Classification Model for predicting mortality of pediatric patients with brain injuries

Na’il

7/16/2021

Overview:

This machine learning data report is for a University of Colorado School of Medicine data predicition Hackathon. The purpose of the hackathon is to develop a classification model for hospital mortality and another model for total functional status score (FSS, restricted integers) for pediatric patients with brain injuries. This report is specific to classifying pediatric mortality, a binary variable. Accuracy is used as the model performance metric. High performance models were developed: a random forest model and a null model. Future work includes calculating other performance metrics, including sensitivity and specificity.

Load data set into R

ds<-read.csv('../csvs/training.csv')

-optional i- Create R environment with training data set.

env\_hptbi<-new.env()  
assign('ds',ds,envir=env\_hptbi)  
save(env\_hptbi,file='env\_hptbi.rdata')

-optional ii: Loads environment with data set-

env\_hptbi\_<-load(file='env\_hptbi.rdata')  
attach(env\_hptbi)

Preview data

dim(ds)

## [1] 300 103

head(ds)

## studyid age female sourceinj injurytoadmit  
## 1 102 1179 0 Other provider note in EMR 0  
## 2 103 90 0 Trauma surgery attending 0  
## 3 110 1164 1 ED attending 0  
## 4 112 1413 1 Trauma surgery attending 0  
## 5 116 5791 0 Other provider note in EMR 0  
## 6 120 6190 1 Other provider note in EMR 0  
## injurymech gcsyned gcseyeed gcsverbaled gcsmotored gcsed  
## 1 Traffic 1 1 1 4 6  
## 2 Known or suspected abuse 1 1 1 2 4  
## 3 Other 1 2 1 4 7  
## 4 Fall 1 4 1 5 10  
## 5 Traffic 1 1 1 5 7  
## 6 Fall 1 1 1 1 3  
## gcsetted gcsseded gcspared gcseyeobed eddisposition admittoct  
## 1 1 1 0 0 Intensive Care Unit 0  
## 2 1 1 0 0 Intensive Care Unit 0  
## 3 1 1 0 0 Intensive Care Unit 0  
## 4 1 1 0 0 Intensive Care Unit 0  
## 5 1 1 0 0 Intensive Care Unit 0  
## 6 1 1 1 0 Operating Room 0  
## ctskullfrac ctce ctmidlineshift ctcompress ctintraparhem ctsubarchhem  
## 1 0 1 0 0 1 0  
## 2 0 0 0 0 0 1  
## 3 0 0 0 0 0 0  
## 4 1 0 1 0 0 0  
## 5 1 0 1 0 0 0  
## 6 0 1 1 1 0 0  
## ctintraventhem ctsubhematoma ctepihematoma sourceicu  
## 1 0 0 0 ICU MD  
## 2 0 1 0 ICU MD  
## 3 0 0 0 ICU MD  
## 4 0 1 1 ICU nursing flowsheet  
## 5 0 1 0 ICU MD  
## 6 0 1 0 ICU MD  
## puplrcticu gcsynicu gcseyeicu gcsverbalicu gcsmotoricu gcsicu gcsetticu  
## 1 Both Reactive 1 1 1 4 6 1  
## 2 Both Fixed 1 2 1 4 7 1  
## 3 Both Reactive 1 2 1 4 7 1  
## 4 Both Reactive 1 4 1 6 11 1  
## 5 Both Reactive 1 1 1 5 7 1  
## 6 Both Fixed 1 1 1 1 3 1  
## gcssedicu gcsparicu gcseyeobicu admittoicudc1 admittoicuadmit2 admittoicudc2  
## 1 1 0 0 18 NA NA  
## 2 1 0 1 10 NA NA  
## 3 1 0 0 2 3 8  
## 4 1 0 0 2 NA NA  
## 5 1 0 0 10 NA NA  
## 6 1 0 0 10 NA NA  
## admittoicuadmit3 admittoicudc3 ventyn admittoint admittoext icpyn1  
## 1 NA NA 1 2 0 0  
## 2 NA NA 1 0 9 0  
## 3 NA NA 1 1 0 0  
## 4 NA NA 1 1 0 1  
## 5 NA NA 1 0 7 1  
## 6 NA NA 1 9 0 1  
## icptype1 icptype2 icptype3  
## 1   
## 2   
## 3   
## 4 Intraparenchymal (Camino or bolt)   
## 5 Intraparenchymal (Camino or bolt)   
## 6 Ventriculostomy (External Ventricular Drain or EVD)   
## admittoicpstart1 admittoicpend1 admittoicpstart2 admittoicpend2  
## 1 NA NA NA NA  
## 2 NA NA NA NA  
## 3 NA NA NA NA  
## 4 0 1 NA NA  
## 5 0 6 NA NA  
## 6 0 10 NA NA  
## admittoicpstart3 admittoicpend3 cathtype1  
## 1 NA NA Central venous catheter  
## 2 NA NA Central venous catheter  
## 3 NA NA Central venous catheter  
## 4 NA NA Arterial catheter  
## 5 NA NA Peripherally inserted central catheter (PICC)  
## 6 NA NA Arterial catheter  
## cathtype2  
## 1   
## 2 Arterial catheter  
## 3   
## 4 Central venous catheter  
## 5 Central venous catheter  
## 6 Peripherally inserted central catheter (PICC)  
## cathtype3 cathtype4 admittocathstart1  
## 1 1  
## 2 Peripherally inserted central catheter (PICC) 1  
## 3 0  
## 4 0  
## 5 Arterial catheter 3  
## 6 0  
## admittocathstart2 admittocathstart3 admittocathstart4 admittocathend1  
## 1 NA NA NA 4  
## 2 0 2 NA 5  
## 3 NA NA NA 2  
## 4 0 NA NA 2  
## 5 0 0 NA 13  
## 6 2 NA NA 9  
## admittocathend2 admittocathend3 admittocathend4 newtrachyn admittotrach  
## 1 NA NA NA 1 15  
## 2 2 11 NA 0 NA  
## 3 NA NA NA 0 NA  
## 4 2 NA NA 0 NA  
## 5 3 10 NA 0 NA  
## 6 9 NA NA 0 NA  
## newgastyn admittogast decomcranyn admittocrani lmbrdrainyn admittolmbdrain  
## 1 1 10 0 NA 0 NA  
## 2 0 NA 0 NA 0 NA  
## 3 0 NA 0 NA 0 NA  
## 4 0 NA 0 NA 0 NA  
## 5 0 NA 0 NA 0 NA  
## 6 0 NA 1 0 0 NA  
## epihemyn admittoedhevac subhemyn admittosdhevac rxhypsal rxmann rxbarb  
## 1 0 NA 0 NA 1 0 0  
## 2 0 NA 0 NA 0 0 1  
## 3 0 NA 0 NA 0 0 0  
## 4 0 NA 0 NA 0 0 0  
## 5 0 NA 0 NA 1 1 0  
## 6 0 NA 1 0 1 1 0  
## rxinotrvas tpnyn admittotpn entnutyn admittoentnut hosplos  
## 1 0 0 NA 1 2 22  
## 2 1 1 2 1 6 24  
## 3 0 0 NA 1 3 9  
## 4 0 0 NA 1 4 6  
## 5 1 0 NA 1 9 36  
## 6 0 0 NA 1 0 23  
## hospdisposition cardiacarrestyn cardiacarrestprehosp cardiacarrested  
## 1 Inpatient Rehab 0 0 0  
## 2 Home, no new supports 1 1 0  
## 3 Home, no new supports 0 0 0  
## 4 Home, no new supports 0 0 0  
## 5 Home, no new supports 0 0 0  
## 6 Other 0 0 0  
## cardiacarrestor cardiacarresticu cardiacarrestother admittofss  
## 1 0 0 0 22  
## 2 0 0 0 16  
## 3 0 0 0 9  
## 4 0 0 0 6  
## 5 0 0 0 16  
## 6 0 0 0 23  
## sourcefss fssmental fsssensory fsscommun fssmotor  
## 1 Other 2 2 2 3  
## 2 Verbal report from floor physician 3 2 4 3  
## 3 Other 1 1 1 2  
## 4 Other 1 1 1 2  
## 5 Verbal report from floor nurse 2 1 2 3  
## 6 Other 4 5 5 5  
## fssfeeding fssresp  
## 1 3 3  
## 2 4 1  
## 3 1 1  
## 4 1 1  
## 5 3 1  
## 6 3 1

Generate automated EDA report before modeling

Warning: package 'DataExplorer' was built under R version 4.0.5

processing file: report.rmd

| | | 0% | |.. | 2%  
 inline R code fragments  
  
 | |... | 5%  
label: global\_options (with options)   
List of 1  
 $ include: logi FALSE  
  
 | |..... | 7%  
 ordinary text without R code  
  
 | |....... | 10%  
label: introduce  
 | |........ | 12%  
 ordinary text without R code  
  
 | |.......... | 14%  
label: plot\_intro

| |............ | 17%  
 ordinary text without R code  
  
 | |............. | 19%  
label: data\_structure  
 | |............... | 21%  
 ordinary text without R code  
  
 | |................. | 24%  
label: missing\_profile

| |.................. | 26%  
 ordinary text without R code  
  
 | |.................... | 29%  
label: univariate\_distribution\_header  
 | |...................... | 31%  
 ordinary text without R code  
  
 | |....................... | 33%  
label: plot\_histogram

| |......................... | 36%  
 ordinary text without R code  
  
 | |........................... | 38%  
label: plot\_density  
 | |............................ | 40%  
 ordinary text without R code  
  
 | |.............................. | 43%  
label: plot\_frequency\_bar

| |................................ | 45%  
 ordinary text without R code  
  
 | |................................. | 48%  
label: plot\_response\_bar  
 | |................................... | 50%  
 ordinary text without R code  
  
 | |..................................... | 52%  
label: plot\_with\_bar  
 | |...................................... | 55%  
 ordinary text without R code  
  
 | |........................................ | 57%  
label: plot\_normal\_qq

| |.......................................... | 60%  
 ordinary text without R code  
  
 | |........................................... | 62%  
label: plot\_response\_qq  
 | |............................................. | 64%  
 ordinary text without R code  
  
 | |............................................... | 67%  
label: plot\_by\_qq  
 | |................................................ | 69%  
 ordinary text without R code  
  
 | |.................................................. | 71%  
label: correlation\_analysis  
 | |.................................................... | 74%  
 ordinary text without R code  
  
 | |..................................................... | 76%  
label: principal\_component\_analysis  
 | |....................................................... | 79%  
 ordinary text without R code  
  
 | |......................................................... | 81%  
label: bivariate\_distribution\_header  
 | |.......................................................... | 83%  
 ordinary text without R code  
  
 | |............................................................ | 86%  
label: plot\_response\_boxplot  
 | |.............................................................. | 88%  
 ordinary text without R code  
  
 | |............................................................... | 90%  
label: plot\_by\_boxplot  
 | |................................................................. | 93%  
 ordinary text without R code  
  
 | |................................................................... | 95%  
label: plot\_response\_scatterplot  
 | |.................................................................... | 98%  
 ordinary text without R code  
  
 | |......................................................................| 100%  
label: plot\_by\_scatterplot

output file: C:/Users/nmitchell3/Desktop/hptbi-hackathon/nm/report.knit.md

"C:/Program Files/RStudio/bin/pandoc/pandoc" +RTS -K512m -RTS "C:/Users/nmitchell3/Desktop/hptbi-hackathon/nm/report.knit.md" --to html4 --from markdown+autolink\_bare\_uris+tex\_math\_single\_backslash --output pandoc2ebc5c423c52.html --lua-filter "C:\Users\nmitchell3\Documents\R\win-library\4.0\rmarkdown\rmarkdown\lua\pagebreak.lua" --lua-filter "C:\Users\nmitchell3\Documents\R\win-library\4.0\rmarkdown\rmarkdown\lua\latex-div.lua" --self-contained --variable bs3=TRUE --standalone --section-divs --table-of-contents --toc-depth 6 --template "C:\Users\nmitchell3\Documents\R\win-library\4.0\rmarkdown\rmd\h\default.html" --no-highlight --variable highlightjs=1 --variable theme=yeti --include-in-header "C:\Users\NMITCH~1\AppData\Local\Temp\RtmpumOM56\rmarkdown-str2ebc2f1333b6.html" --mathjax --variable "mathjax-url:https://mathjax.rstudio.com/latest/MathJax.js?config=TeX-AMS-MML\_HTMLorMML"

Output created: report.html

Based on the report, any variable starting with admitto contains high amounts of missing data. All of these variables must be dropped

library('dplyr')

## Warning: package 'dplyr' was built under R version 4.0.5

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:data.table':  
##   
## between, first, last

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

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

ds2<-ds%>%  
select(-starts\_with('admitto'))  
  
100\*(ncol(ds)-ncol(ds2))/ncol(ds)

## [1] 30.09709

30% of the number of variables in the original data set were removed. Now missing data can be dropped

ds2<-na.omit(ds2)

About one third of the observations were removed. Now a model can be fitted. First, I need to choose a target variable. I’ll choose Mortality. The author had some coding for configuring/initializing the target variable. I’ll paste it below and substitute ds2 for hackathon\_mortality\_data

ds2\_<-ds2  
ds2\_$mortality <-  
 as.integer(ds2\_$hospdisposition == "Mortality")  
  
 ds2\_<-  
 ds2\_[-grep("fss", names(ds2))]

Next, the response variable will be converted to a factor.

ds2\_[['mortality']]<-as.factor(ds2\_[['mortality']])

This dropped another 6 variables, now totaling 66. Now that pre-processing the data has finished, the data can be divided into training set 1 and training set 2. Training set 2 is essentially a test set. The sets will be created as random samples. I will first use the set.seed() function to ensure reproducibility of results

set.seed(1)  
sampleInd<-sample(1:nrow(ds2\_),.8\*nrow(ds2\_))  
  
train\_<-ds2\_[sampleInd,]  
test\_<-ds2\_[-sampleInd,]

Now that training and test sets have been created, a model(s) can be selected. First, I’ll use the table function to identify class proportions of the target variable

table(ds2\_[['mortality']])

##   
## 0 1   
## 196 7

Based on the class proportions, 97% of the observations have a mortality of 0, ie they don’t die. Using the null model, which predicts the most common occurring class (which is 0) all of the time, I’d expect to have a high accuracy model. A vector of those predictions is created below.

pred\_null2\_<-rep(0,41)

The next model that will be tried is random forests, a collection of decision trees; ie bootstrapped decision trees.

set.seed(1)  
library(randomForest)

## Warning: package 'randomForest' was built under R version 4.0.5

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##   
## Attaching package: 'randomForest'

## The following object is masked from 'package:dplyr':  
##   
## combine

rf2\_<-randomForest(mortality~.,data=ds2\_,mtry=6)

Now that the random forest model has been fit with the training data (training\_), predictions can be made on the ‘test’ set (test\_)

pred\_rf2\_<-predict(rf2\_,newdata=subset(test\_,select=-mortality))

Let’s see the proportions.

table(pred\_rf2\_)

## pred\_rf2\_  
## 0 1   
## 40 1

As you can see, 98% of the predictions were 0 (no mortality), 2% off from the true number of no mortality. Finally, accuracy will be computed for each model on the ‘test’ set (test\_), which is one of the metrics in the competition. First, a vector will be created with the true response variable classes, then accuracy will be computed.

test\_mort<-test\_$mortality  
  
print('Accuracy null model on test\_:')

## [1] "Accuracy null model on test\_:"

mean(pred\_null2\_==test\_mort)

## [1] 0.9756098

print('Accuracy random forest model on test\_:')

## [1] "Accuracy random forest model on test\_:"

mean(pred\_rf2\_==test\_mort)

## [1] 1

The null model has 98% accuracy, the random forest model has 100% accuracy on test\_.