```
suppressMessages(library(tidyverse))
suppressMessages(library(pROC))
suppressMessages(library(caret))
```

```
# Set seed for reproducibility
set.seed(1123)
```

```
setwd("~/Downloads/UM - Fall 22/STATS 504/HW5")
test <- read.csv("test_df.csv")
train <- read.csv("train_df.csv")
drop <- c("X")
train = train[,!(names(train) %in% drop)]
test = test[,!(names(test) %in% drop)]
```

Logistic Regression:

```
fullmod <- glm(dv.hypertension1 ~., train, family = binomial)
summary(fullmod)
```

```
## 
## Call:
## glm(formula = dv.hypertension1 ~ ., family = binomial, data = train)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.4780  -0.2401  -0.1581  -0.1075   3.6277  
## 
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)    
## (Intercept)          -1.452e+01  1.365e+00 -10.638  < 2e-16 ***
## age                   6.253e-02  1.678e-02   3.727 0.000194 ***
## racehispanic         -8.583e-01  3.140e-01  -2.733 0.006274 ** 
## racenative           -1.397e+00  1.033e+00  -1.352 0.176382    
## raceother            -1.013e+00  2.956e-01  -3.427 0.000609 ***
## racewhite            -7.234e-01  2.301e-01  -3.144 0.001668 ** 
## emosupportTRUE        1.884e-01  5.658e-01   0.333 0.739118    
## financialsupportTRUE -1.501e-01  3.632e-01  -0.413 0.679335    
## prenatalsupportTRUE  -2.869e-01  2.959e-01  -0.970 0.332171    
## deliverysupportTRUE  -5.727e-02  5.930e-01  -0.097 0.923067    
## psstotal             -1.895e-02  2.416e-02  -0.784 0.432930    
## anxtotal              9.532e-04  1.261e-02   0.076 0.939732    
## worryfambaby          8.660e-02  7.851e-02   1.103 0.270032    
## exerciseTRUE          7.940e-02  1.825e-01   0.435 0.663523    
## systolic              6.043e-02  8.419e-03   7.178 7.08e-13 ***
## diastolic             4.321e-02  1.036e-02   4.173 3.01e-05 ***
## worryhealthcare      -5.098e-02  9.171e-02  -0.556 0.578334    
## worrysymptoms         5.357e-02  4.309e-02   1.243 0.213802    
## ssqmean              -2.937e-02  6.878e-02  -0.427 0.669372    
## prepreglbs            6.418e-03  1.660e-03   3.867 0.000110 ***
## familypreeclampsia   -1.055e-01  1.433e-01  -0.736 0.461704    
## income               -1.740e-02  2.341e-02  -0.743 0.457231    
## kidney1TRUE           1.097e+00  4.317e-01   2.541 0.011059 *  
## lupus1TRUE            1.355e+00  1.092e+00   1.241 0.214638
## collagen1TRUE        -8.619e-01  7.532e-01  -1.144 0.252513
## crohns1TRUE           8.775e-01  6.484e-01   1.353 0.175919
## pcos1TRUE             5.615e-01  2.762e-01   2.033 0.042074 *
## discrimination        8.732e-02  5.701e-02   1.532 0.125580
## bornearly            -2.529e-01  1.350e-01  -1.873 0.061059 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1609.0  on 5553  degrees of freedom
## Residual deviance: 1270.9  on 5525  degrees of freedom
## AIC: 1328.9
##
## Number of Fisher Scoring iterations: 7
```

```
backwards = step(fullmod, trace = 0)
summary(backwards)
```

```
##
## Call:
## glm(formula = dv.hypertension1 ~ age + race + systolic + diastolic +
##     worrysymptoms + prepreglbs + kidney1 + pcos1 + discrimination +
##     bornearly, family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4357  -0.2412  -0.1620  -0.1088   3.5674
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -15.278740   1.056563 -14.461  < 2e-16 ***
## age              0.055829   0.014501   3.850 0.000118 ***
## racehispanic    -0.896660   0.306258  -2.928 0.003414 **
## racenative      -1.551804   1.029600  -1.507 0.131762
## raceother       -1.104829   0.275278  -4.014 5.98e-05 ***
## racewhite       -0.837490   0.209449  -3.999 6.37e-05 ***
## systolic         0.061305   0.008375   7.320 2.49e-13 ***
## diastolic        0.040909   0.010341   3.956 7.62e-05 ***
## worrysymptoms    0.063910   0.036076   1.772 0.076476 .
## prepreglbs       0.006513   0.001628   4.001 6.30e-05 ***
## kidney1TRUE      1.099539   0.428573   2.566 0.010300 *
## pcos1TRUE        0.567517   0.273550   2.075 0.038020 *
## discrimination   0.090650   0.056509   1.604 0.108679
## bornearly       -0.285002   0.130037  -2.192 0.028401 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1609.0  on 5553  degrees of freedom
## Residual deviance: 1280.5  on 5540  degrees of freedom
## AIC: 1308.5
```

```
##
## Number of Fisher Scoring iterations: 7
```

```
nothing <- glm(dv.hypertension1 ~ 1, train, family = binomial)
forwards = step(nothing, trace =  0,
                scope=list(lower=formula(nothing),upper=formula(fullmod)),
                direction="forward")
summary(forwards)
```

```
##
## Call:
## glm(formula = dv.hypertension1 ~ systolic + prepreglbs + diastolic +
##     race + age + kidney1 + bornearly + pcos1 + worryfambaby +
##     discrimination, family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4717  -0.2413  -0.1614  -0.1088   3.5732
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -15.216202   1.044475 -14.568  < 2e-16 ***
## systolic        0.061229   0.008384   7.303 2.82e-13 ***
## prepreglbs      0.006651   0.001626   4.091 4.30e-05 ***
## diastolic       0.041355   0.010306   4.013 6.00e-05 ***
## racehispanic   -0.907287   0.306433  -2.961 0.003068 **
## racenative     -1.565584   1.030295  -1.520 0.128624
## raceother      -1.127447   0.276369  -4.079 4.51e-05 ***
## racewhite      -0.853459   0.209441  -4.075 4.60e-05 ***
## age             0.055162   0.014440   3.820 0.000133 ***
## kidney1TRUE     1.105035   0.426701   2.590 0.009606 **
## bornearly      -0.288030   0.129970  -2.216 0.026682 *
## pcos1TRUE       0.563157   0.273768   2.057 0.039680 *
## worryfambaby    0.111072   0.062944   1.765 0.077630 .
## discrimination  0.086584   0.056691   1.527 0.126687
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1609.0  on 5553  degrees of freedom
## Residual deviance: 1280.5  on 5540  degrees of freedom
## AIC: 1308.5
##
## Number of Fisher Scoring iterations: 7
```

```
# note: Backwards model has one extra variable.
```

```
predLOG <-  predict(fullmod, test, type = "response")
predtrainLOG <- predict(fullmod, train, type = "response")
predLOG = as.numeric(predLOG >=  0.5)
predtrainLOG = as.numeric(predtrainLOG >=  0.5)
truthTest <- ifelse(test$dv.hypertension1 == "TRUE", 1, 0)
truthTrain <- ifelse(train$dv.hypertension1 == "TRUE", 1, 0)
```

```r
table(predicted = predLOG, actual=truthTest)
```

```
##          actual
## predicted    0    1
##         0 2303   69
##         1    6    2
```

```r
testErrorLOG <- mean(predLOG!=truthTest)
testErrorLOG
```

```
## [1] 0.03151261
```

```r
table(predicted = predtrainLOG, actual = truthTrain)
```

```
##          actual
## predicted    0    1
##         0 5365  165
##         1    6   18
```

```r
trainErrorLOG <- mean(predtrainLOG != truthTrain)
trainErrorLOG
```

```
## [1] 0.03078862
```

Backwards Model:

```r
predLOG.b <-  predict(backwards, test, type = "response")
predtrainLOG.b <- predict(backwards, train, type = "response")
predLOG.b = as.numeric(predLOG.b >=  0.5)
predtrainLOG.b = as.numeric(predtrainLOG.b >=  0.5)
truthTest.b <- ifelse(test$dv.hypertension1 == "TRUE", 1, 0)
truthTrain.b <- ifelse(train$dv.hypertension1 == "TRUE", 1, 0)
```

```r
table(predicted = predLOG.b, actual=truthTest.b)
```

```
##          actual
## predicted    0    1
##         0 2302   69
##         1    7    2
```

```r
testErrorLOG.b <- mean(predLOG.b!=truthTest.b)
testErrorLOG.b
```

```
## [1] 0.03193277
```

```r
table(predicted = predtrainLOG.b, actual = truthTrain.b)
```

```
##          actual
## predicted    0    1
##         0 5365  168
##         1    6   15
```

```r
trainErrorLOG.b <- mean(predtrainLOG.b != truthTrain.b)
trainErrorLOG.b
```

```
## [1] 0.03132877
```

Forwards Model:

```r
predLOG.f <-  predict(forwards, test, type = "response")
predtrainLOG.f <- predict(forwards, train, type = "response")
predLOG.f = as.numeric(predLOG.f >=  0.5)
predtrainLOG.f = as.numeric(predtrainLOG.f >=  0.5)
truthTest.f <- ifelse(test$dv.hypertension1 == "TRUE", 1, 0)
truthTrain.f <- ifelse(train$dv.hypertension1 == "TRUE", 1, 0)
```

```r
table(predicted = predLOG.f, actual=truthTest.f)
```

```
##           actual
## predicted    0    1
##         0 2302   69
##         1    7    2
```

```r
testErrorLOG.f <- mean(predLOG.f!=truthTest.f)
testErrorLOG.f
```

```
## [1] 0.03193277
```

```r
table(predicted = predtrainLOG.f, actual = truthTrain.f)
```

```
##           actual
## predicted    0    1
##         0 5366  167
##         1    5   16
```

```r
trainErrorLOG.f <- mean(predtrainLOG.f != truthTrain.f)
trainErrorLOG.f
```

```
## [1] 0.03096867
```

All testing errors:

```r
testErrorLOG
```

```
## [1] 0.03151261
```

```r
# AIC full mod: 1328.863
fullmod$aic
```

```
## [1] 1328.863
```

```
# AIC backwards model: 1308.461
backwards$aic
```

```
## [1] 1308.461
```

```
testErrorLOG.b
```

```
## [1] 0.03193277
```

```
# AIC forwards model: 1308.489
forwards$aic
```

```
## [1] 1308.489
```

```
testErrorLOG.f
```

```
## [1] 0.03193277
```

Test Errors very similar. Going to look at lower AIC.

```
summary <- summary(backwards)
exp(summary$coefficients[,1])
```

```
##    (Intercept)            age    racehispanic     racenative      raceother
##   2.314876e-07   1.057417e+00   4.079297e-01   2.118655e-01   3.312675e-01
##      racewhite       systolic       diastolic  worrysymptoms     prepreglbs
##   4.327954e-01   1.063223e+00   1.041758e+00   1.065996e+00   1.006534e+00
##    kidney1TRUE      pcos1TRUE discrimination      bornearly
##   3.002782e+00   1.763882e+00   1.094886e+00   7.520128e-01
```

Confusion Matrices:

```
print("Backwards model: ")
```

```
## [1] "Backwards model: "
```

```
table(predicted = predLOG.b, actual=truthTest.b)
```

```
##          actual
## predicted    0    1
##         0 2302   69
##         1    7    2
```

```
testErrorLOG.b <- mean(predLOG.b!=truthTest.b)
testErrorLOG.b
```

```
## [1] 0.03193277
```

```r
print("Full model: ")
```

```
## [1] "Full model: "
```

```r
table(predicted = predLOG, actual=truthTest)
```

```
##           actual
## predicted    0    1
##         0 2303   69
##         1    6    2
```

```r
testErrorLOG <- mean(predLOG!=truthTest)
testErrorLOG
```

```
## [1] 0.03151261
```

```r
print("Forwards model: ")
```

```
## [1] "Forwards model: "
```

```r
table(predicted = predLOG.f, actual=truthTest.f)
```

```
##           actual
## predicted    0    1
##         0 2302   69
##         1    7    2
```

```r
testErrorLOG.f <- mean(predLOG.f!=truthTest.f)
testErrorLOG.f
```

```
## [1] 0.03193277
```

AUC:

```r
auc(test$dv.hypertension1, predLOG.f)
```

```
## Setting levels: control = FALSE, case = TRUE
```

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

```
## Area under the curve: 0.5126
```

```r
auc(test$dv.hypertension1, predLOG.b)
```

```
## Setting levels: control = FALSE, case = TRUE
## Setting direction: controls < cases
```

```
## Area under the curve: 0.5126
```

```r
auc(test$dv.hypertension1, predLOG)
```

```
## Setting levels: control = FALSE, case = TRUE
## Setting direction: controls < cases
```

```
## Area under the curve: 0.5128
```

```r
test$dv.hypertension1 <- ifelse(test$dv.hypertension1 == "TRUE", 1, 0)
conf_mat = table("truth" = test$dv.hypertension1, "pred"  = predLOG)
conf_mat = confusionMatrix(conf_mat, mode = "everything", positive = "1")
conf_mat$byClass
```

```
##          Sensitivity          Specificity       Pos Pred Value
##         0.2500000000         0.9709106239         0.0281690141
##       Neg Pred Value            Precision               Recall
##         0.9974014725         0.0281690141         0.2500000000
##                   F1           Prevalence       Detection Rate
##         0.0506329114         0.0033613445         0.0008403361
## Detection Prevalence    Balanced Accuracy
##         0.0298319328         0.6104553120
```

```r
conf_mat
```

```
## Confusion Matrix and Statistics
##
##      pred
## truth    0    1
##     0 2303    6
##     1   69    2
##
##                Accuracy : 0.9685
##                  95% CI : (0.9607, 0.9751)
##     No Information Rate : 0.9966
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0449
##
##  Mcnemar’s Test P-Value : 8.118e-13
##
##             Sensitivity : 0.2500000
##             Specificity : 0.9709106
##          Pos Pred Value : 0.0281690
##          Neg Pred Value : 0.9974015
##               Precision : 0.0281690
##                  Recall : 0.2500000
##                      F1 : 0.0506329
##              Prevalence : 0.0033613
##          Detection Rate : 0.0008403
##    Detection Prevalence : 0.0298319
##       Balanced Accuracy : 0.6104553
##
##        ’Positive’ Class : 1
##
```

```r
conf_mat.b = table("truth" = test$dv.hypertension1, "pred"  = predLOG.b)
conf_mat.b = confusionMatrix(conf_mat.b, mode = "everything", positive = "1")
conf_mat.b$byClass
```

```
##          Sensitivity          Specificity       Pos Pred Value
##         0.2222222222         0.9708983551         0.0281690141
##       Neg Pred Value            Precision               Recall
##         0.9969683846         0.0281690141         0.2222222222
##                   F1           Prevalence       Detection Rate
##         0.0500000000         0.0037815126         0.0008403361
## Detection Prevalence    Balanced Accuracy
##         0.0298319328         0.5965602887
```

```r
conf_mat.b
```

```
## Confusion Matrix and Statistics
##
##      pred
## truth    0    1
##     0 2302    7
##     1   69    2
##
##                Accuracy : 0.9681
##                  95% CI : (0.9602, 0.9748)
##     No Information Rate : 0.9962
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0436
##
##  Mcnemar's Test P-Value : 2.612e-12
##
##             Sensitivity : 0.2222222
##             Specificity : 0.9708984
##          Pos Pred Value : 0.0281690
##          Neg Pred Value : 0.9969684
##               Precision : 0.0281690
##                  Recall : 0.2222222
##                      F1 : 0.0500000
##              Prevalence : 0.0037815
##          Detection Rate : 0.0008403
##    Detection Prevalence : 0.0298319
##       Balanced Accuracy : 0.5965603
##
##        'Positive' Class : 1
##
```