STA 5167 Final Project

Brianna Alston and Lucas Nolting

2023-03-20

# Reading in the Dataset

| age | sex | cp | trestbps | chol | fbs | restecg | thalach | exang | oldpeak | slope | ca | thal | target |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 52 | 1 | 0 | 125 | 212 | 0 | 1 | 168 | 0 | 1.0 | 2 | 2 | 3 | 0 |
| 53 | 1 | 0 | 140 | 203 | 1 | 0 | 155 | 1 | 3.1 | 0 | 0 | 3 | 0 |
| 70 | 1 | 0 | 145 | 174 | 0 | 1 | 125 | 1 | 2.6 | 0 | 0 | 3 | 0 |
| 61 | 1 | 0 | 148 | 203 | 0 | 1 | 161 | 0 | 0.0 | 2 | 1 | 3 | 0 |
| 62 | 0 | 0 | 138 | 294 | 1 | 1 | 106 | 0 | 1.9 | 1 | 3 | 2 | 0 |
| 58 | 0 | 0 | 100 | 248 | 0 | 0 | 122 | 0 | 1.0 | 1 | 0 | 2 | 1 |

## Selecting Variables for the Model

### Categorical Variables

* sex (1=male; 0=female)
* cp: chest pain type
* exang: exercise induced angina (1=yes; 0=no)
* fbs: fasting blood sugar >120 mg/dl (1=true; 0=false)
* ca: number of major vessels colored by flourosopy (0-3)
* target: heart disease (1=yes; 0=no)

### Continuous Variables

* age in years
* trestbps: resting blood pressure
* oldpeak: ST depression
* thalach: maximum heart rate
* chol: cholesterol

# Model Selection

## Fitting the full model

summary(glm(target~age+sex+trestbps+thalach+exang+oldpeak+fbs+ca+chol, data = variables, family=binomial))

##   
## Call:  
## glm(formula = target ~ age + sex + trestbps + thalach + exang +   
## oldpeak + fbs + ca + chol, family = binomial, data = variables)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.6517 -0.5338 0.1596 0.5695 2.6986   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.226512 1.270488 0.965 0.334351   
## age 0.027054 0.012579 2.151 0.031499 \*   
## sex1 -1.942666 0.236193 -8.225 < 2e-16 \*\*\*  
## trestbps -0.022034 0.005602 -3.933 8.39e-05 \*\*\*  
## thalach 0.035484 0.005402 6.568 5.09e-11 \*\*\*  
## exang1 -1.560161 0.210082 -7.426 1.12e-13 \*\*\*  
## oldpeak -0.505671 0.093201 -5.426 5.78e-08 \*\*\*  
## fbs1 0.603560 0.253350 2.382 0.017204 \*   
## ca1 -1.862927 0.237439 -7.846 4.30e-15 \*\*\*  
## ca2 -3.122868 0.341282 -9.150 < 2e-16 \*\*\*  
## ca3 -2.325278 0.437422 -5.316 1.06e-07 \*\*\*  
## ca4 0.354334 0.708426 0.500 0.616955   
## chol -0.007647 0.001979 -3.865 0.000111 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1420.24 on 1024 degrees of freedom  
## Residual deviance: 767.78 on 1012 degrees of freedom  
## AIC: 793.78  
##   
## Number of Fisher Scoring iterations: 5

### Loading Packages

# Packages for ROC  
library(caTools)  
library(ROCR)

## Evaluating Model Performance

We will split the data into training and testing data sets, and evaluate the percentage of correct classifications.

### Creating the Training and Testing Data

set.seed(12345)  
split <- sample.split(variables, SplitRatio = 0.8)  
training\_data <- subset(variables, split == "TRUE")  
test\_data <- subset(variables, split == "FALSE")

### Fitting the model on the training data

logistic\_model<-glm(target~sex+trestbps+thalach+exang+oldpeak+ca+chol, data = training\_data, family="binomial")  
summary(logistic\_model)

### Predicting the likelihood of heart disease on the testing data using a cutoff value of .5

# Predict test data based on model  
predict\_reg <- predict(logistic\_model,   
 test\_data, type = "response")  
  
# Changing probabilities at cutoff value  
predict\_reg <- ifelse(predict\_reg >0.5, 1, 0)

### Creating a Confusion Matrix

|  | 0 | 1 |
| --- | --- | --- |
| 0 | 81 | 15 |
| 1 | 19 | 90 |

### Evaluating Model Accuracy

## [1] "Accuracy = 0.834146341463415"

### Plotting the ROC Curve and Calculating the AUC

# ROC-AUC Curve  
ROCPred <- prediction(predict\_reg, test\_data$target)   
ROCPer <- performance(ROCPred, measure = "tpr",   
 x.measure = "fpr")  
   
auc <- performance(ROCPred, measure = "auc")  
auc <- auc@y.values[[1]]  
  
# Plotting curve  
plot(ROCPer,main="ROC Curve")  
abline(a = 0, b = 1)  
auc <- round(auc, 4)  
legend(.6, .4, auc, title = "AUC", cex = 1)

