ST635 Final Project

Jean Wang

December 7, 2017

Executive Summary

In this project, I used two methods to analyze the data. One is logistic regression and the other one is Bagging. For the first method, I used the P -value as a way to measure the importance of variables and I found out the P-value of tobacco, ldl, famhistpresent, typea and age are significant, and the coefficients are all positive. I reached the conclusion that the increase in cumulative tobacco usage, low density lipoprotein cholesterol, type-A behavior and age at one set, and the present of family history of heart disease will increase the possibility of response, coronary heart disease positive. The overall percentage of correct prediction is 74.2% and accordingly the classification error is only 25.8% by using the logistic regression.

As for the Bagging method, I used variable importance measures plot to help me decide that the variables low density lipoprotein cholesterol, age at onset and cumulative tobacco usage are more important in determining response, coronary heart disease positive. Based on the cross validation, the overall percentage of correct prediction is 71.4% and the classification error is 28.6%. Since I used the whole data set to predict the overall performance in the logistic regression, the classification error is a little bit higher in the Bagging method.

Analysis Part

Firstly, I changed the class of "chd" from integer to factor.

```
CHD[,'chd']=factor(CHD[,'chd'])
class(CHD$chd)

## [1] "factor"
```

Method One: Logistic Regression

a. Fit the Model

Firstly, I tried to fit the model with all variables:

$log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 \quad (X_5 = 1\ if\ famhist\ present, 0\ otherwise)$

```
m3=glm(chd~.,data=CHD,family=binomial)
summary(m3)
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -6.1507209  1.3082600  -4.701 2.58e-06 ***
## sbp             0.0065040  0.0057304   1.135 0.256374
## tobacco         0.0793764  0.0266028   2.984 0.002847 **
## ldl             0.1739239  0.0596617   2.915 0.003555 **
## adiposity       0.0185866  0.0292894   0.635 0.525700
## famhistPresent  0.9253704  0.2278940   4.061 4.90e-05 ***
## typea           0.0395950  0.0123202   3.214 0.001310 **
## obesity        -0.0629099  0.0442477  -1.422 0.155095
## alcohol         0.0001217  0.0044832   0.027 0.978350
## age             0.0452253  0.0121298   3.728 0.000193 ***
```

As we can see from the output, the P-values of tobacco, ldl, present famhist, typea and age are all significant, which means there is indeed a relationship between those significant variables and chd. Then I removed non-significant variables and refitted the model.

```
m7=glm(chd~tobacco+ldl+as.factor(famhist)+typea+age,data=CHD,family=bin
omial)
summary(m7)
##
## Coefficients:
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -6.44644    0.92087  -7.000 2.55e-12 ***
## tobacco                     0.08038    0.02588   3.106  0.00190 **
## ldl                         0.16199    0.05497   2.947  0.00321 **
```

```
## as.factor(famhist)Present  0.90818    0.22576   4.023 5.75e-05 ***
## typea                      0.03712    0.01217   3.051  0.00228 **
## age                        0.05046    0.01021   4.944 7.65e-07 ***
```

Based on the results, the P-values of variables are all significant and the coefficients are all positive. We can conclude that the increases in the cumulative tobacco usage, low density lipoprotein cholesterol, type-A behavior and age will increase the probability of CHD positive. The present of family history of heart disease will also increase the probability of CHD positive. The variables tobacco, ldl, famhist, typea and age are more important in determining the prediction.

b. Predictions

Now, we used the fitted model to predict the probability of CHD positive and overall performance. I didn't use the cross validation to obtain the test error since I want to get the lowest classification error.

```
m7.probs=predict(m7,type="response")
m7.pred=rep("0",nrow(CHD))
m7.pred[m7.probs>.5]="1"
table(m7.pred,CHD$chd)

##
## m7.pred   0   1
##       0 256  73
##       1  46  87

mean(m7.pred==CHD$chd)

## [1] 0.7424242

mean(m7.pred!=CHD$chd)

## [1] 0.2575758
```

Based on the results of the confusion matrix, there are predominance of absent of chd prediction. The overall percentage of correct prediction is 74.2%. We can conclude that when the man does not have chd, the model has 77.8% possibility to predict right, and when the man has positive chd, the model has 66.4% possibility to predict right. The overall classification error is 25.8%.

Method Two: Bagging

a. Fit the Model

For the tree method, I tried classification tree, tree pruning, bagging and RandomForest. I found out that bagging gives me the best prediction. Therefore, I decided to use bagging as my method two and illustrated it in detail.

In order to evaluate the performance of a classification tree, I used cross validation to estimate its test error.

Firstly, we randomly divided the data set into two parts (training data and test data) of roughly the same size.

```
set.seed(1)
train=sample(1:nrow(CHD), as.integer(nrow(CHD)/2))
Chd.test=CHD[-train,]
```

Then we fitted the bagging model to the training data set.

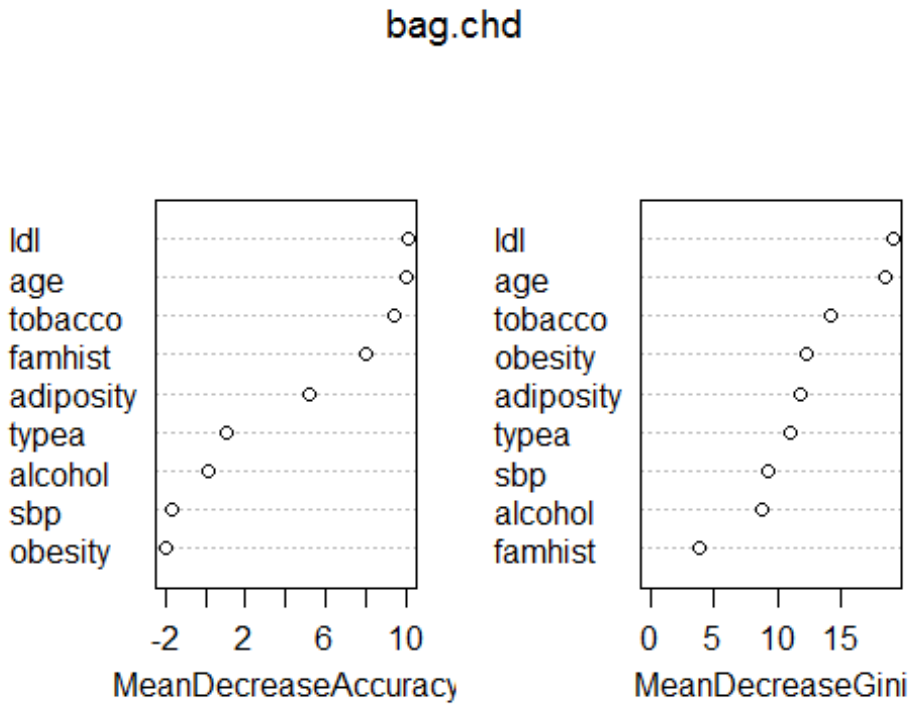
```
set.seed(1)
bag.chd=randomForest(chd~.,data=CHD,subset=train,mtry=ncol(CHD)-1,ntree
=500,importance=TRUE)
bag.chd

##
## Call:
##  randomForest(formula = chd ~ ., data = CHD, mtry = ncol(CHD) -
 1, ntree = 500, importance = TRUE, subset = train)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 9
##
##         OOB estimate of  error rate: 35.06%
## Confusion matrix:
##     0  1 class.error
## 0 109 33   0.2323944
## 1  48 41   0.5393258
```

The overall classification error from the training data set is 35.06%.

Since bagging results in a single tree, I can easily display the results in diagram as follows:

```
varImpPlot(bag.chd)
```



We can obtain an overall summary of the importance of each predictor using the Gini index for bagging classification trees. As we can see from the varImport plot, we find out that the variables with the largest mean decrease in Gini index are ldl, age, and tobacco.

b. Predictions

Now we computed the test error of prediction on the test data set with the fitted bagging classification tree model.

```
bag.pred=predict(bag.chd,newdata=Chd.test)
mean(bag.pred==Chd.test$chd)

## [1] 0.7142857

mean(bag.pred!=Chd.test$chd)

## [1] 0.2857143
```

As we can see from the results, the overall percentage of correct prediction is 71.4% the test error with the bagged classification tree is 28.6%, and the important variables in this method are low density lipoprotein cholesterol, age at onset and cumulative tobacco usage.