Churn_Analysis_worst_case

2024-02-03

Package Load the data set and eliminate

```
churn <- read.csv('Telco_customer_churn_cleaned.csv')

#Eliminated Total.Charges, Churn Label, Churn Score, and CLTV as we are not using it
colnames(churn)[15] <- "Tenure"
colnames(churn)[31] <- "Churn_val"
churn <- churn[ -c(29, 30, 32, 33) ]
```

```
# Create the best case (not churned) and worst case (churned group)

# Case we do not know if customers will churn or not since they have not churned yet and
unknown_churn <- filter(churn, churn$Tenure < 12 & churn$Churn_val== 0)
unknown_churn_best <- unknown_churn
unknown_churn_worst <- unknown_churn

# Known Case 1) churned before 12 month 2) churned after 12 month 3) not churned after 12 month
known_churn <- churn %>%
      filter(! CustomerID %in% unknown_churn$CustomerID) #5973

# Created best and worst case column
unknown_churn_best[ , 'churn_12month'] = 0
unknown_churn_worst[ , 'churn_12month'] = 1
known_churn[ , 'churn_12month'] = known_churn$Churn_val

# Bind known and unknown case
best_case <- rbind(known_churn, unknown_churn_best)
worst_case <- rbind(known_churn, unknown_churn_worst)

# Change churned after 12 month as not churned since a customer hasn't churned yet at 12 month time poi
best_case$churn_12month[best_case$Tenure > 12 & best_case$Churn_val == 1] <- 0
worst_case$churn_12month[worst_case$Tenure > 12 & worst_case$Churn_val == 1] <- 0

# eliminate churn_val and tenure since we substitute them with 12 month churn_val using two cases and u
```

```
best_case <- best_case[-c(15, 29)]
worst_case <- worst_case[-c(15, 29)]
```

Create the best and worst case data set

```r
# divide customer into churned and not churned group
churned_best <- best_case[best_case$churn_12month == 1,]
not_churned_best <- best_case[best_case$churn_12month == 0,]

# Confidence interval for monthly based on churned_12 month or not
t.test(churned_best$Monthly.Charges)$conf
```

Confidence Interval for Monthly Charge

```
## [1] 65.02100 67.96695
## attr(,"conf.level")
## [1] 0.95
```

```r
t.test(not_churned_best$Monthly.Charges)$conf
```

```
## [1] 63.67870 65.24649
## attr(,"conf.level")
## [1] 0.95
```

```r
# divide customer into churned and not churned group
churned_worst <- worst_case[worst_case$churn_12month == 1,]
not_churned_worst <- worst_case[worst_case$churn_12month == 0,]

# Confidence interval for monthly based on churned_12 month or not
t.test(churned_worst$Monthly.Charges)$conf
```

```
## [1] 55.30520 57.56794
## attr(,"conf.level")
## [1] 0.95
```

```r
t.test(not_churned_worst$Monthly.Charges)$conf
```

```
## [1] 67.45510 69.17567
## attr(,"conf.level")
## [1] 0.95
```

```r
print('best case')
```

Contingency table for churn_12 month and non-demographic qualitative variables

```
## [1] "best case"
```

```r
# Best Case
phone_service <- table(best_case$churn_12month, best_case$Phone.Service)
phone_service
```

```
##
##       No  Yes
##   0  579 5427
##   1  103  934
```

```
multi_lines <- table(best_case$churn_12month, best_case$Multiple.Lines)
multi_lines
```

```
##
##       No No phone service  Yes
##   0 2775              579 2652
##   1  615              103  319
```

```
internet_service <- table(best_case$churn_12month, best_case$Internet.Service)
internet_service
```

```
##
##      DSL Fiber optic   No
##   0 2121        2451 1434
##   1  300         645   92
```

```
online_security <- table(best_case$churn_12month, best_case$Online.Security)
online_security
```

```
##
##       No No internet service  Yes
##   0 2645                1434 1927
##   1  853                  92   92
```

```
online_backup <- table(best_case$churn_12month, best_case$Online.Backup)
online_backup
```

```
##
##       No No internet service  Yes
##   0 2306                1434 2266
##   1  782                  92  163
```

```
device_protect <- table(best_case$churn_12month, best_case$Device.Protection)
device_protect
```

```
##
##       No No internet service  Yes
##   0 2334                1434 2238
##   1  761                  92  184
```

```
tech_support <- table(best_case$churn_12month, best_case$Tech.Support)
tech_support
```

```
##
##       No No internet service  Yes
##   0 2635                1434 1937
##   1  838                  92  107
```

```r
payment <- table(best_case$churn_12month, best_case$Payment.Method)
payment
```

```
##
##     Bank transfer (automatic) Credit card (automatic) Electronic check
##   0                      1441                    1441             1759
##   1                       103                      81              606
##
##     Mailed check
##   0         1365
##   1          247
```

```r
paperless <- table(best_case$churn_12month, best_case$Paperless.Billing)
paperless
```

```
##
##       No  Yes
##   0 2574 3432
##   1  298  739
```

```r
contract <- table(best_case$churn_12month, best_case$Contract)
contract
```

```
##
##     Month-to-month One year Two year
##   0           2851     1460     1695
##   1           1024       13        0
```

```r
print('worst case')
```

```
## [1] "worst case"
```

```r
# Worst Case
phone_service <- table(worst_case$churn_12month, worst_case$Phone.Service)
phone_service
```

```
##
##       No  Yes
##   0  474 4462
##   1  208 1899
```

```r
multi_lines <- table(worst_case$churn_12month, worst_case$Multiple.Lines)
multi_lines
```

```
##
##       No No phone service  Yes
##   0 1970              474 2492
##   1 1420              208  479
```

```
internet_service <- table(worst_case$churn_12month, worst_case$Internet.Service)
internet_service
```

```
##
##      DSL Fiber optic   No
##   0 1702        2190 1044
##   1  719         906  482
```

```
online_security <- table(worst_case$churn_12month, worst_case$Online.Security)
online_security
```

```
##
##       No No internet service  Yes
##   0 2131                1044 1761
##   1 1367                 482  258
```

```
online_backup <- table(worst_case$churn_12month, worst_case$Online.Backup)
online_backup
```

```
##
##       No No internet service  Yes
##   0 1788                1044 2104
##   1 1300                 482  325
```

```
device_protect <- table(worst_case$churn_12month, worst_case$Device.Protection)
device_protect
```

```
##
##       No No internet service  Yes
##   0 1787                1044 2105
##   1 1308                 482  317
```

```
tech_support <- table(worst_case$churn_12month, worst_case$Tech.Support)
tech_support
```

```
##
##       No No internet service  Yes
##   0 2114                1044 1778
##   1 1359                 482  266
```

```
payment <- table(worst_case$churn_12month, worst_case$Payment.Method)
payment
```

```
##
##     Bank transfer (automatic) Credit card (automatic) Electronic check
##   0                      1333                    1311             1414
##   1                       211                     211              951
##
##     Mailed check
##   0          878
##   1          734
```

```r
paperless <- table(worst_case$churn_12month, worst_case$Paperless.Billing)
paperless
```

```
##
##       No  Yes
##   0 2009 2927
##   1  863 1244
```

```r
contract <- table(worst_case$churn_12month, worst_case$Contract)
contract
```

```
##
##     Month-to-month One year Two year
##   0           1934     1366     1636
##   1           1941      107       59
```

Separating data set into training and test data sets with stratification using churn_12month column (Response variable)

```r
# Chose 0.7 for convention
train.index <- createDataPartition(best_case$churn_12month, p = .7, list = FALSE)
train_best <- best_case[ train.index,]
test_best  <- best_case[-train.index,]

train.index <- createDataPartition(worst_case$churn_12month, p = .7, list = FALSE)
train_worst <- worst_case[ train.index,]
test_worst  <- worst_case[-train.index,]
```

```r
# model with every possible variable (Total 17, 1 continuous and others categorical)
model_best_case <- glm(I(churn_12month) ~ Gender + Senior.Citizen + Partner + Dependents + Phone.Servic

model_worst_case <- glm(I(churn_12month) ~ Gender + Senior.Citizen + Partner + Dependents + Phone.Servi
```

Set the full model, without interaction terms

```r
# Best model with best case data set
null <- glm(I(churn_12month) ~ 1,  data = train_best, family = "binomial")
step(null, scope = list(lower=null,upper=model_best_case),
     direction="both", criterion = "AIC", trace = FALSE)
```

Fit the full and null model and use AIC to find the model with lowest AIC and BIC, but use AIC for prediction

```
##
## Call:  glm(formula = I(churn_12month) ~ Contract + Online.Backup + Online.Security +
```

```
##     Payment.Method + Dependents + Tech.Support + Multiple.Lines +
##     Device.Protection + Partner + Paperless.Billing + Streaming.TV +
##     Internet.Service, family = "binomial", data = train_best)
##
## Coefficients:
##                             (Intercept)                       ContractOne year
##                                -0.52220                                -2.93299
##                         ContractTwo year       Online.BackupNo internet service
##                                -16.90766                                -1.21923
##                         Online.BackupYes     Online.SecurityNo internet service
##                                -0.88342                                      NA
##                       Online.SecurityYes  Payment.MethodCredit card (automatic)
##                                -1.06812                                -0.35128
##          Payment.MethodElectronic check             Payment.MethodMailed check
##                                 0.47694                                 0.49011
##                            DependentsYes        Tech.SupportNo internet service
##                                -0.71158                                      NA
##                         Tech.SupportYes         Multiple.LinesNo phone service
##                                -0.65107                                 0.04956
##                        Multiple.LinesYes   Device.ProtectionNo internet service
##                                -0.60215                                      NA
##                    Device.ProtectionYes                              PartnerYes
##                                -0.37463                                -0.36134
##                    Paperless.BillingYes        Streaming.TVNo internet service
##                                 0.27328                                      NA
##                          Streaming.TVYes            Internet.ServiceFiber optic
##                                -0.32720                                 0.35341
##                      Internet.ServiceNo
##                                      NA
##
## Degrees of Freedom: 4930 Total (i.e. Null);  4913 Residual
## Null Deviance:       4139
## Residual Deviance: 2833  AIC: 2869
```

```
#named as AIC but k = log(n) makes it calculate BIC
step(null, scope = list(lower=null,upper=model_best_case),
     direction="both", criterion = "BIC", k = log(4931),trace = FALSE)
```

```
##
## Call:  glm(formula = I(churn_12month) ~ Contract + Online.Backup + Online.Security +
##     Dependents + Tech.Support + Payment.Method + Multiple.Lines +
##     Device.Protection + Partner, family = "binomial", data = train_best)
##
## Coefficients:
##                             (Intercept)                       ContractOne year
##                                 -0.2059                                 -3.0065
##                         ContractTwo year       Online.BackupNo internet service
##                                -16.9834                                -1.4229
##                         Online.BackupYes     Online.SecurityNo internet service
##                                 -0.8823                                      NA
##                       Online.SecurityYes                           DependentsYes
##                                 -1.1026                                 -0.7604
##        Tech.SupportNo internet service                         Tech.SupportYes
##                                      NA                                 -0.7240
```

```
## Payment.MethodCredit card (automatic)         Payment.MethodElectronic check
##                                -0.3228                                 0.5132
##            Payment.MethodMailed check         Multiple.LinesNo phone service
##                                 0.4560                                -0.1674
##                     Multiple.LinesYes   Device.ProtectionNo internet service
##                                -0.5368                                    NA
##                  Device.ProtectionYes                             PartnerYes
##                                -0.4093                                -0.3505
##
## Degrees of Freedom: 4930 Total (i.e. Null);  4916 Residual
## Null Deviance:       4139
## Residual Deviance: 2854  AIC: 2884
```

```
null <- glm(I(churn_12month) ~ 1,  data = train_worst, family = "binomial")
step(null, scope = list(lower=null,upper=model_worst_case),
     direction="both", criterion = "AIC", trace = FALSE)
```

```
##
## Call:  glm(formula = I(churn_12month) ~ Contract + Monthly.Charges +
##     Payment.Method + Multiple.Lines + Internet.Service + Partner +
##     Online.Backup + Online.Security + Device.Protection + Tech.Support +
##     Senior.Citizen, family = "binomial", data = train_worst)
##
## Coefficients:
##                            (Intercept)                      ContractOne year
##                                2.76940                               -2.21567
##                       ContractTwo year                        Monthly.Charges
##                               -2.84048                               -0.03853
## Payment.MethodCredit card (automatic)         Payment.MethodElectronic check
##                                0.15030                                0.98541
##            Payment.MethodMailed check         Multiple.LinesNo phone service
##                                1.34262                               -1.09987
##                     Multiple.LinesYes            Internet.ServiceFiber optic
##                               -0.84361                                0.77727
##                    Internet.ServiceNo                             PartnerYes
##                               -1.74765                               -0.78567
##      Online.BackupNo internet service                       Online.BackupYes
##                                    NA                               -0.60235
##    Online.SecurityNo internet service                     Online.SecurityYes
##                                    NA                               -0.61272
##  Device.ProtectionNo internet service                   Device.ProtectionYes
##                                    NA                               -0.38853
##       Tech.SupportNo internet service                        Tech.SupportYes
##                                    NA                               -0.35869
##                     Senior.CitizenYes
##                               -0.30984
##
## Degrees of Freedom: 4930 Total (i.e. Null);  4914 Residual
## Null Deviance:       6064
## Residual Deviance: 3685  AIC: 3719
```

```
#named as AIC but k = log(n) makes it calculate BIC
step(null, scope = list(lower=null,upper=model_worst_case),
     direction="both", criterion = "BIC", k = log(4931),trace = FALSE)
```

```
##
## Call:  glm(formula = I(churn_12month) ~ Contract + Monthly.Charges +
##     Payment.Method + Multiple.Lines + Internet.Service + Partner +
##     Online.Backup + Online.Security + Device.Protection, family = "binomial",
##     data = train_worst)
##
## Coefficients:
##                              (Intercept)                       ContractOne year
##                                  2.92979                                -2.22601
##                         ContractTwo year                         Monthly.Charges
##                                 -2.86489                                -0.04416
## Payment.MethodCredit card (automatic)         Payment.MethodElectronic check
##                                  0.14353                                 0.98657
##             Payment.MethodMailed check         Multiple.LinesNo phone service
##                                  1.33791                                -1.20715
##                       Multiple.LinesYes            Internet.ServiceFiber optic
##                                 -0.82093                                 0.93945
##                      Internet.ServiceNo                              PartnerYes
##                                 -1.78886                                -0.79167
##      Online.BackupNo internet service                      Online.BackupYes
##                                      NA                                -0.58298
##    Online.SecurityNo internet service                    Online.SecurityYes
##                                      NA                                -0.59431
##  Device.ProtectionNo internet service                  Device.ProtectionYes
##                                      NA                                -0.37329
##
## Degrees of Freedom: 4930 Total (i.e. Null);  4916 Residual
## Null Deviance:       6064
## Residual Deviance: 3699  AIC: 3729
```

```r
# Final Models
AIC_best_case <- glm(I(churn_12month) ~ Contract + Online.Security + Partner +
    Online.Backup + Payment.Method + Dependents + Multiple.Lines +
    Tech.Support + Device.Protection + Internet.Service + Monthly.Charges +
    Paperless.Billing + Gender,
    data = train_best, family = "binomial")
summary(AIC_best_case)
```

```
##
## Call:
## glm(formula = I(churn_12month) ~ Contract + Online.Security +
##     Partner + Online.Backup + Payment.Method + Dependents + Multiple.Lines +
##     Tech.Support + Device.Protection + Internet.Service + Monthly.Charges +
##     Paperless.Billing + Gender, family = "binomial", data = train_best)
##
## Coefficients: (4 not defined because of singularities)
##                                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)                          0.425860   0.362383   1.175 0.239929
## ContractOne year                    -2.919759   0.362937  -8.045 8.64e-16
## ContractTwo year                   -16.897768 294.109851  -0.057 0.954184
## Online.SecurityNo internet service  -1.739154   0.254767  -6.826 8.70e-12
```

```
## Online.SecurityYes                     -0.960948   0.157347  -6.107 1.01e-09
## PartnerYes                             -0.364543   0.108032  -3.374 0.000740
## Online.BackupNo internet service              NA         NA      NA       NA
## Online.BackupYes                       -0.779431   0.125879  -6.192 5.94e-10
## Payment.MethodCredit card (automatic)  -0.348481   0.207140  -1.682 0.092502
## Payment.MethodElectronic check          0.487531   0.150026   3.250 0.001155
## Payment.MethodMailed check              0.485701   0.169368   2.868 0.004134
## DependentsYes                          -0.714784   0.163887  -4.361 1.29e-05
## Multiple.LinesNo phone service         -0.341605   0.213260  -1.602 0.109195
## Multiple.LinesYes                      -0.487276   0.119109  -4.091 4.29e-05
## Tech.SupportNo internet service               NA         NA      NA       NA
## Tech.SupportYes                        -0.541764   0.151166  -3.584 0.000338
## Device.ProtectionNo internet service          NA         NA      NA       NA
## Device.ProtectionYes                   -0.256695   0.131715  -1.949 0.051311
## Internet.ServiceFiber optic             0.884526   0.233181   3.793 0.000149
## Internet.ServiceNo                            NA         NA      NA       NA
## Monthly.Charges                        -0.020655   0.006682  -3.091 0.001993
## Paperless.BillingYes                    0.272815   0.108499   2.514 0.011922
## GenderMale                             -0.027748   0.093161  -0.298 0.765814
##
## (Intercept)
## ContractOne year                      ***
## ContractTwo year
## Online.SecurityNo internet service    ***
## Online.SecurityYes                    ***
## PartnerYes                            ***
## Online.BackupNo internet service
## Online.BackupYes                      ***
## Payment.MethodCredit card (automatic) .
## Payment.MethodElectronic check        **
## Payment.MethodMailed check            **
## DependentsYes                         ***
## Multiple.LinesNo phone service
## Multiple.LinesYes                     ***
## Tech.SupportNo internet service
## Tech.SupportYes                       ***
## Device.ProtectionNo internet service
## Device.ProtectionYes                  .
## Internet.ServiceFiber optic           ***
## Internet.ServiceNo
## Monthly.Charges                       **
## Paperless.BillingYes                  *
## GenderMale
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4138.6  on 4930  degrees of freedom
## Residual deviance: 2832.6  on 4912  degrees of freedom
## AIC: 2870.6
##
## Number of Fisher Scoring iterations: 18
```

```r
AIC_worst_case <- glm(I(churn_12month) ~ Contract + Monthly.Charges + Payment.Method +
    Multiple.Lines + Internet.Service + Partner + Online.Backup +
    Online.Security + Senior.Citizen + Tech.Support + Device.Protection, data = train_worst, family = "
summary(AIC_worst_case)
```

```
##
## Call:
## glm(formula = I(churn_12month) ~ Contract + Monthly.Charges +
##     Payment.Method + Multiple.Lines + Internet.Service + Partner +
##     Online.Backup + Online.Security + Senior.Citizen + Tech.Support +
##     Device.Protection, family = "binomial", data = train_worst)
##
## Coefficients: (4 not defined because of singularities)
##                                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)                            2.769400   0.328142   8.440  < 2e-16 ***
## ContractOne year                      -2.215672   0.143080 -15.486  < 2e-16 ***
## ContractTwo year                      -2.840479   0.189614 -14.980  < 2e-16 ***
## Monthly.Charges                       -0.038528   0.006151  -6.264 3.76e-10 ***
## Payment.MethodCredit card (automatic)  0.150296   0.154034   0.976 0.329198
## Payment.MethodElectronic check         0.985408   0.128450   7.672 1.70e-14 ***
## Payment.MethodMailed check             1.342621   0.138012   9.728  < 2e-16 ***
## Multiple.LinesNo phone service        -1.099870   0.194243  -5.662 1.49e-08 ***
## Multiple.LinesYes                     -0.843609   0.104125  -8.102 5.41e-16 ***
## Internet.ServiceFiber optic            0.777274   0.213518   3.640 0.000272 ***
## Internet.ServiceNo                    -1.747651   0.231859  -7.538 4.79e-14 ***
## PartnerYes                            -0.785670   0.086100  -9.125  < 2e-16 ***
## Online.BackupNo internet service             NA         NA      NA       NA
## Online.BackupYes                      -0.602351   0.107116  -5.623 1.87e-08 ***
## Online.SecurityNo internet service           NA         NA      NA       NA
## Online.SecurityYes                    -0.612722   0.121263  -5.053 4.35e-07 ***
## Senior.CitizenYes                     -0.309835   0.114631  -2.703 0.006874 **
## Tech.SupportNo internet service              NA         NA      NA       NA
## Tech.SupportYes                       -0.358691   0.126575  -2.834 0.004599 **
## Device.ProtectionNo internet service         NA         NA      NA       NA
## Device.ProtectionYes                  -0.388533   0.116149  -3.345 0.000822 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6064.0  on 4930  degrees of freedom
## Residual deviance: 3684.7  on 4914  degrees of freedom
## AIC: 3718.7
##
## Number of Fisher Scoring iterations: 6
```

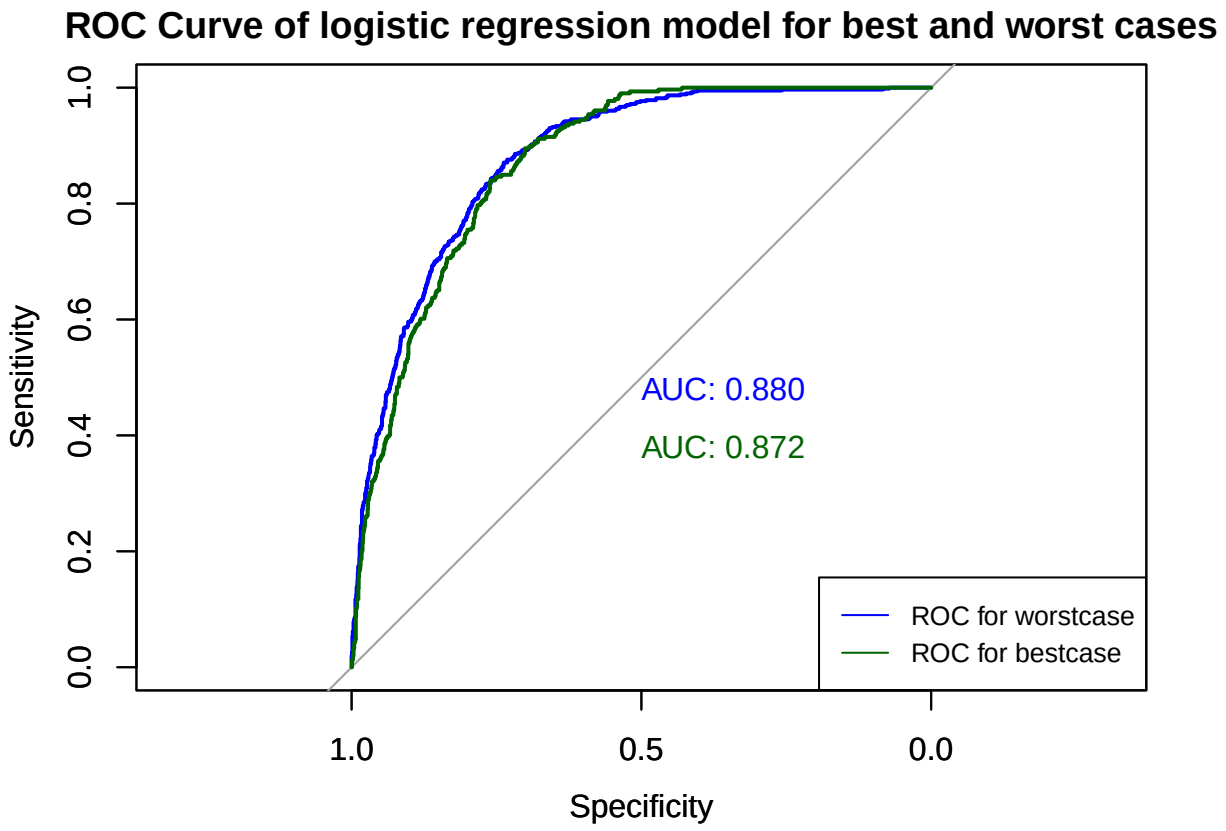
```r
suppressWarnings({logit_P_best = predict(AIC_best_case , newdata = test_best[-test_best$churn_12month]
roc_plot_bestcase=roc(test_best$churn_12month, logit_P_best, level = c(0,1), direction = "<") #AUC scor
auc_bestcase = auc(roc_plot_bestcase)
print(auc_bestcase)
```

```
## Area under the curve: 0.8724
```

```
logit_P_worst = predict(AIC_worst_case, newdata = test_worst[-test_worst$churn_12month] ,type = 'respons
roc_plot_worstcase = roc(test_worst$churn_12month, logit_P_worst, level = c(0,1), direction = "<") #AUC
auc_worstcase= auc(roc_plot_worstcase)
print(auc_worstcase)
```

```
## Area under the curve: 0.8799
```

```
plot(roc_plot_worstcase, print.auc= TRUE, type="l", col="blue" , main = "ROC Curve of logistic regressio
par(new=TRUE)
plot(roc_plot_bestcase, type="l", print.auc=TRUE, print.auc.y = .4,col="darkgreen" )
par(new=TRUE)
legend("bottomright", legend=c("ROC for worstcase", "ROC for bestcase"),
       col=c("blue", "darkgreen"), lty=1:1, cex=0.8)
```



Conduct bootstrap on AUC by resampling the test set everytime. Since we want to check if the same model works well in bootstrapped data set, ##### we do not resample train set or retrain models everytime

```
set.seed(4444)  # for reproducibility

AUC_B <- c()
AUC_W <- c()
w <- nrow(test_worst)
b <- nrow(test_best)
```

```
for (i in 1:1000) {
test_best_B <- as.data.frame(test_best[sample(b, replace = T),])
suppressWarnings({logit_P_best = predict(AIC_best_case, newdata = test_best_B[-test_best_B$churn_12month
roc_bestcase= auc(roc(test_best_B$churn_12month, logit_P_best,level = c(0,1), direction = "<")) #AUC sc
AUC_B[i] <- roc_bestcase
}

quantile(AUC_B, probs = c(0.025, 0.975))
```

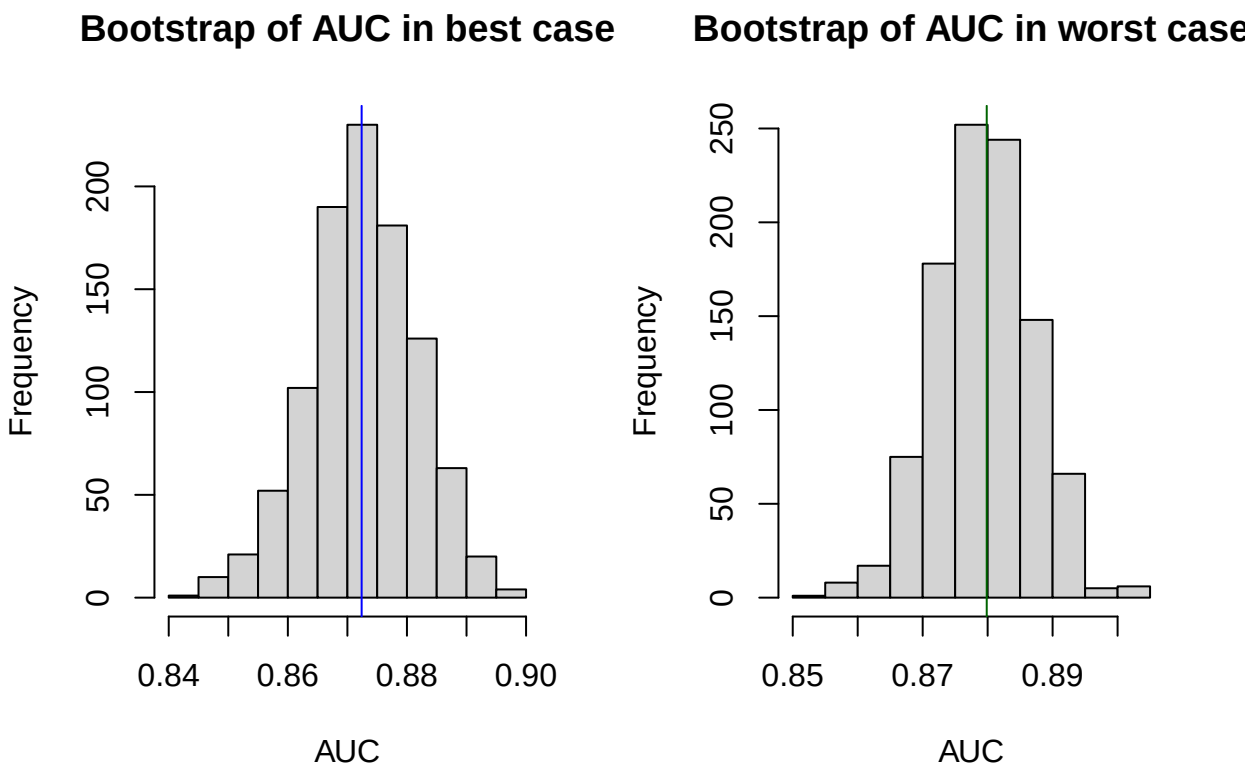
```
##      2.5%     97.5%
## 0.8526066 0.8899145
```

```
for (i in 1:1000) {
test_worst_B <- as.data.frame(test_worst[sample(w, replace = T),])
logit_P_worst = predict(AIC_worst_case, newdata = test_worst_B[-test_worst_B$churn_12month], type = 'res
roc_worstcase= auc(roc(test_worst_B$churn_12month, logit_P_worst,level = c(0,1), direction = "<")) #AUC
AUC_W[i] <- roc_worstcase
}

quantile(AUC_W, probs = c(0.025, 0.975))
```

```
##      2.5%     97.5%
## 0.8649474 0.8925738
```

```
par( mfrow= c(1,2) )

hist(AUC_B, main = 'Bootstrap of AUC in best case', xlab = 'AUC')
abline(v = auc_bestcase, col= 'blue')
hist(AUC_W, main = 'Bootstrap of AUC in worst case', xlab= 'AUC')
abline(v=auc_worstcase, col = 'darkgreen')
```

Crate histogram of AUC from bootstrap with original data's AUC as vertical line for both cases



Not necessary

Code that fits model using one-hot encoding and general and check the result

```
set.seed(5315)
#data subset for testing
encoded_train <- read.csv('preprocessed_train_encoded_no_corr.csv')
encoded_test <-read.csv('preprocessed_test_encoded_no_corr.csv')
worst_case <-read.csv('worst_case_final.csv')

train.index <- createDataPartition(worst_case$churn_12month, p = .9, list = FALSE)
train_worst <- worst_case[ train.index,]
test_worst  <- worst_case[-train.index,]

encoded <- rbind(encoded_train, encoded_test)
encoded_subset <- encoded_train[,c(5,14,15, 16, 17)]
worst_subset <- train_worst[,c(15,16,30)]
```

```
worst_subset_model <- glm(churn_12month ~ ., data = worst_subset,family = "binomial")
encoded_subset_model <- glm(churn~ ., data = encoded_subset,family = "binomial")
summary(encoded_subset_model)
```

```
##
## Call:
## glm(formula = churn ~ ., family = "binomial", data = encoded_subset)
```

```
##
## Coefficients: (2 not defined because of singularities)
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -1.69899    0.05358 -31.708  < 2e-16 ***
## phone_service_noTrue     0.83049    0.10372   8.007 1.17e-15 ***
## phone_service_yesTrue         NA         NA      NA       NA
## multiple_lines_noTrue    1.36163    0.06492  20.975  < 2e-16 ***
## multiple_lines_yesTrue        NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7685.9  on 6337  degrees of freedom
## Residual deviance: 7196.1  on 6335  degrees of freedom
## AIC: 7202.1
##
## Number of Fisher Scoring iterations: 4
```

```
summary(worst_subset_model)
```

```
##
## Call:
## glm(formula = churn_12month ~ ., family = "binomial", data = worst_subset)
##
## Coefficients: (1 not defined because of singularities)
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -0.85269    0.08780  -9.711  < 2e-16 ***
## Phone.ServiceYes                 0.53252    0.09522   5.593 2.24e-08 ***
## Multiple.LinesNo phone service       NA         NA      NA       NA
## Multiple.LinesYes               -1.31441    0.06382 -20.596  < 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: 7736.6  on 6338  degrees of freedom
## Residual deviance: 7269.8  on 6336  degrees of freedom
## AIC: 7275.8
##
## Number of Fisher Scoring iterations: 4
```

```
logit_P_W = predict(worst_subset_model, newdata = test_worst[-test_worst$churn_12month] ,type = 'respons
roc_plot_worst = roc(test_worst$churn_12month, logit_P_W) #AUC score
```

```
## Setting levels: control = 0, case = 1
```

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

```
auc(roc_plot_worst)
```

```
## Area under the curve: 0.6504
```

```r
logit_P_E = predict(encoded_subset_model, newdata = encoded_test[-encoded_test$churn] ,type = 'response
roc_plot_encoded = roc(encoded_test$churn, logit_P_E) #AUC score
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```

```r
auc(roc_plot_encoded)
```

```
## Area under the curve: 0.6405
```