Predicting Heart Disease Risk Using Clinical Variables

Autumn Brinkerhoff

This analysis aims to use logistic regression to predict Heart disease using the listed variables.

```
### We will need for this analysis
library(tidyverse)
```

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

## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.0     v forcats 0.5.1

## Warning: package 'tidyr' was built under R version 4.0.5

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

```
library(caret)
```

```
## Loading required package: lattice

##
## Attaching package: 'caret'

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

```
library(glmnet)
```

```
## Loading required package: Matrix

##
## Attaching package: 'Matrix'

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

```
## Loaded glmnet 4.1
```

```
library(pwr)
library(mgcv)
```

```
## Loading required package: nlme

##
## Attaching package: 'nlme'

## The following object is masked from 'package:dplyr':
##
##     collapse

## This is mgcv 1.8-33. For overview type 'help("mgcv-package")'.
```

```
library(ggplot2)
library(pROC)
```

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

##
## Attaching package: 'pROC'

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

Import the data

```
df <- read.csv(file = 'Heart_Disease_Prediction.csv')

head(df,n=5)
```

```
##   index Age Sex Chest.pain.type  BP Cholesterol FBS.over.120 EKG.results Max.HR
## 1     0  70   1               4 130         322            0           2    109
## 2     1  67   0               3 115         564            0           2    160
## 3     2  57   1               2 124         261            0           0    141
## 4     3  64   1               4 128         263            0           0    105
## 5     4  74   0               2 120         269            0           2    121
##   Exercise.angina ST.depression Slope.of.ST Number.of.vessels.fluro Thallium
## 1               0           2.4           2                      3        3
## 2               0           1.6           2                      0        7
## 3               0           0.3           1                      0        7
## 4               1           0.2           2                      1        7
## 5               1           0.2           1                      1        3
##   Heart.Disease
## 1      Presence
## 2       Absence
## 3      Presence
## 4       Absence
## 5       Absence
```

```
df$Heart.Disease <-ifelse(df$Heart.Disease=="Presence",1,0)
head(df,n=5)
```

```
##   index Age Sex Chest.pain.type  BP Cholesterol FBS.over.120 EKG.results Max.HR
## 1     0  70   1               4 130         322            0           2    109
## 2     1  67   0               3 115         564            0           2    160
## 3     2  57   1               2 124         261            0           0    141
## 4     3  64   1               4 128         263            0           0    105
## 5     4  74   0               2 120         269            0           2    121
##   Exercise.angina ST.depression Slope.of.ST Number.of.vessels.fluro Thallium
## 1               0           2.4           2                      3        3
## 2               0           1.6           2                      0        7
## 3               0           0.3           1                      0        7
## 4               1           0.2           2                      1        7
## 5               1           0.2           1                      1        3
##   Heart.Disease
## 1             1
## 2             0
## 3             1
## 4             0
## 5             0
```

Data inspection & clean

```
summary(df)
```

```
##      index             Age             Sex         Chest.pain.type
##  Min.   :  0.00   Min.   :29.00   Min.   :0.0000   Min.   :1.000
##  1st Qu.: 67.25   1st Qu.:48.00   1st Qu.:0.0000   1st Qu.:3.000
##  Median :134.50   Median :55.00   Median :1.0000   Median :3.000
##  Mean   :134.50   Mean   :54.43   Mean   :0.6778   Mean   :3.174
##  3rd Qu.:201.75   3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:4.000
##  Max.   :269.00   Max.   :77.00   Max.   :1.0000   Max.   :4.000
##        BP          Cholesterol     FBS.over.120     EKG.results
##  Min.   : 94.0   Min.   :126.0   Min.   :0.0000   Min.   :0.000
##  1st Qu.:120.0   1st Qu.:213.0   1st Qu.:0.0000   1st Qu.:0.000
##  Median :130.0   Median :245.0   Median :0.0000   Median :2.000
##  Mean   :131.3   Mean   :249.7   Mean   :0.1481   Mean   :1.022
##  3rd Qu.:140.0   3rd Qu.:280.0   3rd Qu.:0.0000   3rd Qu.:2.000
##  Max.   :200.0   Max.   :564.0   Max.   :1.0000   Max.   :2.000
##      Max.HR      Exercise.angina  ST.depression   Slope.of.ST
##  Min.   : 71.0   Min.   :0.0000   Min.   :0.00   Min.   :1.000
##  1st Qu.:133.0   1st Qu.:0.0000   1st Qu.:0.00   1st Qu.:1.000
##  Median :153.5   Median :0.0000   Median :0.80   Median :2.000
##  Mean   :149.7   Mean   :0.3296   Mean   :1.05   Mean   :1.585
##  3rd Qu.:166.0   3rd Qu.:1.0000   3rd Qu.:1.60   3rd Qu.:2.000
##  Max.   :202.0   Max.   :1.0000   Max.   :6.20   Max.   :3.000
##  Number.of.vessels.fluro    Thallium     Heart.Disease
##  Min.   :0.0000          Min.   :3.000   Min.   :0.0000
##  1st Qu.:0.0000          1st Qu.:3.000   1st Qu.:0.0000
##  Median :0.0000          Median :3.000   Median :0.0000
```

```
##  Mean   :0.6704          Mean   :4.696   Mean   :0.4444
##  3rd Qu.:1.0000          3rd Qu.:7.000   3rd Qu.:1.0000
##  Max.   :3.0000          Max.   :7.000   Max.   :1.0000
```

```
sample_n(df, 3)
```

```
##   index Age Sex Chest.pain.type  BP Cholesterol FBS.over.120 EKG.results Max.HR
## 1   102  49   0               4 130         269            0           0    163
## 2   181  56   0               4 134         409            0           2    150
## 3   124  54   1               3 125         273            0           2    152
##   Exercise.angina ST.depression Slope.of.ST Number.of.vessels.fluro Thallium
## 1               0           0.0           1                       0        3
## 2               1           1.9           2                       2        7
## 3               0           0.5           3                       1        3
##   Heart.Disease
## 1             0
## 2             1
## 3             0
```

Partition the data 80% training and 20% Test

```
set.seed(123)
training.samples <- df$Heart.Disease %>%
  createDataPartition(p = 0.8, list = FALSE)

train.data <- df[training.samples, ]
test.data <- df[-training.samples, ]
```

Logistic Regression

Using Generalized Linear Model(GLM)

```
model <- glm(Heart.Disease ~., data = train.data, family = binomial)

summary(model)
```

```
##
## Call:
## glm(formula = Heart.Disease ~ ., family = binomial, data = train.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5748  -0.4770  -0.1547   0.3844   2.4953
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -9.8729577  3.4497236  -2.862 0.004210 **
## index                  -0.0007223  0.0028122  -0.257 0.797307
```

```
## Age                      -0.0183268  0.0290462  -0.631 0.528071
## Sex                       1.4782197  0.6178922   2.392 0.016740 *
## Chest.pain.type           0.5537920  0.2416921   2.291 0.021945 *
## BP                        0.0293383  0.0129706   2.262 0.023703 *
## Cholesterol               0.0138029  0.0053877   2.562 0.010410 *
## FBS.over.120             -0.5776730  0.6643345  -0.870 0.384546
## EKG.results               0.3813842  0.2250396   1.695 0.090124 .
## Max.HR                   -0.0252001  0.0119909  -2.102 0.035588 *
## Exercise.angina           0.6999530  0.4855330   1.442 0.149410
## ST.depression             0.2295083  0.2654397   0.865 0.387239
## Slope.of.ST               0.8576083  0.4695956   1.826 0.067810 .
## Number.of.vessels.fluro   1.0355024  0.3124193   3.314 0.000918 ***
## Thallium                  0.3423278  0.1260026   2.717 0.006591 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 297.59  on 215  degrees of freedom
## Residual deviance: 140.24  on 201  degrees of freedom
## AIC: 170.24
##
## Number of Fisher Scoring iterations: 6
```

In the model, by looking at the p-value, the most significant impact variables are the Number.of.vessels.fluro, and Thallium.

Evaluating the model

```
probabilities <- model %>%
  predict(test.data, type = "response")
predicted.classes <- ifelse(probabilities > 0.5, "1", "0")

observed.classes <- test.data$Heart.Disease

predictions <- predict(model, test.data)
prediction.probabilities <- predictions


accuracy <- mean(predicted.classes == test.data$Heart.Disease)

accuracy
```

```
## [1] 0.8148148
```

```
### The accuracy (Measure of total error)
```

81% accuracy is good but the accuracy is not the best metric for evaluating how a model performs. we are also going to look at the confusionMatrix

confusionMatrix

```
table(observed.classes, predicted.classes)
```

```
##                 predicted.classes
## observed.classes  0  1
##                0 28  4
##                1  6 16
```

```
confusionMatrix(as.factor(predicted.classes), as.factor(observed.classes))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 28  6
##          1  4 16
##
##                Accuracy : 0.8148
##                  95% CI : (0.6857, 0.9075)
##     No Information Rate : 0.5926
##     P-Value [Acc > NIR] : 0.0004365
##
##                   Kappa : 0.611
##
##  Mcnemar's Test P-Value : 0.7518296
##
##             Sensitivity : 0.8750
##             Specificity : 0.7273
##          Pos Pred Value : 0.8235
##          Neg Pred Value : 0.8000
##              Prevalence : 0.5926
##          Detection Rate : 0.5185
##    Detection Prevalence : 0.6296
##       Balanced Accuracy : 0.8011
##
##        'Positive' Class : 0
##
```

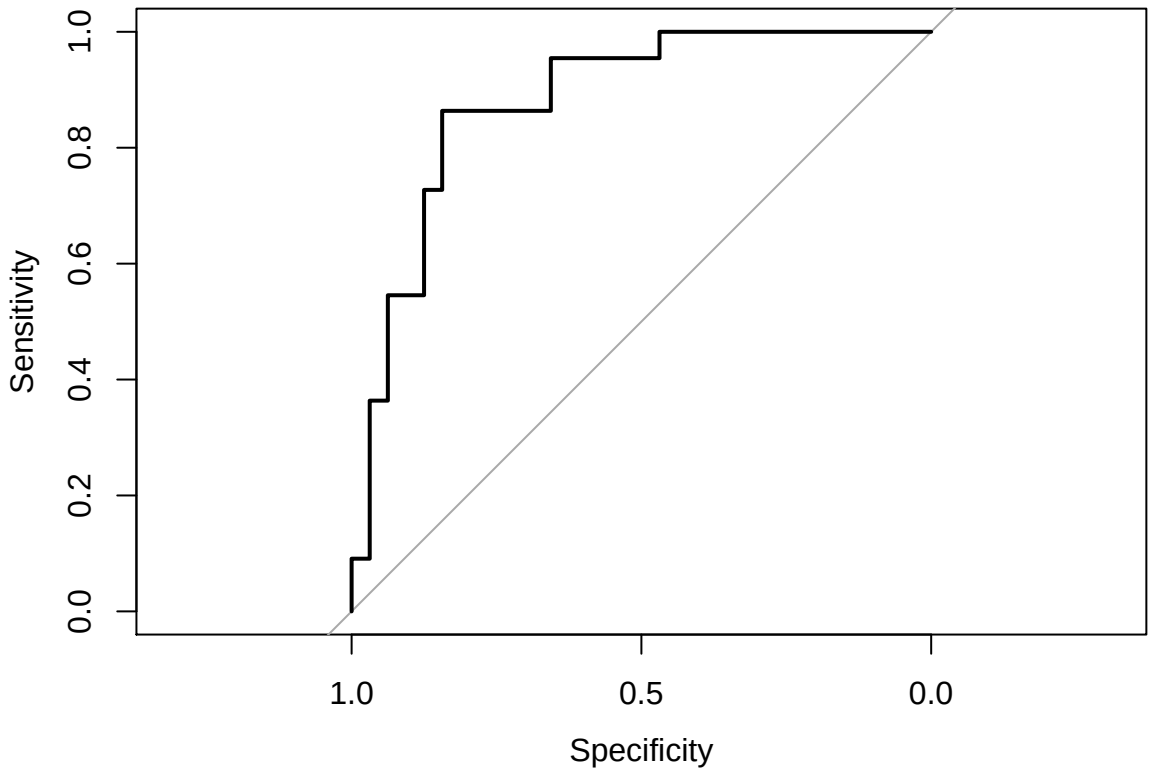
```
## True Positive: 28
## False Negative:: 6
## False Positive: 4
## True Negative: 16

## Recall: 0.82
## Precision: 0.88
## F-score: 0.8489
```

ROC

```
plot(roc(as.numeric(observed.classes), as.numeric(prediction.probabilities)))
```

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

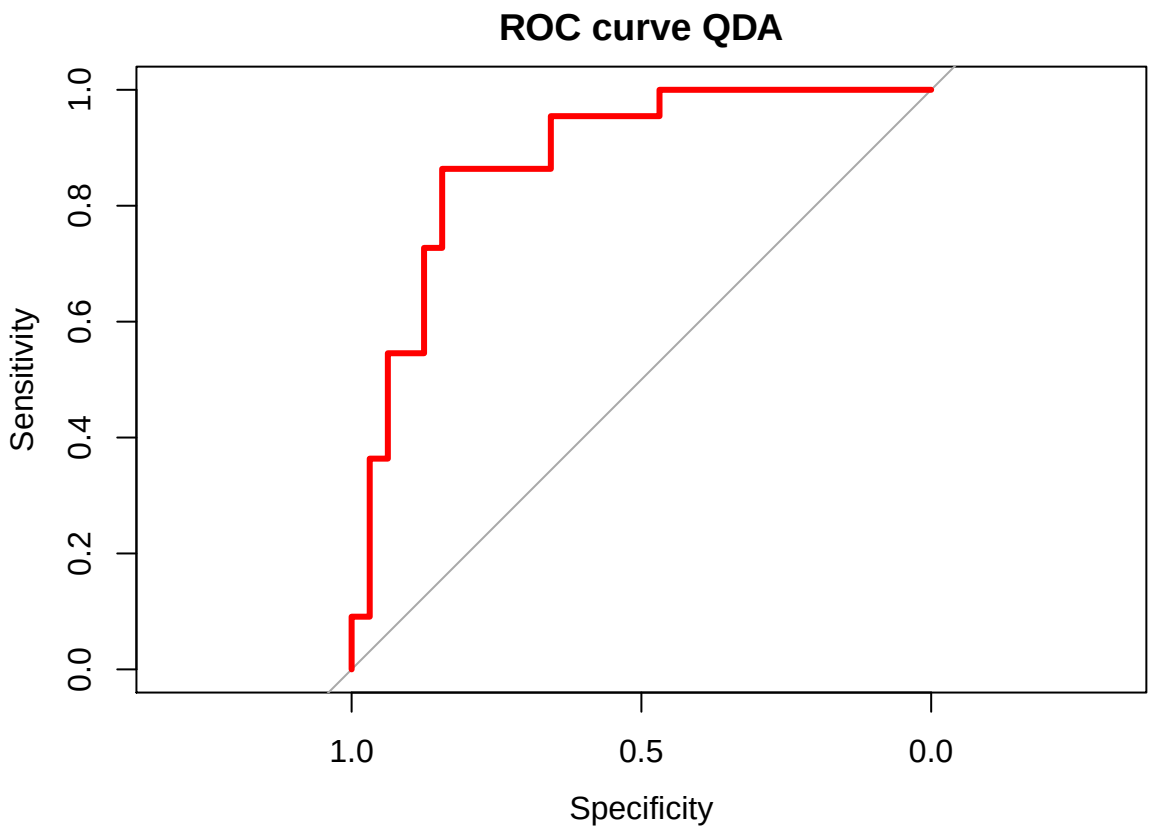
```
## Setting direction: controls < cases
```




```
roc_qda <- roc(response = as.numeric(observed.classes), predictor = as.numeric(prediction.probabilities)
```

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

```
plot(roc_qda, col="red", lwd=3, main="ROC curve QDA")
```




```
auc(roc_qda)
```

```
## Area under the curve: 0.8807
```