consulting project_modeling

2022-11-10

(1) LASSO

a. model

```
library(glmnet)
MUA_data2 = MUA_data1[,c(109,108,6,9,75:78, 80:89,91:104,106,116,118)]
MUA_data2$ASA_C_TKA<-as.factor(MUA_data2$ASA_C_TKA)
y = MUA_data2$C_MUA_bi
# Fac_age_C_TKA(116) is included instead of age_C_TKA(74), Diabetes_cc_C_TKA(90) is removed b/c it's duplicated with Diabetes_no_cc_C_TKA
# 108: MUA_bi,  6 : sex, 9: ethnicity, 75: bmi_C_TKA, 76: tobacco_C_TKA,  77: Insurance_C_TKA, 78: los_C_TKA, 80:ASA_C_TKA, 81: op_time_C_TKA, 82~103: comorbidities, 104:readmit_90d_C_TKA, 106:ed_90d_C_TKA,116:Fac_age_C_TKA, 118: redu_race

d <- as.data.frame(MUA_data2)
options(na.action="na.pass")
m <- model.matrix(C_MUA_bi ~ ., data=d)[,-1]


set.seed(1234)
cv.out = cv.glmnet(m, y, alpha=1,family="binomial", nfolds = nrow(d))
bestlam = cv.out$lambda.min


lasso.mod = glmnet(m, y, alpha= 1, lambda = bestlam,family="binomial")
coef(lasso.mod)
```

```
## 45 x 1 sparse Matrix of class "dgCMatrix"
##                                     s0
## (Intercept)                  -4.663227
## MUA_bi                        1.359357
## sexM                          .       
## ethnicityNon-Hispanic Origin  .       
## bmi_C_TKA                     .       
## tobacco_C_TKAPassive          .       
## tobacco_C_TKAQuit             .       
## tobacco_C_TKAYes              .       
## Insurance_C_TKAMedicare       .       
## Insurance_C_TKAPrivate        .       
## Insurance_C_TKAUninsured      .       
## Insurance_C_TKAWork_Comp      1.233295
## los_C_TKA                     .       
## ASA_C_TKA2                    .       
## ASA_C_TKA3                    .       
## ASA_C_TKA4                    .       
## op_time_C_TKA                 .       
## blood_transfusion_C_TKA       2.117675
## platelet_transfusion_C_TKA    .       
## AIDS_C_TKA                    .       
## Malignancy_C_TKA              .       
## Cerebrovascular_C_TKA         .       
## COPD_C_TKA                    .       
## CHF_C_TKA                     .       
## Dementia_C_TKA                .       
## Diabetes_no_cc_C_TKA          .       
## Hemiplegia_C_TKA              .       
## Metastatic_C_TKA              .       
## Mild_Liver_C_TKA              .       
## Moderate_Liver_C_TKA          .       
## MI_C_TKA                      .       
## Peptic_Ulcer_C_TKA            .       
## PVD_C_TKA                     .       
## CKD_C_TKA                     .       
## Rheumatic_C_TKA               .       
## hematoma_C_TKA                .       
## wound_infection_C_TKA         .       
## knee_infection_C_TKA          .       
## readmit_90d_C_TKA             .       
## ed_90d_C_TKA                  .       
## Fac_age_C_TKA60s              .       
## Fac_age_C_TKAless50           .       
## Fac_age_C_TKAover70s          .       
## redu_raceOther                .       
## redu_raceWhite                .
```

```
paste0(round(coef(lasso.mod)@x,4),"X",coef(lasso.mod)@i, collapse=" + ")
```

```
## [1] "-4.6632X0 + 1.3594X1 + 1.2333X11 + 2.1177X17"
```

```
##### ROC
#lasso.mod = lasso mod using bestlam from cv.glmnet
#m = original model matrix
predict_fit = predict(lasso.mod, m, type = "response")

library(pROC)
```

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

```
##
## Attaching package: 'pROC'
```

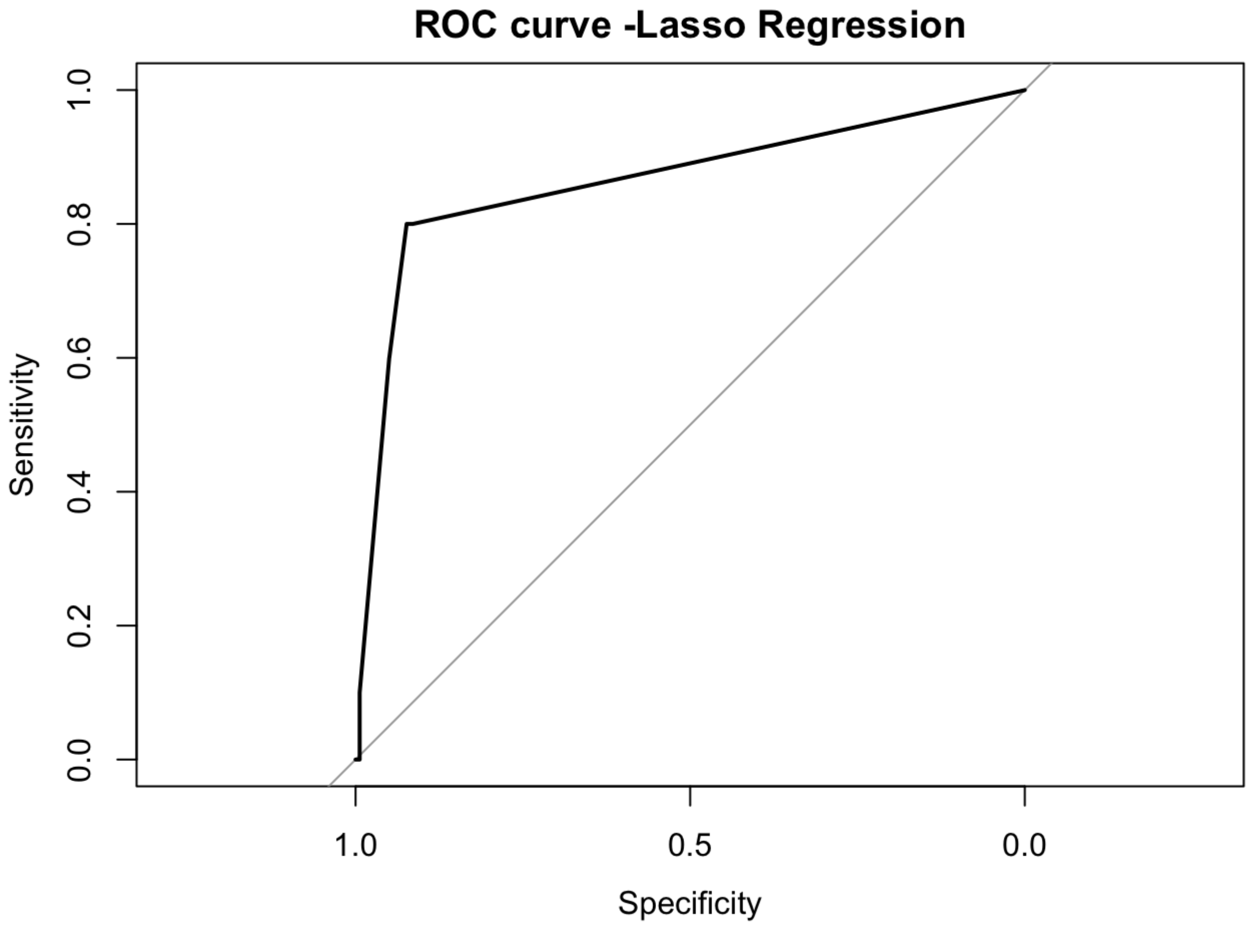
```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
roc_score=roc(predictor=as.vector(predict_fit), response = y )
```

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

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

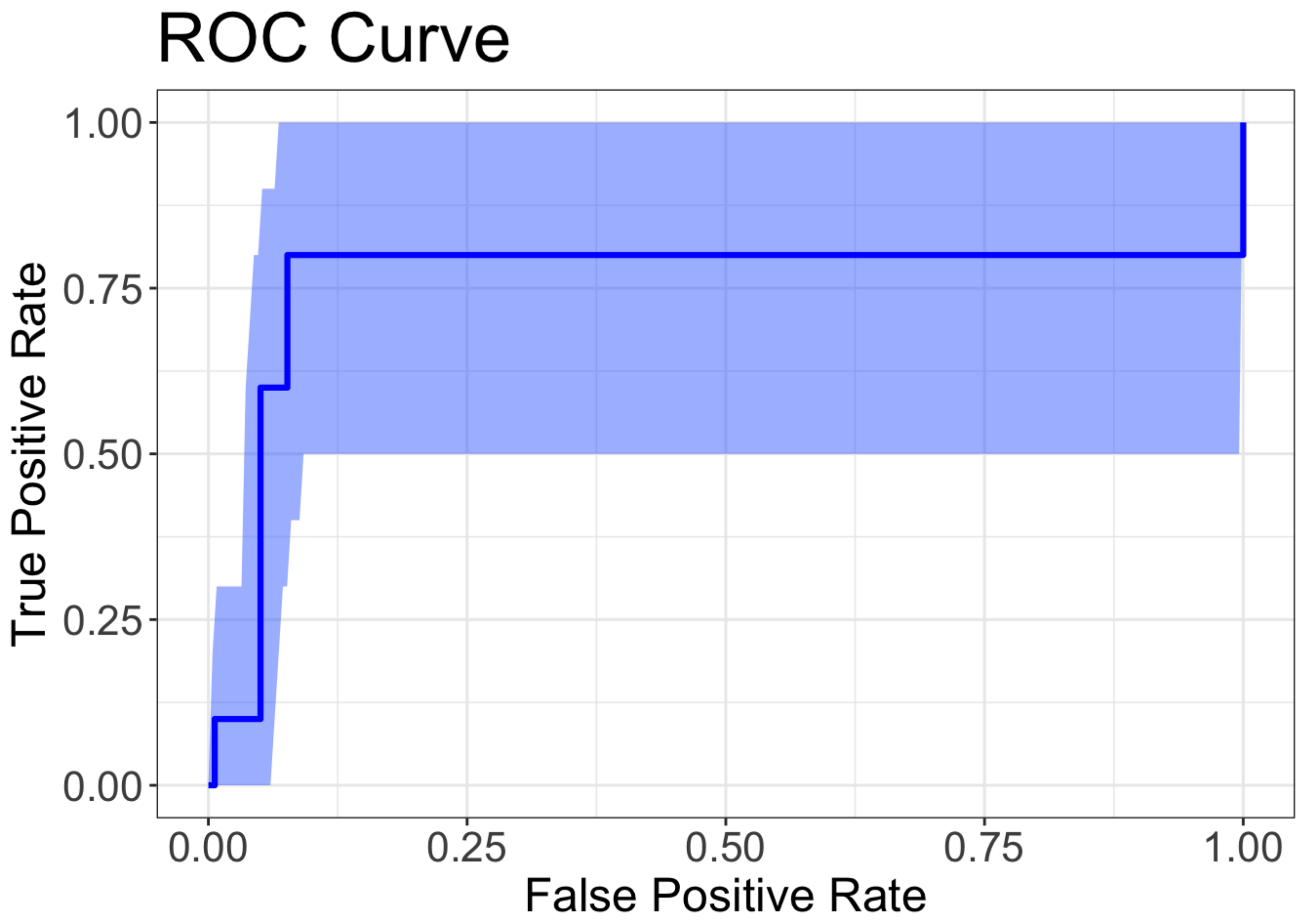
```
#ROC Plot
plot(roc_score ,main ="ROC curve -Lasso Regression")
```



```
roc_score$auc #AUC score
```

```
## Area under the curve: 0.864
```

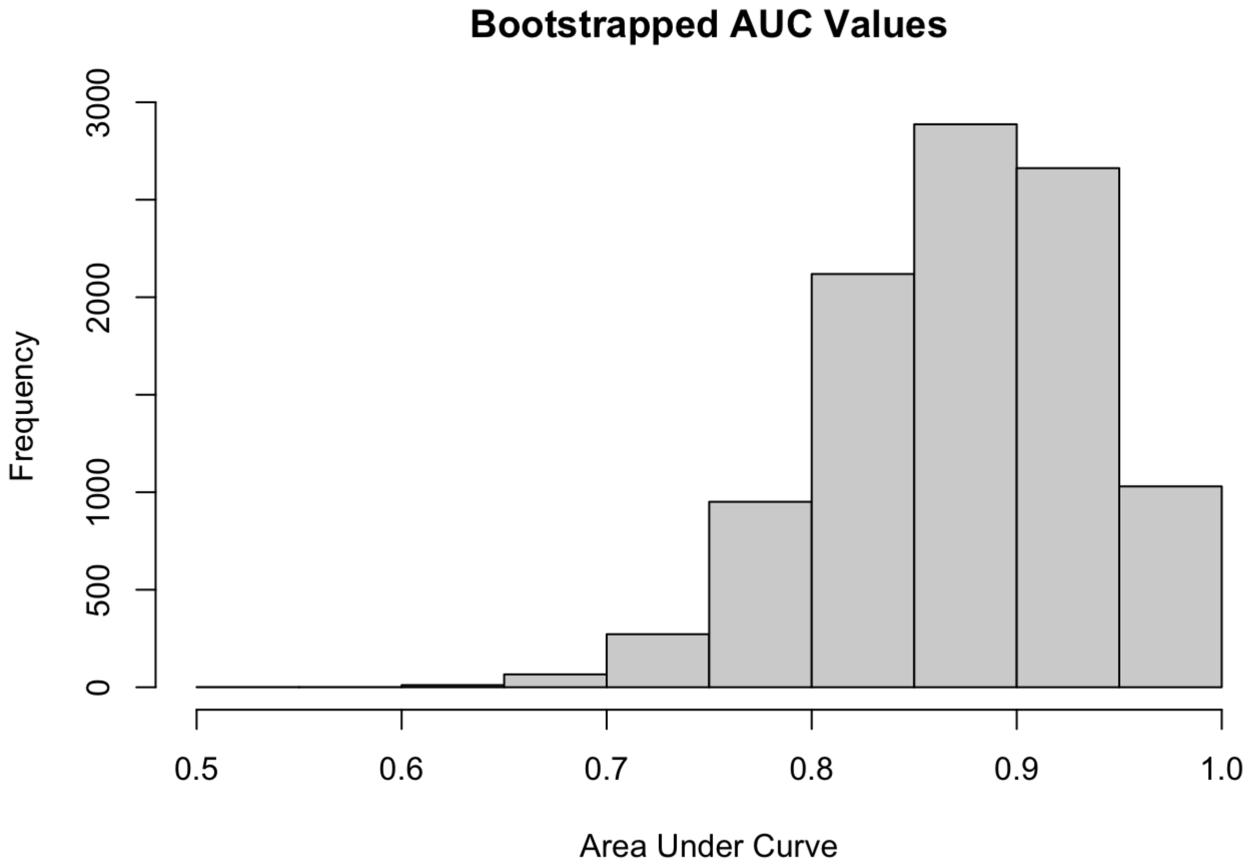
```
##### Bootstrapping AUC to find 95% CI
#Finding Bootstrapped AUC confidence intervals
library(fbroc)
boot_roc = boot.roc(as.numeric(predict_fit), as.logical(as.numeric(y)), n.boot = 10000)
plot(boot_roc)
```



```r
w=perf(boot_roc, "auc") #Measuring performance
w #AUC confidence interval
```

```
##
##
##                    Bootstrapped ROC performance metric
##
## Metric: AUC
## Bootstrap replicates: 10000
## Observed: 0.864
## Std. Error: 0.065
## 95% confidence interval:
## 0.716 0.972
```

```r
hist(w$boot.results, main = "Bootstrapped AUC Values", xlab = "Area Under Curve") #Histo
gram of bootstrapped AUC
```



b. leave one out cross validation

```
#library(glmnet)
#y = MUA_data1$C_MUA_bi
#MUA_data2 = MUA_data1[,c(108, 6, 9,75:78, 80:89,91:104,106,116,118)] # Fac_age_C_TKA(11
6) is included instead of age_C_TKA(74), Diabetes_cc_C_TKA(90) is removed b/c it's dupli
cated with Diabetes_no_cc_C_TKA
# 108: MUA_bi,  6 : sex, 9: ethnicity, 75: bmi_C_TKA, 76: tobacco_C_TKA,  77: Insurance_
C_TKA, 78: los_C_TKA, 80:ASA_C_TKA, 81: op_time_C_TKA, 82~103: comorbidities, 104:readmi
t_90d_C_TKA, 106:ed_90d_C_TKA,116:Fac_age_C_TKA, 118: redu_race
#MUA_data2$ASA_C_TKA<-as.factor(MUA_data2$ASA_C_TKA)

#d <- data.frame(x=MUA_data2, y=y)
#m <- model.matrix(y ~ ., data=d)[,-1]

pred_prob <- c()
for(i in 1:nrow(d)){   #nrow(d)
y_train = d[-i,]$C_MUA_bi
x_train = m[-i,]

y_test = d[i,]$C_MUA_bi
x_test = m[i,]

set.seed(1234)
cv.out = cv.glmnet(x_train, y_train, alpha =1,family="binomial")
bestlam = cv.out$lambda.min

lasso.mod = glmnet(x_train, y_train, alpha= 1, lambda = bestlam,family="binomial")
pred_prob[i] <- predict(lasso.mod,s = bestlam, newx = x_test,type="response")
}

pred_C_MUA<-cbind(pred_prob,y)
pred_C_MUA[y==1,]
```

```
##         pred_prob y
##  [1,] 0.048048738 1
##  [2,] 0.050905980 1
##  [3,] 0.046554476 1
##  [4,] 0.006625576 1
##  [5,] 0.007843469 1
##  [6,] 0.009362077 1
##  [7,] 0.010370889 1
##  [8,] 0.048875925 1
##  [9,] 0.050949329 1
## [10,] 0.500000000 1
```

```
mean(pred_C_MUA[y==0,1])
```

```
## [1] 0.01514417
```

```
mean(pred_C_MUA[y==1,1])
```

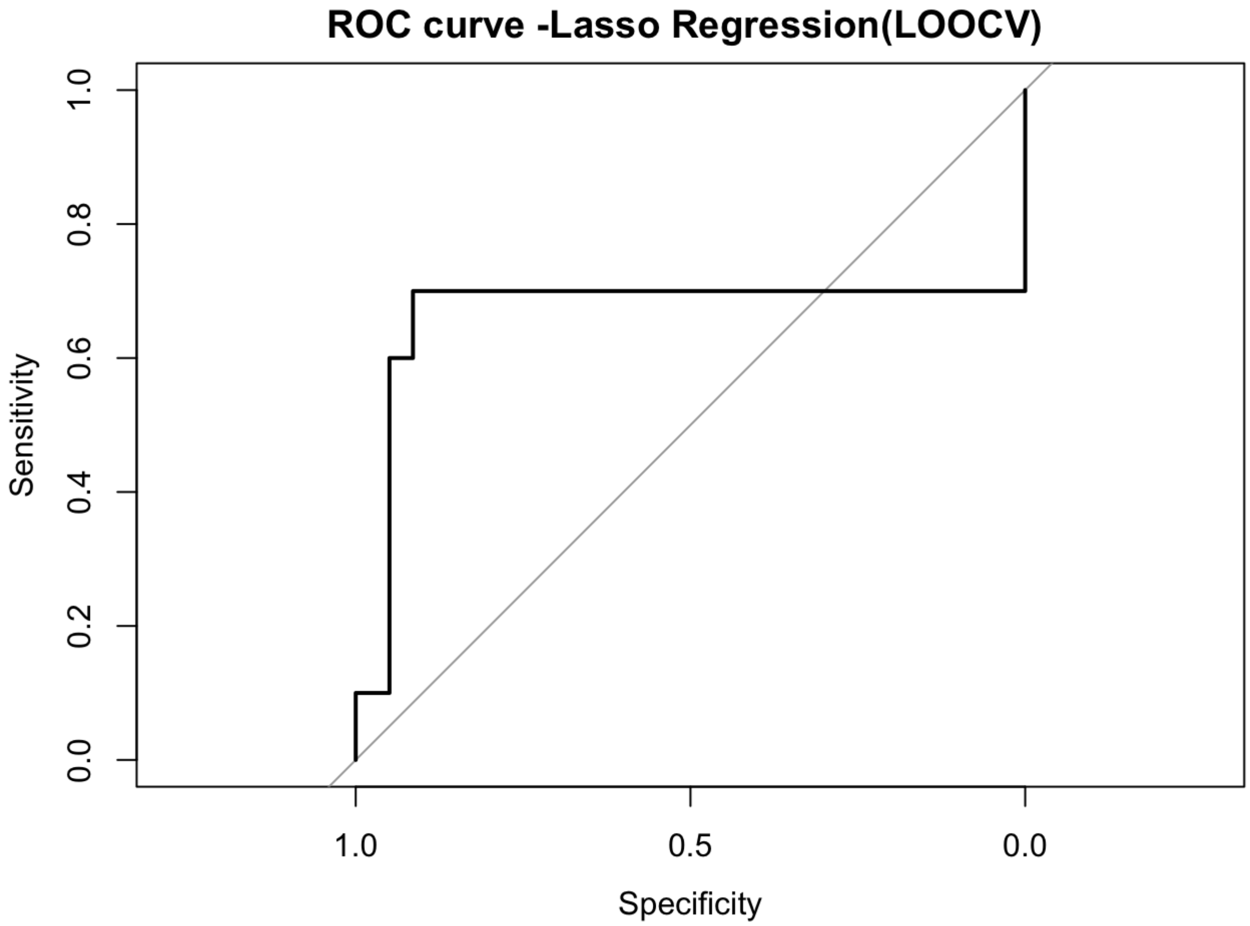
```
## [1] 0.07795365
```

```
library(pROC)
roc_score_predictpower=roc(predictor=as.vector(pred_prob), response = y )
```

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

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

```
#ROC Plot
plot(roc_score_predictpower ,main ="ROC curve -Lasso Regression(LOOCV) ")
```



```
roc_score_predictpower$auc #AUC score
```

```
## Area under the curve: 0.6663
```