Cordero_week9.2

Joaquin Cordero

2024-07-25

```
library(foreign)
library(caTools)
```

Thoraric Surgery

```
ts_data <- read.arff("ThoracicSurgery.arff")

ts_split <- sample.split(ts_data, SplitRatio = 0.8)

ts_train <- subset(ts_data, ts_split == "TRUE")
ts_test <- subset(ts_data, ts_split == "FALSE")
```

1.b.i

```
ts_model1 <- glm(Risk1Yr ~ AGE + PRE17 + PRE19 + PRE30, data = ts_train, family = 'binomial')
summary(ts_model1)
```

```
##
## Call:
## glm(formula = Risk1Yr ~ AGE + PRE17 + PRE19 + PRE30, family = "binomial",
##     data = ts_train)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -2.47033    1.28297  -1.925  0.05417 .
## AGE           -0.01067    0.01842  -0.579  0.56243
## PRE17T         1.45824    0.47241   3.087  0.00202 **
## PRE19T       -13.95361 1029.09687  -0.014  0.98918
## PRE30T         1.48188    0.61810   2.397  0.01651 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 311.20  on 359  degrees of freedom
## Residual deviance: 293.89  on 355  degrees of freedom
```

```
## AIC: 303.89
##
## Number of Fisher Scoring iterations: 14
```

1.b.ii

According to the summary, variables PRE17 and PRE30 had the greatest effect on survival rate based on their p-values

1.b.iii

```
ts_preds <- predict(ts_model1, ts_test, type = "response")

ts_preds <- predict(ts_model1, ts_train, type = "response")

ts_confmatrix <- table(Actual_Value=ts_train$Risk1Yr, Predicted_Value = ts_preds > 0.5)
ts_confmatrix
```

```
##              Predicted_Value
## Actual_Value FALSE
##            F   304
##            T    56
```

```
ts_accuracy <- ts_confmatrix[1,1] / sum(ts_confmatrix)
ts_accuracy
```

```
## [1] 0.8444444
```

Binary Classifier

2.b

```
bc_data <- read.csv("binary-classifier-data.csv")

bc_split <- sample.split(bc_data, SplitRatio = 0.8)

bc_train <- subset(bc_data, bc_split == "TRUE")
bc_test <- subset(bc_data, bc_split == "FALSE")

bc_model1 <- glm(label ~ x + y, data = bc_train, family = 'binomial')
summary(bc_model1)
```

```
##
## Call:
## glm(formula = label ~ x + y, family = "binomial", data = bc_train)
##
## Coefficients:
```

```
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.415200   0.143553   2.892 0.003824 **
## x           -0.002376   0.002237  -1.062 0.288036
## y           -0.007953   0.002302  -3.456 0.000549 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1382.9  on 997  degrees of freedom
## Residual deviance: 1367.6  on 995  degrees of freedom
## AIC: 1373.6
##
## Number of Fisher Scoring iterations: 4
```

```
bc_preds <- predict(bc_model1, bc_test, type = "response")

bc_preds <- predict(bc_model1, bc_train, type = "response")

bc_confmatrix <- table(Actual_Value= bc_train$label, Predicted_Value = bc_preds > 0.5)
bc_confmatrix
```

```
##              Predicted_Value
## Actual_Value FALSE TRUE
##            0   289  222
##            1   202  285
```

```
bc_accuracy <- (bc_confmatrix[1,1] + bc_confmatrix[2,2]) / sum(bc_confmatrix)
```

2.b.i

```
bc_accuracy
```

```
## [1] 0.5751503
```