final189

Kalyani Cauwenberghs

3/7/2020

Part 1

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.

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

```
## [1] "Age"           "Education"     "MaritalStatus" "BMI_WHO"
## [5] "SleepTrouble"  "CompHrsDay"
```

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.

```
## [1] "Final model:"
```

```
##               (Intercept)                       Age   Education9 - 11th Grade
##              -1.767844403              -0.017383837                1.066092225
##      EducationHigh School     EducationSome College     EducationCollege Grad
##                1.499309047               1.698981822                2.393716018
##  MaritalStatusLivePartner      MaritalStatusMarried MaritalStatusNeverMarried
##                0.287987778               0.247213482                0.055079865
##    MaritalStatusSeparated      MaritalStatusWidowed       BMI_WHO18.5_to_24.9
```

```
##              0.216977250             -0.321665887              0.964961151
##       BMI_WHO25.0_to_29.9          BMI_WHO30.0_plus           SleepTroubleYes
##              0.598235207             -0.384106182             -0.895935885
##       CompHrsDay0_to_1_hr           CompHrsDay1_hr           CompHrsDay2_hr
##              0.115407406              0.377522538              0.303841361
##           CompHrsDay3_hr           CompHrsDay4_hr       CompHrsDayMore_4_hr
##              0.001982988             -0.223866436             -0.099279708
```

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.

```
## [1] 2337
```

```
## [1] 2337
```

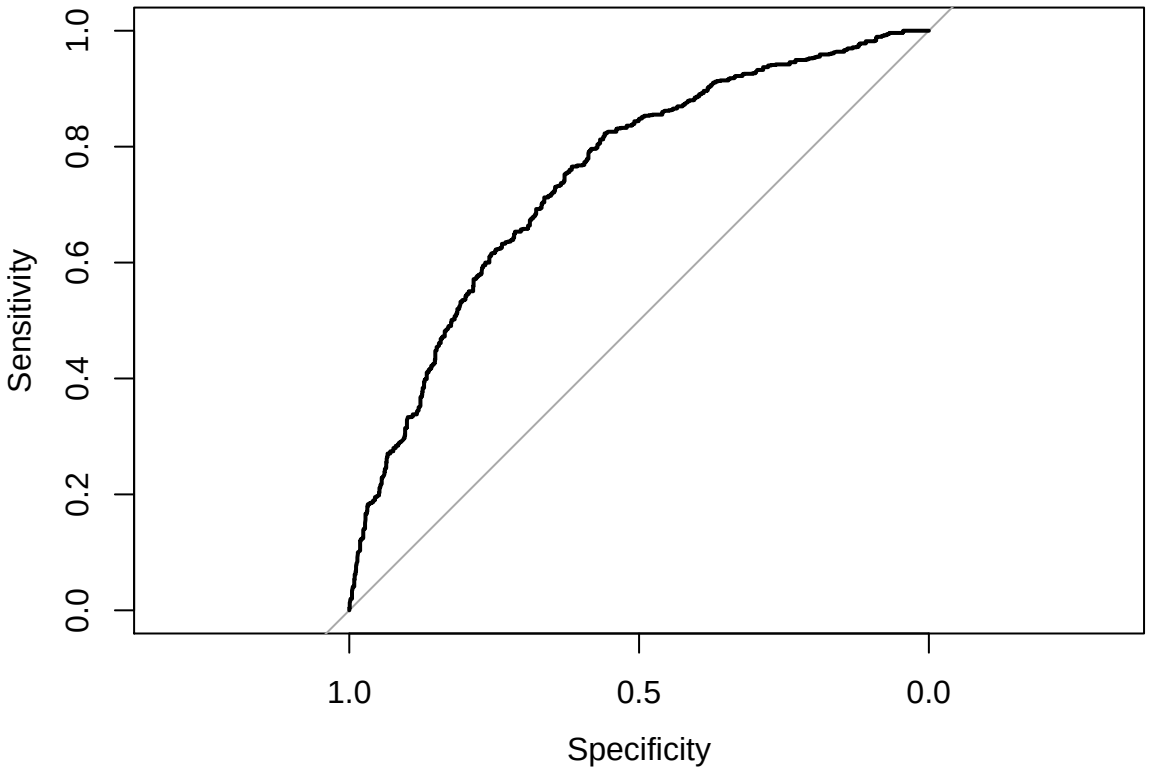
```
## [1] 2337
```

Find generalized R-squared:

```
## ##generalized R-squared:  0.1747409
```

ROC and AUC

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




```
## [1] "AUC of our selected model"
```

```
## Area under the curve: 0.7459
```

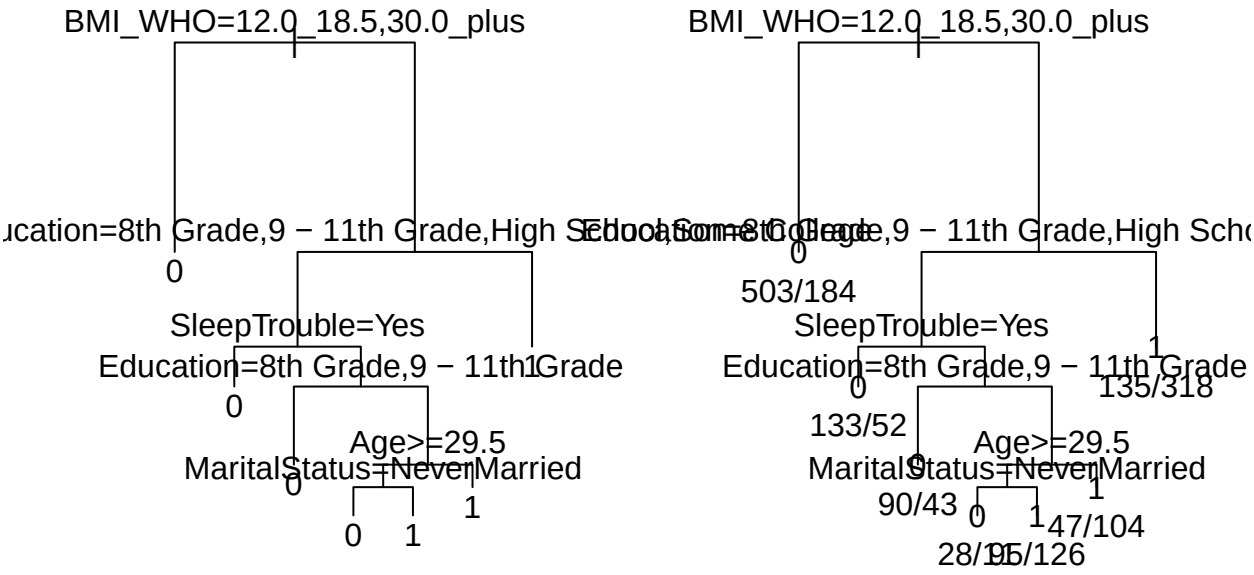
Cross-validated ROC and AUC

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

```
## [1] 0.7374213
```

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.




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

```
##
## Classification tree:
## rpart(formula = formula, data = dat, subset = train)
##
## Variables actually used in tree construction:
## [1] Age           BMI_WHO       Education     MaritalStatus SleepTrouble
##
## Root node error: 838/1869 = 0.44837
##
## n= 1869
##
##         CP nsplit rel error  xerror     xstd
## 1 0.150358      0   1.00000 1.00000 0.025657
## 2 0.068019      1   0.84964 0.87589 0.025194
## 3 0.042363      2   0.78162 0.83413 0.024962
## 4 0.010143      4   0.69690 0.71002 0.024032
## 5 0.010000      6   0.67661 0.72554 0.024169
```

6.)

Discuss any limitations in the analysis.

Bonus

Explore random forest on the data above.