MATH 189 Final Project

Winter 2020

Description of the Dataset

The dataset "dat.rda" is a selected subdataset from the dataset named NHANES. You can load the dat.Rda to get the data for your final project.

If you would like to see the description of each variables in the dataset, you could take a look at the description of the original NHANES dataset. The following link shows a detailed dexcription of the NHANES package https://cran.r-project.org/web/packages/NHANES/NHANES.pdf. You can also find the description of variables using the help function in R after you install and require the package.

```
install.packages("NHANES")
require(NHANES)
? NHANES
```

The dataset "dat.rda" contains 17 variables and 2783 observations selected from the original NHANES dataset. The observations are from the year 2011 and with the age between 18 and 59. The 17 variables are:

```
##       [,1]
##  [1,] "Gender"
##  [2,] "Age"
##  [3,] "Race1"
##  [4,] "Education"
##  [5,] "MaritalStatus"
##  [6,] "HHIncome"
##  [7,] "BMI_WHO"
##  [8,] "HealthGen"
##  [9,] "Depressed"
## [10,] "SleepTrouble"
## [11,] "PhysActiveDays"
## [12,] "TVHrsDay"
## [13,] "CompHrsDay"
## [14,] "AlcoholYear"
## [15,] "SmokeNow"
## [16,] "RegularMarij"
## [17,] "SexOrientation"
```

If you are interested in how the data dat.Rda is obtained from the original NHANES dataset, see the following codes for the data generating process. For the convenience of your analysis, the categorical variable HealthGen has been grouped into two categories Excellent/Vgood, versus Good/Fair/Poor, and are denoted by 1 and 0 repectively.

```
require(NHANES)
dat <- NHANES
SurveyYr <- NHANES$SurveyYr
summary(SurveyYr)
# select the observations that is in year 2011
dat <- dat[SurveyYr=="2011_12", ]
```

```r
age <- dat$Age
#summary(age)
# select the obervations with age between 18 and 59
dat <- dat[(age >= 18) & (age <= 59), ]

#colnames(dat)
# select the variables
ind <- c("Gender","Age","Race1","Education","MaritalStatus","HHIncome",
         "BMI_WHO","HealthGen","Depressed","SleepTrouble","PhysActiveDays",
         "TVHrsDay","CompHrsDay","AlcoholYear","SmokeNow","RegularMarij",
         "SexOrientation")
dat <- dat[,ind]

# dichotomize into two categories: Excellent/Vgood vs others
dat$HealthGen = as.factor(as.numeric((dat$HealthGen == "Excellent" | dat$HealthGen == "Vgood" )))
```

Final Project

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.

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.

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.

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.

6) Discuss any limitations in the analysis.

Bonus (2%): Explore random forest on the data above.

The report should be a pdf file generated by R markdown. Be sure to append all your R codes in the back of the report, using comments to mark each section of the codes in terms of what it does.