Assignment_2.2.R

spoor

Mon Mar 27 06:37:18 2017

```
#Evaluating Logistic Regression Model
#Installing Packages for Statistical Study
library(lattice)  #For visualizing data involving multiple variables
library(vcd)  # For visualizing data involving categorical variables
```

```
## Warning: package 'vcd' was built under R version 3.2.5
```

```
## Loading required package: grid
```

```
library(ROCR)  # For evaluating binary classifiers
```

```
## Warning: package 'ROCR' was built under R version 3.2.5
```

```
## Loading required package: gplots
```

```
## Warning: package 'gplots' was built under R version 3.2.5
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
library(UsingR)
```

```
## Warning: package 'UsingR' was built under R version 3.2.5
```

```
## Loading required package: MASS
```

```
## Loading required package: HistData
```

```
## Warning: package 'HistData' was built under R version 3.2.5
```

```
## Loading required package: Hmisc
```

```
## Warning: package 'Hmisc' was built under R version 3.2.5
```

```
## Loading required package: survival
```

```
## Warning: package 'survival' was built under R version 3.2.5
```

```
## Loading required package: Formula
```

```
## Warning: package 'Formula' was built under R version 3.2.5
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.2.5
```

```
##
## Attaching package: 'Hmisc'
```

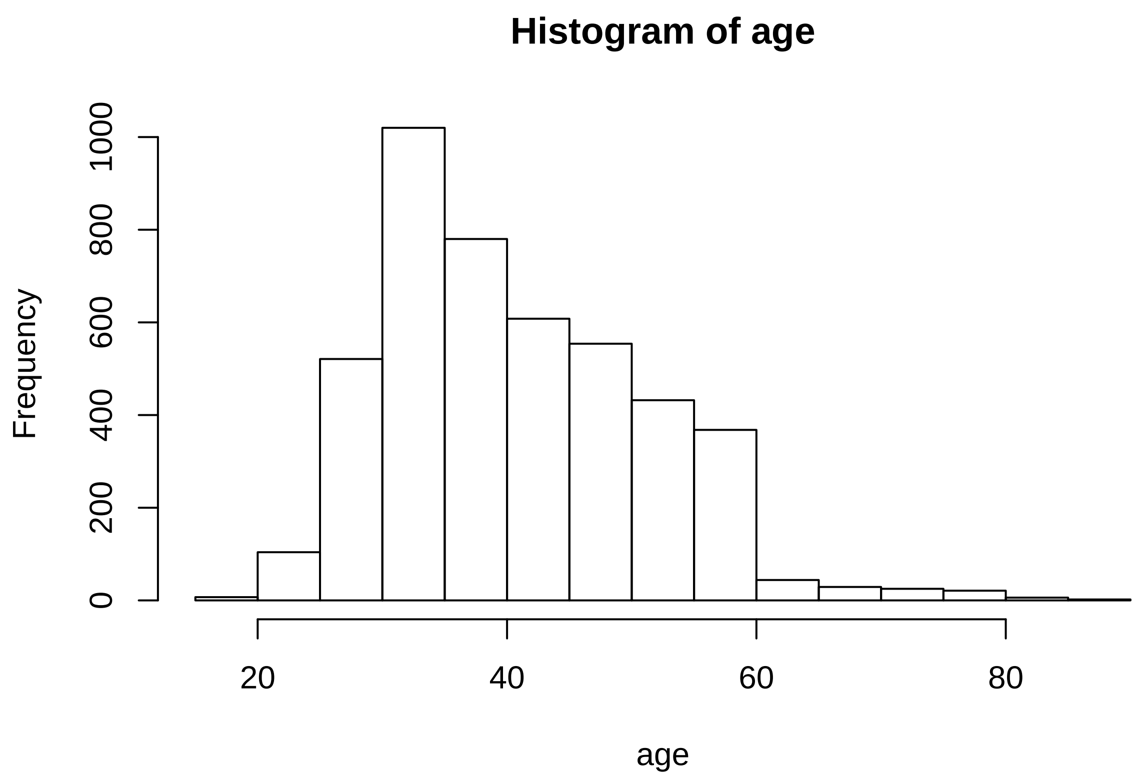
```
## The following objects are masked from 'package:base':
##
##     format.pval, round.POSIXt, trunc.POSIXt, units
```

```
##
## Attaching package: 'UsingR'

## The following object is masked from 'package:survival':
##
##     cancer
```

```
#Reading Data from source file. (CSV File). The dataset is of a Bank
bank_data <- read.csv("C:/Users/spoor/Desktop/Marketing Analytics/bank.csv",sep = ";", stringsAsFactors
#Just to check if the data is loaded correctly and completely
View(bank_data)

#looking at the variables of the dataset
print(names(bank_data))
```

```
##  [1] "age"       "job"       "marital"   "education" "default"
##  [6] "balance"   "housing"   "loan"      "contact"   "day"
## [11] "month"     "duration"  "campaign"  "pdays"     "previous"
## [16] "poutcome"  "response"
```

```
# Let us build a histogram for age
with(bank_data, hist(age))
```




```
#Dispersing the types of jobs into 3 categories, white collar, blue collar and other

white_collar <- c("admin.","entrepreneur","management","self-employed")
blue_collar <- c("blue-collar","services","technician")
bank_data$jobtype <- rep(3, length = nrow(bank_data))
```

```r
bank_data$jobtype <- ifelse((bank_data$job %in% white_collar), 1, bank_data$jobtype)
bank_data$jobtype <- ifelse((bank_data$job %in% blue_collar), 2, bank_data$jobtype)
bank_data$jobtype <- factor(bank_data$jobtype, levels = c(1, 2, 3),
    labels = c("White Collar", "Blue Collar", "Other"))
with(bank_data, table(job, jobtype, useNA = c("always")))  # check definition
```

```
##                  jobtype
## job               White Collar Blue Collar Other <NA>
##   admin.                   478           0     0    0
##   blue-collar                0         946     0    0
##   entrepreneur             168           0     0    0
##   housemaid                  0           0   112    0
##   management               969           0     0    0
##   retired                    0           0   230    0
##   self-employed            183           0     0    0
##   services                   0         417     0    0
##   student                    0           0    84    0
##   technician                 0         768     0    0
##   unemployed                 0           0   128    0
##   unknown                    0           0    38    0
##   <NA>                       0           0     0    0
```

```r
# Similarly dividing other categorical variables into categories
bank_data$marital <- factor(bank_data$marital,
    labels = c("Divorced", "Married", "Single"))
bank_data$education <- factor(bank_data$education,
    labels = c("Primary", "Secondary", "Tertiary", "Unknown"))
bank_data$default <- factor(bank_data$default, labels = c("No", "Yes"))
bank_data$housing <- factor(bank_data$housing, labels = c("No", "Yes"))
bank_data$loan <- factor(bank_data$loan, labels = c("No", "Yes"))
bank_data$response <- factor(bank_data$response, labels = c("No", "Yes"))

# let us check the data where there was no previous contact
# selecting only few variables
bank_work <- subset(bank_data, subset = (previous == 0),
    select = c("response", "age", "jobtype", "marital", "education",
               "default", "balance", "housing", "loan"))

# examine the structure of the bank_work frame and view the frame
print(str(bank_work))
```

```
## 'data.frame':    3705 obs. of  9 variables:
##  $ response : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 2 1 ...
##  $ age      : int  30 30 59 39 41 39 43 36 20 40 ...
##  $ jobtype  : Factor w/ 3 levels "White Collar",..: 3 1 2 2 1 2 1 2 3 1 ...
##  $ marital  : Factor w/ 3 levels "Divorced","Married",..: 2 2 2 2 2 2 2 2 3 2 ...
##  $ education: Factor w/ 4 levels "Primary","Secondary",..: 1 3 2 2 3 2 2 3 2 3 ...
##  $ default  : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ balance  : int  1787 1476 0 147 221 9374 264 1109 502 194 ...
##  $ housing  : Factor w/ 2 levels "No","Yes": 1 2 2 2 2 2 2 1 1 1 ...
##  $ loan     : Factor w/ 2 levels "No","Yes": 1 2 1 1 1 1 1 1 1 2 ...
## NULL
```

```r
View(bank_work)
```

```r
# Let us check the summary statistics for the dataset
print(summary(bank_work))
```

```
##  response       age                 jobtype        marital
##  No :3368   Min.   :19.00   White Collar:1453   Divorced: 443
##  Yes: 337   1st Qu.:33.00   Blue Collar :1776   Married :2305
##             Median :39.00   Other       : 476   Single  : 957
##             Mean   :41.08
##             3rd Qu.:49.00
##             Max.   :87.00
##      education     default       balance      housing     loan
##  Primary  : 580   No :3634   Min.   :-3313   No :1662   No :3113
##  Secondary:1891   Yes:  71   1st Qu.:   60   Yes:2043   Yes: 592
##  Tertiary :1084              Median :  415
##  Unknown  : 150              Mean   : 1375
##                              3rd Qu.: 1412
##                              Max.   :71188
```

```r
#performing the model
bank_spec <- {response ~ age + jobtype + education + marital +
    default + balance + housing + loan}

#Now, we perform the logistic model
bank_data_logit <- glm(bank_spec, family=binomial, data=bank_work)
print(summary(bank_data_logit))
```

```
##
## Call:
## glm(formula = bank_spec, family = binomial, data = bank_work)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8546  -0.4787  -0.3985  -0.3247   2.7165
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -2.250e+00  4.072e-01  -5.526 3.27e-08 ***
## age                 1.004e-02  6.315e-03   1.591 0.111702
## jobtypeBlue Collar -1.435e-01  1.447e-01  -0.992 0.321168
## jobtypeOther        4.139e-01  1.771e-01   2.337 0.019443 *
## educationSecondary  1.036e-01  1.820e-01   0.569 0.569413
## educationTertiary   3.025e-01  2.043e-01   1.481 0.138716
## educationUnknown   -3.338e-01  3.527e-01  -0.946 0.344041
## maritalMarried     -5.717e-01  1.668e-01  -3.428 0.000608 ***
## maritalSingle      -3.509e-02  1.939e-01  -0.181 0.856376
## defaultYes          3.461e-01  3.876e-01   0.893 0.371917
## balance             4.783e-06  1.736e-05   0.276 0.782918
## housingYes         -4.058e-01  1.221e-01  -3.324 0.000888 ***
## loanYes            -6.961e-01  1.997e-01  -3.485 0.000491 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2258.2  on 3704  degrees of freedom
```

```
## Residual deviance: 2177.6  on 3692  degrees of freedom
## AIC: 2203.6
##
## Number of Fisher Scoring iterations: 5
```

```r
print(anova(bank_data_logit, test="Chisq"))
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: response
##
## Terms added sequentially (first to last)
##
##
##           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                       3704     2258.2
## age        1   3.4257      3703     2254.8 0.0641901 .
## jobtype    2  20.1014      3701     2234.7 4.316e-05 ***
## education  3   8.0101      3698     2226.7 0.0458042 *
## marital    2  23.4978      3696     2203.2 7.898e-06 ***
## default    1   0.2848      3695     2202.9 0.5935650
## balance    1   0.2644      3694     2202.6 0.6071299
## housing    1  10.7676      3693     2191.8 0.0010329 **
## loan       1  14.2114      3692     2177.6 0.0001634 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Now we predict the probability
bank_work$Prob_Response <- predict.glm(bank_data_logit, type = "response")

pdf(file = "logit_density_eval.pdf",
    width = 8.5, height = 8.5)
plot_data_work <- densityplot( ~ Prob_Response | response,
               data = bank_work,
               layout = c(1,2), aspect=1, col = "black",
               plot.points = "rug",
               strip=function(...) strip.default(..., style=1),
               xlab="Prediction of Probability")
print(plot_data_work)
dev.off()
```

```
## pdf
##   2
```

```r
# Let us use 50% cut off
bank_work$Pred_Resp <-
    ifelse((bank_work$Prob_Response > 0.5), 2, 1)
bank_work$Pred_Resp <- factor(bank_work$Pred_Resp,
    levels = c(1, 2), labels = c("NO", "YES"))
conf_matrix <- table(bank_work$Pred_Resp, bank_work$response)
cat("\nconf_matrix (rows=Predicted Response, columns=Actual Choice\n")
```

```
##
## conf_matrix (rows=Predicted Response, columns=Actual Choice
```

```r
print(conf_matrix)
```

```
##
##         No  Yes
##   NO  3368  337
##   YES    0    0
```

```r
pred_accuracy <- (conf_matrix[1,1] + conf_matrix[2,2])/
                        sum(conf_matrix)
cat("\nPercent Accuracy: ", round(pred_accuracy * 100, digits = 1))
```

```
##
## Percent Accuracy:  90.9
```

```r
# So, Let us try lower cutoff - 10%
bank_work$Pred_Resp <-
  ifelse((bank_work$Prob_Response > 0.1), 2, 1)
bank_work$Pred_Resp <- factor(bank_work$Pred_Resp,
                              levels = c(1, 2), labels = c("NO", "YES"))
conf_matrix <- table(bank_work$Pred_Resp, bank_work$response)
cat("\nconf_matrix (rows=Predicted Response, columns=Actual Choice\n")
```

```
##
## conf_matrix (rows=Predicted Response, columns=Actual Choice
```

```r
print(conf_matrix)
```

```
##
##         No  Yes
##   NO  2262  159
##   YES 1106  178
```

```r
pred_accuracy <- (conf_matrix[1,1] + conf_matrix[2,2])/
  sum(conf_matrix)
cat("\nPercent Accuracy: ", round(pred_accuracy * 100, digits = 1))
```

```
##
## Percent Accuracy:  65.9
```