580_barrett_quiz4.R

Nick

2021-04-05

```
#AMS580 Quiz 4
#By Nicholas Barrett
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.6.3
```

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

```
## v ggplot2 3.3.3     v purrr   0.3.3
## v tibble  3.1.0     v dplyr   1.0.4
## 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 3.6.3
```

```
## Warning: package 'tibble' was built under R version 3.6.3
```

```
## Warning: package 'tidyr' was built under R version 3.6.3
```

```
## Warning: package 'readr' was built under R version 3.6.3
```

```
## Warning: package 'dplyr' was built under R version 3.6.3
```

```
## Warning: package 'forcats' was built under R version 3.6.3
```

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

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.6.3
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 3.6.3
```

```
##
## Attaching package: 'MASS'
```

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

```
#Q1
data("birthwt", package = "MASS")
#this factoring method was taken from the reference in the homework
bwt <- with(birthwt, {
  race <- factor(race, labels = c("white", "black", "other"))
  ptd <- factor(ptl > 0)
  ftv <- factor(ftv)
  levels(ftv)[-(1:2)] <- "2+"
  data.frame(low = factor(low), age, lwt, race, smoke = (smoke > 0), ptd, ht = (ht >
0), ui = (ui > 0), ftv)
})
options(contrasts = c("contr.treatment", "contr.poly"))

#Split Data
set.seed(123)
training.samples <- bwt$low %>%
  createDataPartition(p = 0.8, list = FALSE)
train.data  <- bwt[training.samples, ]
test.data <- bwt[-training.samples, ]


#Testing function
test <- function(model,test.data){
  probabilities <- model %>% predict(test.data, type = "response")
  predicted.classes <- ifelse(probabilities > 0.5, 1, 0)
  count.acc = 0
  count.tp = 0
  count.fp = 0
  count.fn = 0
  len = length(test.data[ ,1])
  for(i in 1:len){
    if(predicted.classes[i]==test.data$low[i]){ #true positive and true negative
      if(predicted.classes[i]==1){count.tp = count.tp +1}#TP
      count.acc = count.acc + 1
    }
    if(test.data$low[i]==0 && predicted.classes[i]==1){
      count.fp = count.fp +1 #false positive
    }
    if(test.data$low[i]==1 && predicted.classes[i]==0){
      count.fn = count.fn +1 #false negative
    }
  }
  sen = count.tp/(count.tp+count.fn)
  count.tn =(count.acc-count.tp) #TN
  spe = (count.tn)/(count.tn + count.fp)
  acc = count.acc/len
  out = list(Acc = acc, Specif = spe, Sensit = sen)
  conf.mat = matrix(data = c(count.tn,count.fp,count.fn,count.tp),byrow = TRUE,nrow=
2)
  print(conf.mat)
  return(out)
```

```
}

#Full Model
model <- glm(train.data$low ~., data = train.data, family = binomial)
summary(model)
```

```
##
## Call:
## glm(formula = train.data$low ~ ., family = binomial, data = train.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0336  -0.7262  -0.4318   0.8866   2.2005
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.741588   1.563566   1.114  0.26534
## age         -0.040291   0.045521  -0.885  0.37611
## lwt         -0.020592   0.008941  -2.303  0.02127 *
## raceblack    1.032835   0.639413   1.615  0.10625
## raceother    0.509939   0.526140   0.969  0.33244
## smokeTRUE    0.358188   0.491957   0.728  0.46656
## ptdTRUE      1.757537   0.559158   3.143  0.00167 **
## htTRUE       1.911114   0.885964   2.157  0.03100 *
## uiTRUE       1.038787   0.537062   1.934  0.05309 .
## ftv1        -0.864554   0.590892  -1.463  0.14343
## ftv2+        0.265106   0.502134   0.528  0.59753
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 189.59  on 151  degrees of freedom
## Residual deviance: 149.67  on 141  degrees of freedom
## AIC: 171.67
##
## Number of Fisher Scoring iterations: 5
```

```
test(model,test.data)
```

```
##      [,1] [,2]
## [1,]   21    5
## [2,]    9    2
```

```
## $Acc
## [1] 0.6216216
##
## $Specif
## [1] 0.8076923
##
## $Sensit
## [1] 0.1818182
```

```
#Q2.1
step <- stepAIC(model)
```

```
## Start:  AIC=171.67
## train.data$low ~ age + lwt + race + smoke + ptd + ht + ui + ftv
##
##         Df Deviance    AIC
## - smoke  1   150.20 170.20
## - race   2   152.38 170.38
## - age    1   150.47 170.47
## - ftv    2   153.15 171.15
## <none>       149.66 171.66
## - ui     1   153.41 173.41
## - ht     1   154.64 174.64
## - lwt    1   156.11 176.11
## - ptd    1   160.41 180.41
##
## Step:  AIC=170.2
## train.data$low ~ age + lwt + race + ptd + ht + ui + ftv
##
##        Df Deviance    AIC
## - race  2   152.42 168.42
## - age   1   151.08 169.08
## <none>      150.20 170.20
## - ftv   2   154.97 170.97
## - ui    1   153.95 171.95
## - ht    1   154.99 172.99
## - lwt   1   157.05 175.05
## - ptd   1   163.12 181.12
##
## Step:  AIC=168.42
## train.data$low ~ age + lwt + ptd + ht + ui + ftv
##
##        Df Deviance    AIC
## - age   1   154.21 168.21
## <none>      152.42 168.42
## - ftv   2   157.48 169.48
## - ui    1   156.06 170.06
## - ht    1   157.41 171.41
## - lwt   1   159.16 173.16
## - ptd   1   165.99 179.99
##
## Step:  AIC=168.21
## train.data$low ~ lwt + ptd + ht + ui + ftv
##
##        Df Deviance    AIC
## <none>      154.21 168.21
## - ui    1   157.82 169.82
## - ftv   2   160.07 170.07
## - ht    1   159.55 171.55
## - lwt   1   162.11 174.11
## - ptd   1   166.70 178.70
```

```
step$anova
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## train.data$low ~ age + lwt + race + smoke + ptd + ht + ui + ftv
##
## Final Model:
## train.data$low ~ lwt + ptd + ht + ui + ftv
##
##
##      Step Df Deviance Resid. Df Resid. Dev      AIC
## 1                           141   149.6654 171.6654
## 2 - smoke  1 0.530663       142   150.1961 170.1961
## 3  - race  2 2.220214       144   152.4163 168.4163
## 4   - age  1 1.790531       145   154.2068 168.2068
```

```
test(step,test.data)
```

```
##      [,1] [,2]
## [1,]   23    3
## [2,]    9    2
```

```
## $Acc
## [1] 0.6756757
##
## $Specif
## [1] 0.8846154
##
## $Sensit
## [1] 0.1818182
```

```
BIC <- stepAIC(model,k=log(nrow(bwt)))
```

```
## Start:  AIC=207.32
## train.data$low ~ age + lwt + race + smoke + ptd + ht + ui + ftv
##
##         Df Deviance    AIC
## - race   2   152.38 199.56
## - ftv    2   153.15 200.32
## - smoke  1   150.20 202.61
## - age    1   150.47 202.88
## - ui     1   153.41 205.83
## - ht     1   154.64 207.06
## <none>       149.66 207.32
## - lwt    1   156.11 208.53
## - ptd    1   160.41 212.83
##
## Step:  AIC=199.56
## train.data$low ~ age + lwt + smoke + ptd + ht + ui + ftv
##
##         Df Deviance    AIC
## - ftv    2   156.98 193.67
## - smoke  1   152.42 194.35
## - age    1   154.18 196.11
## - ui     1   156.03 197.97
## - ht     1   157.41 199.34
## <none>       152.38 199.56
## - lwt    1   159.09 201.02
## - ptd    1   165.06 206.99
##
## Step:  AIC=193.67
## train.data$low ~ age + lwt + smoke + ptd + ht + ui
##
##         Df Deviance    AIC
## - smoke  1   157.48 188.94
## - age    1   159.49 190.94
## - ui     1   161.25 192.70
## - ht     1   161.81 193.26
## <none>       156.98 193.67
## - lwt    1   162.39 193.84
## - ptd    1   167.33 198.78
##
## Step:  AIC=188.93
## train.data$low ~ age + lwt + ptd + ht + ui
##
##        Df Deviance    AIC
## - age   1   160.07 186.28
## - ui    1   161.78 187.98
## - ht    1   162.06 188.27
## <none>      157.48 188.94
## - lwt   1   162.93 189.14
## - ptd   1   168.95 195.16
##
## Step:  AIC=186.28
```

```
## train.data$low ~ lwt + ptd + ht + ui
##
##        Df Deviance    AIC
## - ui    1   164.21 185.17
## - ht    1   165.08 186.05
## <none>      160.07 186.28
## - lwt   1   166.98 187.95
## - ptd   1   170.21 191.17
##
## Step:  AIC=185.17
## train.data$low ~ lwt + ptd + ht
##
##        Df Deviance    AIC
## - ht    1   168.43 184.16
## <none>      164.21 185.17
## - lwt   1   173.03 188.76
## - ptd   1   175.90 191.63
##
## Step:  AIC=184.16
## train.data$low ~ lwt + ptd
##
##        Df Deviance    AIC
## <none>      168.43 184.16
## - lwt   1   176.05 186.53
## - ptd   1   180.69 191.18
```

```
BIC$anova
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## train.data$low ~ age + lwt + race + smoke + ptd + ht + ui + ftv
##
## Final Model:
## train.data$low ~ lwt + ptd
##
##
##      Step Df  Deviance Resid. Df Resid. Dev      AIC
## 1                           141   149.6654 207.3246
## 2  - race  2 2.7173201       143   152.3827 199.5585
## 3   - ftv  2 4.5981331       145   156.9809 193.6731
## 4 - smoke  1 0.5032579       146   157.4841 188.9346
## 5   - age  1 2.5871158       147   160.0712 186.2800
## 6    - ui  1 4.1332841       148   164.2045 185.1715
## 7    - ht  1 4.2258991       149   168.4304 184.1557
```

```
test(BIC,test.data)
```

```
##      [,1] [,2]
## [1,]   24    2
## [2,]    9    2
```

```
## $Acc
## [1] 0.7027027
##
## $Specif
## [1] 0.9230769
##
## $Sensit
## [1] 0.1818182
```

```
#Q2.2
library(leaps)
```

```
## Warning: package 'leaps' was built under R version 3.6.3
```

```
library(bestglm)
```

```
## Warning: package 'bestglm' was built under R version 3.6.3
```

```
library(dummies)
```

```
## dummies-1.5.6 provided by Decision Patterns
```

```
library(tidyverse)

bwt.move<-bwt[,-1]
bwt.move$low<-bwt$low

race = data.frame(dummy(bwt$race)[,c(1,2)])
```

```
## Warning in model.matrix.default(~x - 1, model.frame(~x - 1), contrasts = FALSE):
## non-list contrasts argument ignored
```

```
ftv = data.frame(dummy(bwt$ftv)[,c(2,3)])
```

```
## Warning in model.matrix.default(~x - 1, model.frame(~x - 1), contrasts = FALSE):
## non-list contrasts argument ignored
```

```
bwt.dummy = bwt[,-c(1,4,9)]
low = bwt$low
bwt.dummy = cbind(bwt.dummy,race,ftv, low)

BIC.sub = bestglm(bwt.dummy,IC="BIC",family=binomial)
```

```
## Morgan-Tatar search since family is non-gaussian.
```

```
test(BIC.sub$BestModel,test.data)
```

```
##      [,1] [,2]
## [1,]   24    2
## [2,]    8    3
```

```
## $Acc
## [1] 0.7297297
##
## $Specif
## [1] 0.9230769
##
## $Sensit
## [1] 0.2727273
```

```
#The subset variable BIC seems to be the best, with the stepwise BIC just behind it
```