Logistic Regression on Telecom Churn Data

Alexander Vaillant

9/7/2021

Environment Setup

Import Necessary Libraries

```
# Load in necessary libraries using library()
library(dplyr)
```

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

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

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

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.1.1
```

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

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v stringr 1.4.0
## v tidyr   1.1.3     v forcats 0.5.1
## v readr   2.0.1
```

```
## Warning: package 'tibble' was built under R version 4.1.1
```

```
## Warning: package 'tidyr' was built under R version 4.1.1
```

```
## Warning: package 'readr' was built under R version 4.1.1
```

```
## Warning: package 'purrr' was built under R version 4.1.1
```

```
## Warning: package 'forcats' was built under R version 4.1.1
```

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

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.1.1
```

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

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

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

```r
library(leaps)
```

```
## Warning: package 'leaps' was built under R version 4.1.1
```

```r
library(reshape2)
```

```
## Warning: package 'reshape2' was built under R version 4.1.1

##
## Attaching package: 'reshape2'

## The following object is masked from 'package:tidyr':
##
##     smiths
```

```r
library(fastDummies) #Create dummy columns easily
```

```
## Warning: package 'fastDummies' was built under R version 4.1.1
```

```r
library(MLmetrics) #Calculate F1_Score
```

```
## Warning: package 'MLmetrics' was built under R version 4.1.1

##
## Attaching package: 'MLmetrics'

## The following objects are masked from 'package:caret':
##
##     MAE, RMSE

## The following object is masked from 'package:base':
##
##     Recall
```

```r
library(plyr) # Rename columns
```

```
## Warning: package 'plyr' was built under R version 4.1.1

## ------------------------------------------------------------------------------

## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)

## ------------------------------------------------------------------------------

##
## Attaching package: 'plyr'

## The following object is masked from 'package:purrr':
##
##     compact

## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

Data Gathering

Load Dataset into Dataframe using read.csv()

```r
# Import the raw dataset using read.csv()
url <- "C:/Users/tedda/Desktop/Data Science Portfolio/Machine Learning/Supervised Learning/Regression/L
churndata <- read.csv(url, header = TRUE)
```

Data Preparation

```r
# Remove customer demographic data by indexing
churn_indexed <- churndata[c(20:50)]

# Transform categorical variables in dummy variable columns by using fastDummies::dummy_cols()
churn_dummies <- dummy_cols(churn_indexed, remove_first_dummy = TRUE, remove_selected_columns = TRUE)
names(churn_dummies)
```

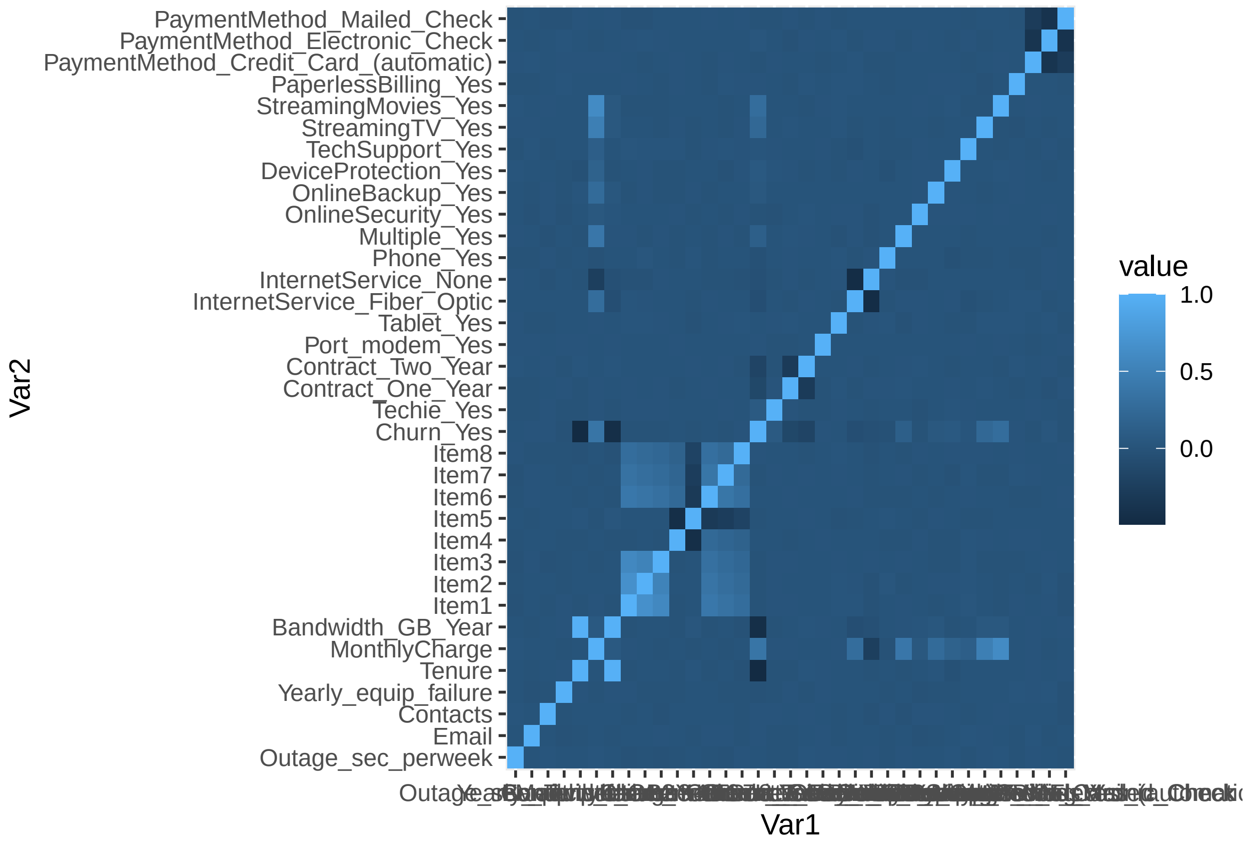
```
##  [1] "Outage_sec_perweek"
##  [2] "Email"
##  [3] "Contacts"
##  [4] "Yearly_equip_failure"
##  [5] "Tenure"
##  [6] "MonthlyCharge"
##  [7] "Bandwidth_GB_Year"
##  [8] "Item1"
##  [9] "Item2"
## [10] "Item3"
## [11] "Item4"
## [12] "Item5"
## [13] "Item6"
## [14] "Item7"
## [15] "Item8"
## [16] "Churn_Yes"
## [17] "Techie_Yes"
## [18] "Contract_One year"
## [19] "Contract_Two Year"
## [20] "Port_modem_Yes"
## [21] "Tablet_Yes"
## [22] "InternetService_Fiber Optic"
## [23] "InternetService_None"
## [24] "Phone_Yes"
## [25] "Multiple_Yes"
## [26] "OnlineSecurity_Yes"
## [27] "OnlineBackup_Yes"
## [28] "DeviceProtection_Yes"
## [29] "TechSupport_Yes"
## [30] "StreamingTV_Yes"
## [31] "StreamingMovies_Yes"
## [32] "PaperlessBilling_Yes"
## [33] "PaymentMethod_Credit Card (automatic)"
## [34] "PaymentMethod_Electronic Check"
## [35] "PaymentMethod_Mailed Check"
```

```r
# Rename any variables with spaces in their names by using plyr::rename()
churn_renamed <- rename(churn_dummies, replace = c("Contract_One year" = "Contract_One_Year"))
churn_renamed <- rename(churn_renamed, replace = c("Contract_Two Year" = "Contract_Two_Year"))
churn_renamed <- rename(churn_renamed, replace = c("InternetService_Fiber Optic" = "InternetService_Fibe
churn_renamed <- rename(churn_renamed, replace = c("PaymentMethod_Credit Card (automatic)" = "PaymentMet
churn_renamed <- rename(churn_renamed, replace = c("PaymentMethod_Electronic Check" = "PaymentMethod_Ele
churn_renamed <- rename(churn_renamed, replace = c("PaymentMethod_Mailed Check" = "PaymentMethod_Mailed_

# Normalize all variables by using caret::preProcess()
preproc <- preProcess(churn_renamed, method = c("range"))
churn_norm <- predict(preproc, churn_renamed)
summary(churn_norm)
```

```
##  Outage_sec_perweek     Email           Contacts      Yearly_equip_failure
##  Min.   :0.0000     Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.3751     1st Qu.:0.4091   1st Qu.:0.0000   1st Qu.:0.00000
##  Median :0.4699     Median :0.5000   Median :0.1429   Median :0.00000
##  Mean   :0.4691     Mean   :0.5007   Mean   :0.1420   Mean   :0.06633
##  3rd Qu.:0.5623     3rd Qu.:0.5909   3rd Qu.:0.2857   3rd Qu.:0.16667
##  Max.   :1.0000     Max.   :1.0000   Max.   :1.0000   Max.   :1.00000
##      Tenure        MonthlyCharge    Bandwidth_GB_Year     Item1
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.0000    Min.   :0.0000
##  1st Qu.:0.09743   1st Qu.:0.2855   1st Qu.:0.1543    1st Qu.:0.3333
##  Median :0.48494   Median :0.4163   Median :0.4461    Median :0.3333
##  Mean   :0.47220   Mean   :0.4408   Mean   :0.4622    Mean   :0.4151
##  3rd Qu.:0.85184   3rd Qu.:0.5745   3rd Qu.:0.7754    3rd Qu.:0.5000
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.0000    Max.   :1.0000
##      Item2            Item3            Item4            Item5
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.3333   1st Qu.:0.2857   1st Qu.:0.3333   1st Qu.:0.3333
##  Median :0.5000   Median :0.2857   Median :0.3333   Median :0.3333
##  Mean   :0.4175   Mean   :0.3553   Mean   :0.4163   Mean   :0.4155
##  3rd Qu.:0.5000   3rd Qu.:0.4286   3rd Qu.:0.5000   3rd Qu.:0.5000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##      Item6            Item7            Item8          Churn_Yes
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000
##  1st Qu.:0.2857   1st Qu.:0.3333   1st Qu.:0.2857   1st Qu.:0.000
##  Median :0.2857   Median :0.5000   Median :0.2857   Median :0.000
##  Mean   :0.3568   Mean   :0.4183   Mean   :0.3565   Mean   :0.265
##  3rd Qu.:0.4286   3rd Qu.:0.5000   3rd Qu.:0.4286   3rd Qu.:1.000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.000
##    Techie_Yes     Contract_One_Year Contract_Two_Year Port_modem_Yes
##  Min.   :0.0000   Min.   :0.0000    Min.   :0.0000    Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000    1st Qu.:0.0000    1st Qu.:0.0000
##  Median :0.0000   Median :0.0000    Median :0.0000    Median :0.0000
##  Mean   :0.1679   Mean   :0.2102    Mean   :0.2442    Mean   :0.4834
##  3rd Qu.:0.0000   3rd Qu.:0.0000    3rd Qu.:0.0000    3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000    Max.   :1.0000    Max.   :1.0000
##    Tablet_Yes     InternetService_Fiber_Optic InternetService_None
##  Min.   :0.0000   Min.   :0.0000              Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000              1st Qu.:0.0000
##  Median :0.0000   Median :0.0000              Median :0.0000
##  Mean   :0.2991   Mean   :0.4408              Mean   :0.2129
##  3rd Qu.:1.0000   3rd Qu.:1.0000              3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000              Max.   :1.0000  
##    Phone_Yes       Multiple_Yes    OnlineSecurity_Yes OnlineBackup_Yes
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000     Min.   :0.0000  
##  1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0.0000     1st Qu.:0.0000  
##  Median :1.0000   Median :0.0000   Median :0.0000     Median :0.0000  
##  Mean   :0.9067   Mean   :0.4608   Mean   :0.3576     Mean   :0.4506  
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000     3rd Qu.:1.0000  
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000     Max.   :1.0000  
##  DeviceProtection_Yes TechSupport_Yes StreamingTV_Yes  StreamingMovies_Yes
##  Min.   :0.0000       Min.   :0.000   Min.   :0.0000   Min.   :0.000      
##  1st Qu.:0.0000       1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.000      
##  Median :0.0000       Median :0.000   Median :0.0000   Median :0.000      
##  Mean   :0.4386       Mean   :0.375   Mean   :0.4929   Mean   :0.489      
##  3rd Qu.:1.0000       3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:1.000      
##  Max.   :1.0000       Max.   :1.000   Max.   :1.0000   Max.   :1.000      
##  PaperlessBilling_Yes PaymentMethod_Credit_Card_(automatic)
##  Min.   :0.0000       Min.   :0.0000                       
##  1st Qu.:0.0000       1st Qu.:0.0000                       
##  Median :1.0000       Median :0.0000                       
##  Mean   :0.5882       Mean   :0.2083                       
##  3rd Qu.:1.0000       3rd Qu.:0.0000                       
##  Max.   :1.0000       Max.   :1.0000                       
##  PaymentMethod_Electronic_Check PaymentMethod_Mailed_Check
##  Min.   :0.0000                 Min.   :0.000             
##  1st Qu.:0.0000                 1st Qu.:0.000             
##  Median :0.0000                 Median :0.000             
##  Mean   :0.3398                 Mean   :0.229             
##  3rd Qu.:1.0000                 3rd Qu.:0.000             
##  Max.   :1.0000                 Max.   :1.000             
```

```
# Create a correlation matrix and heatmap to identify multicollinearity by using cor(), ggplot2::ggplot
cormatrix <- round(cor(churn_norm),2)
melted_cormatrix <- melt(cormatrix)
ggplot(melted_cormatrix, aes(x = Var1, y= Var2, fill = value)) + geom_tile()
```



```
cormatrix[,"Churn_Yes"]
```

```
##                  Outage_sec_perweek                               Email
##                                0.00                                0.01
##                            Contacts                Yearly_equip_failure
##                                0.01                               -0.02
##                              Tenure                       MonthlyCharge
##                               -0.49                                0.37
##                   Bandwidth_GB_Year                               Item1
##                               -0.44                               -0.01
##                               Item2                               Item3
##                               -0.01                               -0.01
##                               Item4                               Item5
##                                0.00                               -0.01
##                               Item6                               Item7
##                                0.00                               -0.01
##                               Item8                           Churn_Yes
##                                0.01                                1.00
##                          Techie_Yes                   Contract_One_Year
##                                0.07                               -0.14
##                   Contract_Two_Year                      Port_modem_Yes
##                               -0.18                                0.01
##                          Tablet_Yes         InternetService_Fiber_Optic
##                                0.00                               -0.06
##                InternetService_None                           Phone_Yes
##                               -0.04                               -0.03
##                          Multiple_Yes                    OnlineSecurity_Yes
##                                  0.13                                 -0.01
##                      OnlineBackup_Yes                  DeviceProtection_Yes
##                                  0.05                                  0.06
##                       TechSupport_Yes                       StreamingTV_Yes
##                                  0.02                                  0.23
##                   StreamingMovies_Yes                  PaperlessBilling_Yes
##                                  0.29                                  0.01
## PaymentMethod_Credit_Card_(automatic)        PaymentMethod_Electronic_Check
##                                 -0.01                                  0.03
##            PaymentMethod_Mailed_Check
##                                 -0.01
```

```
write.csv(cormatrix,"C:/Users/tedda/Desktop/Data Science Portfolio/Machine Learning/Supervised Learning

# Remove Bandwidth_GB_Year from analysis as it is highly correlated with Tenure
churn_norm <- churn_norm[c(1:6,8:35)]

# Export the prepared dataset as a .csv file using write.csv()
write.csv(churn_norm,"C:/Users/tedda/Desktop/Data Science Portfolio/Machine Learning/Supervised Learnin
```

Exploratory Data Analysis on Initial Model

```
# Create the Gross "Initial" Model
LG_GrossModel <- glm(Churn_Yes ~ ., data = churn_norm, family = "binomial")
summary(LG_GrossModel)
```

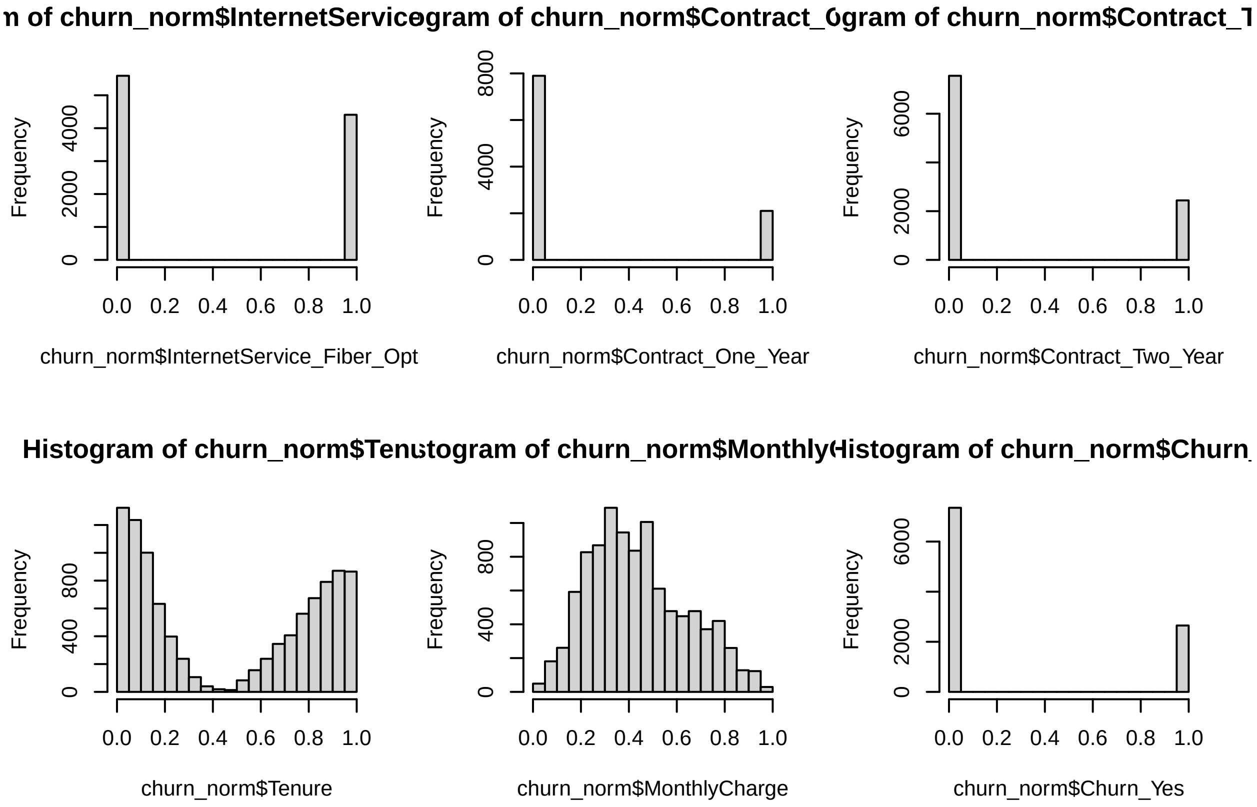
```
##
## Call:
## glm(formula = Churn_Yes ~ ., family = "binomial", data = churn_norm)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8049  -0.2637  -0.0537   0.0693   3.4602
##
## Coefficients:
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                              -1.48829    0.36592  -4.067 4.76e-05
## Outage_sec_perweek                       -0.04107    0.27385  -0.150   0.8808
## Email                                    -0.18939    0.27857  -0.680   0.4966
## Contacts                                  0.44355    0.27164   1.633   0.1025
## Yearly_equip_failure                     -0.23628    0.36524  -0.647   0.5177
## Tenure                                   -8.26115    0.21235 -38.904  < 2e-16
## MonthlyCharge                             8.69342    0.99863   8.705  < 2e-16
## Item1                                    -0.14457    0.32655  -0.443   0.6580
## Item2                                    -0.02660    0.31103  -0.086   0.9318
## Item3                                     0.15994    0.32849   0.487   0.6263
## Item4                                    -0.20618    0.25119  -0.821   0.4117
## Item5                                    -0.21835    0.26581  -0.821   0.4114
## Item6                                    -0.14864    0.31561  -0.471   0.6377
## Item7                                    -0.01918    0.25766  -0.074   0.9407
## Item8                                    -0.07534    0.28144  -0.268   0.7889
## Techie_Yes                                1.08883    0.10248  10.625  < 2e-16
## Contract_One_Year                        -3.40212    0.12821 -26.536  < 2e-16
## Contract_Two_Year                       -3.48297    0.12554 -27.744  < 2e-16
## Port_modem_Yes                           0.14244    0.07716   1.846   0.0649
## Tablet_Yes                              -0.06279    0.08425  -0.745   0.4561
## InternetService_Fiber_Optic             -2.20966    0.13396 -16.495  < 2e-16
## InternetService_None                    -0.94394    0.12484  -7.561 3.99e-14
## Phone_Yes                               -0.29739    0.13251  -2.244   0.0248
## Multiple_Yes                             0.33259    0.17183   1.936   0.0529
## OnlineSecurity_Yes                      -0.24728    0.08112  -3.048   0.0023
## OnlineBackup_Yes                        -0.11254    0.13015  -0.865   0.3872
## DeviceProtection_Yes                    -0.07800    0.09772  -0.798   0.4248
## TechSupport_Yes                         -0.24275    0.09937  -2.443   0.0146
## StreamingTV_Yes                          1.16832    0.22130   5.279 1.30e-07
## StreamingMovies_Yes                      1.28346    0.26401   4.861 1.17e-06
## PaperlessBilling_Yes                     0.17592    0.07841   2.243   0.0249
## `PaymentMethod_Credit_Card_(automatic)`  0.20921    0.11755   1.780   0.0751
## PaymentMethod_Electronic_Check           0.62916    0.10564   5.955 2.59e-09
## PaymentMethod_Mailed_Check               0.23138    0.11583   1.998   0.0458
##                                            
## (Intercept)                             ***
## Outage_sec_perweek                         
## Email                                      
## Contacts                                   
## Yearly_equip_failure                       
## Tenure                                  ***
## MonthlyCharge                           ***
## Item1                                      
## Item2                                      
## Item3                                      
## Item4                                      
## Item5                                      
## Item6                                      
## Item7                                      
## Item8                                      
## Techie_Yes                              ***
## Contract_One_Year                       ***
## Contract_Two_Year                       ***
## Port_modem_Yes                          .  
## Tablet_Yes                                 
## InternetService_Fiber_Optic             ***
## InternetService_None                    ***
## Phone_Yes                               *  
## Multiple_Yes                            .  
## OnlineSecurity_Yes                      ** 
## OnlineBackup_Yes                           
## DeviceProtection_Yes                       
## TechSupport_Yes                         *  
## StreamingTV_Yes                         ***
## StreamingMovies_Yes                     ***
## PaperlessBilling_Yes                    *  
## `PaymentMethod_Credit_Card_(automatic)` .  
## PaymentMethod_Electronic_Check          ***
## PaymentMethod_Mailed_Check              *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11564.4  on 9999  degrees of freedom
## Residual deviance:  4362.5  on 9966  degrees of freedom
## AIC: 4430.5
##
## Number of Fisher Scoring iterations: 7
```

```
# Subset Regression to identify the top 5 variables affecting Churn using leaps::regsubsets()
subsets <- regsubsets(Churn_Yes ~ ., data = churn_norm, nvmax = 5)
summary(subsets)
```

```
## Subset selection object
## Call: regsubsets.formula(Churn_Yes ~ ., data = churn_norm, nvmax = 5)
## 33 Variables  (and intercept)
##                                          Forced in Forced out
## Outage_sec_perweek                           FALSE      FALSE
## Email                                        FALSE      FALSE
## Contacts                                     FALSE      FALSE
## Yearly_equip_failure                         FALSE      FALSE
## Tenure                                       FALSE      FALSE
## MonthlyCharge                                FALSE      FALSE
## Item1                                        FALSE      FALSE
## Item2                                        FALSE      FALSE
## Item3                                        FALSE      FALSE
## Item4                                        FALSE      FALSE
## Item5                                        FALSE      FALSE
## Item6                                        FALSE      FALSE
## Item7                                        FALSE      FALSE
## Item8                                        FALSE      FALSE
## Techie_Yes                                   FALSE      FALSE
## Contract_One_Year                            FALSE      FALSE
## Contract_Two_Year                            FALSE      FALSE
## Port_modem_Yes                               FALSE      FALSE
## Tablet_Yes                                   FALSE      FALSE
## InternetService_Fiber_Optic                  FALSE      FALSE
## InternetService_None                         FALSE      FALSE
## Phone_Yes                                    FALSE      FALSE
## Multiple_Yes                                 FALSE      FALSE
## OnlineSecurity_Yes                           FALSE      FALSE
## OnlineBackup_Yes                             FALSE      FALSE
## DeviceProtection_Yes                         FALSE      FALSE
## TechSupport_Yes                              FALSE      FALSE
## StreamingTV_Yes                              FALSE      FALSE
## StreamingMovies_Yes                          FALSE      FALSE
## PaperlessBilling_Yes                         FALSE      FALSE
## `PaymentMethod_Credit_Card_(automatic)`      FALSE      FALSE
## PaymentMethod_Electronic_Check               FALSE      FALSE
## PaymentMethod_Mailed_Check                   FALSE      FALSE
## 1 subsets of each size up to 5
## Selection Algorithm: exhaustive
##          Outage_sec_perweek Email Contacts Yearly_equip_failure Tenure
## 1  ( 1 ) " "                " "   " "      " "                  "*"
## 2  ( 1 ) " "                " "   " "      " "                  "*"
## 3  ( 1 ) " "                " "   " "      " "                  "*"
## 4  ( 1 ) " "                " "   " "      " "                  "*"
## 5  ( 1 ) " "                " "   " "      " "                  "*"
##          MonthlyCharge Item1 Item2 Item3 Item4 Item5 Item6 Item7 Item8
## 1  ( 1 ) " "           " "   " "   " "   " "   " "   " "   " "   " "
## 2  ( 1 ) "*"           " "   " "   " "   " "   " "   " "   " "   " "
## 3  ( 1 ) "*"           " "   " "   " "   " "   " "   " "   " "   " "
## 4  ( 1 ) "*"           " "   " "   " "   " "   " "   " "   " "   " "
## 5  ( 1 ) "*"           " "   " "   " "   " "   " "   " "   " "   " "
##          Techie_Yes Contract_One_Year Contract_Two_Year Port_modem_Yes
## 1  ( 1 ) " "        " "               " "               " "
## 2  ( 1 ) " "        " "               " "               " "
## 3  ( 1 ) " "        " "               " "               " "
## 4  ( 1 ) " "        "*"               "*"               " "
## 5  ( 1 ) " "        "*"               "*"               " "
##          Tablet_Yes InternetService_Fiber_Optic InternetService_None Phone_Yes
## 1  ( 1 ) " "        " "                         " "                  " "
## 2  ( 1 ) " "        " "                         " "                  " "
## 3  ( 1 ) " "        "*"                         " "                  " "
## 4  ( 1 ) " "        " "                         " "                  " "
## 5  ( 1 ) " "        "*"                         " "                  " "
##          Multiple_Yes OnlineSecurity_Yes OnlineBackup_Yes DeviceProtection_Yes
## 1  ( 1 ) " "          " "                " "              " "
## 2  ( 1 ) " "          " "                " "              " "
## 3  ( 1 ) " "          " "                " "              " "
## 4  ( 1 ) " "          " "                " "              " "
## 5  ( 1 ) " "          " "                " "              " "
##          TechSupport_Yes StreamingTV_Yes StreamingMovies_Yes
## 1  ( 1 ) " "             " "             " "
## 2  ( 1 ) " "             " "             " "
## 3  ( 1 ) " "             " "             " "
## 4  ( 1 ) " "             " "             " "
## 5  ( 1 ) " "             " "             " "
##          PaperlessBilling_Yes `PaymentMethod_Credit_Card_(automatic)`
## 1  ( 1 ) " "                  " "
## 2  ( 1 ) " "                  " "
## 3  ( 1 ) " "                  " "
## 4  ( 1 ) " "                  " "
## 5  ( 1 ) " "                  " "
##          PaymentMethod_Electronic_Check PaymentMethod_Mailed_Check
## 1  ( 1 ) " "                            " "
## 2  ( 1 ) " "                            " "
## 3  ( 1 ) " "                            " "
## 4  ( 1 ) " "                            " "
## 5  ( 1 ) " "                            " "
```

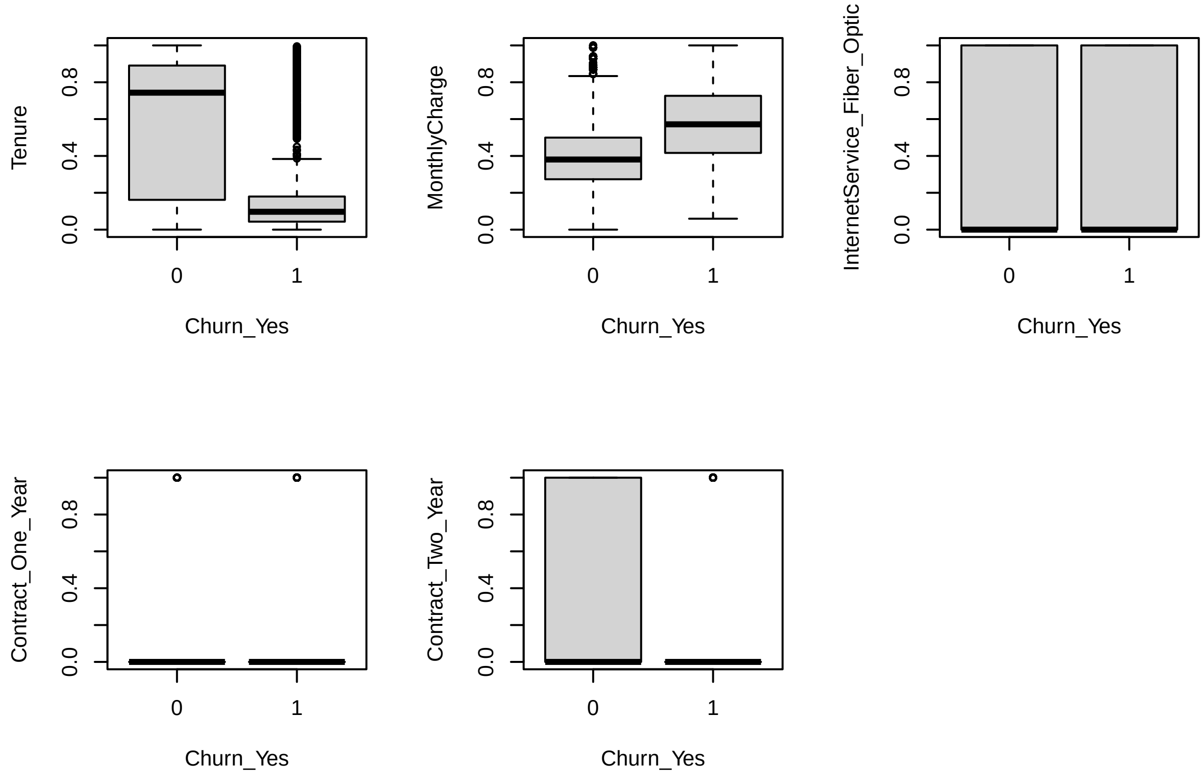
```r
# Create Univariate Distributions using histograms
par(mfrow = c(2,3))
InternetServiceFiberOptic_hist <- hist(churn_norm$InternetService_Fiber_Optic)
ContractOneYear_hist <- hist(churn_norm$Contract_One_Year)
ContractTwoYear_hist <- hist(churn_norm$Contract_Two_Year)
Tenure_hist <- hist(churn_norm$Tenure)
MonthlyCharge_hist <- hist(churn_norm$MonthlyCharge)
ChurnYes_hist <- hist(churn_norm$Churn_Yes)
```



```
# Create Bivariate Distrbutions using boxplot()
par(mfrow = c(2,3))
Tenure_boxplot <- boxplot(Tenure ~ Churn_Yes, data = churn_norm)
MonthlyCharge_boxplot <- boxplot(MonthlyCharge ~ Churn_Yes, data = churn_norm)
InternetServiceFiberOptic_boxplot <- boxplot(InternetService_Fiber_Optic ~ Churn_Yes, data = churn_norm)
ContractOneYear_boxplot <- boxplot(Contract_One_Year~ Churn_Yes, data = churn_norm)
ContractTwoYear_boxplot <- boxplot(Contract_Two_Year~ Churn_Yes, data = churn_norm)

# Reduced Correlation Matrix of only top 5 variables
reduced_data <- churn_norm[c(16,5:6,18:19,22)]
reduced_cormatrix <- round(cor(reduced_data),2)
reduced_cormatrix
```

```
##                       Techie_Yes Tenure MonthlyCharge Contract_Two_Year
## Techie_Yes                  1.00  -0.01          0.01             -0.01
## Tenure                     -0.01   1.00          0.00              0.02
## MonthlyCharge               0.01   0.00          1.00              0.00
## Contract_Two_Year          -0.01   0.02          0.00              1.00
## Port_modem_Yes             -0.01   0.01          0.00              0.00
## InternetService_None       -0.01  -0.01         -0.24             -0.01
##                       Port_modem_Yes InternetService_None
## Techie_Yes                     -0.01                -0.01
## Tenure                          0.01                -0.01
## MonthlyCharge                   0.00                -0.24
## Contract_Two_Year               0.00                -0.01
## Port_modem_Yes                  1.00                 0.00
## InternetService_None           0.00                 1.00
```



Model Building

```
# Create the Adjusted "Reduced" Model based on the 5 variables found above
LG_AdjustedModel <- glm(Churn_Yes ~ Tenure + MonthlyCharge + Contract_One_Year + Contract_Two_Year + In
summary(LG_AdjustedModel)
```

```
##
## Call:
## glm(formula = Churn_Yes ~ Tenure + MonthlyCharge + Contract_One_Year +
##     Contract_Two_Year + InternetService_Fiber_Optic, family = "binomial",
##     data = churn_norm)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8039  -0.3017  -0.0709   0.1003   3.5673
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -2.07003    0.09877  -20.96   <2e-16 ***
## Tenure                      -7.56228    0.18855  -40.11   <2e-16 ***
## MonthlyCharge               11.56168    0.29583   39.08   <2e-16 ***
## Contract_One_Year           -3.11287    0.11917  -26.12   <2e-16 ***
## Contract_Two_Year           -3.19047    0.11659  -27.36   <2e-16 ***
## InternetService_Fiber_Optic -2.11462    0.08854  -23.88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11564.4  on 9999  degrees of freedom
## Residual deviance:  4737.7  on 9994  degrees of freedom
## AIC: 4749.7
##
## Number of Fisher Scoring iterations: 7
```

```
# Extract Coefficients of the Adjusted Model using coef()
coef(LG_AdjustedModel)
```

```
##                 (Intercept)                      Tenure
##                   -2.070027                   -7.562276
##               MonthlyCharge           Contract_One_Year
##                   11.561685                   -3.112873
##           Contract_Two_Year InternetService_Fiber_Optic
##                   -3.190465                   -2.114616
```

```
# Create the 4-variable reduced model based on the subsets found
LG_Reduced4Model <- glm(Churn_Yes ~ Tenure + MonthlyCharge + Contract_One_Year + Contract_Two_Year, chu
summary(LG_Reduced4Model)
```

```
##
## Call:
## glm(formula = Churn_Yes ~ Tenure + MonthlyCharge + Contract_One_Year +
##     Contract_Two_Year, family = "binomial", data = churn_norm)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8416  -0.3692  -0.1087   0.1564   3.2145
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -2.16951    0.09365  -23.17   <2e-16 ***
## Tenure            -6.46717    0.15632  -41.37   <2e-16 ***
## MonthlyCharge      8.93594    0.23467   38.08   <2e-16 ***
## Contract_One_Year -2.65906    0.10555  -25.19   <2e-16 ***
## Contract_Two_Year -2.73856    0.10230  -26.77   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11564.4  on 9999  degrees of freedom
## Residual deviance:  5444.3  on 9995  degrees of freedom
## AIC: 5454.3
##
## Number of Fisher Scoring iterations: 7
```

Save and Load Model

```
# Save and Load 5-variable Model
var5_model_url <- "C:/Users/tedda/Desktop/Data Science Portfolio/Machine Learning/Supervised Learning/Re
saveRDS(LG_AdjustedModel, var5_model_url)
LG_AdjustedModel <- readRDS(var5_model_url)

# Save and Load 4-variable Model
var4_model_url <- "C:/Users/tedda/Desktop/Data Science Portfolio/Machine Learning/Supervised Learning/Re
saveRDS(LG_Reduced4Model, var4_model_url)
LG_Reduced4Model <- readRDS(var4_model_url)
```

Model Evaluation

```
# Confusion Matrix for Gross Model with all variables
LGmodelGPred <- round(predict(LG_GrossModel, churn_norm, type = "response"))
LGmodelG <- confusionMatrix(as.factor(LGmodelGPred), as.factor(churn_norm$Churn))
LGmodelG
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 6904  518
##          1  446 2132
##
##                Accuracy : 0.9036
##                  95% CI : (0.8976, 0.9093)
##     No Information Rate : 0.735
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.7504
##
##  Mcnemar's Test P-Value : 0.02221
##
##             Sensitivity : 0.9393
##             Specificity : 0.8045
##          Pos Pred Value : 0.9302
##          Neg Pred Value : 0.8270
##              Prevalence : 0.7350
##          Detection Rate : 0.6904
##    Detection Prevalence : 0.7422
##       Balanced Accuracy : 0.8719
##
##        'Positive' Class : 0
##
```

```
# Confusion Matrix for Adjusted Model with all variables
LGmodelAPred <- round(predict(LG_AdjustedModel, churn_norm, type = "response"))
LGmodelA <- confusionMatrix(as.factor(LGmodelAPred), as.factor(churn_norm$Churn_Yes))
LGmodelA
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 6871  579
##          1  479 2071
##
##                Accuracy : 0.8942
##                  95% CI : (0.888, 0.9002)
##     No Information Rate : 0.735
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.7251
##
##  Mcnemar's Test P-Value : 0.002337
##
##             Sensitivity : 0.9348
##             Specificity : 0.7815
##          Pos Pred Value : 0.9223
##          Neg Pred Value : 0.8122
##              Prevalence : 0.7350
##          Detection Rate : 0.6871
##    Detection Prevalence : 0.7450
##       Balanced Accuracy : 0.8582
##
##        'Positive' Class : 0
##
```

```
# Confusion Matrix for Reduced-4 variable Model with all variables
LGmodel4Pred <- round(predict(LG_Reduced4Model, churn_norm, type = "response"))
LGmodel4 <- confusionMatrix(as.factor(LGmodel4Pred), as.factor(churn_norm$Churn_Yes))
LGmodel4
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 6815  690
##          1  535 1960
##
##                Accuracy : 0.8775
##                  95% CI : (0.8709, 0.8839)
##     No Information Rate : 0.735
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.6795
##
##  Mcnemar's Test P-Value : 1.083e-05
##
##             Sensitivity : 0.9272
##             Specificity : 0.7396
##          Pos Pred Value : 0.9081
##          Neg Pred Value : 0.7856
##              Prevalence : 0.7350
##          Detection Rate : 0.6815
##    Detection Prevalence : 0.7505
##       Balanced Accuracy : 0.8334
##
##        'Positive' Class : 0
```

```
##
```

```
# Calculate F1_Score of Gross Model
pred <- ifelse(LG_GrossModel$fitted.values < 0.5, 0, 1)
F1_Score(y_pred = pred, y_true = churn_norm$Churn_Yes, positive = "0")
```

```
## [1] 0.9347414
```

```
# Calculate F1_Score of Adjusted Model
pred <- ifelse(LG_AdjustedModel$fitted.values < 0.5, 0, 1)
F1_Score(y_pred = pred, y_true = churn_norm$Churn_Yes, positive = "0")
```

```
## [1] 0.9285135
```

```
# Calculate F1_Score of Reduced-4 variable Model
pred <- ifelse(LG_Reduced4Model$fitted.values < 0.5, 0, 1)
F1_Score(y_pred = pred, y_true = churn_norm$Churn_Yes, positive = "0")
```

```
## [1] 0.9175362
```