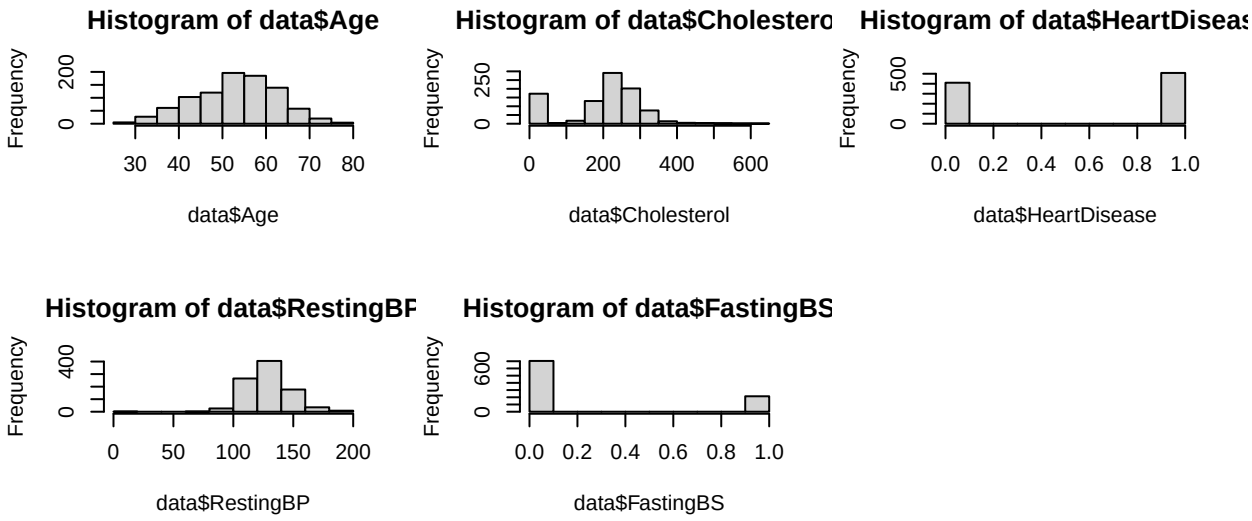
GLM Project Heart Disease Prediction

Heart diseases are the number one cause of death around the world. Approximatrely 18 million people lose their lives to these diseases each year. Early detection of heart disease can greatly impact the likelyhood of long term survival and can lower the risk of death if remediation actions are taken (medication, lowering stress, diet change, etc.) Building a model that can detect heart disease can give doctors a clear path to reducing the damage that it can cause to a population.

This heart disease data set is the combination of several different data sets. It countains 11 features (5 numerical and 6 categorical) and a binary target variable. The target variable has a binomial distribution so the logit link function will be used to link the response to the linear combination of weighted feature variables. With this data, I was able to train a logisitc regression model that could predict the probability of heart disease given some data. The final model had a misclassification rate of 14.5%, and AUC of 91.8%. According to the Hoslem test, this model is well fit to the data. Overall, this model performed reasonably well, though the missclassification rate is a little high if this were to be used as a diagnostics tool.

```
########## DATA Cleaning

##Imputiing NA values
#fit.mice <- mice(data, m=1, maxit=50, method='pmm', seed=5474, printFlag=FALSE)
#data <- complete(fit.mice, 1)



############## EDA ####################################

par(mfrow = c(3,3))
hist(data$Age)
hist(data$Cholesterol)
hist(data$HeartDisease)
hist(data$RestingBP)
hist(data$FastingBS)

datanumeric <- subset(data, select =c(-2,-3,-5,-6,-7,-9,-11,-12))


## check multicolinearty
cor(datanumeric)
```

```
##                  Age  RestingBP      MaxHR    Oldpeak
## Age        1.0000000  0.2543994 -0.3820447  0.2586115
## RestingBP  0.2543994  1.0000000 -0.1121350  0.1648030
## MaxHR     -0.3820447 -0.1121350  1.0000000 -0.1606906
## Oldpeak    0.2586115  0.1648030 -0.1606906  1.0000000
```



Looking at the response variable (Heart Disease), one can see that this is a fairly balanced classification problem. The main assumptions of logistic GLM are that the feature varaibles are independently sampled and that there is no multicolinearity. According to the correlation matrix of the numerical data, there does not seem to be any multicolinearity present in the dataset. All numerical variables seem to be relatively normally distributed, except for oldpeak, which seems to be right skewed. I decided to use all the variables in the first model and remove the variables that weren't statistically significant in the second model.

```
###### First Model ###################################

logitmodel <- glm(as.factor(HeartDisease) ~ ., family = binomial(link='logit'), train)
# Checking the model
summary(logitmodel)
```

```
##
## Call:
## glm(formula = as.factor(HeartDisease) ~ ., family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6556  -0.3762   0.1608   0.4137   2.4456
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.702325   1.744330  -0.976  0.32910
## Age               0.021588   0.016544   1.305  0.19195
```

```
## SexM              1.731824   0.346153   5.003 5.64e-07 ***
## ChestPainTypeATA -1.628721   0.396292  -4.110 3.96e-05 ***
## ChestPainTypeNAP -1.461269   0.319856  -4.569 4.91e-06 ***
## ChestPainTypeTA  -1.815178   0.554376  -3.274  0.00106 **
## RestingBP         0.006322   0.007289   0.867  0.38576
## Cholesterol      -0.004440   0.001350  -3.290  0.00100 **
## FastingBS         0.954742   0.339304   2.814  0.00490 **
## RestingECGNormal -0.171476   0.333481  -0.514  0.60711
## RestingECGST     -0.364998   0.435215  -0.839  0.40166
## MaxHR            -0.006680   0.006166  -1.083  0.27864
## ExerciseAnginaY   0.871273   0.303691   2.869  0.00412 **
## Oldpeak           0.419658   0.142352   2.948  0.00320 **
## ST_SlopeFlat      1.619690   0.519163   3.120  0.00181 **
## ST_SlopeUp       -0.938794   0.542648  -1.730  0.08363 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 885.97  on 642  degrees of freedom
## Residual deviance: 406.78  on 627  degrees of freedom
## AIC: 438.78
##
## Number of Fisher Scoring iterations: 6
```

```
pred <- predict(logitmodel,test, type = "response")

pred1 <- ifelse(pred>0.5, 1, 0)
table(pred1)
```

```
## pred1
##   0   1
## 116 159
```

```
## Missclasification rate
(miss.rate <- mean(yobs != pred1))
```

```
## [1] 0.1381818
```

```
##### Finidng MSE
MSE.a <- mean((yobs-pred)^2)
MSE.a
```

```
## [1] 0.1054969
```

```
#Plotting ROC curve of the fit.best model.
AUC <- ci.cvAUC(predictions = pred, labels = yobs, folds=1:NROW(test), confidence = 0.95)
AUC
```

```
## $cvAUC
## [1] 0.9181151
```

```
##
## $se
## [1] 0.01723922
##
## $ci
## [1] 0.8843268 0.9519033
##
## $confidence
## [1] 0.95
```

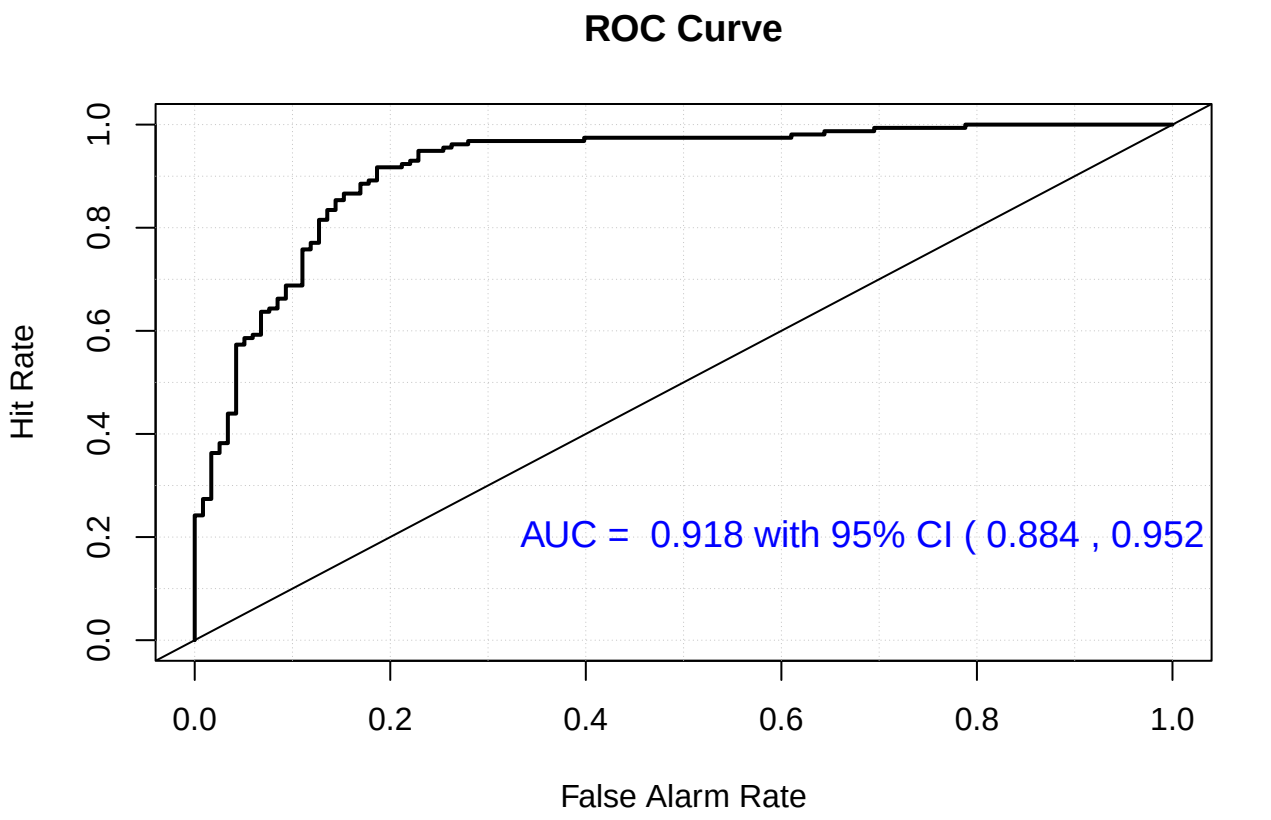
```
(auc.ci <- round(AUC$ci, digits = 3))
```

```
## [1] 0.884 0.952
```

```
logit.glm <- verify(obs = yobs, pred = pred)
```

```
## If baseline is not included, baseline values  will be calculated from the  sample obs.
```

```
roc.plot(logit.glm, plot.thres=NULL)
text(x=0.7, y=0.2, paste("AUC = ", round(AUC$cvAUC, digits = 3), "with 95% CI (",
                         auc.ci[1], ",", auc.ci[2], ").", sep = " "), col="blue", cex =1.2)
```



```r
ytest <-as.factor(yobs)
predtest <- as.numeric(pred1)
predtest <- as.factor(pred1)
#confusion matrix
confusionMatrix(as.factor(yobs), as.factor(pred1))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0  98  20
##          1  18 139
##
##                Accuracy : 0.8618
##                  95% CI : (0.8153, 0.9003)
##     No Information Rate : 0.5782
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.7174
##
##  Mcnemar's Test P-Value : 0.8711
##
##             Sensitivity : 0.8448
##             Specificity : 0.8742
##          Pos Pred Value : 0.8305
##          Neg Pred Value : 0.8854
##              Prevalence : 0.4218
##          Detection Rate : 0.3564
##    Detection Prevalence : 0.4291
##       Balanced Accuracy : 0.8595
##
##        'Positive' Class : 0
##
```

```r
###### STATS TEST GLOBAL NULL, NULL DEVIANCE<< ETC

### Global Null ###

C=logitmodel$null.deviance - logitmodel$deviance
pchisq(C,df=399-394,lower.tail = F) # Small p-value. We reject the null hypothesis.
```

```
## [1] 2.480479e-101
```

```r
#### Goodness of fit

#Hoslem Test

h <- hoslem.test(logitmodel$y, fitted(logitmodel), g=3)
h #Very high p-value. Model fits the data well
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
```

```
##
## data:  logitmodel$y, fitted(logitmodel)
## X-squared = 1.0115, df = 1, p-value = 0.3145
```

This first model has an AUC of 91.9%, a misclassification rate ofof 14.9% and Mean Square error of 10.8%. According the confusion matrix, the model has an accuracy of 85.1%. The globul noll hypothesis is rejected, suggesting the atleast some of the feature variables are statitically significant to the model. According to the model summary, RestingBP, RestingECG, and ST_Slope Flat are not statitically significant, so we will remove these variables and train a new model for comparison.

```
### Drop insignificant coefficients and compare (Resting BP, RestingECG, MaxHR)
logitmodelrefined <- glm(as.factor(HeartDisease) ~ Age + Sex + ChestPainType
                         + Cholesterol + FastingBS
                         + ExerciseAngina + ST_Slope, family = binomial(link='logit'), train)

summary(logitmodelrefined)
```

```
##
## Call:
## glm(formula = as.factor(HeartDisease) ~ Age + Sex + ChestPainType +
##     Cholesterol + FastingBS + ExerciseAngina + ST_Slope, family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5756  -0.3982   0.1765   0.4349   2.3356
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -2.179988   1.054462  -2.067 0.038697 *
## Age               0.039165   0.014377   2.724 0.006446 **
## SexM              1.760912   0.339445   5.188 2.13e-07 ***
## ChestPainTypeATA -1.707190   0.383105  -4.456 8.34e-06 ***
## ChestPainTypeNAP -1.498097   0.314276  -4.767 1.87e-06 ***
## ChestPainTypeTA  -1.722380   0.543157  -3.171 0.001519 **
## Cholesterol      -0.003708   0.001260  -2.944 0.003242 **
## FastingBS         0.876206   0.335535   2.611 0.009018 **
## ExerciseAnginaY   1.034970   0.286030   3.618 0.000296 ***
## ST_SlopeFlat      1.149272   0.493706   2.328 0.019920 *
## ST_SlopeUp       -1.682350   0.492957  -3.413 0.000643 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 885.97  on 642  degrees of freedom
## Residual deviance: 418.59  on 632  degrees of freedom
## AIC: 440.59
##
## Number of Fisher Scoring iterations: 6
```

```r
pred <- predict(logitmodelrefined,test, type = "response")

pred1 <- ifelse(pred>0.5, 1, 0)
table(pred1)
```

```
## pred1
##   0   1
## 113 162
```

```r
## Missclasification rate
(miss.rate <- mean(yobs != pred1))
```

```
## [1] 0.1418182
```

```r
##### Finidng MSE
MSE.a <- mean((yobs-pred)^2)
MSE.a
```

```
## [1] 0.1072785
```

```r
#Plotting ROC curve of the fit.best model.
AUC <- ci.cvAUC(predictions = pred, labels = yobs, folds=1:NROW(test), confidence = 0.95)
AUC
```

```
## $cvAUC
## [1] 0.9189248
##
## $se
## [1] 0.01689283
##
## $ci
## [1] 0.8858154 0.9520341
##
## $confidence
## [1] 0.95
```

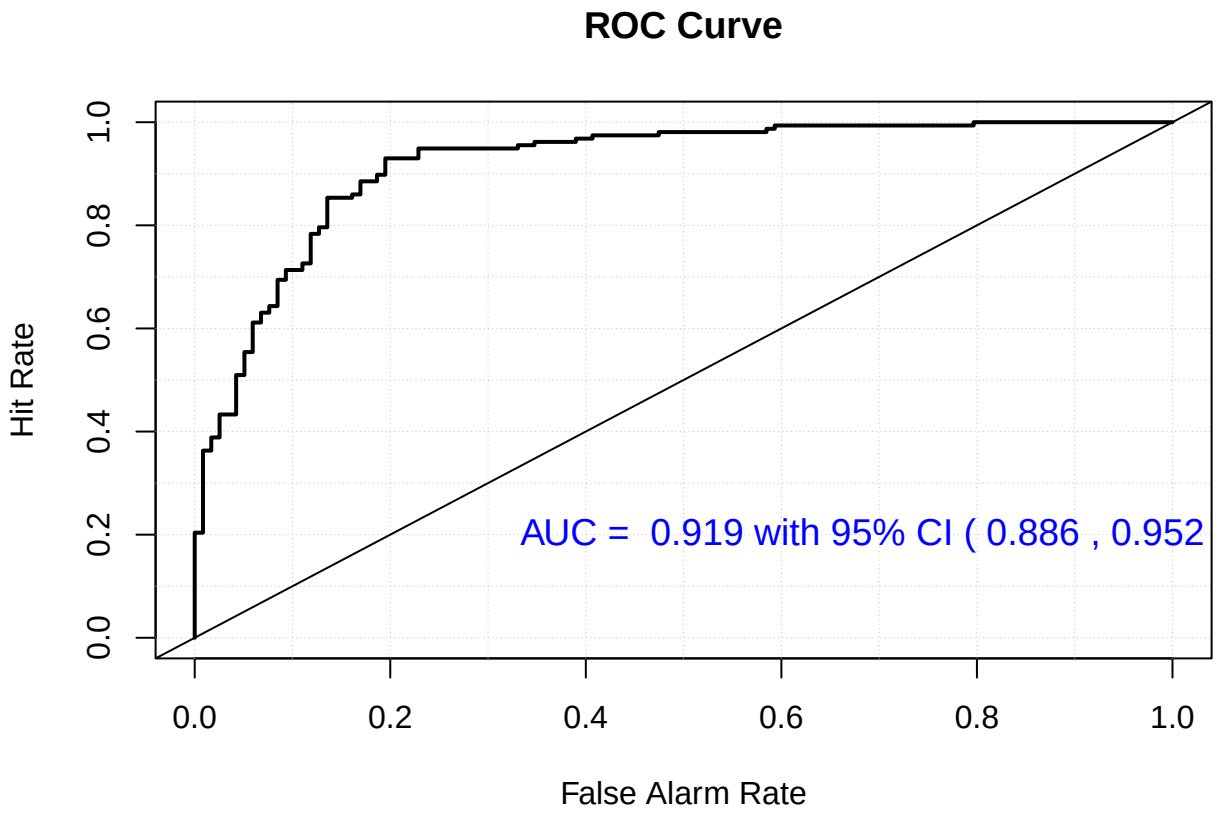
```r
(auc.ci <- round(AUC$ci, digits = 3))
```

```
## [1] 0.886 0.952
```

```r
logit.glm <- verify(obs = yobs, pred = pred)
```

```
## If baseline is not included, baseline values  will be calculated from the  sample obs.
```

```r
roc.plot(logit.glm, plot.thres=NULL)
text(x=0.7, y=0.2, paste("AUC = ", round(AUC$cvAUC, digits = 3), "with 95% CI (",
                         auc.ci[1], ",", auc.ci[2], ").", sep = " "), col="blue", cex =1.2)
```




```
ytest <-as.factor(yobs)
predtest <- as.numeric(pred1)
predtest <- as.factor(pred1)
#confusion matrix
confusionMatrix(as.factor(yobs), as.factor(pred1))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0  96  22
##          1  17 140
##
##                Accuracy : 0.8582
##                  95% CI : (0.8113, 0.8972)
##     No Information Rate : 0.5891
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.709
##
##  Mcnemar's Test P-Value : 0.5218
##
##             Sensitivity : 0.8496
##             Specificity : 0.8642
##          Pos Pred Value : 0.8136
```

```
##          Neg Pred Value : 0.8917
##              Prevalence : 0.4109
##          Detection Rate : 0.3491
##    Detection Prevalence : 0.4291
##       Balanced Accuracy : 0.8569
##
##        'Positive' Class : 0
##
```

```
###### STATS TEST GLOBAL NULL, NULL DEVIANCE<< ETC

### Global Null ###

C=logitmodelrefined$null.deviance - logitmodelrefined$deviance
pchisq(C,df=399-394,lower.tail = F) # Small p-value. We reject the null hypothesis.
```

```
## [1] 8.773806e-99
```

```
#### Goodness of fit

#Hoslem Test

h <- hoslem.test(logitmodelrefined$y, fitted(logitmodelrefined), g=3)
h #Very high p-value. Model fits the data well
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  logitmodelrefined$y, fitted(logitmodelrefined)
## X-squared = 2.5564, df = 1, p-value = 0.1098
```

```
anova(logitmodel,logitmodelrefined)
```

```
## Analysis of Deviance Table
##
## Model 1: as.factor(HeartDisease) ~ Age + Sex + ChestPainType + RestingBP +
##     Cholesterol + FastingBS + RestingECG + MaxHR + ExerciseAngina +
##     Oldpeak + ST_Slope
## Model 2: as.factor(HeartDisease) ~ Age + Sex + ChestPainType + Cholesterol +
##     FastingBS + ExerciseAngina + ST_Slope
##   Resid. Df Resid. Dev Df Deviance
## 1       627     406.78
## 2       632     418.59 -5  -11.812
```

The refined model has an AUC of 91.8%, a missclassification rate of 14.5% and a Mean Square Error of 10.9%. These are fairly similar to the previous model. According to the confusion matrix, the model has an accuracy of 85.5%. The global null hypothesis was rejected as well.

The AIC's for model 1 and 2 were 438.93 and 435.22 respectively. An anova test was conducted to compare model deviance, showing that the differene between the two was -6.5297. This suggests that model 2 has slightly more deviance. Looking at all of these model metrics, it is not very clear which model is the best. When in doubt, one should always choose lest complicatied model that fully explains the data. In this case, model 2 has less feature variables so we will choose this model.

Overall these models performed reasonably well, though they certainly could not be used to make any concrete medical decisions at this point in time. Creating this model was a great learning experience as I had a few hiccups along the way. Originally, I was going to train the model on a different dataset, but the training model accuracy was attrocious. At first, I could not figure out what the issues was, until I remembered that I neglected to check the distribution of the response variable. When I took a look, i saw the there was a severe imbalance in the binary outcomes of the dataset, which would be difficult to model. The lesson to always fully explore the data before training a model was definitely reinforced after this experience.

Bibiliography

Fedesoriano. "Heart Failure Prediction Dataset." Kaggle, 10 Sept. 2021, https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction.