Midterm\_test\_part2.R

DavidAarhus

2020-04-16

rm(list = ls()) #removing all variables  
library("ISLR")  
library("ggplot2")  
library("corrplot")

## corrplot 0.84 loaded

library("caret")

## Loading required package: lattice

library("tidyverse")

## ── Attaching packages ──────────────────────────────────────────────────── tidyverse 1.3.0 ──

## ✓ tibble 2.1.3 ✓ dplyr 0.8.3  
## ✓ tidyr 1.0.0 ✓ stringr 1.4.0  
## ✓ readr 1.3.1 ✓ forcats 0.5.0  
## ✓ purrr 0.3.3

## ── Conflicts ─────────────────────────────────────────────────────── tidyverse\_conflicts() ──  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()  
## x purrr::lift() masks caret::lift()

library("leaps")  
library("doBy")  
library("glmnet")

## Loading required package: Matrix

##   
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':  
##   
## expand, pack, unpack

## Loaded glmnet 3.0-2

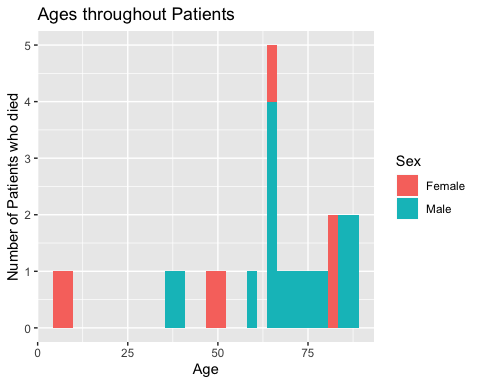
library("glmnetUtils")

##   
## Attaching package: 'glmnetUtils'

## The following objects are masked from 'package:glmnet':  
##   
## cv.glmnet, glmnet

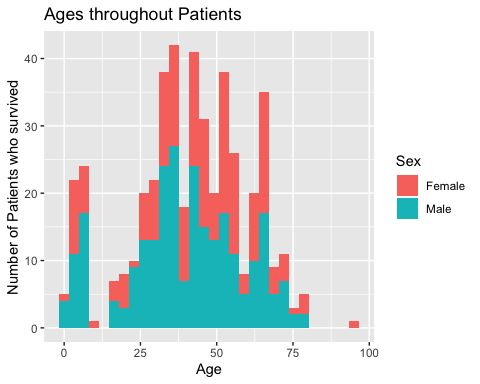
library("plotROC")  
  
covid\_train <- read.csv("/Users/DavidAarhus/Documents/310 R/Datasets/patients\_train\_A.csv")  
covid\_test <- read.csv("/Users/DavidAarhus/Documents/310 R/Datasets/patients\_test\_A.csv")  
  
#Question 2  
# Create histograms of Age for patients who died and survived.   
# Points will be given for creativity and visual aesthetics.   
# Do you see any patterns?  
  
covid\_deathnum <- covid\_train$Death==1  
covid\_death <- covid\_train[covid\_deathnum, ]  
  
covid\_survivenum <- covid\_train$Death==0  
covid\_survive <- covid\_train[covid\_survivenum, ]  
  
ggplot(covid\_death, aes(covid\_death$Age)) + geom\_histogram(aes(fill = Sex)) +   
 labs(x = "Age",  
 y = "Number of Patients who died",  
 title = "Ages throughout Patients")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



ggplot(covid\_survive, aes(covid\_survive$Age)) + geom\_histogram(aes(fill = Sex)) +   
 labs(x = "Age",  
 y = "Number of Patients who survived",  
 title = "Ages throughout Patients")

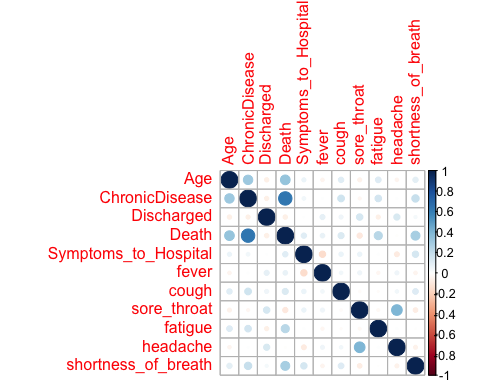
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



#Question 3  
# Print the correlation matrix for all numeric variables.   
# (use="complete.obs" option in cor() function may be used.)   
# Pick one variable and explain its correlation with Death and   
# explain why this correlation makes sense.  
  
cormat <- cor(covid\_train %>% select\_if(is.numeric) %>% drop\_na())  
print(cormat[,"Death"])

## Age ChronicDisease Discharged   
## 0.332304626 0.639732727 -0.068583171   
## Death Symptoms\_to\_Hospital fever   
## 1.000000000 0.100227632 0.073554381   
## cough sore\_throat fatigue   
## 0.129130241 -0.096817067 0.232307577   
## headache shortness\_of\_breath   
## -0.007022323 0.283407621

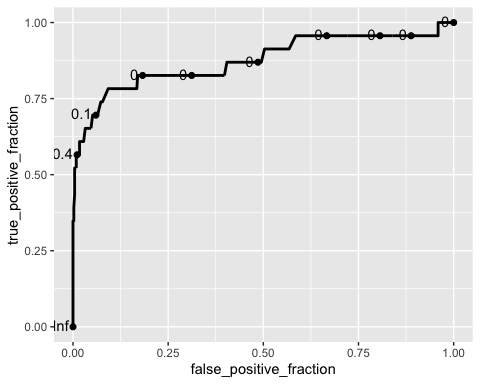
corrplot(cormat)



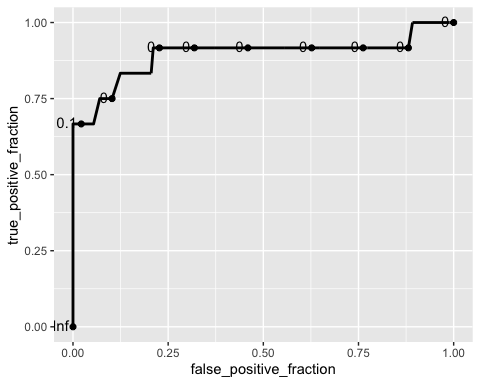
#Question 4  
# Estimate a logistic regression model to   
# predict Death as a function of age, sex, and chronic disease.   
# Be sure to copy the summary of your model to your output word file.   
logit\_fit <- glm(Death ~ Age + Sex + ChronicDisease,  
 family = binomial,  
 data = covid\_train)  
summary(logit\_fit)

##   
## Call:  
## glm(formula = Death ~ Age + Sex + ChronicDisease, family = binomial,   
## data = covid\_train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.7056 -0.2234 -0.1686 -0.1301 3.4234   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -6.16167 0.99750 -6.177 6.53e-10 \*\*\*  
## Age 0.04354 0.01596 2.727 0.00639 \*\*   
## SexMale 0.26341 0.58251 0.452 0.65113   
## ChronicDisease 3.90860 0.60697 6.440 1.20e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 185.42 on 487 degrees of freedom  
## Residual deviance: 108.77 on 484 degrees of freedom  
## AIC: 116.77  
##   
## Number of Fisher Scoring iterations: 7

#Question 5  
# In a sentence, interpret the effect of "ChronicDisease" in the model above.  
  
#Question 6  
# Generate predicted scores from this model and use them to produce   
# ROC plots for the test and training sets data.  
  
preds\_train <- data.frame(scores = predict(logit\_fit, type = "response"), covid\_train)  
preds\_train <- preds\_train %>% mutate(class\_pred05 = ifelse(scores>0.5,1,0))  
  
preds\_test <- data.frame(scores = predict(logit\_fit, newdata = covid\_test, type = "response"), covid\_test)  
preds\_test <- preds\_test %>% mutate(class\_pred05 = ifelse(scores>0.5,1,0))  
  
train\_ROC <- ggplot(preds\_train, aes(m = scores, d = Death)) +  
 geom\_roc()  
train\_ROC



test\_ROC <- ggplot(preds\_test, aes(m = scores, d = Death)) +  
 geom\_roc()  
test\_ROC



#Question 7  
# How good is this model? How do the ROC plots suggest we   
# should alter our cutoff probabilities when assigning predicted classes?  
calc\_auc(train\_ROC)

## PANEL group AUC  
## 1 1 -1 0.8748948

calc\_auc(test\_ROC)

## PANEL group AUC  
## 1 1 -1 0.8941441

#Question 8  
# Use a reasonable cutoff probability (justified by ROC plot),   
# generate class predictions.   
# Copy the confusion matrices in the output file and calculate sensitivity   
# and specificity for the model.  
  
preds\_trainm <- data.frame(scores = predict(logit\_fit, type = "response"), covid\_train)  
preds\_trainm <- data.frame(class\_preds05 = ifelse(preds\_train$scores > 0.5, 1, 0), preds\_trainm)  
  
preds\_testm <- data.frame(scores = predict(logit\_fit, newdata = covid\_test, type = "response"), covid\_test)  
preds\_testm <- data.frame(class\_preds05 = ifelse(preds\_test$scores > 0.5, 1, 0), preds\_testm)  
  
#Train confusion matrix  
table(preds\_trainm$Death,preds\_trainm$class\_preds05)

##   
## 0 1  
## 0 461 4  
## 1 11 12

#Test confusion matrix  
table(preds\_testm$Death, preds\_testm$class\_preds05)

##   
## 0 1  
## 0 185 0  
## 1 6 6

#Question 9  
# Suppose you are an expert in health data science.   
# What is your main takeaway from the models above?   
# What are your suggestions to improve the prediction power?