Homework12

Dima Mikhaylov

12/4/2021

```
library(palmerpenguins)
Data<-penguins
str(Data)
```

```
## tibble [344 × 8] (S3: tbl_df/tbl/data.frame)
##  $ species          : Factor w/ 3 levels "Adelie","Chinstrap",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ island           : Factor w/ 3 levels "Biscoe","Dream",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ bill_length_mm   : num [1:344] 39.1 39.5 40.3 NA 36.7 39.3 38.9 39.2 34.1 42 ...
##  $ bill_depth_mm    : num [1:344] 18.7 17.4 18 NA 19.3 20.6 17.8 19.6 18.1 20.2 ...
##  $ flipper_length_mm: int [1:344] 181 186 195 NA 193 190 181 195 193 190 ...
##  $ body_mass_g      : int [1:344] 3750 3800 3250 NA 3450 3650 3625 4675 3475 4250 ...
##  $ sex              : Factor w/ 2 levels "female","male": 2 1 1 NA 1 2 1 2 NA NA ...
##  $ year             : int [1:344] 2007 2007 2007 2007 2007 2007 2007 2007 2007 2007 ...
```

We will focus on using the four measurement variables (bill length, bill depth, flipper length, body mass) to model the gender of the penguins. Since there are three species involved, we also want to control for species in the logistic regression. We will not consider the island and year in this logistic regression. When you read the data in, notice that there are a number of penguins with missing values for gender. Remove these observations from the data frame.

```
## Remove penguins with gender missing
Data<-Data[complete.cases(Data[ , 7]),-c(2,8)]
```

From the last homework, you should have dropped flipper length from the model, while keeping bill length, bill depth, body mass, and species as predictors.

```
# Drop flipper length
Data <- subset( Data, select = -c(flipper_length_mm))

# Convert sex to 0 and 1 factor for output variable
Data$sex <- factor(ifelse(Data$sex=="male",1,0))
head(Data)
```

```
## # A tibble: 6 × 5
##   species bill_length_mm bill_depth_mm body_mass_g sex
##   <fct>            <dbl>         <dbl>       <int> <fct>
## 1 Adelie            39.1          18.7        3750 1
## 2 Adelie            39.5          17.4        3800 0
## 3 Adelie            40.3          18          3250 0
## 4 Adelie            36.7          19.3        3450 0
## 5 Adelie            39.3          20.6        3650 1
## 6 Adelie            38.9          17.8        3625 0
```

Then, randomly split your data into a training and test set (80-20 split respectively). For reproducibility, use set.seed(1) while performing the split.

```
##80-20 split
set.seed(1)
sample<-sample.int(nrow(Data), floor(.80*nrow(Data)), replace = F)
train<-Data[sample, ]
test<-Data[-sample, ]
head(train)
```

```
## # A tibble: 6 × 5
##   species   bill_length_mm bill_depth_mm body_mass_g sex
##   <fct>              <dbl>         <dbl>       <int> <fct>
## 1 Chinstrap           50.2          18.8        3800 1
## 2 Gentoo              50.2          14.3        5700 1
## 3 Adelie              38.1          17.6        3425 0
## 4 Chinstrap           51            18.8        4100 1
## 5 Chinstrap           52.7          19.8        3725 1
## 6 Gentoo              49.6          16          5700 1
```

```
full_model<-glm(sex ~ ., family="binomial", data=train)
summary(full_model)
```

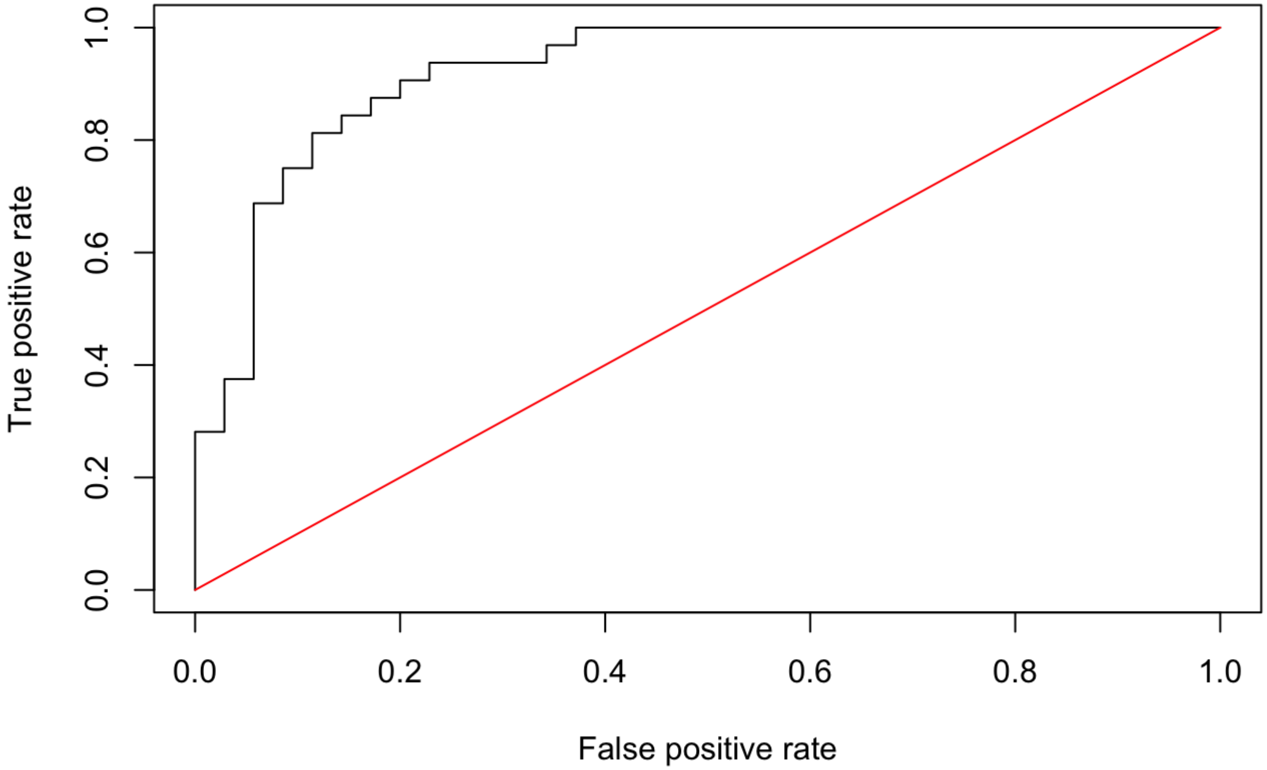
```
## 
## Call:
## glm(formula = sex ~ ., family = "binomial", data = train)
## 
## Deviance Residuals: 
##      Min        1Q    Median        3Q       Max  
## -2.52269  -0.11388   0.00063   0.06524   3.01858  
## 
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)    
## (Intercept)      -1.032e+02  1.706e+01  -6.051 1.44e-09 ***
## speciesChinstrap -1.042e+01  2.544e+00  -4.096 4.20e-05 ***
## speciesGentoo    -1.238e+01  3.383e+00  -3.661 0.000251 ***
## bill_length_mm    9.513e-01  2.210e-01   4.303 1.68e-05 ***
## bill_depth_mm     2.099e+00  4.684e-01   4.481 7.41e-06 ***
## body_mass_g       7.714e-03  1.625e-03   4.746 2.07e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 368.619  on 265  degrees of freedom
## Residual deviance:  70.172  on 260  degrees of freedom
## AIC: 82.172
## 
## Number of Fisher Scoring iterations: 8
```

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

```
library(ROCR)
# Get predictions based on test holdout set
preds <- predict(full_model, newdata=test, type='response')
rates <- prediction(preds, test$sex)
roc_result <- performance(rates, measure = 'tpr', x.measure = 'fpr')
roc_result
```

```
## A performance instance
##   'False positive rate' vs. 'True positive rate' (alpha: 'Cutoff')
##   with 68 data points
```

```
# Plot the ROC curve and random guess 50% line
plot(roc_result, title="The ROC curve")+
lines(x = c(0,1), y = c(0, 1), col="red")
```



```
## integer(0)
```

The ROC curve shows TPR over FPR for various thresholds. Overall it is a promising model becuase ROC increases sharply and achieves good TPR of 0.8-0.9 at relatively small FPR of 0.1-0.3.

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

```
# Get AUC from test performance
auc <- performance(rates, measure="auc")
auc@y.values
```

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

Good AUC of 92% compared to 100% theoretical ideal classifier.

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?

```
prop.table(table(test$sex))
```

```
## 
##         0         1 
## 0.5223881 0.4776119
```

Balanced sample should produce reliable confusion matrix with 0.5 threshold

```
cf <- table(test$sex, preds>0.5)
cf
```

```
## 
##     FALSE TRUE
##   0    28    7
##   1     4   28
```

What is the false positive rate? FP/FP+TN

```
cat("False positive (type 1 error):", cf[3]/(cf[3]+cf[1]))
```

```
## False positive (type 1 error): 0.2
```

What is the false negative rate? FN/(TP+FN)

```
cat("False negative (type 2 error):", cf[2]/(cf[2]+cf[4]))
```

```
## False negative (type 2 error): 0.125
```

What is error rate? (FP+FN)/(TP+TN+FP+FN)

```
cat("Error rate:", (cf[3]+cf[2])/(cf[1]+cf[2]+cf[3]+cf[4]))
```

```
## Error rate: 0.1641791
```

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

The threshold could have been changes debending on the context. For example, it is more important to correctly pick males at the costs of sometimes misclassifying females as males. In this case threshold can be decreased to 40% or 30% so even a lower probability of male observation will be marked as 1.