Week 8 Exercise 13

Nosky, Christopher

2/19/2021

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:

Include all of your answers in a R Markdown report. Here is an example R Markdown report that you can use as a guide.

1. 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.

# Model with all binary variables included  
LogModel.1 <- glm(Risk1Yr ~ PRE6 + PRE7 + PRE8 + PRE9 + PRE10 + PRE11 + PRE14 + PRE17 + PRE19 + PRE25 + PRE30, data = surgery\_train, family = 'binomial')  
summary(LogModel.1)

##   
## Call:  
## glm(formula = Risk1Yr ~ PRE6 + PRE7 + PRE8 + PRE9 + PRE10 + PRE11 +   
## PRE14 + PRE17 + PRE19 + PRE25 + PRE30, family = "binomial",   
## data = surgery\_train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.3009 -0.5820 -0.4883 -0.3154 2.5895   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.31713 0.60289 -5.502 3.75e-08 \*\*\*  
## PRE6PRZ1 0.08484 0.54654 0.155 0.8766   
## PRE6PRZ2 0.04963 0.84725 0.059 0.9533   
## PRE7T 0.45897 0.58192 0.789 0.4303   
## PRE8T 0.29885 0.42263 0.707 0.4795   
## PRE9T 1.04079 0.54214 1.920 0.0549 .   
## PRE10T 0.25632 0.50631 0.506 0.6127   
## PRE11T 0.42326 0.43083 0.982 0.3259   
## PRE14OC12 0.19134 0.36213 0.528 0.5972   
## PRE14OC13 1.31147 0.63835 2.054 0.0399 \*   
## PRE14OC14 1.16772 0.64616 1.807 0.0707 .   
## PRE17T 0.91675 0.47343 1.936 0.0528 .   
## PRE19T -13.11407 882.74348 -0.015 0.9881   
## PRE25T 0.57241 0.89465 0.640 0.5223   
## PRE30T 0.90937 0.52228 1.741 0.0817 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 297.30 on 338 degrees of freedom  
## Residual deviance: 271.74 on 324 degrees of freedom  
## AIC: 301.74  
##   
## Number of Fisher Scoring iterations: 13

# Model with significant varibles only  
LogModel.2 <- glm(Risk1Yr ~ PRE9 + PRE14 + PRE17, data = surgery\_train, family = 'binomial')  
summary(LogModel.2)

##   
## Call:  
## glm(formula = Risk1Yr ~ PRE9 + PRE14 + PRE17, family = "binomial",   
## data = surgery\_train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.3277 -0.5262 -0.5262 -0.4734 2.1186   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.1321 0.2823 -7.554 4.23e-14 \*\*\*  
## PRE9T 1.0438 0.5294 1.971 0.0487 \*   
## PRE14OC12 0.2250 0.3500 0.643 0.5203   
## PRE14OC13 1.4350 0.6237 2.301 0.0214 \*   
## PRE14OC14 1.3439 0.6331 2.123 0.0338 \*   
## PRE17T 0.8708 0.4616 1.886 0.0592 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 297.30 on 338 degrees of freedom  
## Residual deviance: 280.88 on 333 degrees of freedom  
## AIC: 292.88  
##   
## Number of Fisher Scoring iterations: 4

1. According to the summary, which variables had the greatest effect on the survival rate?

* According to the summary of LogModel.1 the most significant variables were PRE9, PRE14 and PRE17 though for PRE14 only the tumor size OC13 appears to be significant. As shown above I also created LogModel.2 to compare the results.

1. 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?

* LogModel.1 Confusion matrix and Accuracy

surgery.1\_train\_pred <- predict(LogModel.1, surgery\_train, type = 'response')  
  
surgery.1\_test\_pred <- predict(LogModel.1, surgery\_test, type= 'response')   
  
# Model 1 Train confusion matrix  
confmatrix.1 <- table(Actual\_Value=surgery\_train$Risk1Yr, Predicted\_Value = surgery.1\_train\_pred > 0.5)  
confmatrix.1

## Predicted\_Value  
## Actual\_Value FALSE TRUE  
## F 281 4  
## T 54 0

# Model 2 Test confusion matrix  
confmatrix.2 <- table(Actual\_Value=surgery\_test$Risk1Yr, Predicted\_Value = surgery.1\_test\_pred > 0.5)  
confmatrix.2

## Predicted\_Value  
## Actual\_Value FALSE TRUE  
## F 114 1  
## T 16 0

# Model 1 Train acc  
(confmatrix.1[[1,1]] + confmatrix.1[[2,2]]) / sum(confmatrix.1)

## [1] 0.8289086

# Model 1 Test acc  
(confmatrix.2[[1,1]] + confmatrix.2[[2,2]]) / sum(confmatrix.2)

## [1] 0.870229

* LogModel.2 Confusion matrix and Accuracy

# Model 2 test and train prediction   
surgery.2\_train\_pred <- predict(LogModel.2, surgery\_train, type = 'response')  
  
surgery.2\_test\_pred <- predict(LogModel.2, surgery\_test, type = 'response')  
  
# Model 2 Train confusion matrix  
confmatrix.3 <- table(Actual\_Value=surgery\_train$Risk1Yr, Predicted\_Value = surgery.2\_train\_pred > 0.5)  
confmatrix.3

## Predicted\_Value  
## Actual\_Value FALSE TRUE  
## F 281 4  
## T 51 3

# Model 2 Train Confusion matrix  
confmatrix.4 <- table(Actual\_Value=surgery\_test$Risk1Yr, Predicted\_Value = surgery.2\_test\_pred > 0.5)  
confmatrix.4

## Predicted\_Value  
## Actual\_Value FALSE TRUE  
## F 114 1  
## T 16 0

# Model 2 Train acc  
(confmatrix.3[[1,1]] + confmatrix.3[[2,2]]) / sum(confmatrix.3)

## [1] 0.8377581

# Model 2 Test acc  
(confmatrix.4[[1,1]] + confmatrix.4[[2,2]]) / sum(confmatrix.4)

## [1] 0.870229

* It would appear that there isn’t a significant difference in Accuracy between the two models as shown. Both models present an accuracy of 84-85%.