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.

```
load("dat (1).rda")
#remove columms with more than 310 missing
dat<-dat[,-which(colSums(is.na(dat)) > 310)]
#screening with univariate regression
screen<-function(var) {
  no_miss<-which(!is.na(var)& !is.na(dat$HealthGen))
  m<-glm(dat$HealthGen[no_miss]~var[no_miss], family = binomial())
  #true if p value < 0.05
  return(coef(summary(m))[2,4]<0.05)
}
var_indices<-which(colnames(dat) !="HealthGen")
#screen each column of dat excluding HealthGen
for (i in var_indices) {
  if(!screen(dat[,i])) {
    var_indices<-var_indices[which(var_indices!=i)]
  }
}
print("our final screened variables:")
```

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

```
colnames(dat)[var_indices]
```

```
## [1] "Age"           "Education"     "MaritalStatus" "BMI_WHO"
```

```
## [5] "SleepTrouble"  "CompHrsDay"
```

2.)

Include “Table 1”

```
#do later.
```

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

Explore random forest on the data above.