# NickBlackfordWeek10

Nick Blackford

2024-02-18

## Thoracic Surgery Dataset

```
library(foreign)

mydata <- read.arff("/Users/nickblackford/Downloads/ThoracicSurgery.arff")

head(data)
```

```
##
## 1 function (..., list = character(), package = NULL, lib.loc = NULL,
## 2     verbose = getOption("verbose"), envir = .GlobalEnv, overwrite = TRUE)
## 3 {
## 4     fileExt <- function(x) {
## 5         db <- grepl("\\\\.[^.]+\\\\.(gz|bz2|xz)$", x)
## 6         ans <- sub(".*\\\\.", "", x)
```

```
library(MASS)
# Multivariate logistic regression with stepwise selection
full_model <- glm(Risk1Yr ~ PRE7 + PRE8 + PRE9 +PRE10 + PRE11 + PRE17 +PRE19 +PRE25 + PRE30 + PRE32, fa
step_model <- stepAIC(full_model, direction = "both")
```

```
## Start:  AIC=395.68
## Risk1Yr ~ PRE7 + PRE8 + PRE9 + PRE10 + PRE11 + PRE17 + PRE19 +
##     PRE25 + PRE30 + PRE32
##
##         Df Deviance    AIC
## - PRE25  1   373.75 393.75
## - PRE32  1   374.02 394.02
## - PRE8   1   374.11 394.11
## - PRE19  1   374.39 394.39
## - PRE7   1   374.83 394.83
## - PRE10  1   374.97 394.97
## - PRE11  1   375.61 395.61
## <none>       373.68 395.68
## - PRE30  1   377.29 397.29
## - PRE9   1   378.47 398.47
## - PRE17  1   378.58 398.58
##
## Step:  AIC=393.75
## Risk1Yr ~ PRE7 + PRE8 + PRE9 + PRE10 + PRE11 + PRE17 + PRE19 +
##     PRE30 + PRE32
##
##         Df Deviance    AIC
```

```
## - PRE32  1   374.08 392.08
## - PRE8   1   374.23 392.23
## - PRE19  1   374.46 392.46
## - PRE7   1   374.86 392.86
## - PRE10  1   375.02 393.02
## - PRE11  1   375.68 393.68
## <none>       373.75 393.75
## - PRE30  1   377.46 395.46
## + PRE25  1   373.68 395.68
## - PRE9   1   378.68 396.68
## - PRE17  1   378.70 396.70
##
## Step:  AIC=392.08
## Risk1Yr ~ PRE7 + PRE8 + PRE9 + PRE10 + PRE11 + PRE17 + PRE19 +
##     PRE30
##
##         Df Deviance    AIC
## - PRE8   1   374.57 390.57
## - PRE19  1   374.79 390.79
## - PRE7   1   375.21 391.21
## - PRE10  1   375.39 391.39
## - PRE11  1   376.03 392.03
## <none>       374.08 392.08
## + PRE32  1   373.75 393.75
## - PRE30  1   377.89 393.89
## + PRE25  1   374.02 394.02
## - PRE9   1   379.05 395.05
## - PRE17  1   379.07 395.07
##
## Step:  AIC=390.57
## Risk1Yr ~ PRE7 + PRE9 + PRE10 + PRE11 + PRE17 + PRE19 + PRE30
##
##         Df Deviance    AIC
## - PRE19  1   375.31 389.31
## - PRE10  1   376.00 390.00
## - PRE7   1   376.13 390.13
## <none>       374.57 390.57
## - PRE11  1   376.67 390.67
## + PRE8   1   374.08 392.08
## + PRE32  1   374.23 392.23
## - PRE30  1   378.25 392.25
## + PRE25  1   374.45 392.45
## - PRE17  1   379.51 393.51
## - PRE9   1   379.95 393.95
##
## Step:  AIC=389.31
## Risk1Yr ~ PRE7 + PRE9 + PRE10 + PRE11 + PRE17 + PRE30
##
##         Df Deviance    AIC
## - PRE10  1   376.70 388.70
## - PRE7   1   376.88 388.88
## <none>       375.31 389.31
## - PRE11  1   377.32 389.32
## + PRE19  1   374.57 390.57
```

```
## + PRE8   1   374.79 390.79
## - PRE30  1   378.96 390.96
## + PRE32  1   374.97 390.97
## + PRE25  1   375.19 391.19
## - PRE17  1   380.34 392.34
## - PRE9   1   380.74 392.74
##
## Step:  AIC=388.7
## Risk1Yr ~ PRE7 + PRE9 + PRE11 + PRE17 + PRE30
##
##         Df Deviance    AIC
## - PRE7   1   378.17 388.17
## <none>       376.70 388.70
## + PRE10  1   375.31 389.31
## - PRE11  1   379.49 389.49
## + PRE19  1   376.00 390.00
## + PRE8   1   376.06 390.06
## + PRE32  1   376.32 390.32
## + PRE25  1   376.59 390.59
## - PRE30  1   381.38 391.38
## - PRE17  1   381.76 391.76
## - PRE9   1   382.61 392.61
##
## Step:  AIC=388.17
## Risk1Yr ~ PRE9 + PRE11 + PRE17 + PRE30
##
##         Df Deviance    AIC
## <none>       378.17 388.17
## - PRE11  1   380.68 388.68
## + PRE7   1   376.70 388.70
## + PRE10  1   376.88 388.88
## + PRE8   1   377.06 389.06
## + PRE19  1   377.45 389.45
## + PRE32  1   377.77 389.77
## + PRE25  1   378.10 390.10
## - PRE30  1   382.62 390.62
## - PRE17  1   383.42 391.42
## - PRE9   1   384.62 392.62
```

```
summary(step_model)
```

```
##
## Call:
## glm(formula = Risk1Yr ~ PRE9 + PRE11 + PRE17 + PRE30, family = binomial,
##     data = mydata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0088  -0.5212  -0.5212  -0.3490   2.3786
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.7681     0.4224  -6.554 5.61e-11 ***
## PRE9T         1.1754     0.4330   2.715  0.00664 **
## PRE11T        0.5232     0.3215   1.627  0.10368
```

```
## PRE17T        0.9940     0.4096   2.427  0.01523 *
## PRE30T        0.8406     0.4319   1.946  0.05164 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 395.61  on 469  degrees of freedom
## Residual deviance: 378.17  on 465  degrees of freedom
## AIC: 388.17
##
## Number of Fisher Scoring iterations: 5
```

```
#best model per stepwise selection
model1 <- glm(formula = Risk1Yr ~ PRE9 + PRE11 + PRE17 + PRE30, family = binomial,
    data = mydata)
summary(model1)
```

```
##
## Call:
## glm(formula = Risk1Yr ~ PRE9 + PRE11 + PRE17 + PRE30, family = binomial,
##     data = mydata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0088  -0.5212  -0.5212  -0.3490   2.3786
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.7681     0.4224  -6.554 5.61e-11 ***
## PRE9T         1.1754     0.4330   2.715  0.00664 **
## PRE11T        0.5232     0.3215   1.627  0.10368
## PRE17T        0.9940     0.4096   2.427  0.01523 *
## PRE30T        0.8406     0.4319   1.946  0.05164 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 395.61  on 469  degrees of freedom
## Residual deviance: 378.17  on 465  degrees of freedom
## AIC: 388.17
##
## Number of Fisher Scoring iterations: 5
```

PRE9T (1.1754): This variable has the largest coefficient, indicating that it has the strongest positive effect on the survival rate. It is statistically significant at the 0.01 level

PRE17T (0.9940): This variable has the second-largest coefficient, suggesting a strong positive effect on the survival rate. It is statistically significant at the 0.05 level PRE30T (0.8406): This variable has a positive effect on the survival rate, but it is only marginally significant at the 0.05 level.

PRE11T (0.5232): This variable has the smallest coefficient among the significant predictors, indicating a weaker positive effect on the survival rate. It is not statistically significant at the 0.05 level.

```
#  Predict probabilities
predicted_probabilities <- predict(model1, type = "response")
```

```r
# Convert probabilities to binary predictions
threshold <- 0.5
predicted_outcomes <- ifelse(predicted_probabilities > threshold, 1, 0)

# Calculate accuracy
actual_outcomes <- mydata$Risk1Yr
actual_outcomes_numeric <- as.numeric(actual_outcomes) - 1
correct_predictions <- sum(predicted_outcomes == actual_outcomes_numeric)
total_predictions <- length(predicted_outcomes)
accuracy <- correct_predictions / total_predictions

# Print the accuracy
print(accuracy)
```

```
## [1] 0.8510638
```

Accuracy for my model was 85% with a threshold of 0.5 to shift my probabilities into binary results.

## Binary Classifier Data

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:MASS':
##
##     select

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v readr     2.1.4
## v ggplot2   3.4.4     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::select() masks MASS::select()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
## Rows: 1498 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (3): label, x, y
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
#logistic regression model
log_model <- glm(formula = label ~ x + y, data = classifier_data)
```

```r
summary(log_model)
```

```
##
## Call:
## glm(formula = label ~ x + y, data = classifier_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6108  -0.4956  -0.3664   0.4925   0.6253
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.6049208  0.0287423  21.046  < 2e-16 ***
## x           -0.0006309  0.0004481  -1.408    0.159
## y           -0.0019662  0.0004578  -4.295 1.86e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.2464139)
##
##     Null deviance: 374.28  on 1497  degrees of freedom
## Residual deviance: 368.39  on 1495  degrees of freedom
## AIC: 2157.8
##
## Number of Fisher Scoring iterations: 2
```

```r
#  Predict probabilities
predicted_probabilities_2 <- predict(log_model, type = "response")

# Convert probabilities to binary predictions
threshold_2 <- 0.5
predicted_outcomes_2 <- ifelse(predicted_probabilities_2 > threshold_2, 1, 0)

# Calculate accuracy
actual_outcomes_2<- classifier_data$label
correct_predictions_2 <- sum(predicted_outcomes_2 == actual_outcomes_2)
total_predictions_2 <- length(predicted_outcomes_2)
accuracy_2 <- correct_predictions_2 / total_predictions_2

# Print the accuracy
print(accuracy_2)
```

```
## [1] 0.5834446
```

This model had an accuracy of 58% when comparing predicted to actual values.