Week9_Assignment9.2_KoppulaVeera

Veera Koppula

August 04 2021

##Fit a Logistic Regression Model to Thoracic Surgery Binary Dataset. #For this problem, you will be working with the thoracic surgery data set from the University of California Irvine machine learning repository. This dataset contains information on life expectancy in lung cancer patients after surgery. The underlying thoracic surgery data is in ARFF format. This is a text-based format with information on each of the attributes. You can load this data using a package such as foreign or by cutting and pasting the data section into a CSV file.

##Assignment Instructions: #Fit a binary logistic regression model to the data set that predicts whether or not the patient survived for one year (the Risk1Y variable) after the surgery. Use the glm() function to perform the logistic regression. See Generalized Linear Models for an example. Include a summary using the summary() function in your results.

```
##
## Call:
## glm(formula = Risk1Yr ~ ., family = binomial, data = thoracic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4929   0.2762   0.4199   0.5439   1.6084
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.604e+01  2.333e+03   0.011 0.991093
## DGNDGN2     -5.557e-01  4.128e-01  -1.346 0.178199
## DGNDGN4     -4.278e-01  4.733e-01  -0.904 0.366122
## DGNDGN6      1.377e+01  1.178e+03   0.012 0.990671
## DGNDGN5     -2.201e+00  6.113e-01  -3.600 0.000318 ***
## DGNDGN8     -3.852e+00  1.550e+00  -2.485 0.012959 *
## DGNDGN1      1.418e+01  2.400e+03   0.006 0.995285
## PRE4         2.272e-01  1.849e-01   1.229 0.219094
## PRE5         3.030e-02  1.786e-02   1.697 0.089715 .
## PRE6PRZ1     1.490e-01  5.783e-01   0.258 0.796647
## PRE6PRZ0    -2.937e-01  7.907e-01  -0.371 0.710303
## PRE7F        7.153e-01  5.556e-01   1.288 0.197884
## PRE8F        1.743e-01  3.892e-01   0.448 0.654188
## PRE9F        1.368e+00  4.868e-01   2.811 0.004942 **
## PRE10F       5.770e-01  4.826e-01   1.196 0.231855
## PRE11F       5.162e-01  3.965e-01   1.302 0.192948
## PRE14OC14   -1.653e+00  6.094e-01  -2.713 0.006675 **
## PRE14OC12   -4.394e-01  3.301e-01  -1.331 0.183177
## PRE14OC13   -1.179e+00  6.165e-01  -1.913 0.055799 .
## PRE17F       9.266e-01  4.445e-01   2.085 0.037092 *
```

```
## PRE19F      -1.466e+01  1.654e+03  -0.009 0.992928
## PRE25F      -9.789e-02  1.003e+00  -0.098 0.922273
## PRE30F       1.084e+00  4.990e-01   2.172 0.029840 *
## PRE32F      -1.398e+01  1.645e+03  -0.008 0.993219
## AGE          9.506e-03  1.810e-02   0.525 0.599442
## ---
## 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: 341.19  on 445  degrees of freedom
## AIC: 391.19
##
## Number of Fisher Scoring iterations: 15
```

#According to the summary, which variables had the greatest effect on the survival rate? Variables that have greatest effect on survival rate are - AGE, Type 2 DM - diabetes mellitus(PRE17), PAD - peripheral arterial diseases(PRE25)

#To compute the accuracy of your model, use the dataset to predict the outcome variable. The percent of correct predictions is the accuracy of your model. What is the accuracy of your model?

```
## [1] 0.2281786
```

Accuracy prediction for my model is about 23%

##Fit a Logistic Regression Model

```
## ResourceSelection 0.3-5   2019-07-22
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   label = col_double(),
##   x = col_double(),
##   y = col_double()
## )
```

```
## spec_tbl_df [1,498 x 3] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ label: num [1:1498] 0 0 0 0 0 0 0 0 0 0 ...
##  $ x    : num [1:1498] 70.9 75 73.8 66.4 69.1 ...
##  $ y    : num [1:1498] 83.2 87.9 92.2 81.1 84.5 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   label = col_double(),
##   ..   x = col_double(),
##   ..   y = col_double()
##   .. )
```

#Fit a logistic regression model to the binary-classifier-data.csv dataset

```
##
## Call:
## glm(formula = label ~ x, family = "binomial", data = binclass)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.246  -1.159  -1.065   1.184   1.293
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.137369   0.095119   1.444   0.1487
## x           -0.004119   0.001775  -2.321   0.0203 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2075.8  on 1497  degrees of freedom
## Residual deviance: 2070.4  on 1496  degrees of freedom
## AIC: 2074.4
##
## Number of Fisher Scoring iterations: 3
```

#The dataset (found in binary-classifier-data.csv) contains three variables; label, x, and y. The label variable is either 0 or 1 and is the output we want to predict using the x and y variables.

```
##
## Call:
## glm(formula = label ~ x + y, family = "binomial", data = binclass)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3728  -1.1697  -0.9575   1.1646   1.3989
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.424809   0.117224   3.624  0.00029 ***
## x           -0.002571   0.001823  -1.411  0.15836
## y           -0.007956   0.001869  -4.257 2.07e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2075.8  on 1497  degrees of freedom
## Residual deviance: 2052.1  on 1495  degrees of freedom
## AIC: 2058.1
##
## Number of Fisher Scoring iterations: 4
```

#What is the accuracy of the logistic regression classifier?

```
## [1] 0.5732267
```

Accuracy using ROC model is 57% - This is the accuracy of the logical regression classifier

#Keep this assignment handy, as you will be comparing your results from this week to next week.