final189

Kalyani Cauwenberghs

3/7/2020

Part 1

Instructions: Compile a collection of tips for best data presentation, including the illustration and R script for each TODO: what does she mean for R script and illustration? resolved one should boxplot next to each other –> provide sample code how that's done use Rmd, show R code and plot. #Part 2 Instructions: For the final project, look back at all the analysis approaches you have used throughout the quarter. Consider HealthGen as the outcome, grouped into Excellent/Vgood, versus Good/Fair/Poor. (In the dataset dat.Rda, the HealthGen variable is assigned value 1 if its original value is Excellent/Vgood and is assigned value 0 otherwise.) Consider all other variables as potential predictors. Develop a comprehensive and reproducible analysis report, to explore the relationship between these variables and the outcome. Pay attention to (but not limited to) the following:

1.)

Missing data: do not remove observations with any missing data from the start; after screening you might reduce to a smaller set of variables, therefore remove fewer observations at that point. Also you may consider removing variables with too much missing.

First step: remove the two columns that have missing data for more than half its observations

```
load("dat (1).rda")
#remove columns with more than half the observations missing
dat<-dat[,-which(colSums(is.na(dat)) > nrow(dat)/2)]
```

Second step: for all categorical variables, reduce the number of categories by combining sparse categories

```
for(i in 1:ncol(dat)){
  if(class(dat[,i])=="factor"){
    print(colnames(dat)[i])
    print(table(dat[,i])/length(dat[,i]))
  }
}
```

Third step: Univariately screen all the remaining predictors and get rid of those whose p-value > 0.2.

```
## [1] "our final screened variables:"
```

```
##  [1] "Gender"        "Age"           "Race1"         "Education"
##  [5] "MaritalStatus" "HHIncome"      "BMI_WHO"       "Depressed"
##  [9] "SleepTrouble"  "TVHrsDay"      "AlcoholYear"   "RegularMarij"
```

2.)

Include “Table 1”

3.)

After univariate screening, building a multiple logistic regression model to predict the general health outcome of very good or excellent versus otherwise. State clearly your criteria at each step in the narrative.

General algorithm: TODO

At each step, we will keep track of the inidces of the variables

```
##
## p:  0.010075 3.174261e-05 0.0006538256 0.0007645511 0.7278755 3.712056e-09 1.634751e-29 1.711103e-10
##
## ind:  1 2 3 4 6 7 9 10
## p:  0.007422199 6.011993e-06 0.0004398826 0.0002094175 1.877886e-10 2.991672e-31 2.979079e-11 4.0281
```

4.)

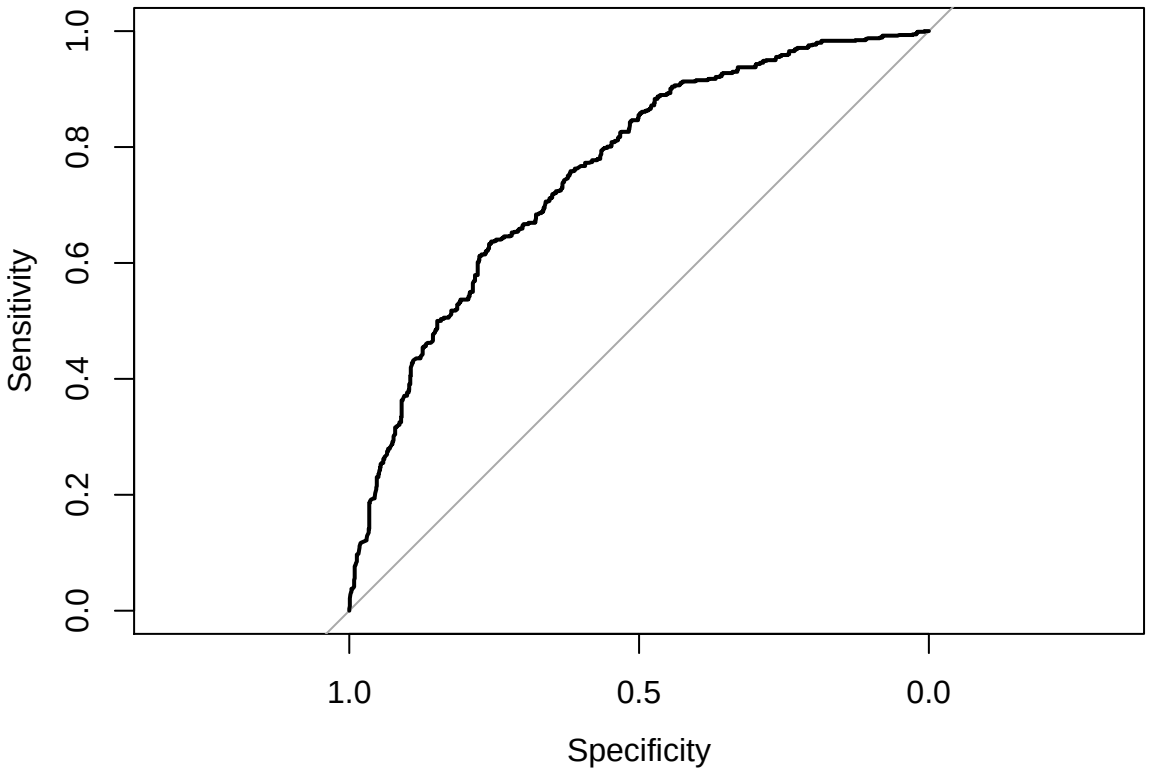
Assess the predictability of the model by computing the (generalized) R-squared and the area under the ROC curve (AUC), as well as the cross-validated AUC.

Find generalized R-squared:

```
## ##generalized R-squared:  0.1943529
```

ROC and AUC

```
## [1] "Plotting ROC of our selected model"
```



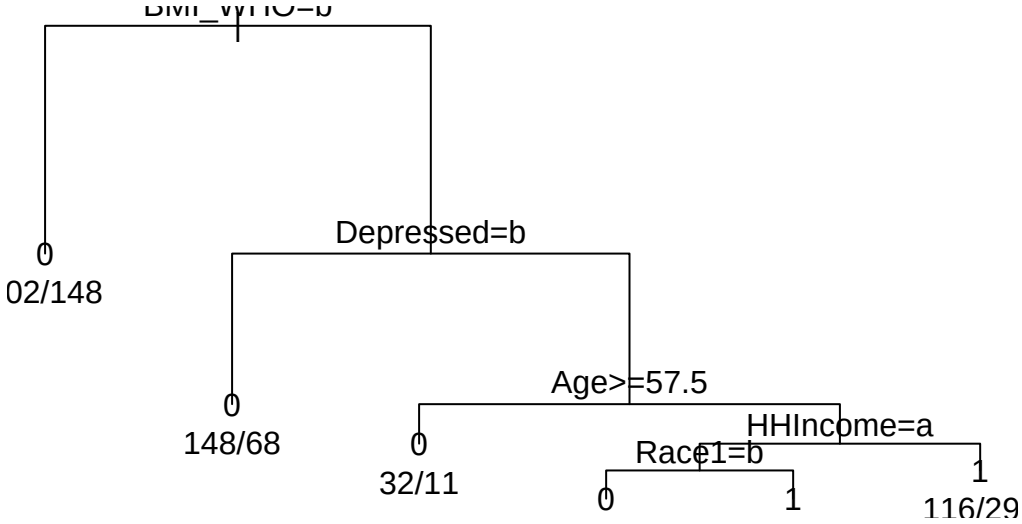
```
## [1] "AUC of our selected model"
## Area under the curve: 0.7585
```

Cross-validated ROC and AUC

```
## [1] "final AUC:"
## [1] 0.7588496
```

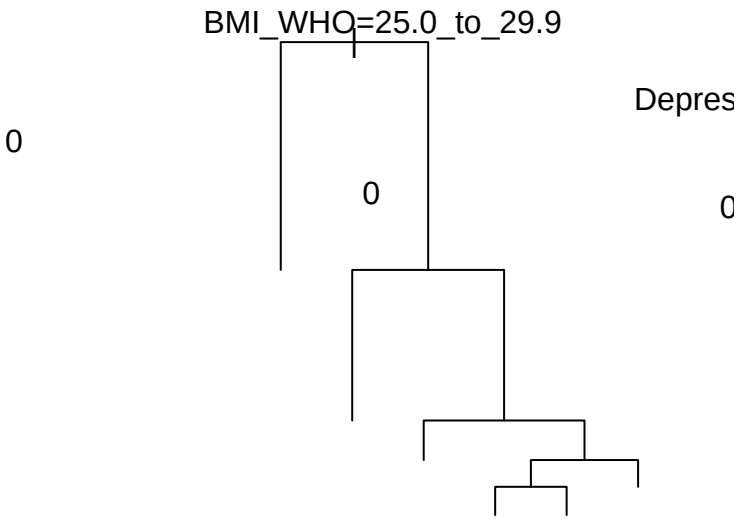
5.)

Use the variables that have passed the univariate screening, to build a classification tree. Describe clearly how you arrive at the final tree. Compute the error rate of your classification tree.



Approach 1:

```
par(mfrow = c(1,2), xpd = NA)
plot.new()
text(fit, pretty=0)
plot(fit)
```




```
plot(fit)
text(fit, pretty=0, use.n = TRUE)
```

```
dev.off()
```

```
## null device
##           1
```

```
printcp(fit)
```

```
##
## Classification tree:
## rpart(formula = form(new_ind), data = dat, subset = train)
##
## Variables actually used in tree construction:
## [1] Age       BMI_WHO   Depressed HHIncome  Race1
##
## Root node error: 722/1577 = 0.45783
##
## n= 1577
##
##         CP nsplit rel error  xerror     xstd
## 1 0.167590      0   1.00000 1.00000 0.027403
## 2 0.110803      1   0.83241 0.83241 0.026712
## 3 0.029086      2   0.72161 0.72161 0.025870
## 4 0.010388      3   0.69252 0.69252 0.025594
## 5 0.010000      5   0.67175 0.70360 0.025702
```

Approach 2:

Testing error

6.)

Discuss any limitations in the analysis.

Bonus

Explore random forest on the data above.