : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1 1 2 1 ...
## - attr(*, ".internal.selfref")=<externalptr>
```

```
summary(dt)
```

```
##    gender     SeniorCitizen    Partner    Dependents     tenure
##  Male  :3488   Min.   :0.0000   No :3641   No :4933   Min.   : 0.00
##  Female:3555   1st Qu.:0.0000   Yes:3402   Yes:2110   1st Qu.: 9.00
##                Median :0.0000                         Median :29.00
##                Mean   :0.1621                         Mean   :32.37
##                3rd Qu.:0.0000                         3rd Qu.:55.00
##                Max.   :1.0000                         Max.   :72.00
##
##  PhoneService          MultipleLines     InternetService
##  No : 682     No              :3390   No         :2421
##  Yes:6361     Yes             : 682   DSL        :3096
##               No phone service:2971   Fiber optic:1526
##
##
##
##
##             OnlineSecurity             OnlineBackup
##  No                 :3498   No                 :3088
##  Yes                :1526   Yes                :1526
##  No internet service:2019   No internet service:2429
##
##
##
##
##            DeviceProtection              TechSupport
##  No                 :3095    No                 :3473
##  Yes                :1526    Yes                :1526
##  No internet service:2422    No internet service:2044
##
##
##
##
##               StreamingTV              StreamingMovies           Contract
##  No                 :2810   No                 :2785   Month-to-month:3875
##  Yes                :1526   Yes                :1526   One year      :1473
```

```
##  No internet service:2707   No internet service:2732   Two year      :1695
##
##
##
##
##  PaperlessBilling                   PaymentMethod  MonthlyCharges
##  No :2872         Electronic check          :1544   Min.   : 18.25
##  Yes:4171         Mailed check              :1522   1st Qu.: 35.50
##                   Bank transfer (automatic):2365   Median : 70.35
##                   Credit card (automatic)  :1612   Mean   : 64.76
##                                                    3rd Qu.: 89.85
##                                                    Max.   :118.75
##
##   TotalCharges    Churn
##  Min.   :  18.8   No :5174
##  1st Qu.: 401.4   Yes:1869
##  Median :1397.5
##  Mean   :2283.3
##  3rd Qu.:3794.7
##  Max.   :8684.8
##  NA's   :11
```

```
#remove NAs
#first, we'll just list how many NA's are present
which(is.na(dt$gender))
```

```
## integer(0)
```

```
which(is.na(dt$SeniorCitizen))
```

```
## integer(0)
```

```
which(is.na(dt$Partner))
```

```
## integer(0)
```

```
which(is.na(dt$Dependents))
```

```
## integer(0)
```

```
which(is.na(dt$tenure))
```

```
## integer(0)
```

```
which(is.na(dt$PhoneService))
```

```
## integer(0)
```

```
which(is.na(dt$MultipleLines))
```

```
## integer(0)
```

```
which(is.na(dt$InternetService))
```

```
## integer(0)
```

```
which(is.na(dt$OnlineSecurity))
```

```
## integer(0)
```

```r
which(is.na(dt$OnlineBackup))
```

```
## integer(0)
```

```r
which(is.na(dt$DeviceProtection))
```

```
## integer(0)
```

```r
which(is.na(dt$TechSupport))
```

```
## integer(0)
```

```r
which(is.na(dt$StreamingTV))
```

```
## integer(0)
```

```r
which(is.na(dt$StreamingMovies))
```

```
## integer(0)
```

```r
which(is.na(dt$Contract))
```

```
## integer(0)
```

```r
which(is.na(dt$PaperlessBilling))
```

```
## integer(0)
```

```r
which(is.na(dt$PaymentMethod))
```

```
## integer(0)
```

```r
which(is.na(dt$MonthlyCharges))
```

```
## integer(0)
```

```r
which(is.na(dt$TotalCharges))
```

```
##  [1]  489  754  937 1083 1341 3332 3827 4381 5219 6671 6755
```

```r
which(is.na(dt$Churn))
```

```
## integer(0)
```

```r
#now we'll remove the NA's and check to make sure they're gone
dt <- dt[complete.cases(dt), ]
which(is.na(dt$customerID))
```

```
## integer(0)
```

```r
which(is.na(dt$gender))
```

```
## integer(0)
```

```r
which(is.na(dt$SeniorCitizen))
```

```
## integer(0)
```

```r
which(is.na(dt$Partner))
```

```
## integer(0)
```

```r
which(is.na(dt$Dependents))
```

```
## integer(0)
```

```r
which(is.na(dt$tenure))
```

```
## integer(0)
```

```r
which(is.na(dt$PhoneService))
```

```
## integer(0)
```

```r
which(is.na(dt$MultipleLines))
```

```
## integer(0)
```

```r
which(is.na(dt$InternetService))
```

```
## integer(0)
```

```r
which(is.na(dt$OnlineSecurity))
```

```
## integer(0)
```

```r
which(is.na(dt$OnlineBackup))
```

```
## integer(0)
```

```r
which(is.na(dt$DeviceProtection))
```

```
## integer(0)
```

```r
which(is.na(dt$TechSupport))
```

```
## integer(0)
```

```r
which(is.na(dt$StreamingTV))
```

```
## integer(0)
```

```r
which(is.na(dt$StreamingMovies))
```

```
## integer(0)
```

```r
which(is.na(dt$Contract))
```

```
## integer(0)
```

```r
which(is.na(dt$PaperlessBilling))
```

```
## integer(0)
```

```r
which(is.na(dt$PaymentMethod))
```

```
## integer(0)
```

```r
which(is.na(dt$MonthlyCharges))
```

```
## integer(0)
```

```r
which(is.na(dt$TotalCharges))
```

```
## integer(0)
```

```r
which(is.na(dt$Churn))
```

```
## integer(0)
```

```r
#NA's have been removed from the dataset!
```

Methods Now we'll split the data up into train and test data and create a multi linear binomial logistic regression mode. We will then try to improve the model to a level which can be called significant. We will check for collinearity, run StepAIC analyses, plot correlation plots. Our null hypothesis is that there is no relationship between any of the data and the Churn column. The alternate hypothesis is that there is indeed a significant correlation between the data and the Churn column. The methods used in the assignment will hope to disprove the null hypothesis and find a correlation in the data. Let's begin:

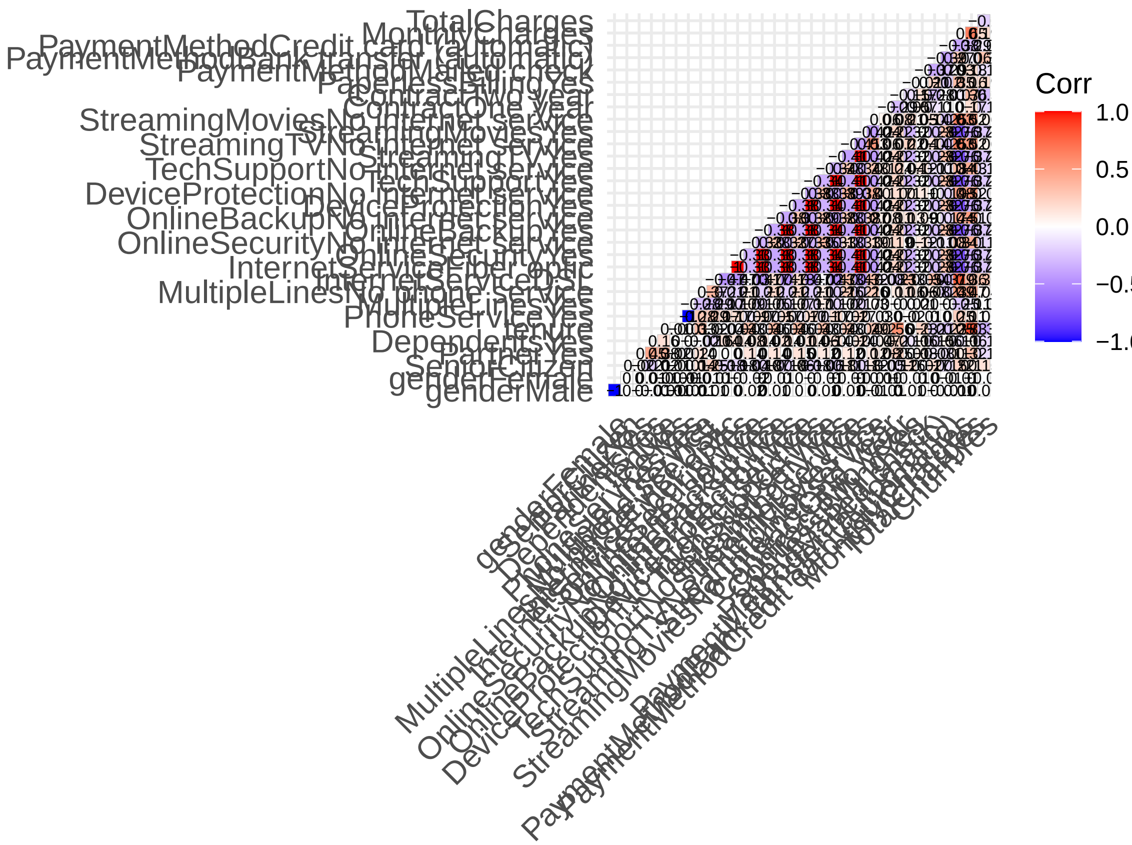
```
#Now time to split the data into a train and test set

#split the data into a train and test set
samp <- sample.split(dt$Churn, SplitRatio = 0.8)
train <- subset(dt, samp == TRUE)
test <- subset(dt, samp == FALSE)

# Create a multi linear binomial logisitc regression
model <- glm(Churn ~ ., data = train, family = "binomial")

# Look at the model summary
summary(model)
```

```
##
## Call:
## glm(formula = Churn ~ ., family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8401  -0.6860  -0.2862   0.7494   3.4760
##
## Coefficients: (7 not defined because of singularities)
##                                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)                           9.985e-01  9.063e-01   1.102  0.27058
## genderFemale                         -4.499e-02  7.240e-02  -0.621  0.53438
## SeniorCitizen                         1.631e-01  9.440e-02   1.728  0.08396
## PartnerYes                            1.053e-01  8.762e-02   1.202  0.22926
## DependentsYes                        -1.442e-01  9.972e-02  -1.446  0.14824
## tenure                               -6.406e-02  7.011e-03  -9.137  < 2e-16
## PhoneServiceYes                      -1.805e-01  7.233e-01  -0.250  0.80292
## MultipleLinesYes                             NA         NA      NA       NA
## MultipleLinesNo phone service         3.852e-01  1.970e-01   1.955  0.05058
## InternetServiceDSL                    1.445e+00  8.857e-01   1.631  0.10286
## InternetServiceFiber optic           -1.421e+00  8.981e-01  -1.582  0.11369
## OnlineSecurityYes                            NA         NA      NA       NA
## OnlineSecurityNo internet service    -2.615e-01  1.979e-01  -1.322  0.18632
## OnlineBackupYes                              NA         NA      NA       NA
## OnlineBackupNo internet service       2.705e-02  1.953e-01   0.139  0.88981
## DeviceProtectionYes                          NA         NA      NA       NA
## DeviceProtectionNo internet service   1.444e-01  1.971e-01   0.732  0.46396
## TechSupportYes                               NA         NA      NA       NA
## TechSupportNo internet service       -2.749e-01  2.013e-01  -1.366  0.17201
## StreamingTVYes                               NA         NA      NA       NA
## StreamingTVNo internet service        4.327e-01  3.640e-01   1.189  0.23450
## StreamingMoviesYes                           NA         NA      NA       NA
## StreamingMoviesNo internet service    4.195e-01  3.637e-01   1.154  0.24869
## ContractOne year                     -7.535e-01  1.212e-01  -6.217 5.05e-10
## ContractTwo year                     -1.391e+00  1.961e-01  -7.095 1.29e-12
```

```
## PaperlessBillingYes                     3.302e-01  8.316e-02   3.971 7.17e-05
## PaymentMethodMailed check              -1.275e-01  1.282e-01  -0.995  0.31972
## PaymentMethodBank transfer (automatic)  2.728e-01  1.055e-01   2.585  0.00974
## PaymentMethodCredit card (automatic)   -7.687e-02  1.277e-01  -0.602  0.54714
## MonthlyCharges                         -2.806e-02  3.533e-02  -0.794  0.42711
## TotalCharges                            3.633e-04  7.937e-05   4.577 4.72e-06
##
## (Intercept)
## genderFemale
## SeniorCitizen                          .
## PartnerYes
## DependentsYes
## tenure                                 ***
## PhoneServiceYes
## MultipleLinesYes
## MultipleLinesNo phone service          .
## InternetServiceDSL
## InternetServiceFiber optic
## OnlineSecurityYes
## OnlineSecurityNo internet service
## OnlineBackupYes
## OnlineBackupNo internet service
## DeviceProtectionYes
## DeviceProtectionNo internet service
## TechSupportYes
## TechSupportNo internet service
## StreamingTVYes
## StreamingTVNo internet service
## StreamingMoviesYes
## StreamingMoviesNo internet service
## ContractOne year                       ***
## ContractTwo year                       ***
## PaperlessBillingYes                    ***
## PaymentMethodMailed check
## PaymentMethodBank transfer (automatic) **
## PaymentMethodCredit card (automatic)
## MonthlyCharges
## TotalCharges                           ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6513.9  on 5624  degrees of freedom
## Residual deviance: 4667.6  on 5601  degrees of freedom
## AIC: 4715.6
##
## Number of Fisher Scoring iterations: 6
```

```
# Check for colinearity
###########vif(model)
#hmmmm, getting the error: "Error in vif.default(model) : there are aliased coefficients in the model"
#it means 2+ variables are very closely related
#let's plot a correlation matrix to see which ones
```

```r
model.matrix(~0+., data=dt) %>%
  cor(use="pairwise.complete.obs") %>%
  ggcorrplot(show.diag = F, type="lower", lab=TRUE, lab_size=2)
```




```r
# I will remove features which have a correlation of 1.
# this is InternetServiceFiber optic -- OnlineSecurity Yes
#   InternetServiceFiber optic -- Online Backup Yes
#   InternetServiceFiber optic -- DeviceProtection Yees
#   InternetServiceFiber optic -- TechSupport Yes
#   InternetServiceFiber optic -- StreamingTV Yes
#   InternetServiceFiber optic -- StreamingMovies Yes

# So looks like I just need to keep one of these 7 features. I will keep InternetService

dt <- dt[, !"OnlineSecurity"]
dt <- dt[, !"OnlineBackup"]
dt <- dt[, !"DeviceProtection"]
dt <- dt[, !"TechSupport"]
dt <- dt[, !"StreamingTV"]
dt <- dt[, !"StreamingMovies"]

samp <- sample.split(dt$Churn, SplitRatio = 0.8)
train <- subset(dt, samp == TRUE)
test <- subset(dt, samp == FALSE)

# Create a multi linear binomial logisitc regression
```

```r
model <- glm(Churn ~ ., data = train, family = "binomial")

# Look at the model summary
summary(model)
```

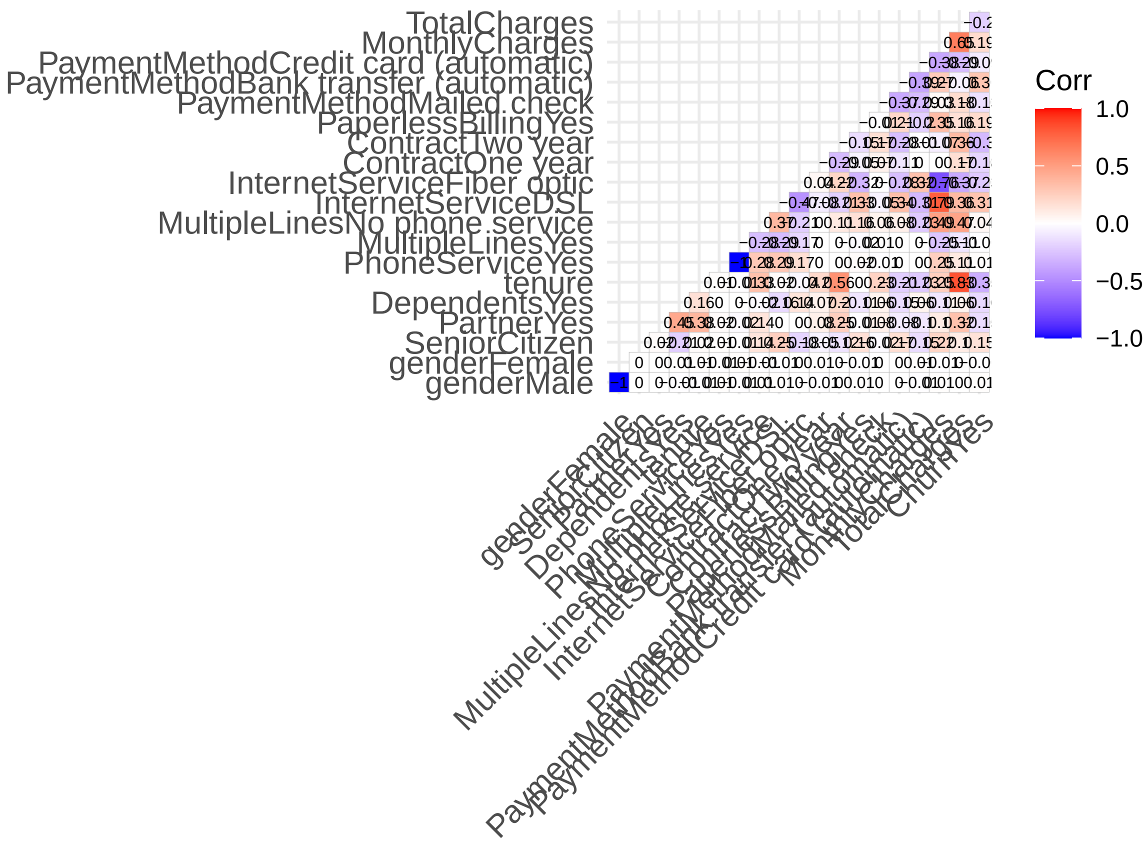
```
##
## Call:
## glm(formula = Churn ~ ., family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7990  -0.6794  -0.2891   0.7589   3.5007
##
## Coefficients: (1 not defined because of singularities)
##                                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)                          -0.0099665  0.2525107  -0.039 0.968516
## genderFemale                         -0.0289637  0.0719761  -0.402 0.687385
## SeniorCitizen                         0.1977683  0.0939381   2.105 0.035265
## PartnerYes                           -0.0007583  0.0863913  -0.009 0.992997
## DependentsYes                        -0.1781361  0.1001908  -1.778 0.075409
## tenure                               -0.0622956  0.0070358  -8.854  < 2e-16
## PhoneServiceYes                      -0.8210581  0.1622100  -5.062 4.16e-07
## MultipleLinesYes                             NA         NA      NA       NA
## MultipleLinesNo phone service         0.3047067  0.0908327   3.355 0.000795
## InternetServiceDSL                    0.8734110  0.1515980   5.761 8.34e-09
## InternetServiceFiber optic           -0.4980240  0.2097727  -2.374 0.017591
## ContractOne year                     -0.7925936  0.1207220  -6.565 5.19e-11
## ContractTwo year                     -1.6539365  0.2025975  -8.164 3.25e-16
## PaperlessBillingYes                   0.4119122  0.0823706   5.001 5.71e-07
## PaymentMethodMailed check            -0.0398969  0.1272724  -0.313 0.753919
## PaymentMethodBank transfer (automatic)  0.4534164  0.1042357   4.350 1.36e-05
## PaymentMethodCredit card (automatic)    0.0042422  0.1285780   0.033 0.973680
## MonthlyCharges                        0.0003440  0.0046201   0.074 0.940654
## TotalCharges                          0.0003417  0.0000798   4.283 1.85e-05
##
## (Intercept)
## genderFemale
## SeniorCitizen                        *
## PartnerYes
## DependentsYes                        .
## tenure                               ***
## PhoneServiceYes                      ***
## MultipleLinesYes
## MultipleLinesNo phone service        ***
## InternetServiceDSL                   ***
## InternetServiceFiber optic           *
## ContractOne year                     ***
## ContractTwo year                     ***
## PaperlessBillingYes                  ***
## PaymentMethodMailed check
## PaymentMethodBank transfer (automatic) ***
## PaymentMethodCredit card (automatic)
## MonthlyCharges
## TotalCharges                         ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6513.9  on 5624  degrees of freedom
## Residual deviance: 4698.2  on 5607  degrees of freedom
## AIC: 4734.2
##
## Number of Fisher Scoring iterations: 6
```

```r
# Check for colinearity
############vif(model)

#Hmm, still getting the error

str(dt)
```

```
## Classes 'data.table' and 'data.frame':   7032 obs. of  14 variables:
##  $ gender          : Factor w/ 2 levels "Male","Female": 1 2 2 2 1 1 2 1 1 2 ...
##  $ SeniorCitizen   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Partner         : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 1 1 2 1 ...
##  $ Dependents      : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 2 1 1 2 ...
##  $ tenure          : int  1 34 2 45 2 8 22 10 28 62 ...
##  $ PhoneService    : Factor w/ 2 levels "No","Yes": 1 2 2 1 2 2 2 1 2 2 ...
##  $ MultipleLines   : Factor w/ 3 levels "No","Yes","No phone service": 2 1 1 2 1 3 3 2 3 1 ...
##  $ InternetService : Factor w/ 3 levels "No","DSL","Fiber optic": 1 1 1 1 2 2 2 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 "Electronic check",..: 3 4 4 1 3 3 2 4 3 1 ...
##  $ MonthlyCharges  : num  29.9 57 53.9 42.3 70.7 ...
##  $ TotalCharges    : num  29.9 1889.5 108.2 1840.8 151.7 ...
##  $ Churn           : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 2 1 1 2 1 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```r
model.matrix(~0+., data=dt) %>%
  cor(use="pairwise.complete.obs") %>%
  ggcorrplot(show.diag = F, type="lower", lab=TRUE, lab_size=2)
```



```r
# Looks like I should remove the -1 correlations too
# This is:
# PhoneService Yes -- MultipleLines Yes
# genderMale -- genderFemale

# Try #3
dt <- dt[, !"gender"]
dt <- dt[, !"MultipleLines"]

samp <- sample.split(dt$Churn, SplitRatio = 0.8)
train <- subset(dt, samp == TRUE)
test <- subset(dt, samp == FALSE)

# Create a multi linear binomial logisitc regression
model <- glm(Churn ~ ., data = train, family = "binomial")

# Look at the model summary
summary(model)
```

```
##
## Call:
## glm(formula = Churn ~ ., family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7446  -0.6740  -0.2996   0.7723   3.4532
```

```
##
## Coefficients:
##                                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)                          -0.2016147  0.2454320  -0.821 0.411379
## SeniorCitizen                         0.2670696  0.0931126   2.868 0.004128
## PartnerYes                            0.0099583  0.0863885   0.115 0.908229
## DependentsYes                        -0.1713647  0.0994842  -1.723 0.084973
## tenure                               -0.0579179  0.0069562  -8.326  < 2e-16
## PhoneServiceYes                      -0.6866813  0.1613268  -4.256 2.08e-05
## InternetServiceDSL                    0.8555871  0.1489703   5.743 9.28e-09
## InternetServiceFiber optic           -0.5125597  0.2102015  -2.438 0.014752
## ContractOne year                     -0.7483763  0.1192483  -6.276 3.48e-10
## ContractTwo year                     -1.4432775  0.1940119  -7.439 1.01e-13
## PaperlessBillingYes                   0.3883977  0.0827071   4.696 2.65e-06
## PaymentMethodMailed check            -0.0775943  0.1268089  -0.612 0.540604
## PaymentMethodBank transfer (automatic)  0.3579683  0.1049645   3.410 0.000649
## PaymentMethodCredit card (automatic)   -0.0771601  0.1270454  -0.607 0.543623
## MonthlyCharges                        0.0042909  0.0044991   0.954 0.340230
## TotalCharges                          0.0002690  0.0000788   3.414 0.000641
##
## (Intercept)
## SeniorCitizen                        **
## PartnerYes
## DependentsYes                        .
## tenure                               ***
## PhoneServiceYes                      ***
## InternetServiceDSL                   ***
## InternetServiceFiber optic           *
## ContractOne year                     ***
## ContractTwo year                     ***
## PaperlessBillingYes                  ***
## PaymentMethodMailed check
## PaymentMethodBank transfer (automatic) ***
## PaymentMethodCredit card (automatic)
## MonthlyCharges
## TotalCharges                         ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6513.9  on 5624  degrees of freedom
## Residual deviance: 4713.7  on 5609  degrees of freedom
## AIC: 4745.7
##
## Number of Fisher Scoring iterations: 6
```

```r
# Check for colinearity
vif(model)
```

```
##                        GVIF Df GVIF^(1/(2*Df))
## SeniorCitizen      1.126337  1        1.061290
## Partner            1.371010  1        1.170901
## Dependents         1.287565  1        1.134709
## tenure            15.694136  1        3.961582
```

```
## PhoneService      1.828328  1        1.352157
## InternetService   8.728208  2        1.718823
## Contract          1.546760  2        1.115208
## PaperlessBilling  1.121476  1        1.058998
## PaymentMethod     1.348379  3        1.051079
## MonthlyCharges   11.147864  1        3.338842
## TotalCharges     20.212642  1        4.495847
```

```r
# Whew okay, looks good


# Looks like tenure, InternetService, MonthlyCharges, and TotalCharges have a GVIF over 5.
# Let's see the correlation matrix again
model.matrix(~0+., data=dt) %>%
  cor(use="pairwise.complete.obs") %>%
  ggcorrplot(show.diag = F, type="lower", lab=TRUE, lab_size=2)
```


```r
#Yeah, the 3 variables above have correlations above |0.75|.
# AND PartnerNo has a correlation of 1 to PartnerYes. I will remove them all
dt <- dt[, !"Partner"]
dt <- dt[, !"InternetService"]
dt <- dt[, !"MonthlyCharges"]
dt <- dt[, !"TotalCharges"]

samp <- sample.split(dt$Churn, SplitRatio = 0.8)
train <- subset(dt, samp == TRUE)
test <- subset(dt, samp == FALSE)
```

```r
# Create a multi linear binomial logisitc regression
model <- glm(Churn ~ ., data = train, family = "binomial")

# Look at the model summary
summary(model)
```

```
##
## Call:
## glm(formula = Churn ~ ., family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6799  -0.7525  -0.3235   0.8043   3.0110
##
## Coefficients:
##                                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)                           -0.70996    0.15685  -4.526 6.00e-06 ***
## SeniorCitizen                          0.41245    0.09002   4.582 4.61e-06 ***
## DependentsYes                         -0.44014    0.08949  -4.918 8.73e-07 ***
## tenure                                -0.02379    0.00214 -11.115  < 2e-16 ***
## PhoneServiceYes                        0.16881    0.11870   1.422  0.15499
## ContractOne year                      -0.92132    0.11172  -8.247  < 2e-16 ***
## ContractTwo year                      -1.87572    0.18644 -10.061  < 2e-16 ***
## PaperlessBillingYes                    0.70597    0.07790   9.063  < 2e-16 ***
## PaymentMethodMailed check             -0.13891    0.12389  -1.121  0.26219
## PaymentMethodBank transfer (automatic)  0.57791    0.10121   5.710 1.13e-08 ***
## PaymentMethodCredit card (automatic)  -0.34468    0.11969  -2.880  0.00398 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6513.9  on 5624  degrees of freedom
## Residual deviance: 4982.9  on 5614  degrees of freedom
## AIC: 5004.9
##
## Number of Fisher Scoring iterations: 6
```

```r
# Check for colinearity
vif(model)
```

```
##                      GVIF Df GVIF^(1/(2*Df))
## SeniorCitizen    1.099099  1        1.048379
## Dependents       1.049989  1        1.024690
## tenure           1.530814  1        1.237261
## PhoneService     1.000901  1        1.000451
## Contract         1.425926  2        1.092759
## PaperlessBilling 1.070223  1        1.034516
## PaymentMethod    1.183126  3        1.028423
```

```r
# Whew okay, looks even better. No GVIF above 5 (should I be using GVIF^(1/(2*Df))???)

# Perform stepAIC to remove high p-values
stepAIC(model, direction = 'both')
```

```
## Start:  AIC=5004.93
## Churn ~ SeniorCitizen + Dependents + tenure + PhoneService +
##     Contract + PaperlessBilling + PaymentMethod
##
##                    Df Deviance    AIC
## <none>                  4982.9 5004.9
## - PhoneService      1   4985.0 5005.0
## - SeniorCitizen     1   5003.9 5023.9
## - Dependents        1   5007.8 5027.8
## - PaperlessBilling  1   5067.5 5087.5
## - PaymentMethod     3   5101.2 5117.2
## - tenure            1   5112.5 5132.5
## - Contract          2   5139.8 5157.8

##
## Call:  glm(formula = Churn ~ SeniorCitizen + Dependents + tenure + PhoneService +
##     Contract + PaperlessBilling + PaymentMethod, family = "binomial",
##     data = train)
##
## Coefficients:
##                           (Intercept)                           SeniorCitizen
##                              -0.70995                                 0.41245
##                         DependentsYes                                  tenure
##                              -0.44014                                -0.02379
##                       PhoneServiceYes                         ContractOne year
##                               0.16881                                -0.92132
##                      ContractTwo year                     PaperlessBillingYes
##                              -1.87572                                 0.70597
##            PaymentMethodMailed check  PaymentMethodBank transfer (automatic)
##                              -0.13891                                 0.57791
##   PaymentMethodCredit card (automatic)
##                              -0.34468
##
## Degrees of Freedom: 5624 Total (i.e. Null);  5614 Residual
## Null Deviance:       6514
## Residual Deviance: 4983  AIC: 5005
```

```
model <- glm(formula = Churn ~ SeniorCitizen + Dependents + tenure + PhoneService + Contract + Paperless

#Check model summary again
summary(model)
```

```
##
## Call:
## glm(formula = Churn ~ SeniorCitizen + Dependents + tenure + PhoneService +
##     Contract + PaperlessBilling + PaymentMethod, family = "binomial",
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6799  -0.7525  -0.3235   0.8043   3.0110
##
## Coefficients:
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                              -0.70996    0.15685  -4.526 6.00e-06 ***
```

```
## SeniorCitizen                           0.41245    0.09002   4.582 4.61e-06 ***
## DependentsYes                          -0.44014    0.08949  -4.918 8.73e-07 ***
## tenure                                 -0.02379    0.00214 -11.115  < 2e-16 ***
## PhoneServiceYes                         0.16881    0.11870   1.422  0.15499    
## ContractOne year                       -0.92132    0.11172  -8.247  < 2e-16 ***
## ContractTwo year                       -1.87572    0.18644 -10.061  < 2e-16 ***
## PaperlessBillingYes                     0.70597    0.07790   9.063  < 2e-16 ***
## PaymentMethodMailed check              -0.13891    0.12389  -1.121  0.26219    
## PaymentMethodBank transfer (automatic)  0.57791    0.10121   5.710 1.13e-08 ***
## PaymentMethodCredit card (automatic)   -0.34468    0.11969  -2.880  0.00398 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 6513.9  on 5624  degrees of freedom
## Residual deviance: 4982.9  on 5614  degrees of freedom
## AIC: 5004.9
## 
## Number of Fisher Scoring iterations: 6
```

Results The results of these tests turned out to be pretty straightforward. I wanted to split some of the code up between the Methods section and the Results section so some of what is mentioned below may be contained in the Methods section, but the results of the tests showed a similar result between the train and the test data, both of which had high accuracy scores of ~78%. This shows that the model had a good fit. All final confusion matrices had Mcnemar's Test P-Value of under <0.05 which shows that the null hypothesis could be rejected and there actually was a significant correlation between the data and the Churn information. This is to be expected as there should be some relationship between the data of an employee and if they will "churn" or not, it cannot be completely random.

```r
# predict on the train data
trainpreds <- predict(model, type = 'response', train)

# Round prediction values at 0.5 cutoff factor and change lables
trainp <- factor(trainpreds >= 0.5, labels = c('No', 'Yes'))

# Build a confusion matrix and see results
trainCM <- confusionMatrix(train$Churn, trainp)
trainCM
```

```
## Confusion Matrix and Statistics
## 
##           Reference
## Prediction   No  Yes
##        No  3735  395
##        Yes  829  666
##                                           
##                Accuracy : 0.7824          
##                  95% CI : (0.7714, 0.7931)
##     No Information Rate : 0.8114          
##     P-Value [Acc > NIR] : 1               
##                                           
##                   Kappa : 0.3855          
##                                           
##  Mcnemar's Test P-Value : <2e-16          
```

```
##
##             Sensitivity : 0.8184
##             Specificity : 0.6277
##          Pos Pred Value : 0.9044
##          Neg Pred Value : 0.4455
##              Prevalence : 0.8114
##          Detection Rate : 0.6640
##    Detection Prevalence : 0.7342
##       Balanced Accuracy : 0.7230
##
##        'Positive' Class : No
##
```

```
# predict on the test data
testpreds <- predict(model, type = 'response', test)

# Round prediction values at 0.5 cutoff factor and change labels
testp <- factor(testpreds >= 0.5, labels = c('No', 'Yes'))

# Build a confusion matrix and see results
testCM <- confusionMatrix(test$Churn, testp)
testCM
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  938  95
##        Yes 218 156
##
##                Accuracy : 0.7775
##                  95% CI : (0.7549, 0.799)
##     No Information Rate : 0.8216
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.3633
##
##  Mcnemar's Test P-Value : 5.354e-12
##
##             Sensitivity : 0.8114
##             Specificity : 0.6215
##          Pos Pred Value : 0.9080
##          Neg Pred Value : 0.4171
##              Prevalence : 0.8216
##          Detection Rate : 0.6667
##    Detection Prevalence : 0.7342
##       Balanced Accuracy : 0.7165
##
##        'Positive' Class : No
##
```

```
# Create a Roc curve and and view ROC results for the Train data
train_roc_curve <- roc(train$Churn, trainpreds)
```

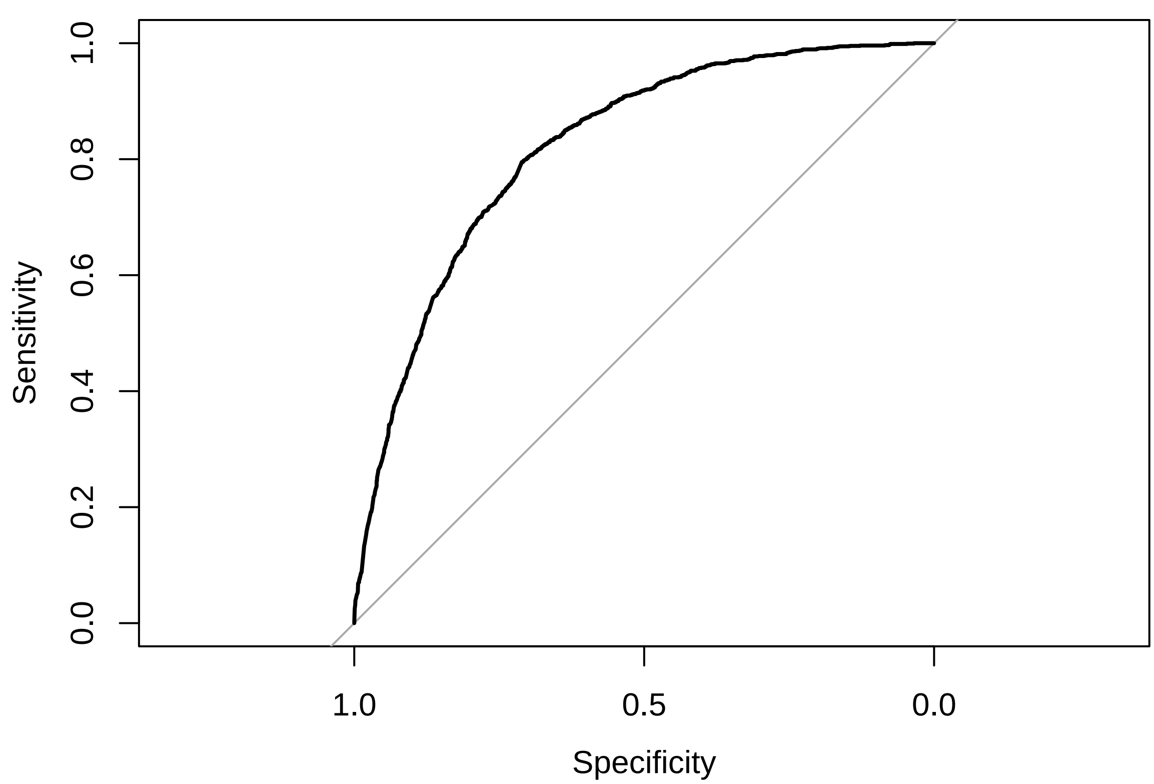
```
## Setting levels: control = No, case = Yes
```

```
## Setting direction: controls < cases
```

```r
train_roc_curve
```

```
##
## Call:
## roc.default(response = train$Churn, predictor = trainpreds)
##
## Data: trainpreds in 4130 controls (train$Churn No) < 1495 cases (train$Churn Yes).
## Area under the curve: 0.8209
```

```r
plot(train_roc_curve)
```




```r
train_rocc <- coords(roc=train_roc_curve, x = 'best', best.method = 'closest.topleft')
train_rocc
```

```
##   threshold specificity sensitivity
## 1 0.2869388   0.7113801   0.7946488
```

```r
# Create a Roc curve and view results for the Test data
test_roc_curve <- roc(test$Churn, testpreds)
```

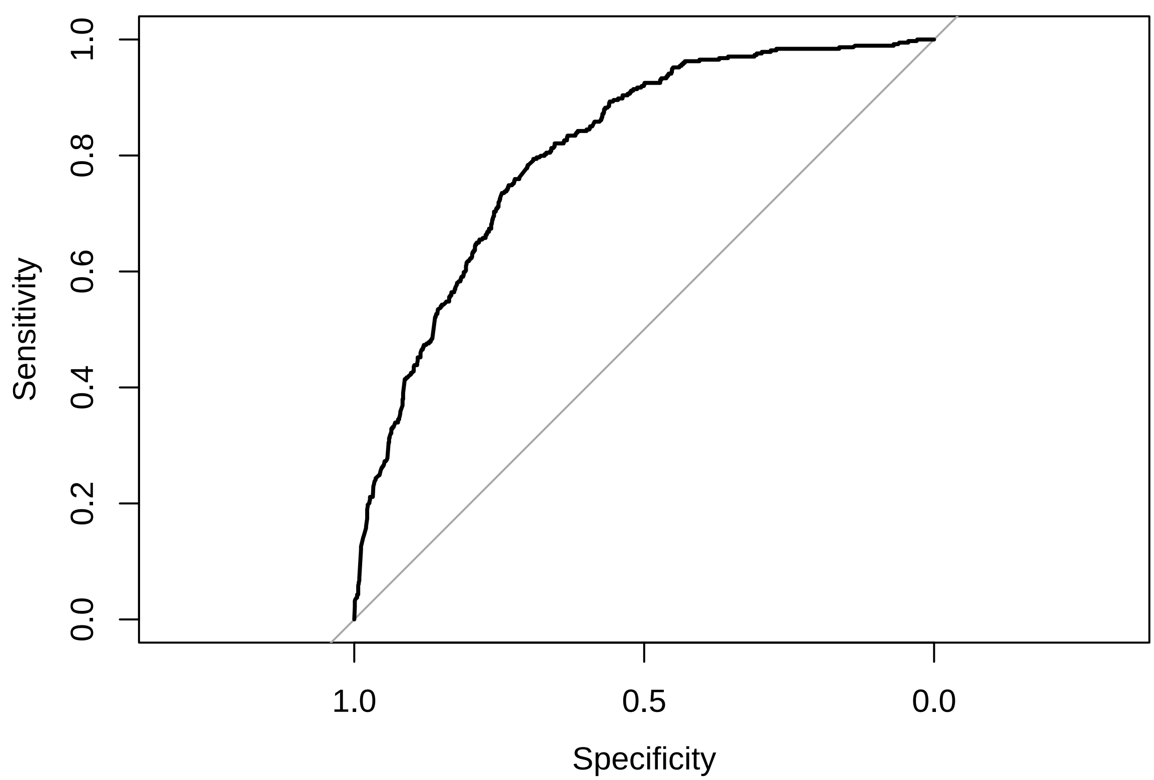
```
## Setting levels: control = No, case = Yes
## Setting direction: controls < cases
```

```r
test_roc_curve
```

```
##
## Call:
## roc.default(response = test$Churn, predictor = testpreds)
```

```
##
## Data: testpreds in 1033 controls (test$Churn No) < 374 cases (test$Churn Yes).
## Area under the curve: 0.8083
```

```r
plot(test_roc_curve)
```




```r
test_rocc <- coords(roc=test_roc_curve, x = 'best', best.method = 'closest.topleft')
test_rocc
```

```
##   threshold specificity sensitivity
## 1 0.3073134   0.7337851   0.7486631
```

```r
# predict on the train data using the ROC cutoff
# Round prediction values at threshold level and change labels
trainrocp <- factor(trainpreds >= as.numeric(train_rocc[1]), labels = c('No', 'Yes'))

# Build a confusion matrix to see results
trainROCCM <- confusionMatrix(train$Churn, trainrocp)
trainROCCM
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   No  Yes
##        No  2938 1192
##        Yes  307 1188
##
##                Accuracy : 0.7335
```

```
##                    95% CI : (0.7217, 0.745)
##       No Information Rate : 0.5769
##       P-Value [Acc > NIR] : < 2.2e-16
##
##                     Kappa : 0.4256
##
##    Mcnemar's Test P-Value : < 2.2e-16
##
##               Sensitivity : 0.9054
##               Specificity : 0.4992
##            Pos Pred Value : 0.7114
##            Neg Pred Value : 0.7946
##                Prevalence : 0.5769
##            Detection Rate : 0.5223
##      Detection Prevalence : 0.7342
##         Balanced Accuracy : 0.7023
##
##          'Positive' Class : No
##
```

```
# predict on the test data using the ROC cutoff
#  Round prediction values at threshold level and change labels
testp <- factor(testpreds >= as.numeric(test_rocc[1]), labels = c('No', 'Yes'))

# Buld a confusion matrix to see results
testROCCM <- confusionMatrix(test$Churn, testp)
testROCCM
```

```
## Confusion Matrix and Statistics
##
##             Reference
## Prediction   No Yes
##        No   758 275
##        Yes   94 280
##
##                  Accuracy : 0.7377
##                    95% CI : (0.7139, 0.7606)
##       No Information Rate : 0.6055
##       P-Value [Acc > NIR] : < 2.2e-16
##
##                     Kappa : 0.4179
##
##    Mcnemar's Test P-Value : < 2.2e-16
##
##               Sensitivity : 0.8897
##               Specificity : 0.5045
##            Pos Pred Value : 0.7338
##            Neg Pred Value : 0.7487
##                Prevalence : 0.6055
##            Detection Rate : 0.5387
##      Detection Prevalence : 0.7342
##         Balanced Accuracy : 0.6971
##
##          'Positive' Class : No
##
```

```r
#View all the Confusion matrices
trainCM
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   No  Yes
##        No  3735  395
##        Yes  829  666
##
##                Accuracy : 0.7824
##                  95% CI : (0.7714, 0.7931)
##     No Information Rate : 0.8114
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.3855
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.8184
##             Specificity : 0.6277
##          Pos Pred Value : 0.9044
##          Neg Pred Value : 0.4455
##              Prevalence : 0.8114
##          Detection Rate : 0.6640
##    Detection Prevalence : 0.7342
##       Balanced Accuracy : 0.7230
##
##        'Positive' Class : No
##
```

```r
trainROCCM
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   No  Yes
##        No  2938 1192
##        Yes  307 1188
##
##                Accuracy : 0.7335
##                  95% CI : (0.7217, 0.745)
##     No Information Rate : 0.5769
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.4256
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9054
##             Specificity : 0.4992
##          Pos Pred Value : 0.7114
##          Neg Pred Value : 0.7946
##              Prevalence : 0.5769
##          Detection Rate : 0.5223
```

```
##    Detection Prevalence : 0.7342
##       Balanced Accuracy : 0.7023
##
##        'Positive' Class : No
##
```

```
testCM
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  938  95
##        Yes 218 156
##
##                Accuracy : 0.7775
##                  95% CI : (0.7549, 0.799)
##     No Information Rate : 0.8216
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.3633
##
##  Mcnemar's Test P-Value : 5.354e-12
##
##             Sensitivity : 0.8114
##             Specificity : 0.6215
##          Pos Pred Value : 0.9080
##          Neg Pred Value : 0.4171
##              Prevalence : 0.8216
##          Detection Rate : 0.6667
##    Detection Prevalence : 0.7342
##       Balanced Accuracy : 0.7165
##
##        'Positive' Class : No
##
```

```
testROCCM
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  758 275
##        Yes  94 280
##
##                Accuracy : 0.7377
##                  95% CI : (0.7139, 0.7606)
##     No Information Rate : 0.6055
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.4179
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.8897
##             Specificity : 0.5045
```

```
##          Pos Pred Value : 0.7338
##          Neg Pred Value : 0.7487
##              Prevalence : 0.6055
##          Detection Rate : 0.5387
##    Detection Prevalence : 0.7342
##       Balanced Accuracy : 0.6971
##
##        'Positive' Class : No
##
```

Conclusion From the results of the tests, we can confidently conclude that there was a relationship between customer churn, and the significant non-collinear variables which were collected of the customer, such as tensure, PaymentMethod, and SeniorCitizen. This means that the regression model can be used to predict, albeit only ~75% of the time correctly, if a customer will "churn" or not, at the 95% confidence level. The testing was mostly straightforward and the big job with this dataset was removing the collinear features. There were many of them! In the future, one way to improve this effort is to find more datafields to collect from the customers that are not collinear. I found a lot of collinear features in this dataset and maybe that is normal behavior of an organic dataset. However, maybe it is abnormal and more significant datafields should be sought from the customers.

Thank you!

Jeremy Beard