Hw12

Hyun Suk (Max) Ryoo (hr2ee)

11/30/2021

Set up

```r
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.2
```

```
## -- Attaching packages --------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

```
## Warning: package 'ggplot2' was built under R version 4.0.2
```

```
## Warning: package 'tidyr' was built under R version 4.0.2
```

```
## Warning: package 'readr' was built under R version 4.0.2
```

```
## Warning: package 'dplyr' was built under R version 4.0.2
```

```
## Warning: package 'stringr' was built under R version 4.0.2
```

```
## Warning: package 'forcats' was built under R version 4.0.2
```

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

```r
library(palmerpenguins)
```

```
## Warning: package 'palmerpenguins' was built under R version 4.0.2
```

```r
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 4.0.2
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```r
library(ROCR)
```

```
## Warning: package 'ROCR' was built under R version 4.0.2
```

```r
Data<-penguins
##remove penguins with gender missing
```

```r
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, ]
head(train)
```

```
## # A tibble: 6 x 6
##   species   bill_length_mm bill_depth_mm flipper_length_mm body_mass_g sex
##   <fct>              <dbl>         <dbl>             <int>       <int> <fct>
## 1 Chinstrap           50.2          18.8               202        3800 male
## 2 Gentoo              50.2          14.3               218        5700 male
## 3 Adelie              38.1          17.6               187        3425 female
## 4 Chinstrap           51            18.8               203        4100 male
## 5 Chinstrap           52.7          19.8               197        3725 male
## 6 Gentoo              49.6          16                 225        5700 male
```

1 - A

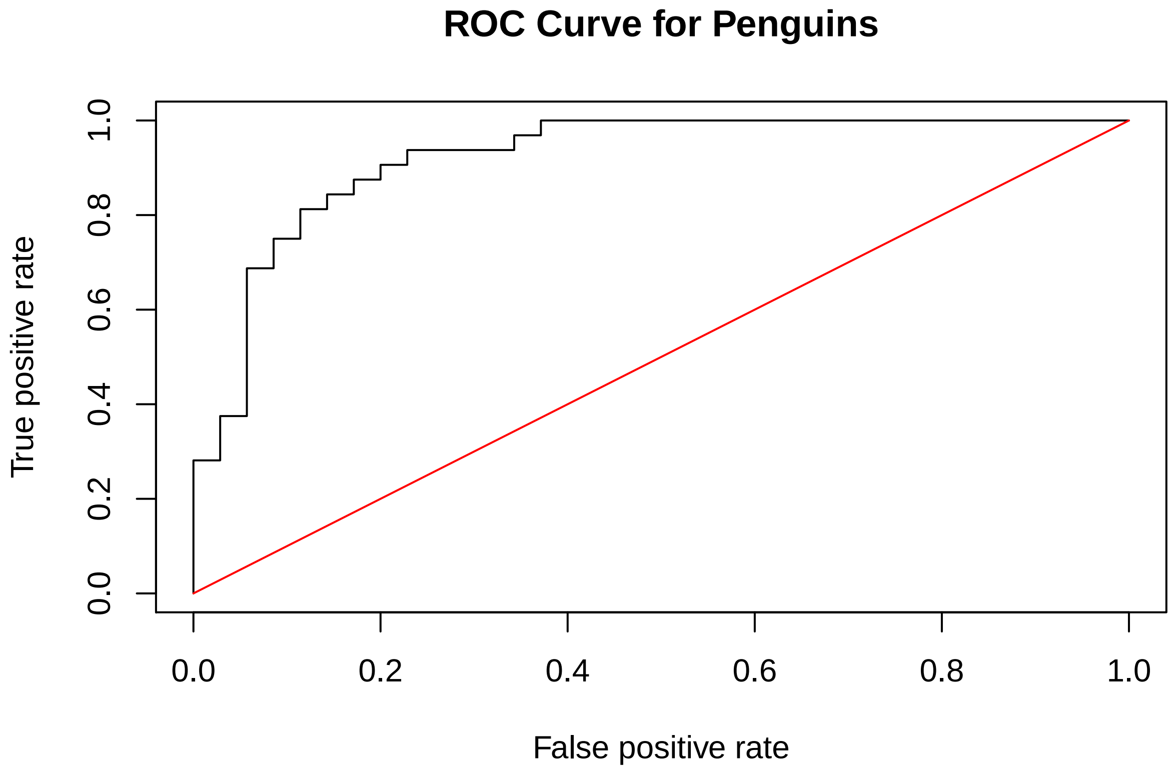
We first need to recreate the model.

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

```
##
## Call:
## glm(formula = sex ~ . - flipper_length_mm, 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
```

```r
##predicted survival rate for test data based on training data
preds<-predict(result,newdata=test, type="response")
```

```r
##transform the input data into a format that is suited for the
##performance() function
rates<-prediction(preds, test$sex)
##store the true positive and false positive rates
roc_result<-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 Penguins")
lines(x = c(0,1), y = c(0,1), col="red")
```



1 - B

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

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

The AUC of our ROC curve is 0.9214286, which measn our logitic regressino does better than random guessing.

1 - C

```r
table(test$sex, preds>0.5)
```

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

From the table above we can compute the false positive rate, false negative rate, and error rate.

- The false positive rate (FRP) is $\frac{7}{7+28} \rightarrow \frac{7}{35} \rightarrow 0.2$.
- The false negative rate (FNR) is $\frac{4}{4+28} \rightarrow \frac{4}{32} \rightarrow 0.125$

1 - D