Heart Data Analysis

Albert Wiryawan (avw2@illinois.edu)

12/9/2020

Table of Contents

Abstract

Heart disease is an illness that effects many people in the US and internationally. However, prevention and reduction treatments can be employed to minimize the damage at which this disease can cause. As such, this data exploration will focus on determining the number of major vessels with more than 50% diameter narrowing. This create 5 categoies in total with v0 indicating none, v1 indicating 1, v2, indicating 2, v3 indicating 3, and v4 indicating 4 major vessels narrowing. Through the manufacturing of a machine learning pipeline three different learning methods of decision tree, k nearest neighbor, and random forest were able to be tested. With a five fold cross validation of the data set accuracy metrics were able to be obtained which was 58.77 for random forest, 54.95 for decision tree, and 54.17 for k nearest neighbor. Although this was able to be concluded, these results are not effective enough especially in the medical space. As such a different approach was suggested for the improvement of this that would clump groups v1-v4 together as a single category of having heart disease whereas v0 is not having heart disease. ***

Introduction

According to a census collected by the Center for Disease Control and Prevention in 2015, heart disease has been determined as a leading cause of death deriving to 23.4% of the total deaths in the United States. Luckily, there are some habitual actions that can help prevent and control heart disease. In this data exploration, a tool will be created to screen

for heart disease – essentially classifying a person as having five different blood types: v0 being 0 majors vessels with greater than 50% diameter narrowing (no heart disease) and v1-v4 which is having 1-4 major vessels with greater than 50% diameter narrowing.

Methods

In order to create this model, the statistical learning technique of: k-nearest neighbor (knn), decision tree (tree), and random forest (rf) will be used and validated through 5-fold cross validation on data that was not missing on data entry values. Each model will be trained and evaluated with one of the three learning techniques. The machine learning pipeline will enable us to obtain accuracies and train the model with less code using the 'caret package.

Data

By checking the train heart disease data set, it can be seen that there are features that contain values of 'na' or not available for values in some of their persons observations. As such, a general rule of thumb that will be applied in this analysis includes the elimination of data features that are missing more than 30% of their data. This will be done for the train and test data sets that were created. Still, there are observations that will have NA for certain columns.

In order to create suitable data for modeling, the rows that contain an NA will be stripped from the data set.

```
hd_trn_no_missing = na.omit(hd_trn)

set.seed(42)
est_idx = createDataPartition(hd_trn_no_missing$num, p= 0.8, list =
TRUE)
hd_est = hd_trn_no_missing[est_idx$Resample1, ]
hd_val = hd_trn_no_missing[-est_idx$Resample1, ]
```

Modeling

A five fold cross validation will be used over the data set that is missing no data (having no NAs) to train and test the three different model creation methods. A machine learning pipeline is employed to reduce the amount of code necessary to test different model hyper parameters and acquire prediction accuracy metrics. The code for the creation of this pipeline can be seen in the appendix below.

Results

It was determined that the model that produced the best result was the random forest learning model as it was able to guess the correct category of heart disease based on a persons non-invasive data roughly 58.77% of the time. This is slightly larger than the decision tree and knn model receiving 54.95% and 54.17% accuracy respectively.

Discussion

Although it was found that random forest learning model was the best for producing a model to predict someones heart disease type of the four categories, it does not seem to produce reassuring– especially to a patient. As such, one way to improve the model to screen heart disease would bet to lump together the types of heart disease into one in order to have a binary classifier to predict whether someone has heart disease or not. Since both types of errors can be harmful (type I and type II) in the medical realm, this recommendation can drastically improve prediction as to better give treatment and recommendation to patients.

Appendix

```
##         age         sex          cp    trestbps        chol         fbs
## 0.000000000 0.000000000 0.000000000 0.065040650 0.031165312 0.105691057
##     restecg     thalach       exang     oldpeak       slope          ca
## 0.002710027 0.059620596 0.059620596 0.069105691 0.327913279 0.662601626
##        thal         num    location
## 0.528455285 0.000000000 0.000000000

#create a 5-fold cross-validation
cv_5= trainControl(method = "cv" , number =  5)

hd_tree_tune = expand.grid(
  cp = c(0, 0.0001, 0.001, 0.01, 0.1, 1)
)

hd_knn_tune = expand.grid(
  k = 1:100
)



hd_tree_mod = train(
  form = num ~.,
  data = hd_trn_no_missing,
  method = "rpart",
```

```
  trControl = cv_5,
  tuneLength = 10
)

hd_knn_mod = train(
  form = num ~.,
  data = hd_trn_no_missing,
  method = "knn",
  trControl = cv_5,
  tuneLength = 100
)



hd_rf_mod = train(
  form = num ~.,
  data = hd_trn_no_missing,
  method= "rf",
  trControl = cv_5,
  verbose = FALSE
)

hd_tree_mod #54.95%
hd_knn_mod #54.17%
hd_rf_mod #58.77%
```