Heart disease capstone

David Williams

June 17, 2019

Introduction

According to the CDC, heart disease is the leading cause of death in the United States, accounting for 635,260 deaths and 23.1 % of recorded deaths in 2016 [1]. To understand how to best prevent heart disease from occurring, different factors can be measured. A study dating from 1988 describes 76 attributes, but in experiments published, 14 attributes are examined [2]. The data from these experiments can be found on Kaggle: https://www.kaggle.com/johnsmith88/heart-disease-dataset, with further details in their publication [3]. This study estimated the probability of coronary disease among different groups of patients in Cleveland, OH, Long Beach, CA, Zurich and Basel, Switzerland, and Budapest, Hungary. The goal of the experiments was to find a link between chest pain levels and heart disease occurrence. The researchers had the patients perform a series of exercises, such as bike riding, and measured their 76 attributes. The 14 attributes from the published experiments will be used in this machine learning capstone project to attempt to find a link between chest pain, other factors, and heart disease occurrence.

The 14 attributes available in the dataset are measured for each patient and are as follows: age, sex, chest pain type, resting blood pressure, serum cholestoral in mg/dL, fasting blood sugar above 120 mg/dL, resting electrocardiographic (ECG) results, maximum heart rate achieved during exercise, exercise induced angina, oldpeak = ST depression induced by exercise relative to rest, slope of the peak exercise ST segment, number of major vessels colored by fluoroscopy, thal, and heart disease occurrence. Age is an integer in years. For sex, value 1 indicates male and 0 indicates female. Chest pain type is an integer rating from 0 (no pain) to 3 (severe pain). Resting blood pressure is in mm Hg on admission to the hospital, and serum cholestoral is in mg/dL. The fbs above 120 mg/dL indicates whether a patient is at risk of having diabetes. Resting ECG (restecg) results indicate three separate levels from 0 (normal) to 2 (abnormal). Maximum heart rate achieved (thalach) is measured in mm Hg. The exercise induced angina (exang) indicates if a patient developed chest pains due to constricting blood flow in the heart during exercise. ST depression after exercise (oldpeak) indicates a feature in ECG readings where the ECG level is below baseline level, and the slope of the peak exercise ST segment (slope) indicates how steep this reading is after exercise on a scale from 0 (normal) to 2 (abnormal). One experiment performed was fluoroscopy of the major vessels, where the number of major vessels colored was measured (ca), ranging from 0 to 3. thal indicates the type of heart defect measured, where 1 is normal, 2 is fixed defect, and 3 is a reversible defect. Finally, heart disease occurrence (target) is measured from the patients.

These various attributes and their relationships to each other are plotted and explored. After seeing how attributes are related, an ensemble method is used to combine several machine learning methods, including generalized linear models, k-Nearest Neighbors, naive Bayes, and random forest. Variable importance is explored for the random forest model. Finally, the accuracy of this ensemble model is evaluated.

```
if(!require(tidyverse)) install.packages("tidyverse", repos = "http://cran.us.r-project.org")
```

```
## Loading required package: tidyverse
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.1.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.1
## v tidyr   0.8.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------
```

```
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
if(!require(caret)) install.packages("caret", repos = "http://cran.us.r-project.org")
```

```
## Loading required package: caret
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
if(!require(broom)) install.packages("broom", repos = "http://cran.us.r-project.org")
```

```
## Loading required package: broom
```

```
# Heart Disease dataset:
# https://www.kaggle.com/johnsmith88/heart-disease-dataset
# https://www.kaggle.com/johnsmith88/heart-disease-dataset/downloads/heart-disease-dataset.zip/2
heart <- read.table("./heart.csv", header = TRUE, sep = ",", quote = "\"")
```

Methods/analysis

Data cleaning

The data here is generally considered tidy, but in order to utilize the dplyr package in R for this, several columns of the heart dataset need to be converted to factors. Here, sex, cp, fbs, restecg, thal, exang, slope, ca, and target are the columns to be converted.

```
heart = heart %>% mutate(sex = as.factor(sex),
                  cp = as.factor(cp),
                  fbs = as.factor(fbs),
                  restecg = as.factor(restecg),
                  thal = as.factor(thal),
                  exang = as.factor(exang),
                  slope = as.factor(slope),
                  ca = as.factor(ca),
                  target = as.factor(target))
```

Test and training

Since this data is relatively small, a test and training set are partitioned here in a 20:80 ratio.

```
set.seed(1) # if using R 3.6.0: set.seed(1, sample.kind = "Rounding")
test_index <- createDataPartition(y = heart$target, times = 1, p = 0.2, list = FALSE)
train_set <- heart[-test_index,]
test_set <- heart[test_index,]
```

Data exploration

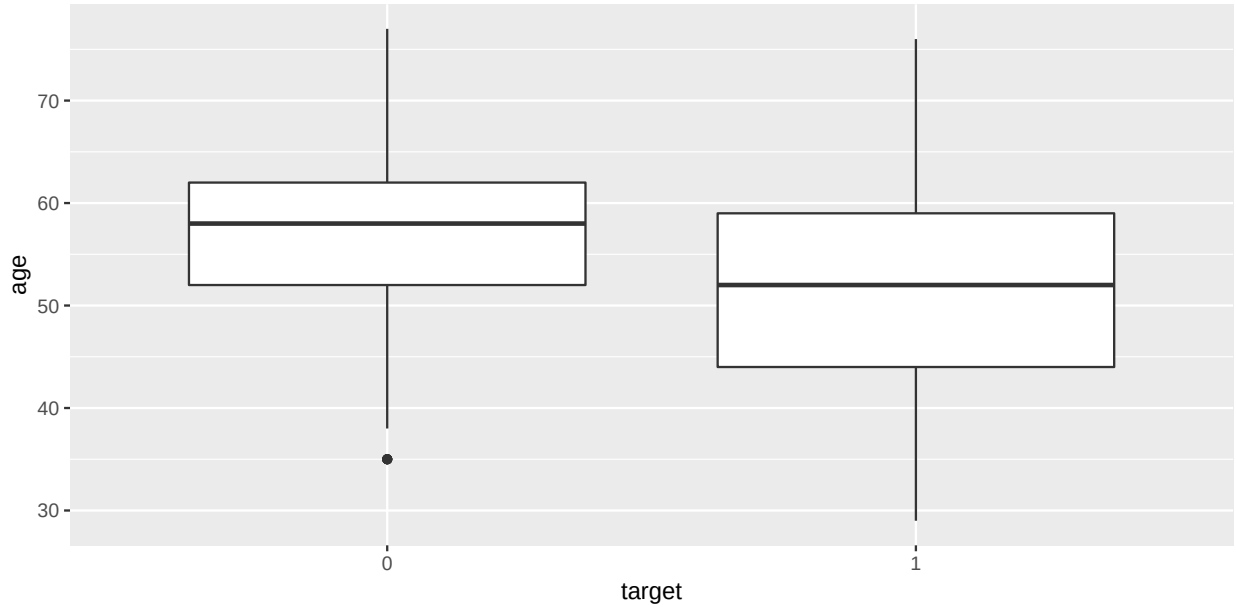
To get a general feel of the dataset, gender counts, ages, heart disease occurrence, chest pains, and heart defects are explored in the plots below.

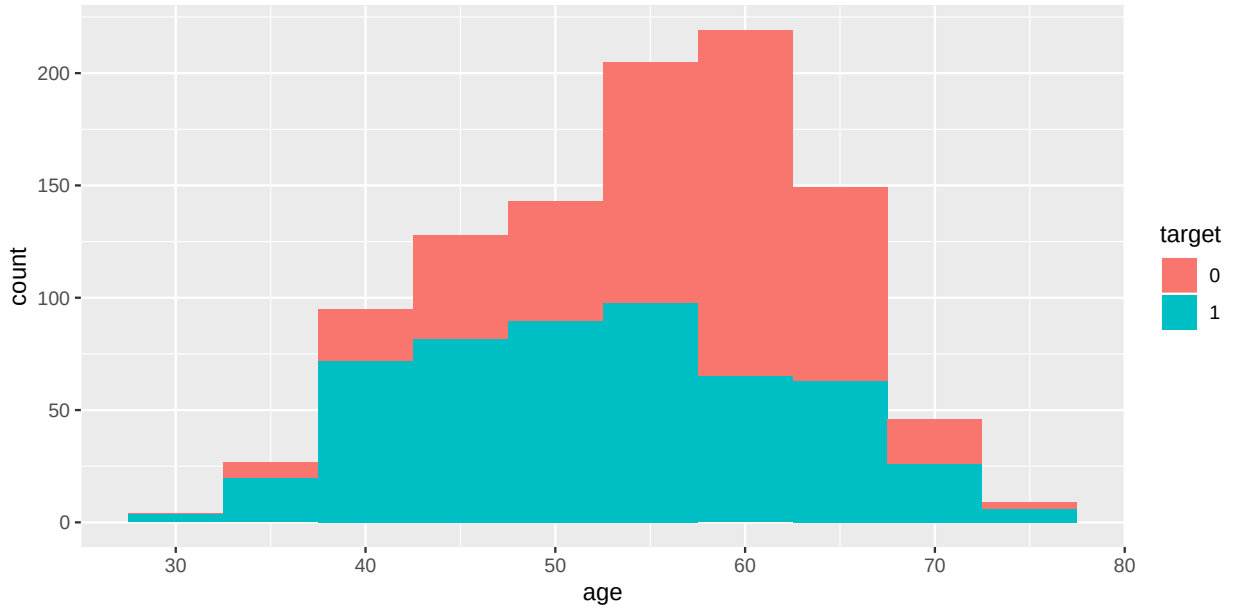
```
heart %>% group_by(sex) %>% summarize(count = n())
```

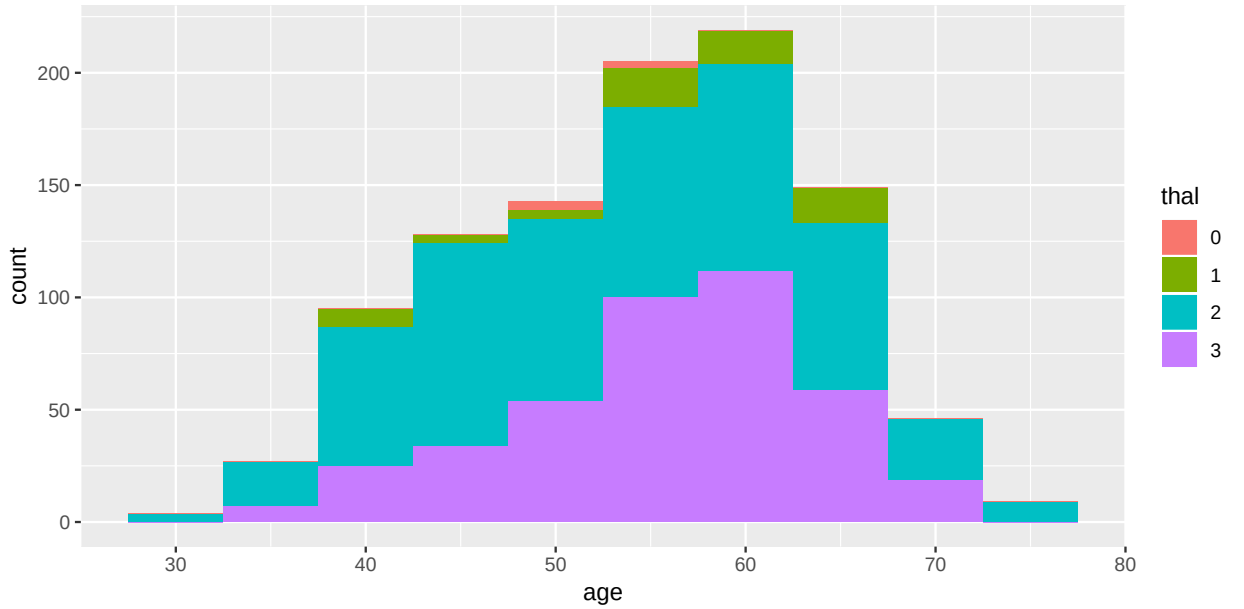
```
## # A tibble: 2 x 2
##   sex   count
##   <fct> <int>
## 1 0       312
## 2 1       713
```

```
summary(heart$age)
```

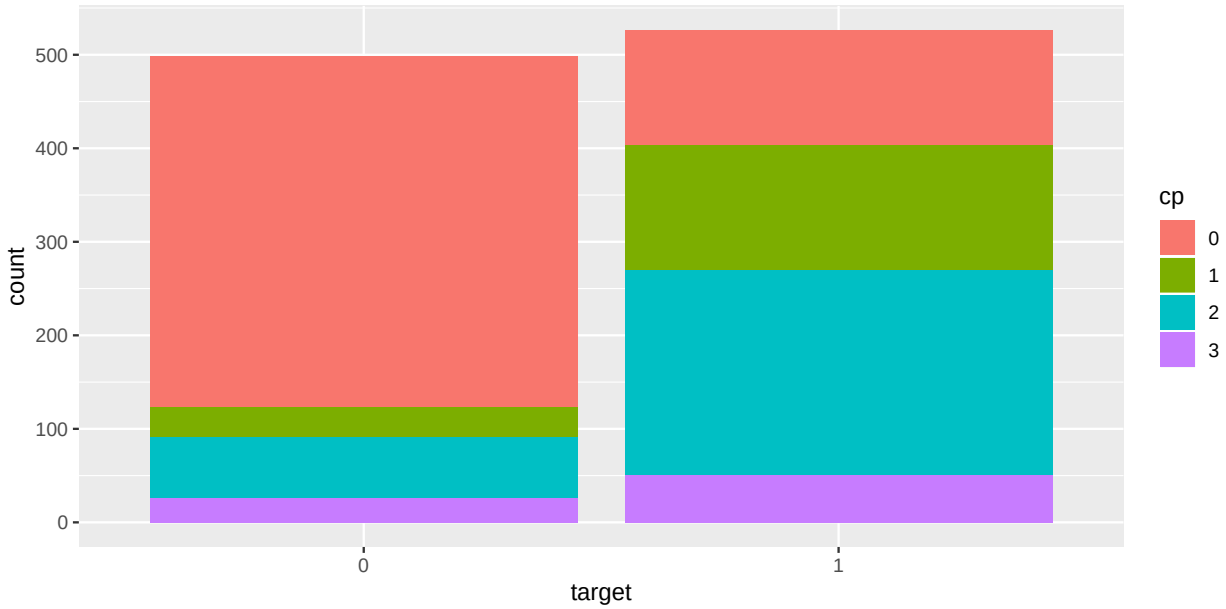
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   29.00   48.00   56.00   54.43   61.00   77.00
```

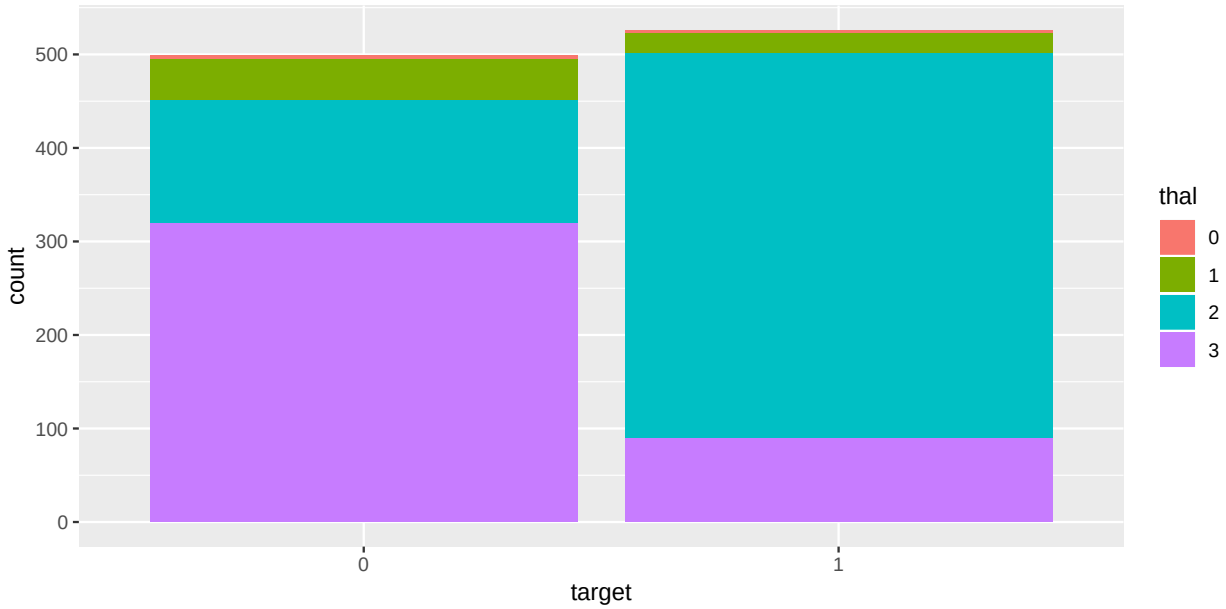


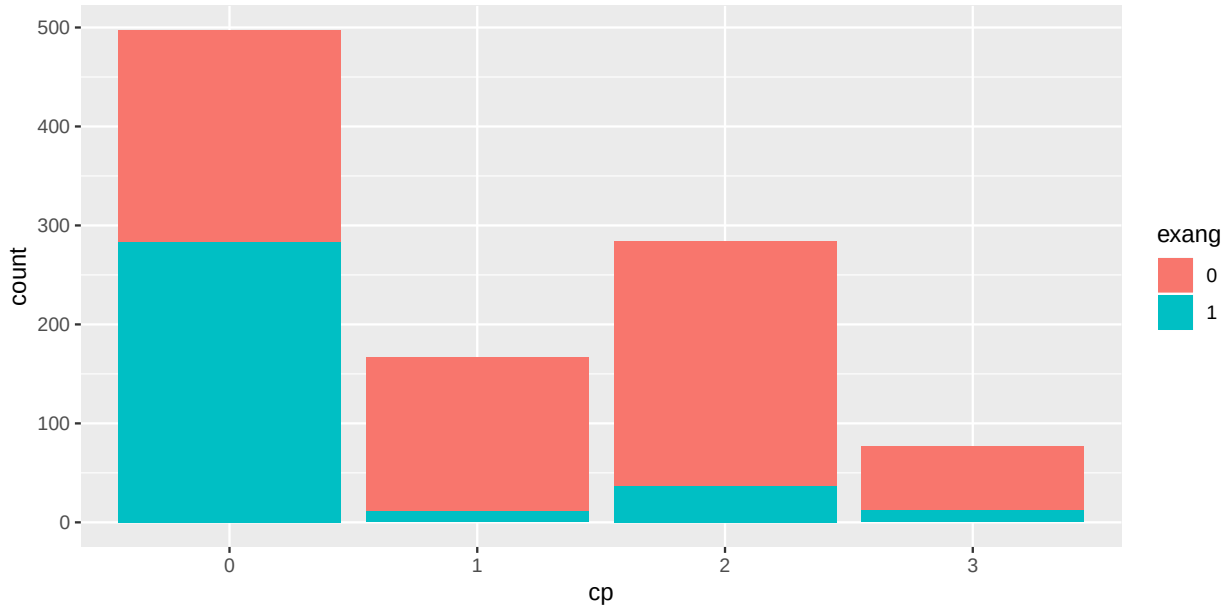


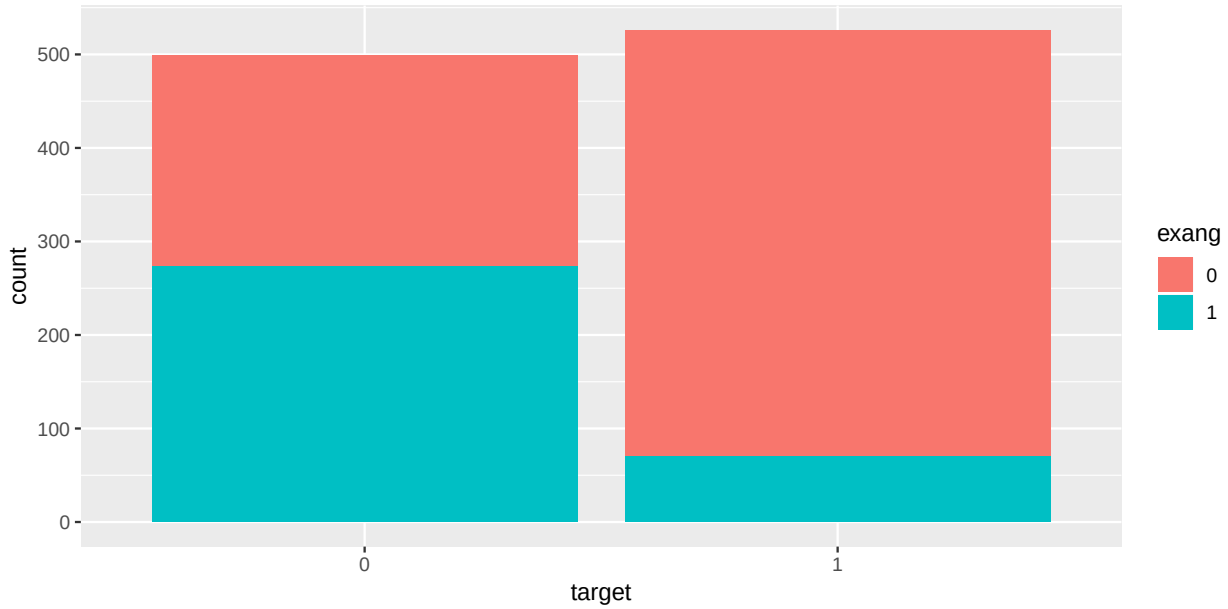


We can see that there are almost twice as many males (1) as there are females (0). Additionally, in the boxplot and histogram above, heart disease appears more prevalent in the younger population than it does older. In fact, the histogram implies that heart disease is more prevalent within age groups ages 40 down and ages 70 and up. This might be a little confusing, but looking at the numbers, however, the mean and median ages are in the mid-50s. This model might not make sense for a general population, especially if looking at a college town where the median and mean ages might be in the 20's. Instead of looking at age alone, other factors such as cholesterol levels, heart defects, and chest pain might make more sense. In the histogram with age and heart defects, we see that at each age, there is a high degree of heart defects within each age group. The dominant type of heart defect appears to be the fixed and reversible heart defects, where at lower ages, there is more fixed defect, while 50's to 60's have more reversible defects.

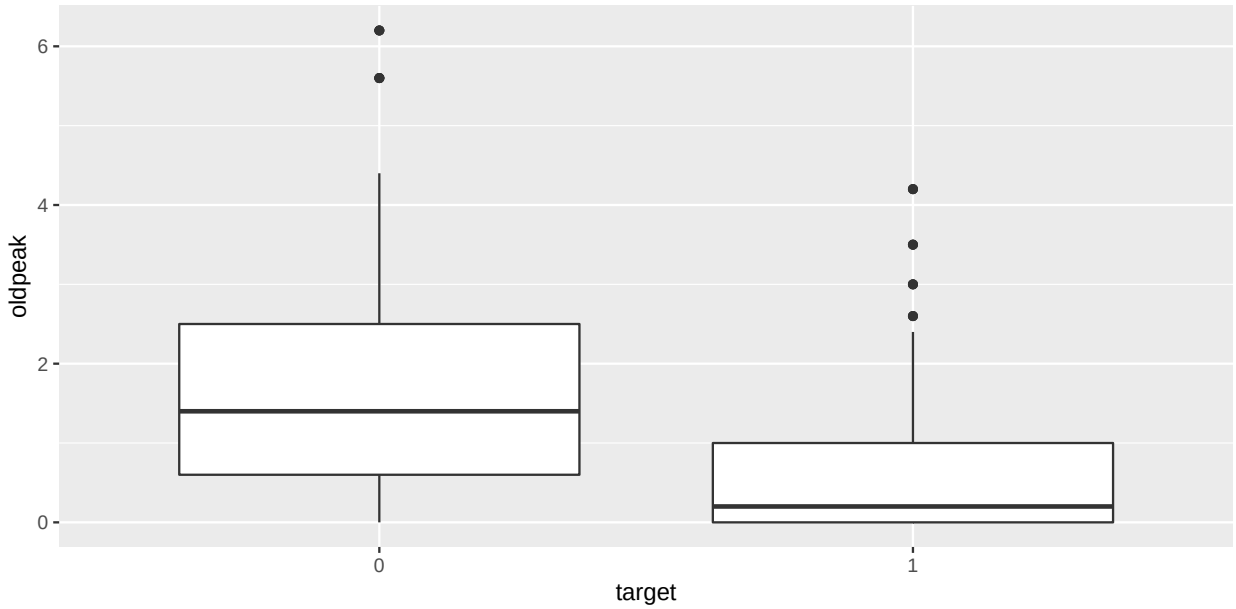

500
400
300
200
100
0
count
0
1
target
cp
0
1
2
3

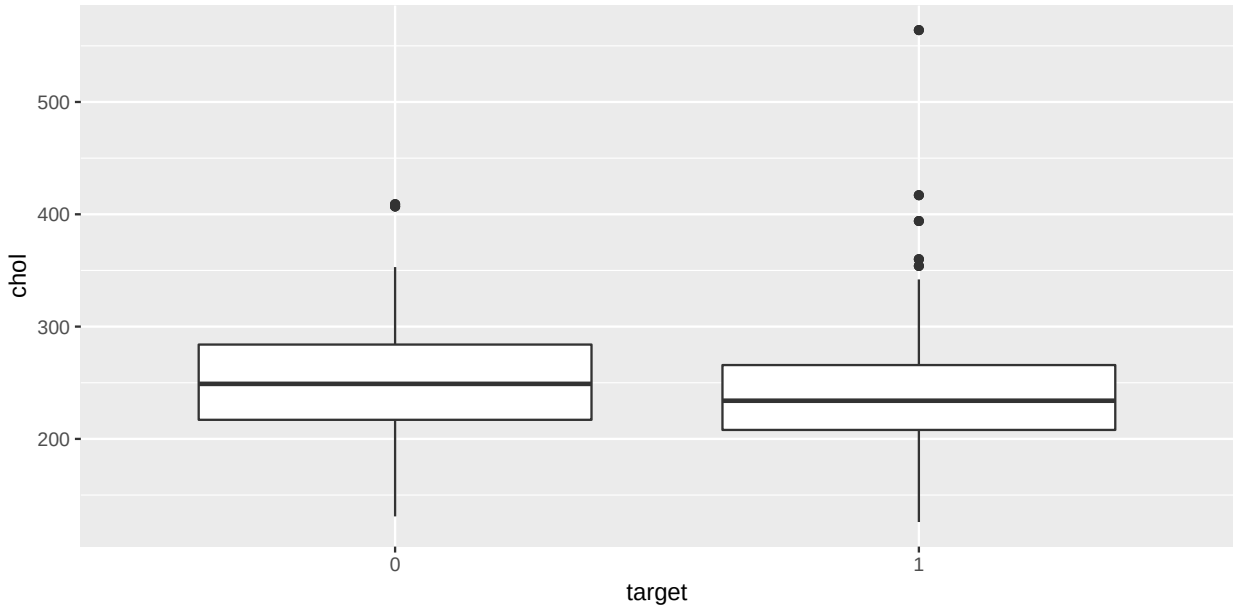

500
400
300
200
100
0
count
0
1
target
thal
0
1
2
3





Looking at the plots for chest pain and heart defects, we see that these have clear correlations with heart disease prevalence, particularly any amount of chest pain (1-3) and a high degree of fixed heart defects (2). When looking at exercise-induced angina, however, there appears to be less association with heart disease prevalence. At no chest pain (0), there is a high degree of angina, but this doesn't make sense. Most lists of heart attack symptoms will include angina [4]. Perhaps there is less correlation with this form of angina.





A clear trend can be seen between heart disease and oldpeak, where a higher oldpeak ST depression induced by rest is more associated with no heart disease. Additionally, there seems to be no correlation between resting cholesterol levels and heart disease in this dataset. Like with age above, it is possible that the entire population of this dataset has generally higher cholesterol levels.

Results

The methods used for the ensemble model are shown below in the code. These are then trained to the training set data using the caret package in R. Results from these fits are converted to a prediction matrix with the test set data. Note that building this ensemble model can take up to an hour to run.

```
models <- c("glm", "lda",  "naive_bayes",  "svmLinear",
            "knn", "kknn", "gam",
```

```
            "rf", "ranger",  "wsrf",
            "avNNet", "mlp", "monmlp",
            "adaboost", "gbm",
            "svmRadial", "svmRadialCost", "svmRadialSigma")

set.seed(1)
fits <- lapply(models, function(model){
  print(model)
  train(target ~ ., method = model, data = train_set)
})
names(fits) <- models

pred <- sapply(fits, function(object)
  predict(object, newdata = test_set))
```

After generating the model, variable importance is quantified by the random forest model and shown here.

```
rf = fits$rf
varImp(rf)
```

```
## rf variable importance
##
##   only 20 most important variables shown (out of 22)
##
##          Overall
## thal2    100.000
## oldpeak   52.951
## age       43.402
## chol      40.929
## thalach   35.718
## trestbps  32.835
## ca1       18.203
## cp3       15.601
## sex1      12.743
## exang1    11.410
## cp2        8.380
## slope2     6.953
## ca3        4.115
## thal3      4.104
## ca2        3.903
## restecg1   3.626
## slope1     3.340
## cp1        2.667
## thal1      2.058
## fbs1       1.632
```

Important factors from the data include fixed heart defects (2), oldpeak ST depression, age, cholesterol, max heart rates reached during exercise, resting heart rates upon admittance, and chest pain. Now the accuracy for each method is shown below.

```
acc <- colMeans(pred == test_set$target)
acc
```

```
##            glm            lda    naive_bayes      svmLinear            knn
##      0.9126214      0.8883495      0.8349515      0.8883495      0.6893204
##           kknn            gam             rf         ranger           wsrf
```

```
##      1.0000000      0.8980583      0.9854369      1.0000000      1.0000000
##          avNNet            mlp          monmlp       adaboost            gbm
##      0.9126214      0.4854369      0.9660194      1.0000000      0.9320388
##      svmRadial  svmRadialCost svmRadialSigma
##      0.9271845      0.9271845      0.9368932
```

```
mean(acc)
```

```
## [1] 0.899137
```

```
ind1 = which(round(acc,6)!=1)
mean(acc[ind1])
```

```
## [1] 0.870319
```

```
new_fits = fits[ind1]
new_pred = pred[,ind1]
new_models = models[ind1]
```

Some of these methods generated an accuracy of 100%, which clearly shows either a method did not work or it overfit to the data. Methods that gave such an accuracy were filtered out, and ensemble accuracy is re-evaluated without them. Ensemble accuracy is evaluated using a majority rules voting with the test set data.

```
votes <- rowMeans(new_pred == 1)
y_hat <- ifelse(votes > 0.5, 1, 0)
mean(y_hat == test_set$target)
```

```
## [1] 0.9417476
```

```
ind <- acc[ind1] > mean(y_hat == test_set$target)
new_models[ind]
```

```
## [1] "rf"     "monmlp"
```

Two methods (random forest and monmlp) gave accuracies above the ensemble accuracy. Below, rather than using ensemble accuracy, an accuracy of 80% is used as a benchmark for individual methods. Now, more than 10 methods are included with the ensemble accuracy and give an accuracy that does not appear to overfit.

```
ind <- (acc_hat >= 0.8)
new_models[ind]
```

```
##  [1] "glm"            "lda"            "naive_bayes"    "svmLinear"
##  [5] "gam"            "rf"             "monmlp"         "gbm"
##  [9] "svmRadial"      "svmRadialCost"  "svmRadialSigma"
```

```
votes <- rowMeans(new_pred[,ind] == 1)
y_hat <- ifelse(votes>=0.5, 1, 0)
mean(y_hat == test_set$target)
```

```
## [1] 0.9271845
```

Conclusion

Using a variety of machine learning methods to make an ensemble method, heart disease was predicted given the 14 attributes of interest with good accuracy. Internal attributes (e.g., ECG baseline, heart defects) were shown to be important in predictions, but other attributes such as chest pain, cholesterol levels, and

exercise-induced angina were not shown to be major predictors. Further analysis would include examining correlations between these attributes to find an attribute where chest pain can be correlated to heart disease.

References cited

1. Heron, M. *Deaths: Leading Causes for 2016.* National vital statistics reports; volume 67, number 6. Hyattsville, MD: National Center for Health Statistics. 2018
2. Janosi, A., Steinbrunn, W., Pfisterer, M., and Detrano, R. (1988). *Heart Disease Dataset.* Retrieved from https://www.kaggle.com/johnsmith88/heart-disease-dataset
3. Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J., Sandhu, S., Guppy, K., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. American Journal of Cardiology, 64, 304-310.
4. Mayo Clinic (2018, March 22). *Heart disease.* Retrieved from https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118