uwa6xv_M12_HW

Alanna Hazlett

2024-04-20

Problem 1

```
library(palmerpenguins)
Data<-penguins
##remove penguins with gender missing
Data<-Data[complete.cases(Data[ , 7]),-c(2,8)]
##80-20 split
set.seed(1)
sample<-sample.int(nrow(Data), floor(.80*nrow(Data)), replace = F)
train<-Data[sample, ]
test<-Data[-sample, ]
#fitting model
result<- glm(sex~bill_length_mm+bill_depth_mm+body_mass_g+species,family=binomial,data=train)
```

(a)

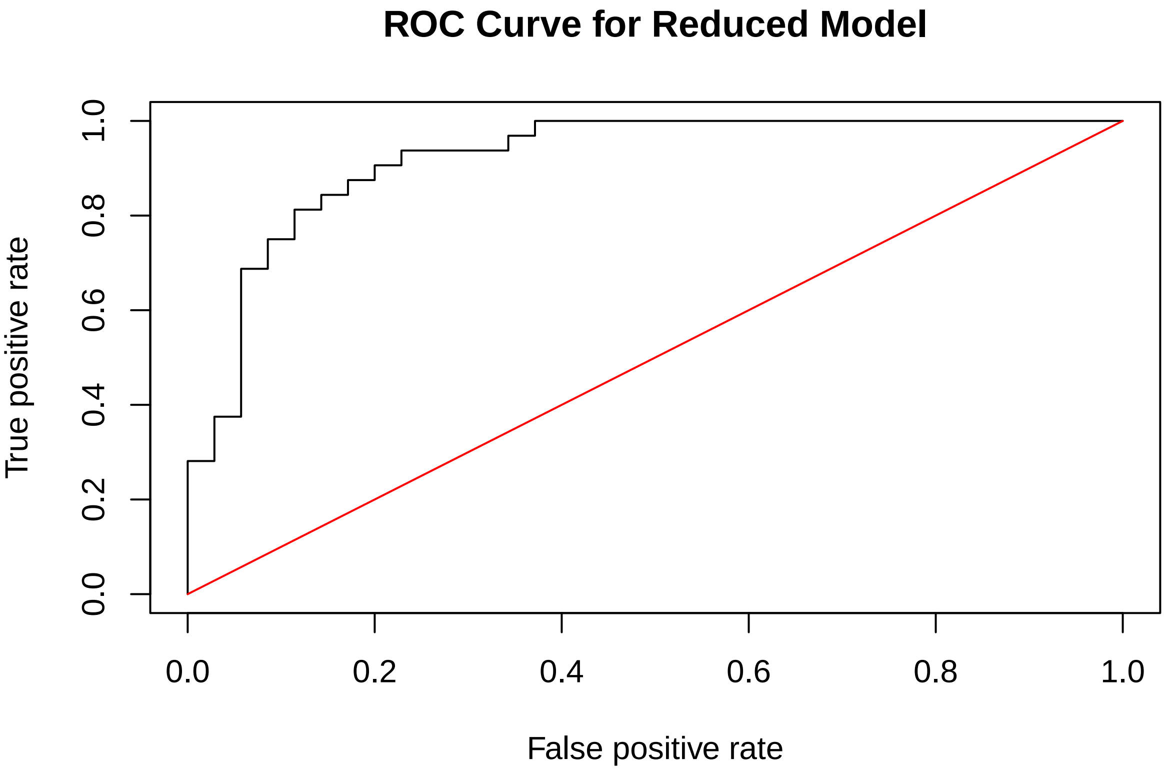
Validate your model on the test data by creating an ROC curve. What does your ROC curve tell you?

```
##predicted probs for test data
preds<-predict(result,newdata=test, type="response")

##produce the numbers associated with classification table
rates<-ROCR::prediction(preds, test$sex)

##store the true positive and false positive rates
roc_result<-ROCR::performance(rates,measure="tpr", x.measure="fpr")

##plot ROC curve and overlay the diagonal line for random guessing
plot(roc_result, main="ROC Curve for Reduced Model")
lines(x = c(0,1), y = c(0,1), col="red")
```



The model performs better than random guessing as the ROC line is above the red diagonal line. The model after a certain threshold point has a TPR of 1, the last third or so of the model line.

(b)

Find the AUC associated with your ROC curve. What does your AUC tell you?

```r
##compute the AUC
auc<-ROCR::performance(rates, measure = "auc")
auc@y.values
```

```
## [[1]]
## [1] 0.9214286
```

This shows that the model does a vary good job of classifying the observations correctly, as this value is close to one.

(c)

Create a confusion matrix using a threshold of 0.5. What is the false positive rate? What is the false negative rate? What is error rate?

```r
##confusion matrix with threshold of 0.5
table(test$sex, preds>0.5)
```

```
##
##          FALSE TRUE
##   female    28    7
##   male       4   28
```

$$FPR = \frac{FP}{TN + FP} = \frac{7}{(28+7)} = 0.2$$

$$FNR = \frac{FN}{FN + TP} = \frac{4}{(4+28)} = 0.125$$

$$error\ rate\ = \frac{FP + FN}{n} = \frac{(7+4)}{67} = 0.1642$$

(d)
Discuss if the threshold should be changed. If it should be changed, explain why, and create another confusion matrix with a different threshold.

Ideally we would consult with a subject matter expert prior to running our statistics and drawing a conclusion to determine how precise our model should be, which error we are trying to minimize the FPR or FNR.

Based on the information we have here it appears that our model does a good job at this threshold. We have a low error rate, indicating that the model incorrectly classifies an observation about 16% of the time. The FPR, false positive rate, indicates that the model classifies as true incorrectly 20% of the time. The FNR, false negative rate, indicates that the model classifies as false incorrectly about 16% of the time.
Lowering our threshold hold would increase our FPR, which means we would increase the number of penguins we classify as male, resulting in more observations being classified as male when they are not truly male. The FNR would decrease, meaning the number of penguins incorrectly classified as female when they are truly male would decrease.
Raising our threshold would decrease our FPR, which means that we would decrease the number of penguins we classify as male when they are not truly male. And it would increase our FNR, meaning the number of penguins incorrectly classified as female when they are truly male would increase.