Heart Disease Prediction Modeling - Final Report

Sam Seatt

5/15/2019

Executive Summary

This project attempts to predict the risk and existence, respectively, of heart disease using two separate datasets, respectively, from Kaggle. The datasets are described, extracted/downloaded, sanitized for ML modeling, partitioned, explored and visualized. This is done to better understand the data, determine that data can be most effectively used for modeling, and to prepare the data for training and testing the models. Then modelling (and hyper-parameter tuning where applicable) is performed.

This document describes my analysis, presents the findings, and attempts to formulate some preliminary recommendations.

Heart disease is affecting or will affect hundreds of millions of people alive today around the world. According to the US CDC (https://www.cdc.gov/heartdisease/facts.htm): "About 610,000 people die of heart disease in the United States every year – that's 1 in every 4 deaths. Heart disease is the leading cause of death for both men and women." with coronary Heart Disease (CHD) leading the charts. This makes predicting heart disease, and especially CHD, a medically and socially important area of study.

Heart disease is also a complex disease with multiple contributing factors that lead to the buildup of plaque in the arteries; this plaque buildup is highly correlated to the symptoms normally attributed to heart disease of heart failure. The factors that are considered to contribute to the progression of heart disease are quite diverse, and range: from diet and lifestyle, to stress levels and other environmental factors, to genetics and family history. Each of these factors often has a partial effect to the disease progression and onset.

Finally, heart diseases is also predictable and "early action" has very significant preventative benefit. Thus, this makes the problem of heart disease prediction, or heart disease risk prediction, a tractable and useful one to (partly) address through machine learning models.

About this Project

Project GitHub Repository: https://github.com/samseatt/fram-heart

The Readme file (https://raw.githubusercontent.com/samseatt/fram-heart/master/Readme.txt) in the GithHub repository describes how to set up this project (if you are interested beyond the artifact attached with this submission). Basically you can download my RStudio project file in the GitHub repository, or you may simply Run model.R file or Knit fram-heart-report.Rmd file from any RStudio project, provided you have some basic R libraries installed.

The project analyzes and runs models on the two datasets listed in the next section.

Goals and Objectives

The goal of this modeling exercise is to be able to obtain better healthcare outcomes when it comes to predicting heart disease and designing appropriate and suitable treatment plans for patients being screened/tested. Two additional proposed benefits include (a) understanding the predictors and their impact on the progression of this disease; and (b) possibly modeling heart disease more comprehensively with elements of precision medicine and stratified healthcare in mind (e.g. incorporating genomic data and IoT/Fitbit-like health metrics that will get ubiquitously sequenced/collected).

Of course, a more informally obvious reason is to demonstrate and consolidate my learning of data science, especially of machine learning. In this last vein I will demonstrate some ideas and slightly alter the flow of the document to over explain some of my under-the-hood reasoning that will often be omitted in the more formal and more final research write-ups.

The objective of my models is binary classification.

I will test several ML models. This will not only allow me to understand and apply various classification models we studied, but will also provide a better understanding to the relative benefits and deficiencies of each model through these two use cases, and, as well, to determine which model(s) should best be used for further training and prediction in real life.

Using two separate data sets to predict two similar but subtly different outcomes should assist in better understanding the type of prediction tasks we may encounter, the technical (ML, data collection) and non-technical challenges (for example when predicting qualitative outcomes) involved, and to understand the relevance to the data collection campaigns as well as on the timely availability (or not) of the same data at prediction time. Some if this lies beyond the scope of this project, but this is definitely a good starting point to start thinking about all that.

Datasets

As mentioned, two separate datasets were used for this analysis. ML models were run on each of these with appropriate modifications.

The Framingham Study data set (the first dataset analyzed) drives my primary investigation and modeling; it predicts the risk of coronary heart disease (CHD), specifically the chance of the onset of CHD in 10 years), and hence is medically more relevant as preventative treatments could be appropriately devised ahead of time for the subjects evaluated as high-risk by the properly selected and trained ML model.

The UCI (Cleveland Study) Heart dataset, on the other hand, tracks already present heart disease. It thus supports the relatively easier and more structured task of predicting existing heart disease.

This additional dataset is trained in order to gain further comparative insights around data collection, modeling, and results interpretation.

Framingham Heart Study dataset:

```
https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset
https://raw.githubusercontent.com/samseatt/fram-heart/master/data/framingham.csv
```

- Output (Dependent Variable):
  - TenYearCHD: Chance of getting Coronary Heart Disease within 10 years (0: no, 1: yes)
- Inputs (Independent Variable):
  - male: 0 = Female; 1 = Male
  - age: Age at exam time.
  - education: 1 = Some High School; 2 = High School or GED; 3 = Some College or Vocational School; 4 = college
  - currentSmoker: 0 = non-smoker; 1 = smoker
  - cigsPerDay: Number of cigarettes smoked per day (estimated average)
  - BPMeds: 0 = Not on Blood Pressure medications; 1 = Is on Blood Pressure medications
  - prevalentStroke: Prevalent stroke - 0 = No; 1 = Yes
  - prevalentHyp: Prevalent hypertension - 0 = No; 1 = Yes
  - diabetes: 0 = No; 1 = Yes
  - totChol: Total cholesterol (mg/dL)
  - sysBP: Systolic blood pressure (mmHg)
  - diaBP: Diastolic blood pressure (mmHg)
  - BMI: Body Mass Index calculated - Weight (kg) / Height(meter-squared)
  - heartRate: Heart rate - ventricular (Beats/Min)
  - glucose: Blood glucose level (mg/dL)

UCI (Cleveland) Heart Disease dataset:

`https://www.kaggle.com/ronitf/heart-disease-uci`
`https://raw.githubusercontent.com/samseatt/fram-heart/master/data/uci.csv`

- Output (Dependent Variable):
  - target: Presence of (existing) heart disease in the patient
- Inputs (Independent Variable):
  - age | Age in years
  - sex | (1 = male; 0 = female)
  - cp: Chest pain type
  - trestbps: Resting blood pressure (in mm Hg on admission to the hospital)
  - chol: Serum cholesterol in mg/dl
  - fbs: Fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
  - restecg: Resting electrocardiographic results
  - thalach: Maximum heart rate achieved
  - exang: Exercise induced angina (1 = yes; 0 = no)
  - oldpeak: ST depression induced by exercise relative to rest
  - slope: The slope of the peak exercise ST segment
  - ca: Number of major vessels (0-3) colored by fluoroscopy
  - thal: Thallium heart scan / stress test: 3 = normal; 6 = fixed defect; 7 = reversible defect

Key Steps Performed

This subsection highlights the key steps performed during this analysis. The output of these steps is furnished in the section that follows.

- Framingham Heart Study dataset (described first above) is loaded and checked for consistency usint tgsks like input inspection, data wrangling and exploratory plotting.
- The data is then cleaned and separated into training (train) and testing (test) data sets.
- The data is analyzed in details to understand relationships and help in proper input selection and model selection.
- Six binary classification different models are run on the curated Framingham Heart Study dataset. Model hyper-parameters are tuned further where necessary. The models are evaluated include:
  - Logistic regression
  - K-Nearest Neighbors (KNN)
  - Quadratic Discriminant Analysis (QDA)
  - Linear Discriminant Analysis (LDA)
  - Decision Tree (CART) - using rpart
  - Random Forest
- The results are then evaluated primarily for specificity and precision, with higher emphasis on specificity
- Final models is selected and further improved.
- Results are repeated on the UCI/Cleveland Heart dataset for two-folds purpose: (a) analyze how prediction confidence is affected when we are predicting an outcome vs. when we are predicting the onset of the outcome (the latter involves other factors that are outside the domain of how strong a model is), and (b) gain some initial insight into what makes good data and a good data collection campaign.

Methods and Analysis

Process and techniques used are described in following sub-sections in more or less the sequence in which they are applied.

Reading/loading and understanding the data

The data is first loaded from and external CSV file, into an R dataframe object:

$$”spec_{t}bl_{d}f””tbl_{d}f””tbl””data.frame”$$

Used head() and str() to check the data columns and their format. This information is useful in determining what data transformation will be required (during the data cleaning step) for proper ML modeling.

```
## # A tibble: 6 x 16
##    male   age education currentSmoker cigsPerDay BPMeds prevalentStroke
##   <dbl> <dbl>     <dbl>         <dbl>      <dbl>  <dbl>           <dbl>
## 1     1    39         4             0          0      0               0
## 2     0    46         2             0          0      0               0
## 3     1    48         1             1         20      0               0
## 4     0    61         3             1         30      0               0
## 5     0    46         3             1         23      0               0
## 6     0    43         2             0          0      0               0
## # ... with 9 more variables: prevalentHyp <dbl>, diabetes <dbl>,
## #   totChol <dbl>, sysBP <dbl>, diaBP <dbl>, BMI <dbl>, heartRate <dbl>,
## #   glucose <dbl>, TenYearCHD <dbl>
```

```
## Classes 'spec_tbl_df', 'tbl_df', 'tbl' and 'data.frame': 4240 obs. of  16 variables:
##  $ male           : num  1 0 1 0 0 0 0 0 1 1 ...
##  $ age            : num  39 46 48 61 46 43 63 45 52 43 ...
##  $ education      : num  4 2 1 3 3 2 1 2 1 1 ...
##  $ currentSmoker  : num  0 0 1 1 1 0 0 1 0 1 ...
##  $ cigsPerDay     : num  0 0 20 30 23 0 0 20 0 30 ...
##  $ BPMeds         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ prevalentStroke: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ prevalentHyp   : num  0 0 0 1 0 1 0 0 1 1 ...
##  $ diabetes       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ totChol        : num  195 250 245 225 285 228 205 313 260 225 ...
##  $ sysBP          : num  106 121 128 150 130 ...
##  $ diaBP          : num  70 81 80 95 84 110 71 71 89 107 ...
##  $ BMI            : num  27 28.7 25.3 28.6 23.1 ...
##  $ heartRate      : num  80 95 75 65 85 77 60 79 76 93 ...
##  $ glucose        : num  77 76 70 103 85 99 85 78 79 88 ...
##  $ TenYearCHD     : num  0 0 0 1 0 0 1 0 0 0 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   male = col_double(),
##   ..   age = col_double(),
##   ..   education = col_double(),
##   ..   currentSmoker = col_double(),
##   ..   cigsPerDay = col_double(),
##   ..   BPMeds = col_double(),
##   ..   prevalentStroke = col_double(),
##   ..   prevalentHyp = col_double(),
##   ..   diabetes = col_double(),
##   ..   totChol = col_double(),
##   ..   sysBP = col_double(),
##   ..   diaBP = col_double(),
##   ..   BMI = col_double(),
##   ..   heartRate = col_double(),
##   ..   glucose = col_double(),
##   ..   TenYearCHD = col_double()
##   .. )
```

See the summary of the data frame:

```
##       male             age          education     currentSmoker
##  Min.   :0.0000   Min.   :32.00   Min.   :1.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:42.00   1st Qu.:1.000   1st Qu.:0.0000
##  Median :0.0000   Median :49.00   Median :2.000   Median :0.0000
##  Mean   :0.4292   Mean   :49.58   Mean   :1.979   Mean   :0.4941
##  3rd Qu.:1.0000   3rd Qu.:56.00   3rd Qu.:3.000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :70.00   Max.   :4.000   Max.   :1.0000
##                                   NA's   :105
##    cigsPerDay         BPMeds        prevalentStroke     prevalentHyp
##  Min.   : 0.000   Min.   :0.00000   Min.   :0.000000   Min.   :0.0000
##  1st Qu.: 0.000   1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.0000
##  Median : 0.000   Median :0.00000   Median :0.000000   Median :0.0000
##  Mean   : 9.006   Mean   :0.02962   Mean   :0.005896   Mean   :0.3106
##  3rd Qu.:20.000   3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:1.0000
##  Max.   :70.000   Max.   :1.00000   Max.   :1.000000   Max.   :1.0000
##  NA's   :29       NA's   :53
##     diabetes          totChol          sysBP           diaBP
##  Min.   :0.00000   Min.   :107.0   Min.   : 83.5   Min.   : 48.0
##  1st Qu.:0.00000   1st Qu.:206.0   1st Qu.:117.0   1st Qu.: 75.0
##  Median :0.00000   Median :234.0   Median :128.0   Median : 82.0
##  Mean   :0.02571   Mean   :236.7   Mean   :132.4   Mean   : 82.9
##  3rd Qu.:0.00000   3rd Qu.:263.0   3rd Qu.:144.0   3rd Qu.: 90.0
##  Max.   :1.00000   Max.   :696.0   Max.   :295.0   Max.   :142.5
##                    NA's   :50
##       BMI          heartRate         glucose         TenYearCHD
##  Min.   :15.54   Min.   : 44.00   Min.   : 40.00   Min.   :0.0000
##  1st Qu.:23.07   1st Qu.: 68.00   1st Qu.: 71.00   1st Qu.:0.0000
##  Median :25.40   Median : 75.00   Median : 78.00   Median :0.0000
##  Mean   :25.80   Mean   : 75.88   Mean   : 81.96   Mean   :0.1519
##  3rd Qu.:28.04   3rd Qu.: 83.00   3rd Qu.: 87.00   3rd Qu.:0.0000
##  Max.   :56.80   Max.   :143.00   Max.   :394.00   Max.   :1.0000
##  NA's   :19      NA's   :1        NA's   :388
```

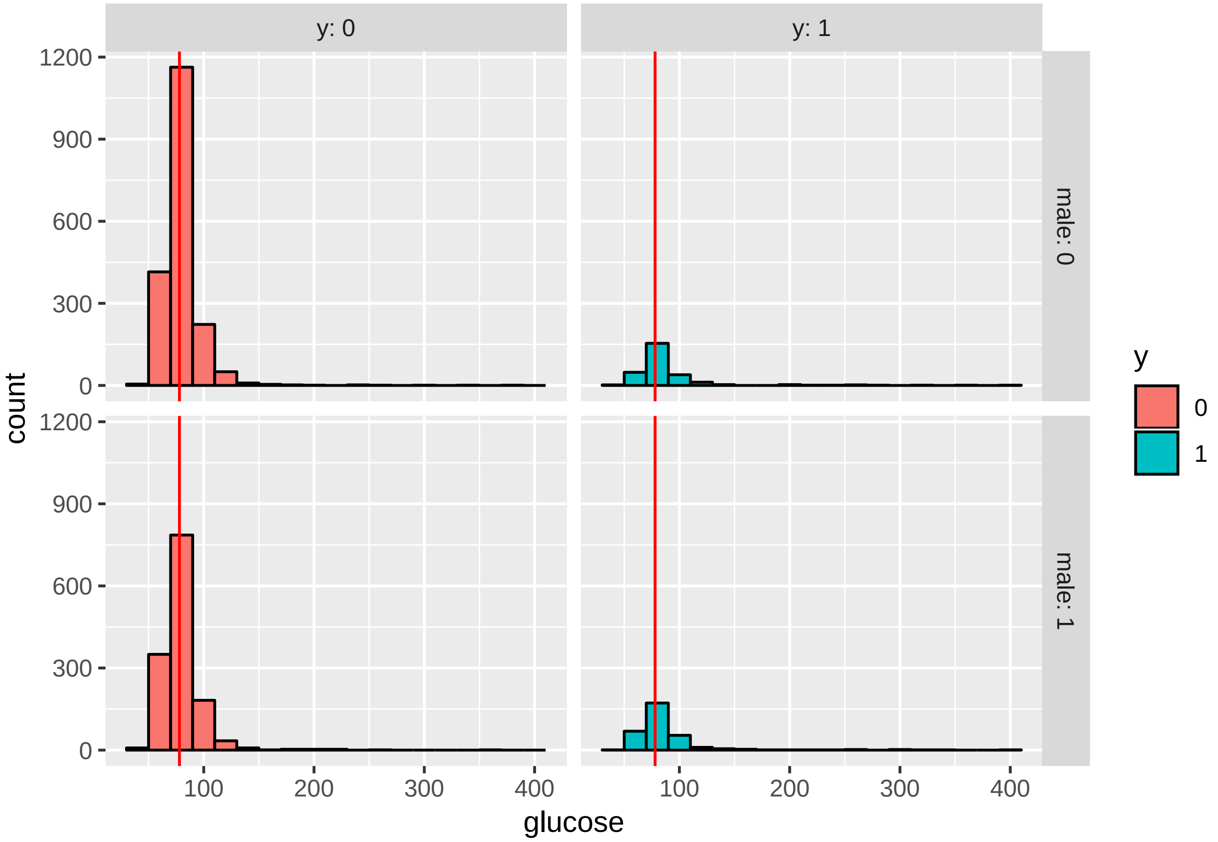
This gives us the ranges and means of each input parameter (and the output). The ranges appear to not vary too much from each other (all values are within a couple of hundred), so normalization is not necessary. I will thus leave the inputs un-normalized so their graphs and stats are more intuitive to follow.

Data cleaning

First, keeping with tradition, I rename the output column to y. I also change this output to be a factor as classification is about predicting between competing outcomes and not predicting numeric values.

In my data wrangling exercise, I do see several NA values (645 in total). I would need to remove these as part of data cleaning.

I also see that 388 of these missing values are for glucose. Removing all these rows will be a significant (but not overwhelming) loss of 9.2% of the data. So, let’s first see how useful glucose level is in predicting heart disease.



```
## # A tibble: 4 x 3
## # Groups:   male [2]
##    male y         n
##   <dbl> <fct> <int>
## 1     0 0      2119
## 2     0 1       301
## 3     1 0      1477
## 4     1 1       343
```

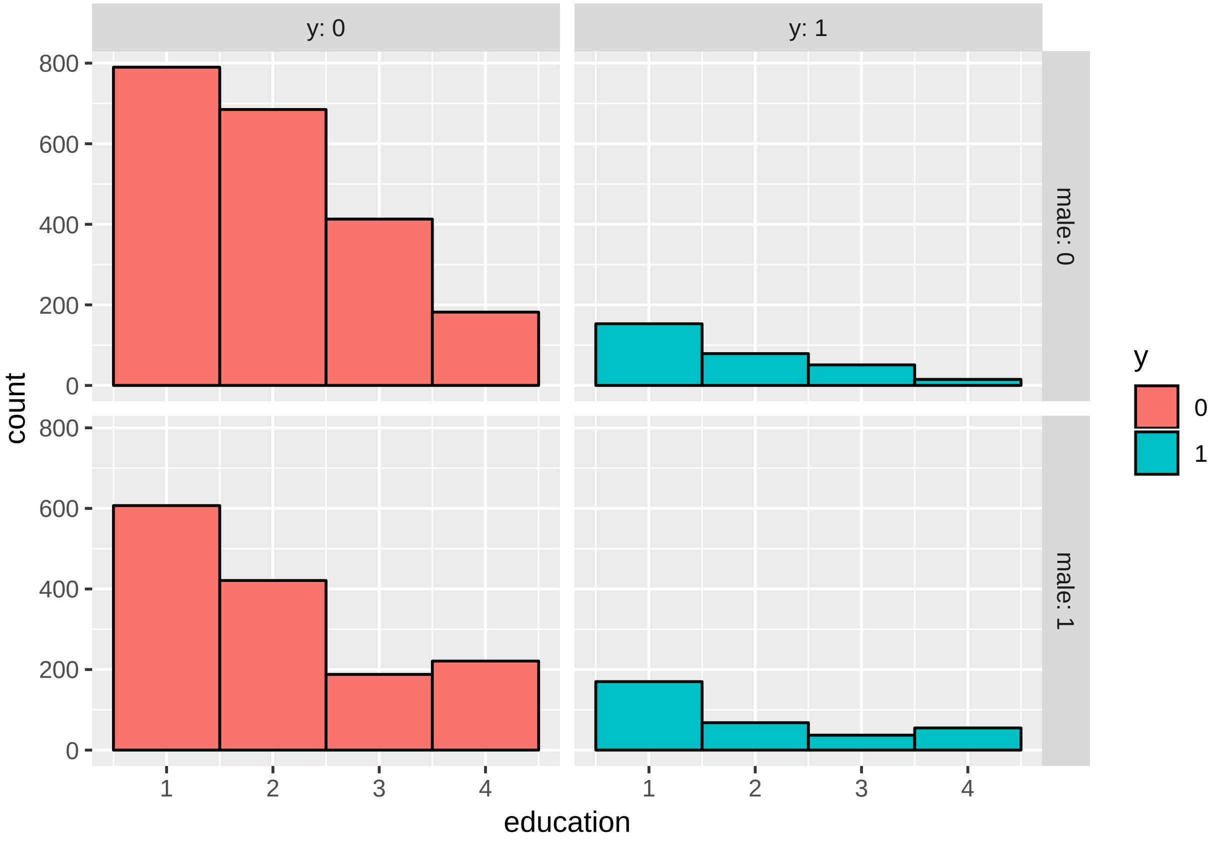
Well, it seems fairly even key for both healthy and diseased outcomes, but healthy subjects slightly favor towards lower glucose levels (bigger red bars on the left side, for both genders). The contribution is very slight, and therefore, I decide to save the rows and simply drop the glucose column.

Another secondary reason for removing this column with high NA is that it will also be likely to missing more often in real life and affect prediction. This would be a good time to ask the subject matter experts to see why the blood glucose information is not readily available in patient data. (One possible reason would be that it is considered more related to diabetes or diabetes mediate heart disease, and may not be always ordered in the blood test. So, it is not likely that much tied to heart disease even from the point of view of the medical professionals)

Remove possible Human Bias from the dataset

I also drop education column as it is medically non-relevant (though it could potentially be indirectly correlated to nutritional habits). It appears that this field has the potential of introducing training and social bias in the ML models (when considering a patient's treatment options during the prediction phase).

Below visualization confirms that education trends more or less the same for both healthy and at-risk subjects.



Now we have relatively few rows with NAs and can afford to throw them away. So, I now get rid of all the rows in which NA values still remain.

```
## [1] 4090
```

Now I have 4090 rows to train and validate. This seems just about right to get a good representative training in reasonable time, allowing to run the models relatively quickly and comparing multiple models.

```
## [1] 611
```

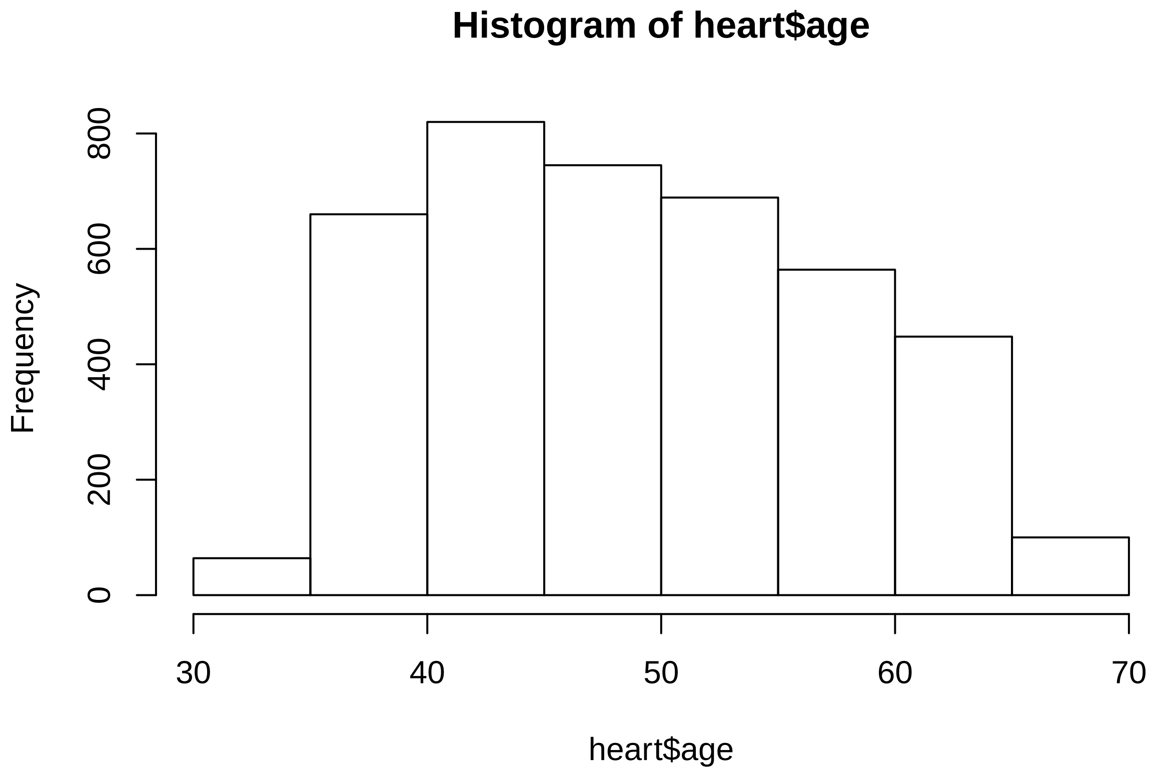
```
## [1] 3479
```

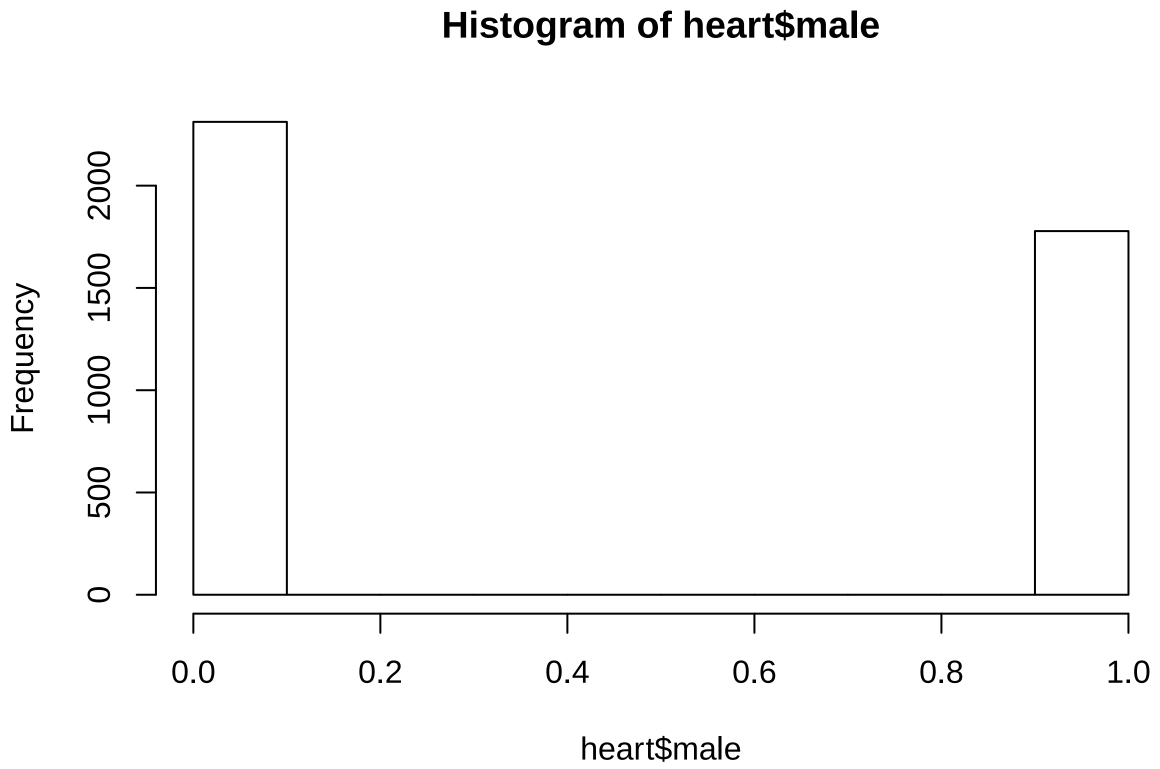
I also have a reasonable fraction of positive prediction i.e. I'm not starved of one outcome in my training data.

So the data looks clean and I can proceed with further data exploration.

Data exploration and visualization, and insights gained

First do quick checks of age and gender demographics, just to see who we are collecting this data from.

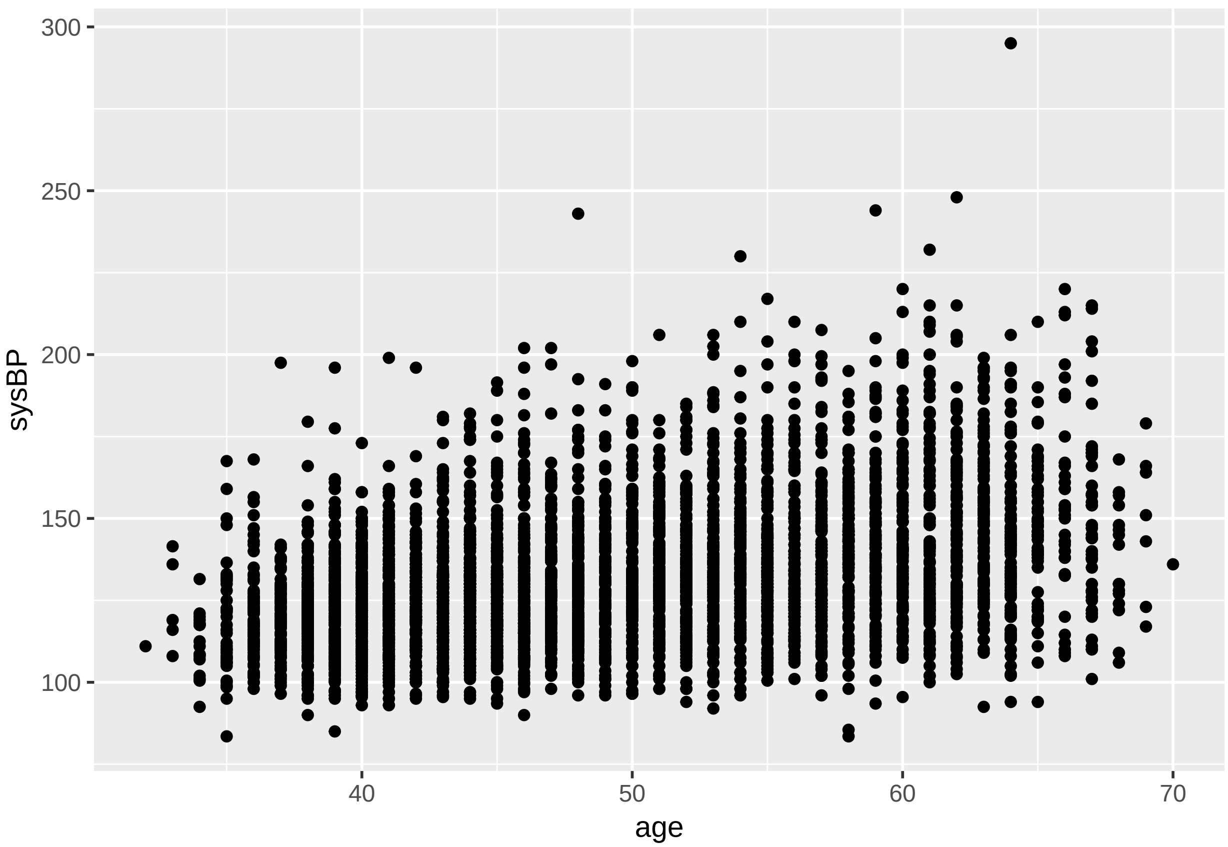

Histogram of heart$age
Frequency
0
200
400
600
800
30
40
50
60
70
heart$age



```
## # A tibble: 4 x 3
## # Groups:   male [2]
##    male y         n
##   <dbl> <fct> <int>
## 1     0 0      2036
## 2     0 1       276
## 3     1 0      1443
## 4     1 1       335
```

Check correlations between inputs (independent variables that ideally should not be correlated in order to get the maximum benifit of each) and between inputs and output (the dependant variable, where correlation should exist).

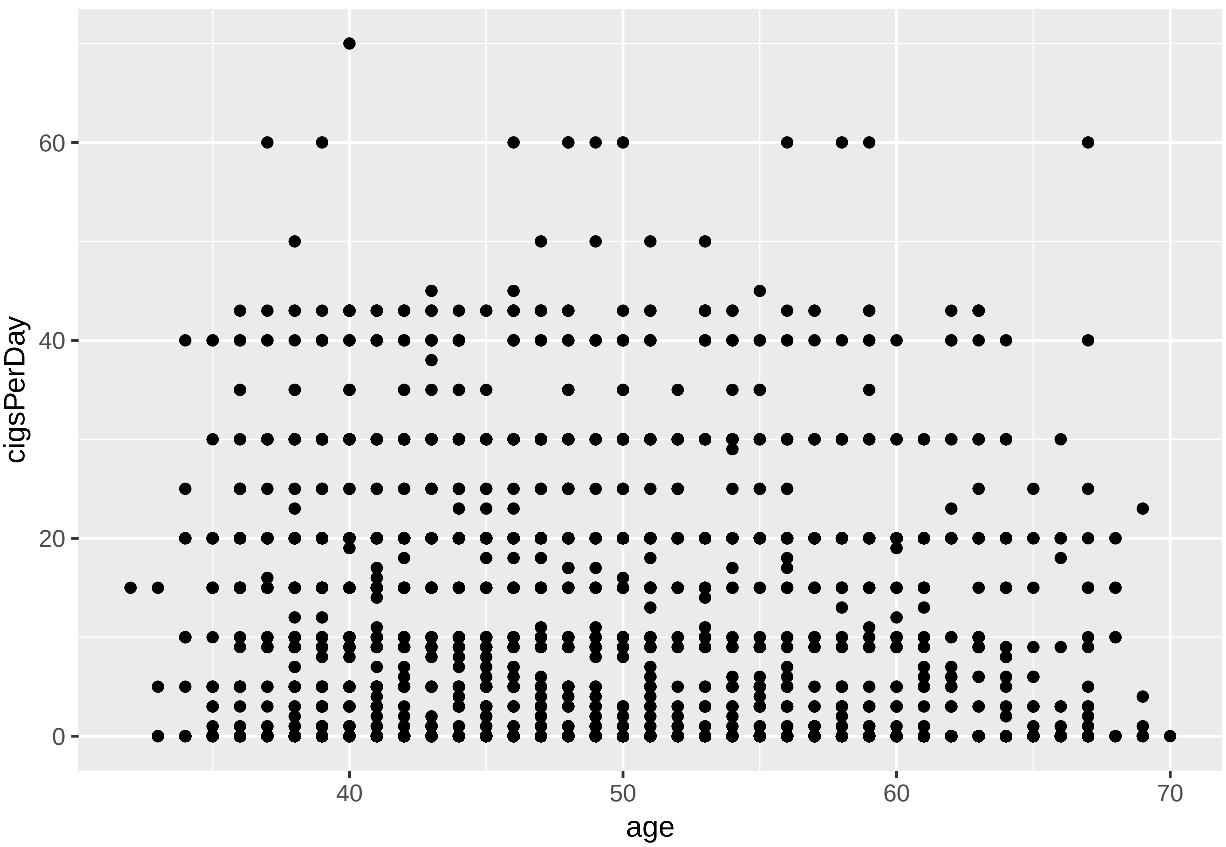
Correlation between age and systolic blood pressure:



```
## [1] 0.3896845
```

As expected, there is some correlation showing blood pressure deteriorates with age. But not entirely, as the correlation is not very strong. So, we can consider BP to be an independent variable along with age for the purpose of heart risk prediction.

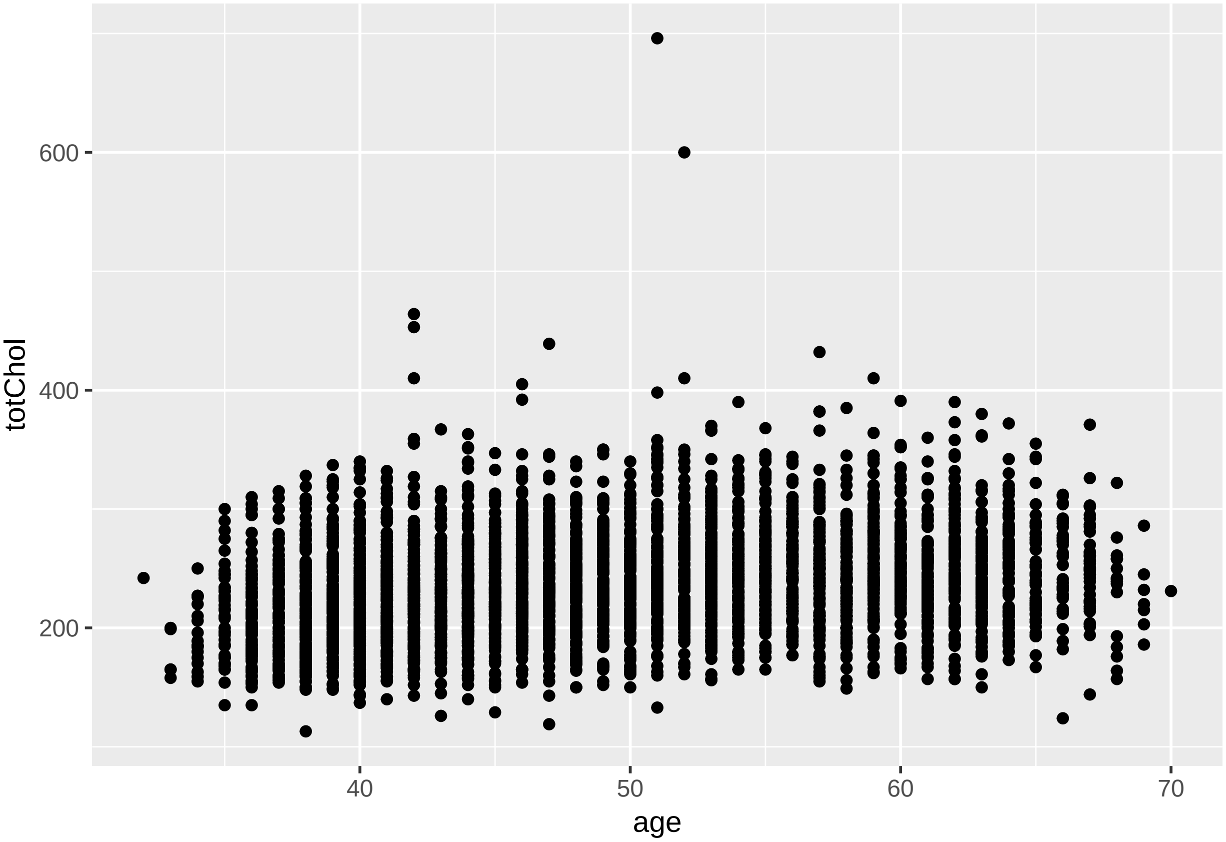
Correlation between age and cigarettes per day:



```
## [1] -0.1905578
```

There is almost no correlation: all ages smoke almost equally. So age is definitely not a confounder for smoking.

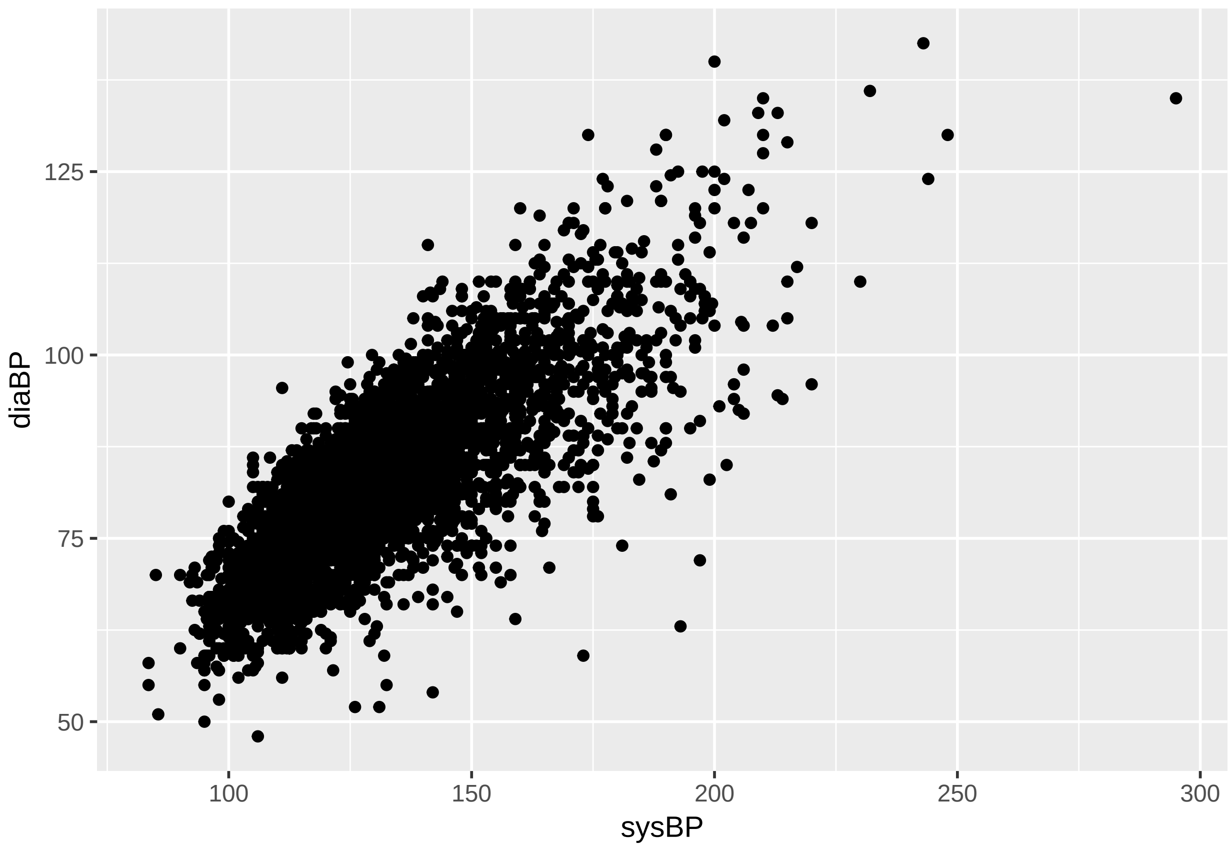
Correlation between age and total cholesterol:



```
## [1] 0.2633436
```

Somehow, we don't have a strong positive correlation between age and total cholesterol. In other words, total cholesterol does not increase drastically with age - this could be because younger people have higher HDL (good cholesterol i.e. we should have been seeking cholesterol ratio instead), or older patients are more likely to be on cholesterol-lowering medicines, or younger people in the study are those who had a reason to get their heart metrics checked, or some other medical reason like the dimensions of the cholesterol molecules. In any case, whether or not total cholesterol (rather than cholesterol ratio) is a good predictor, we can safely include it as a variable independent of age.

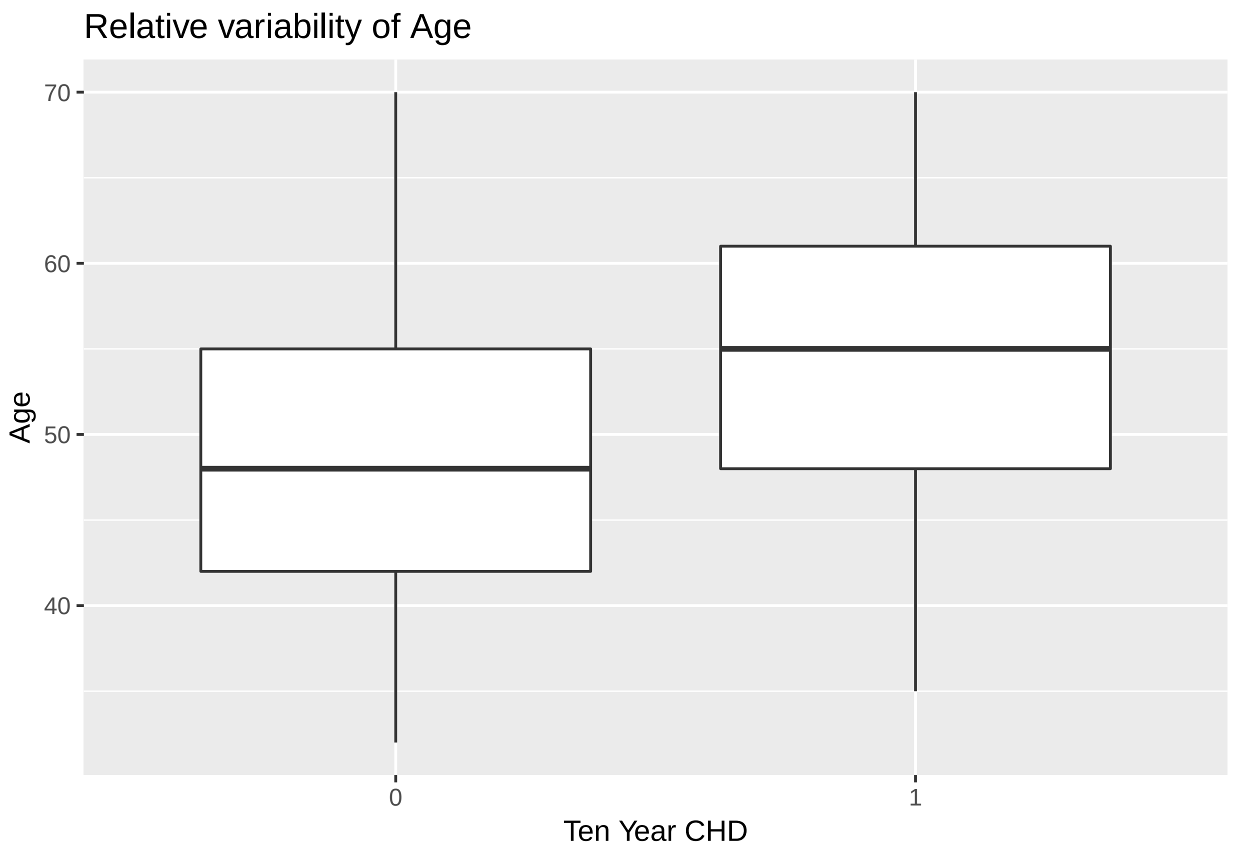
Correlation between systolic BP and diastolic BP

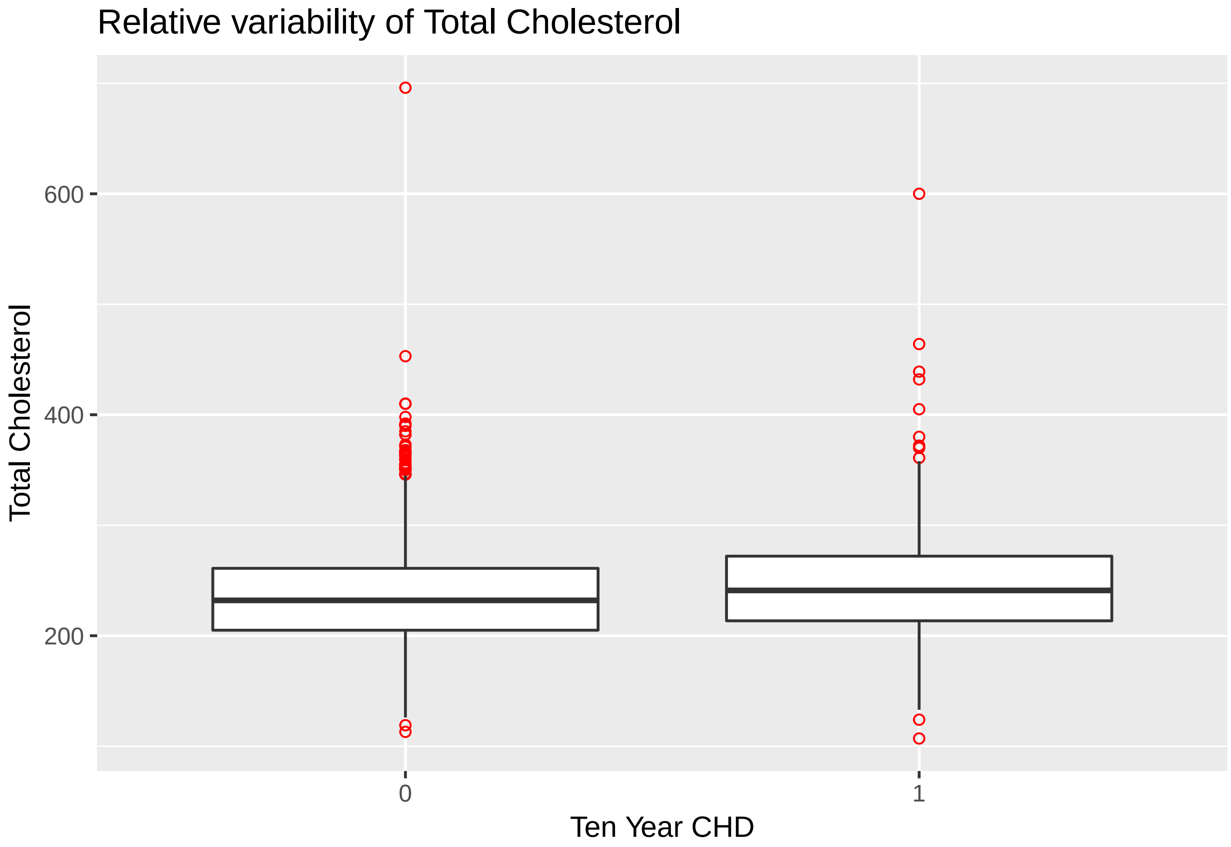


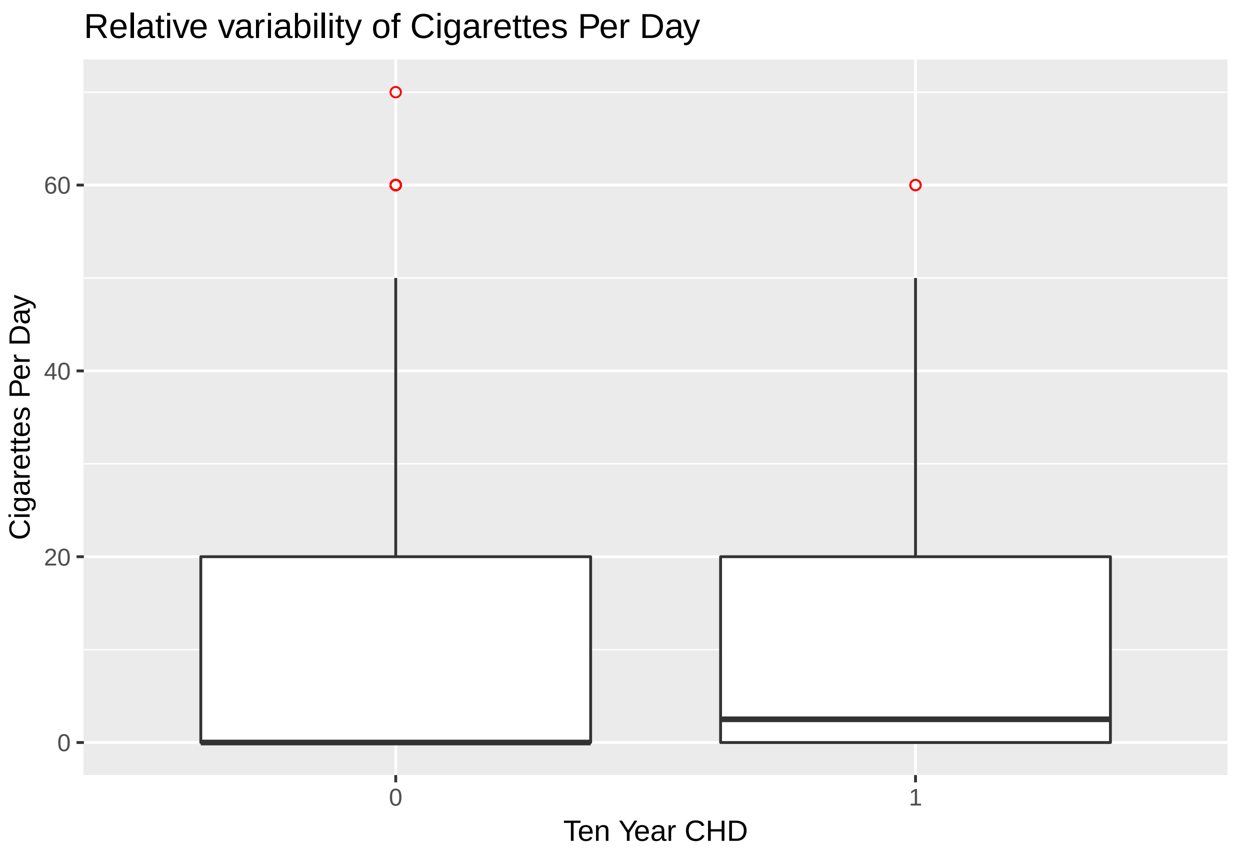
```
## [1] 0.7846247
```

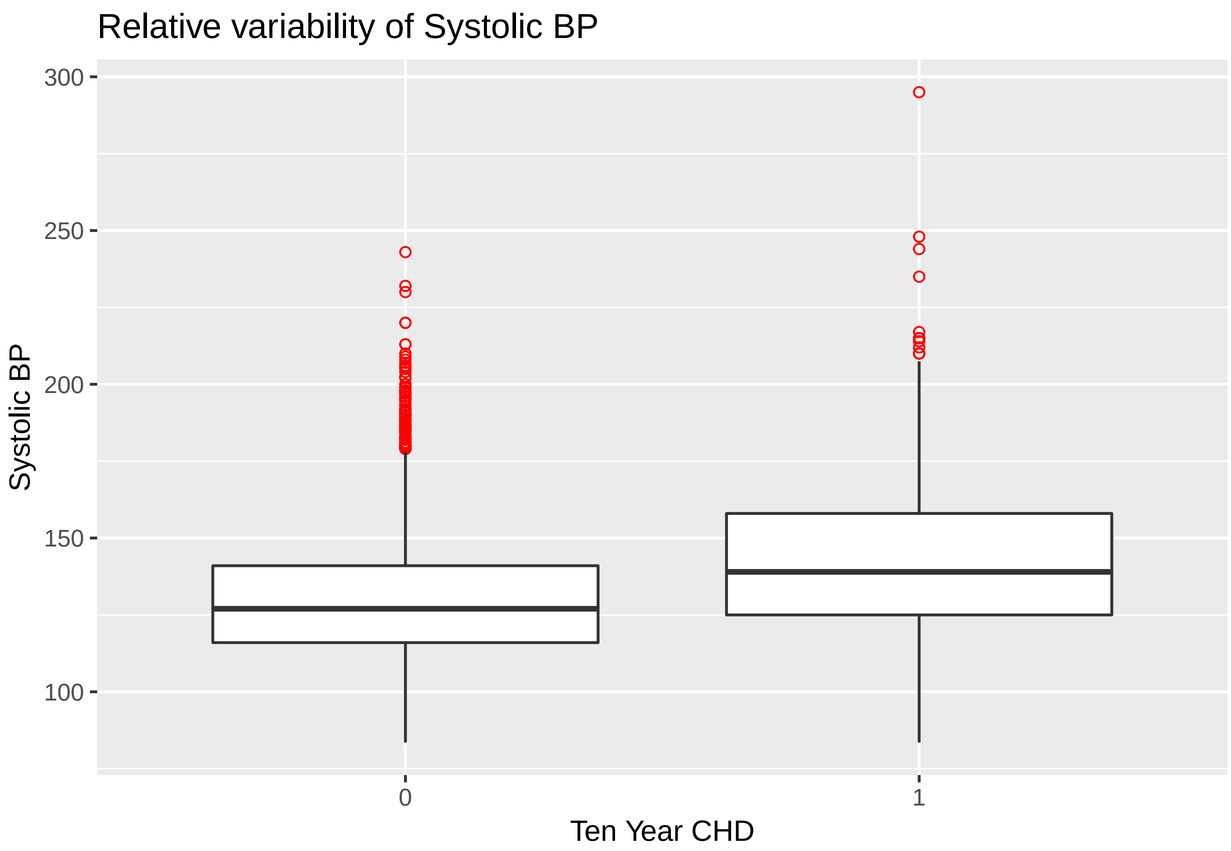
As seen from the above graph and the correlation values, the two blood pressures are correlated, however the correlation is not complete, so I will keep both these inputs for the analysis as the second input can still provide additional training information.

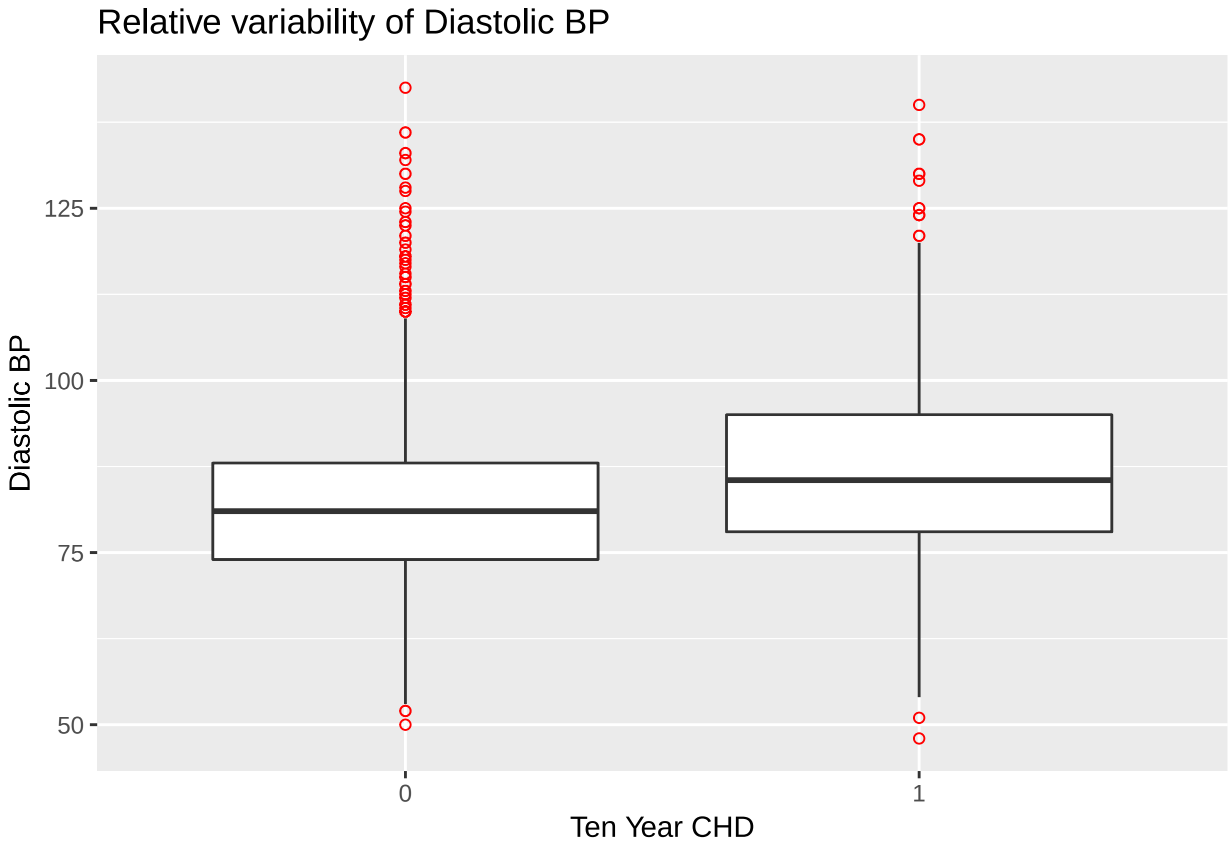
I visualize all the inputs using boxplots to see the relative contribution of each on healthy vs. diseased outcomes (after 10 years)

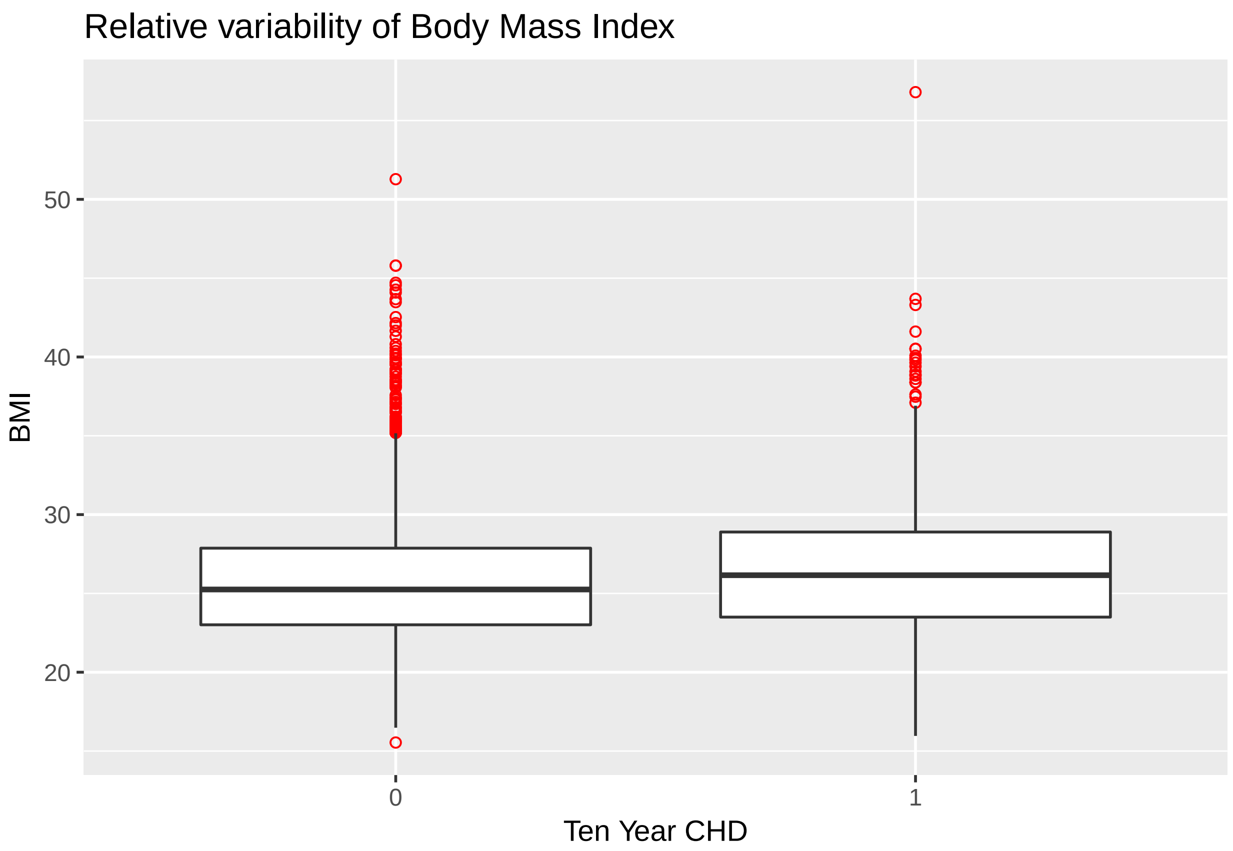

Relative variability of Age
Age
70
60
50
40
0
1
Ten Year CHD

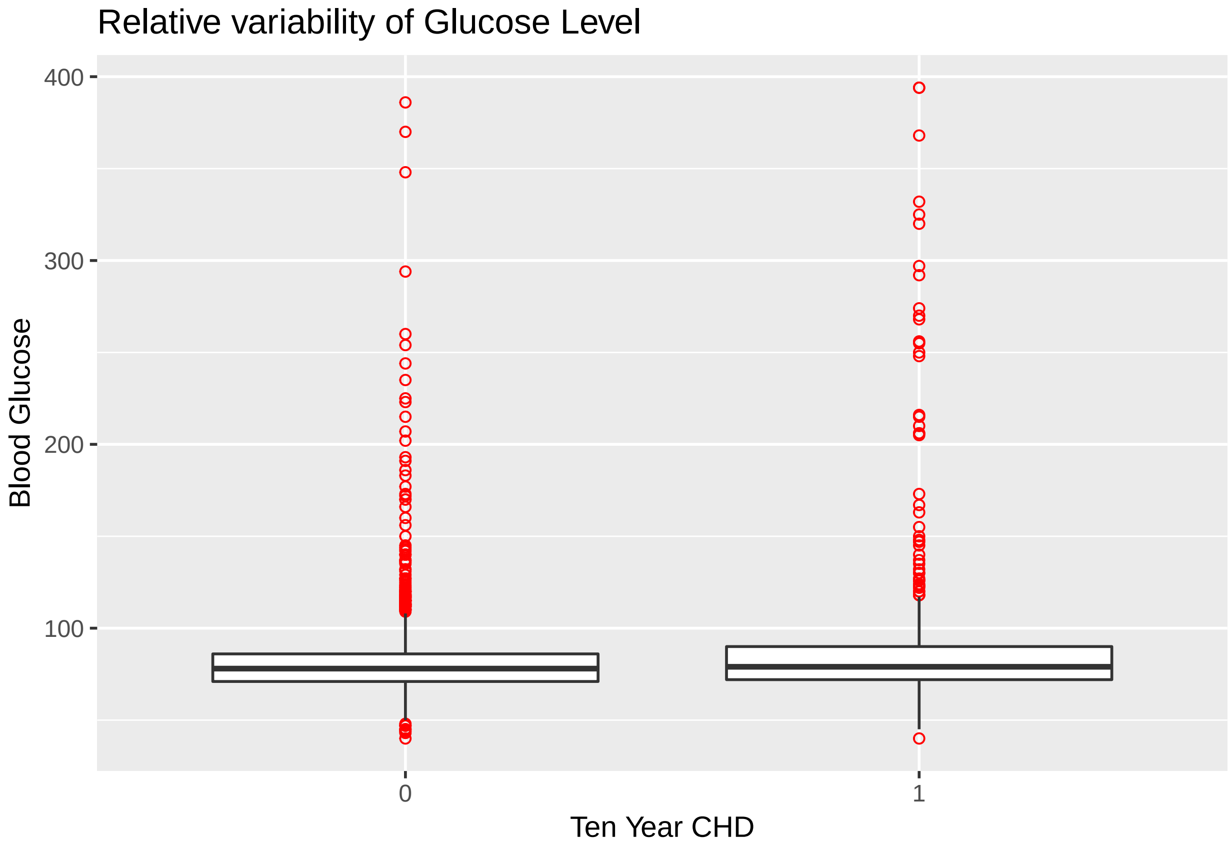

Relative variability of Total Cholesterol
Total Cholesterol
600
400
200
0
1
Ten Year CHD


Relative variability of Cigarettes Per Day
Cigarettes Per Day
0
20
40
60
0
1
Ten Year CHD


Relative variability of Systolic BP
Systolic BP
300
250
200
150
100
0
1
Ten Year CHD


Relative variability of Diastolic BP
Diastolic BP
125
100
75
50
0
1
Ten Year CHD

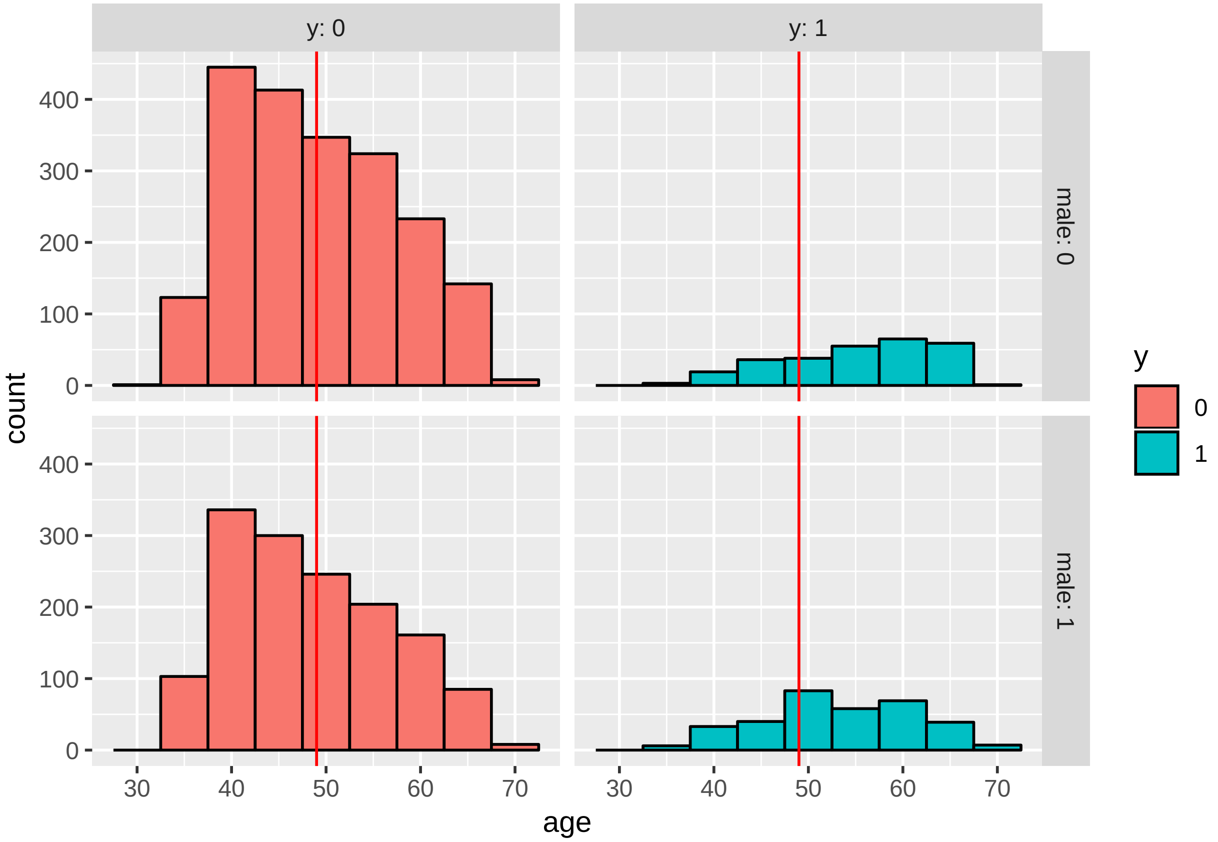

Relative variability of Body Mass Index
BMI
50
40
30
20
0
1
Ten Year CHD

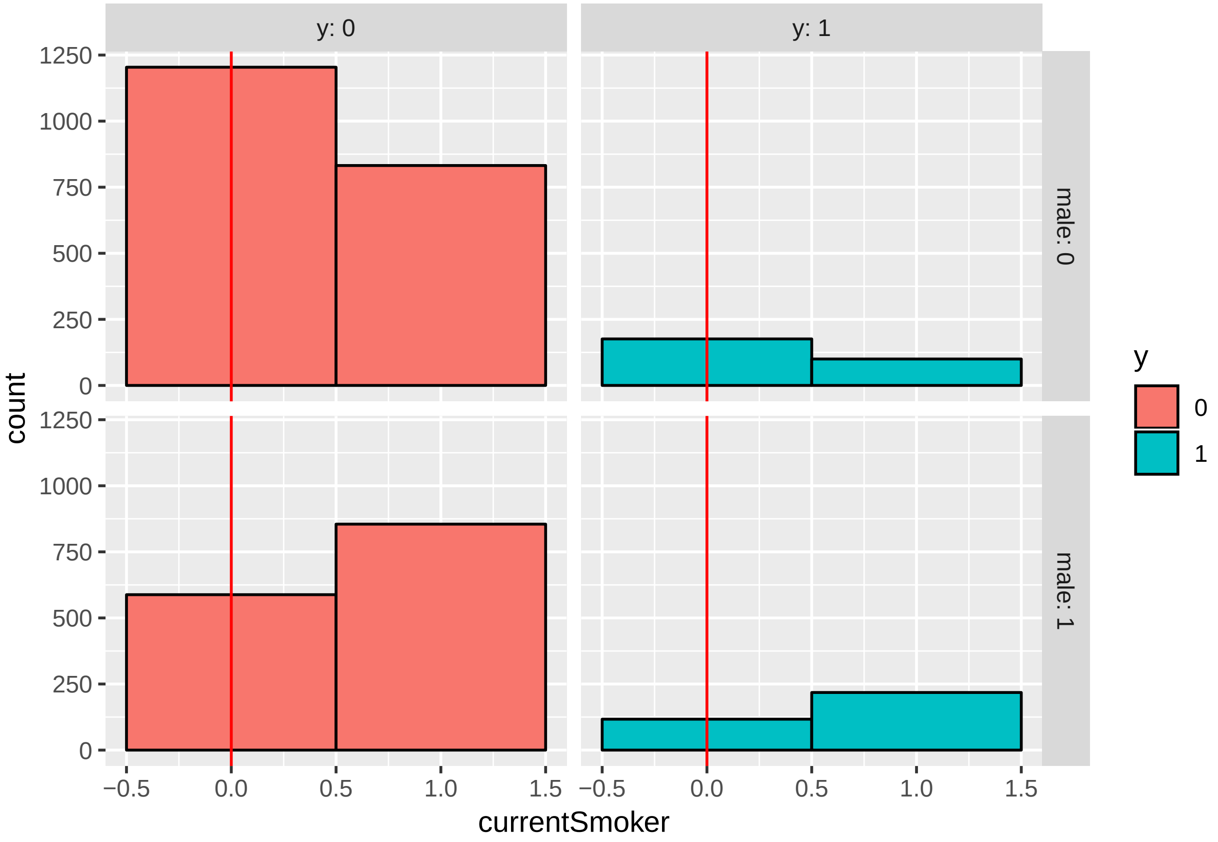


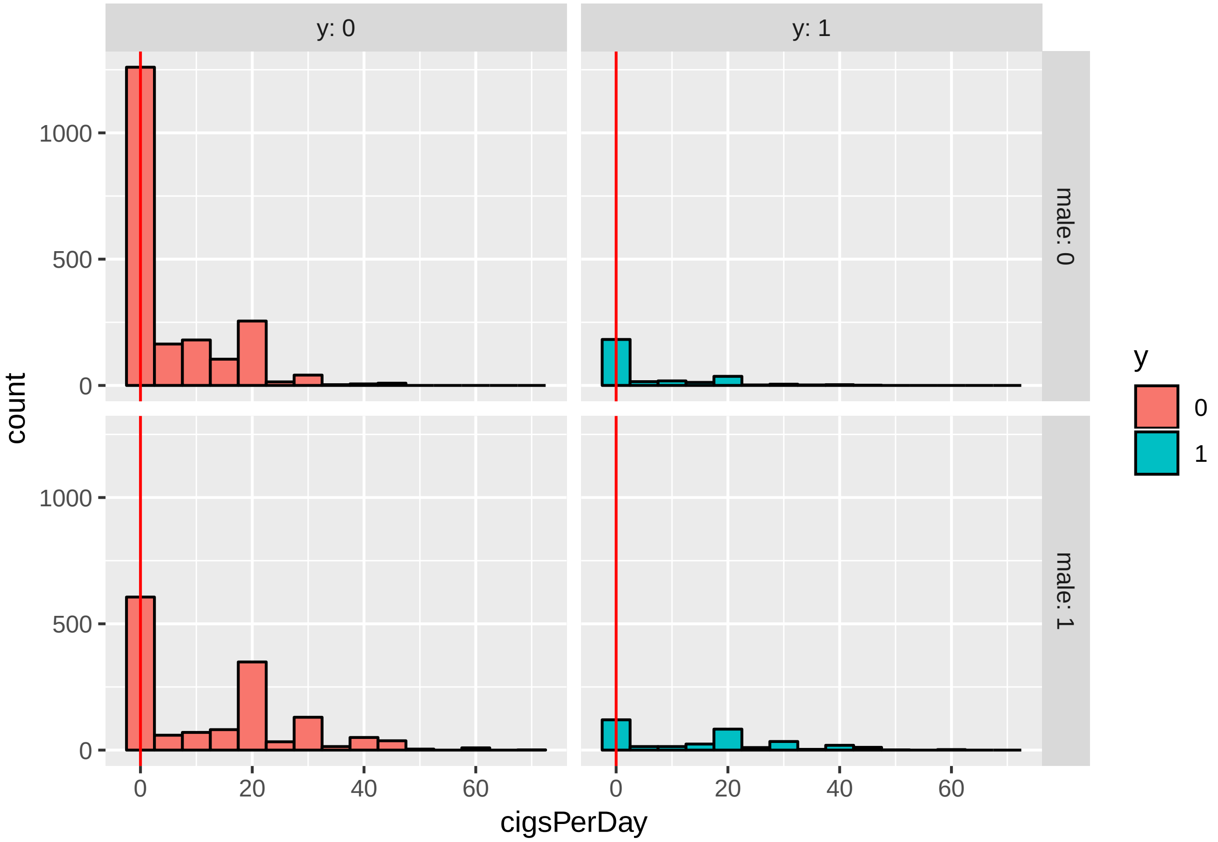
We see, among other things, that the average cigarettes consumption is close to zero cigarettes for healthy vs. higher for those who will end up with CHD.

Also, heart rate does not seem to be a good indicator of the 10-year onset of CHD.

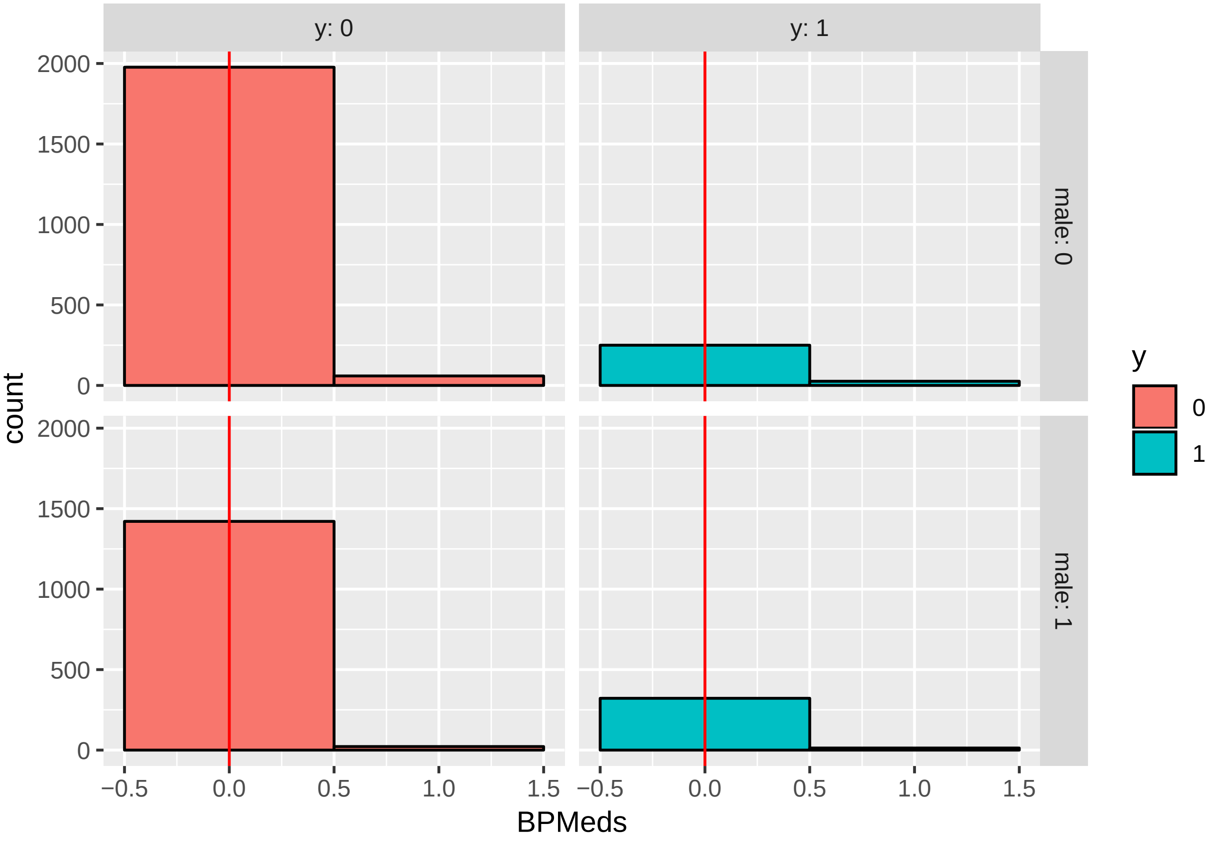
Now, for full visualization, I go ahead and plot histograms for each independent variable, this time segregated into male/female) to see its relative correlation to the 10-year onset of coronary heart disease for each gender.

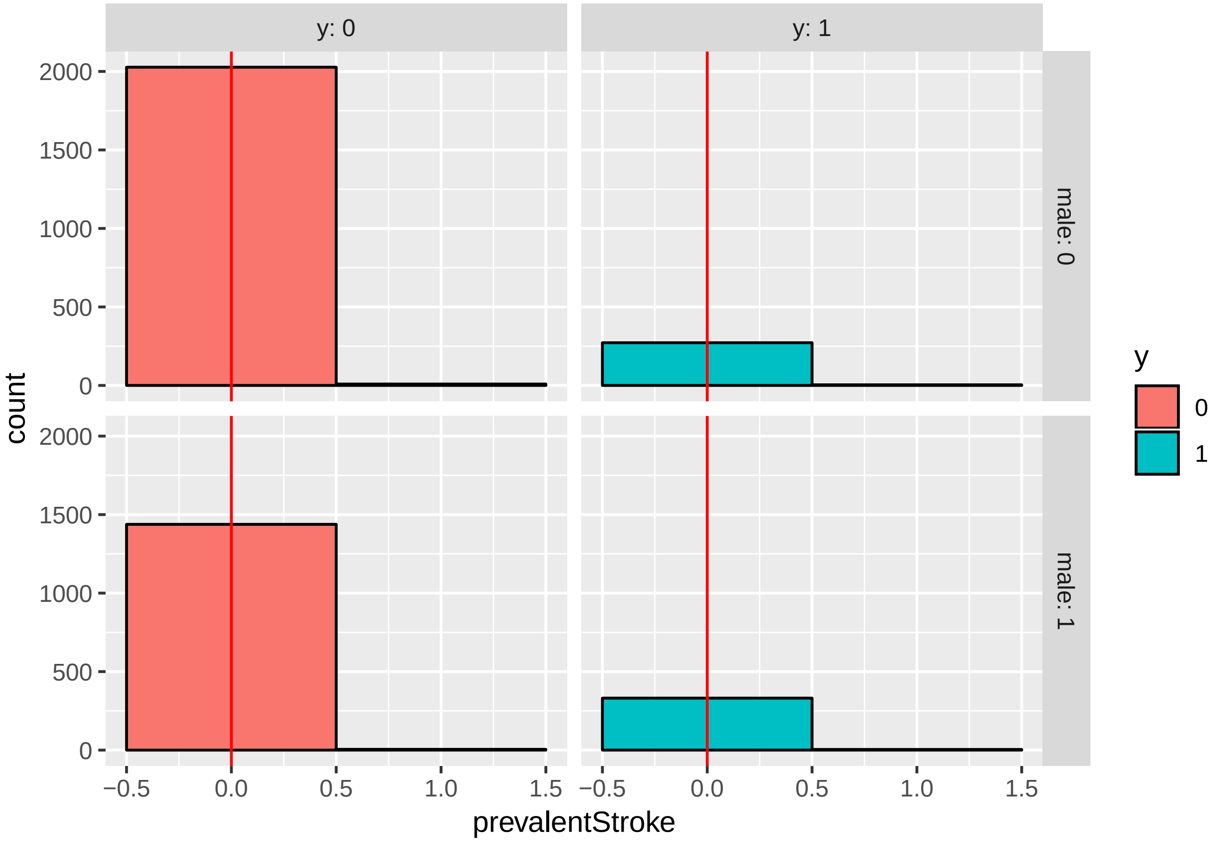

y: 0
y: 1
male: 0
male: 1
count
400
300
200
100
0
30
40
50
60
70
age
y
0
1

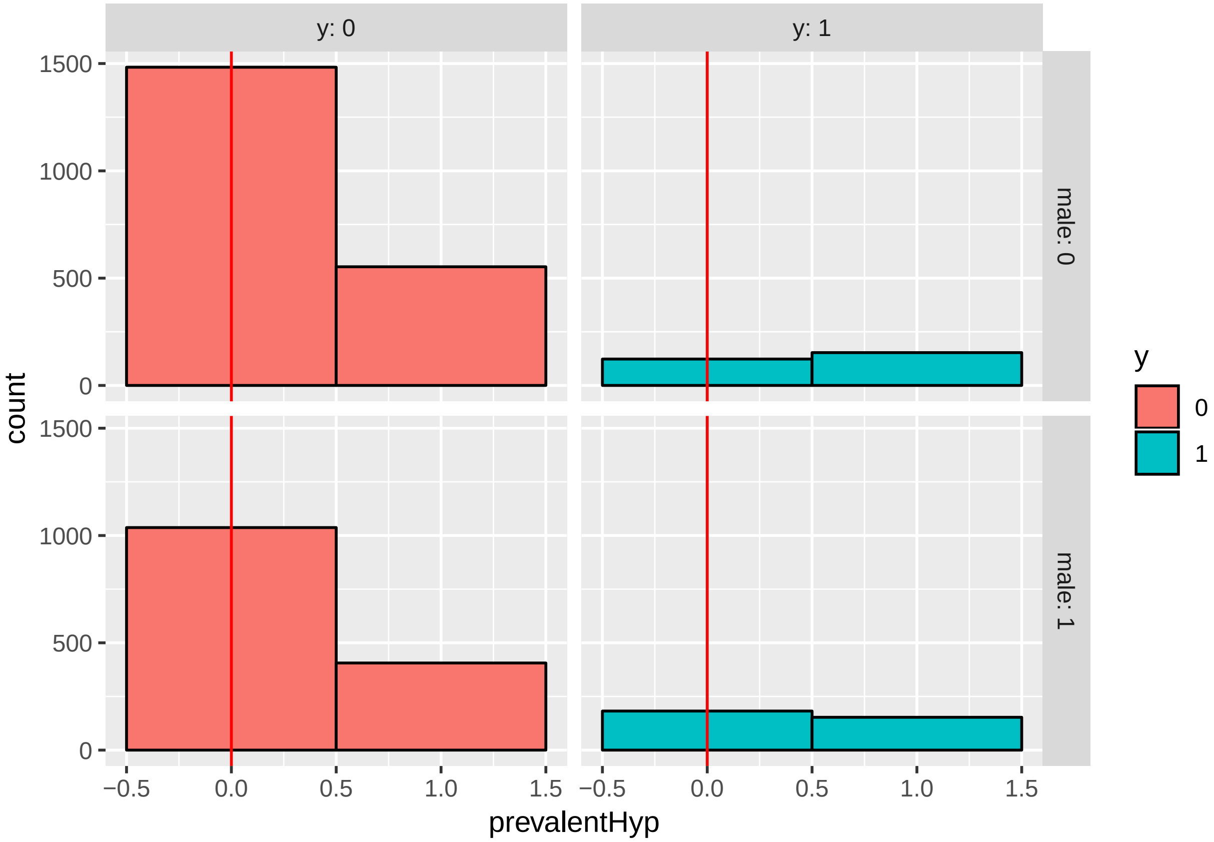

y: 0
y: 1
male: 0
male: 1
count
1250
1000
750
500
250
0
-0.5
0.0
0.5
1.0
1.5
currentSmoker
y
0
1

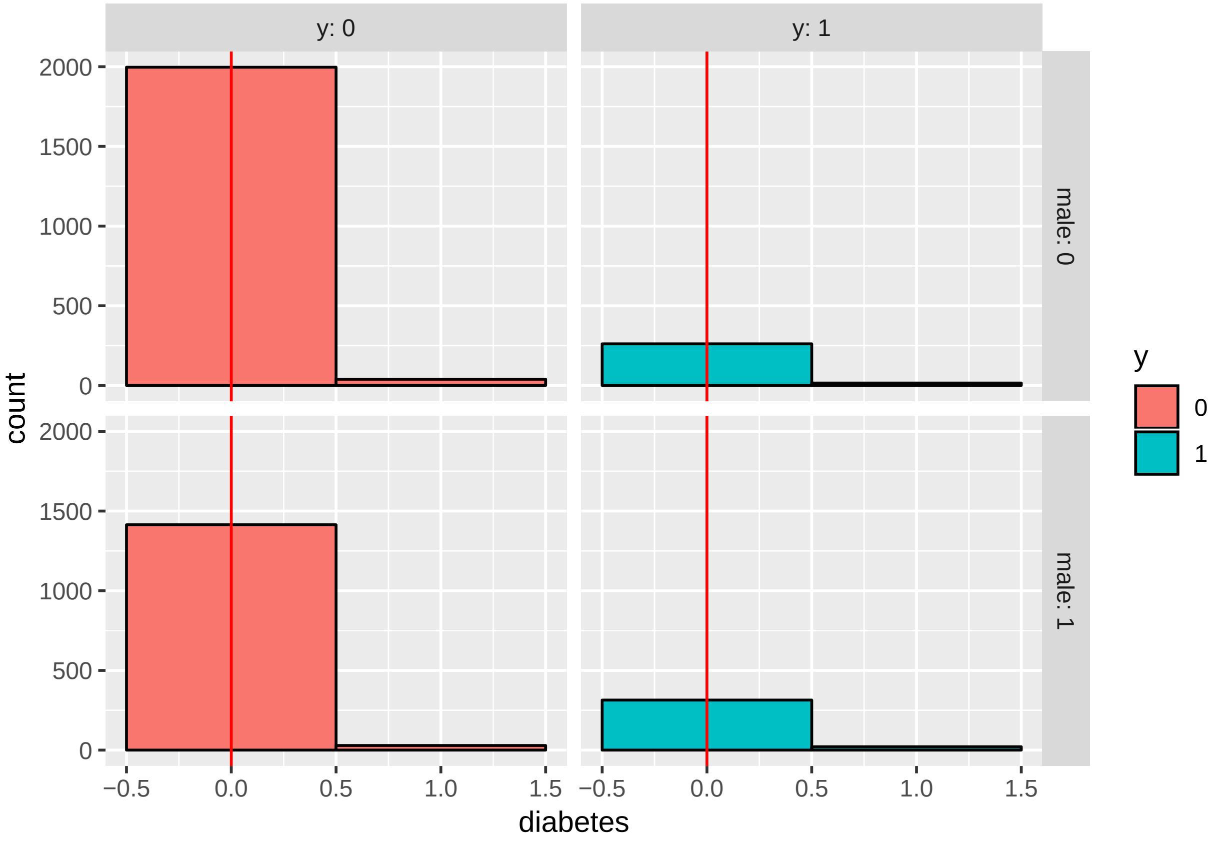


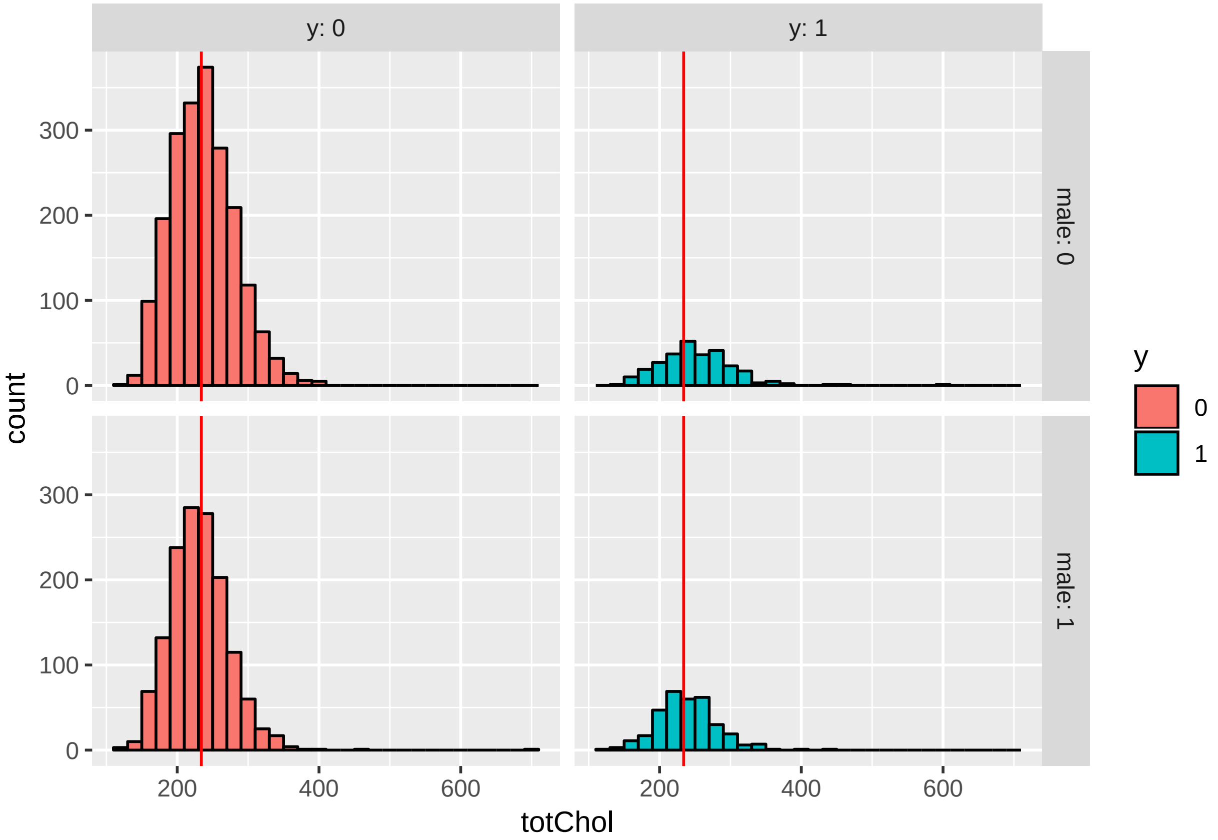
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   8.995  20.000  70.000
```


y: 0
y: 1
male: 0
male: 1
count
2000
1500
1000
500
0
−0.5
0.0
0.5
1.0
1.5
BPMeds
y
0
1

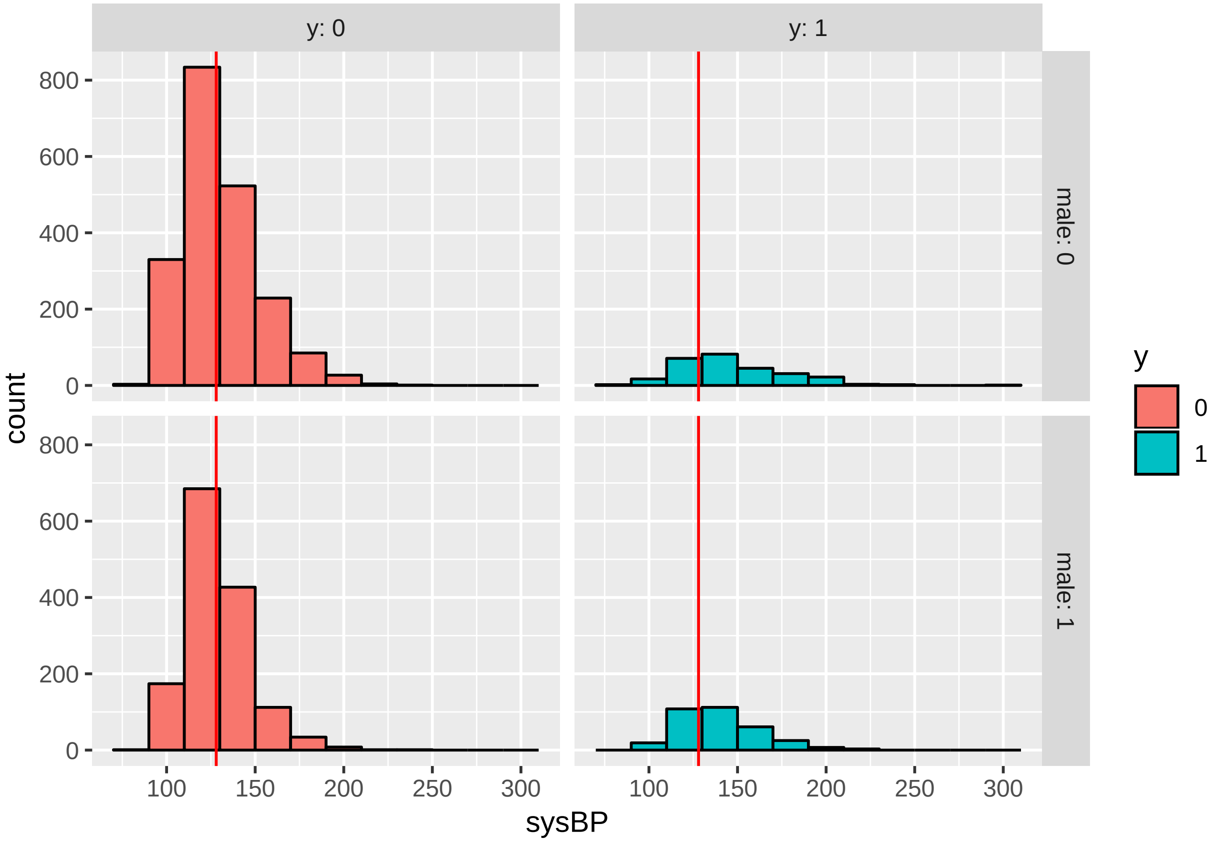

y: 0
y: 1
male: 0
male: 1
count
2000
1500
1000
500
0
-0.5
0.0
0.5
1.0
1.5
prevalentStroke
y
0
1


y: 0
y: 1
male: 0
male: 1
count
1500
1000
500
0
1500
1000
500
0
−0.5
0.0
0.5
1.0
1.5
−0.5
0.0
0.5
1.0
1.5
prevalentHyp
y
0
1

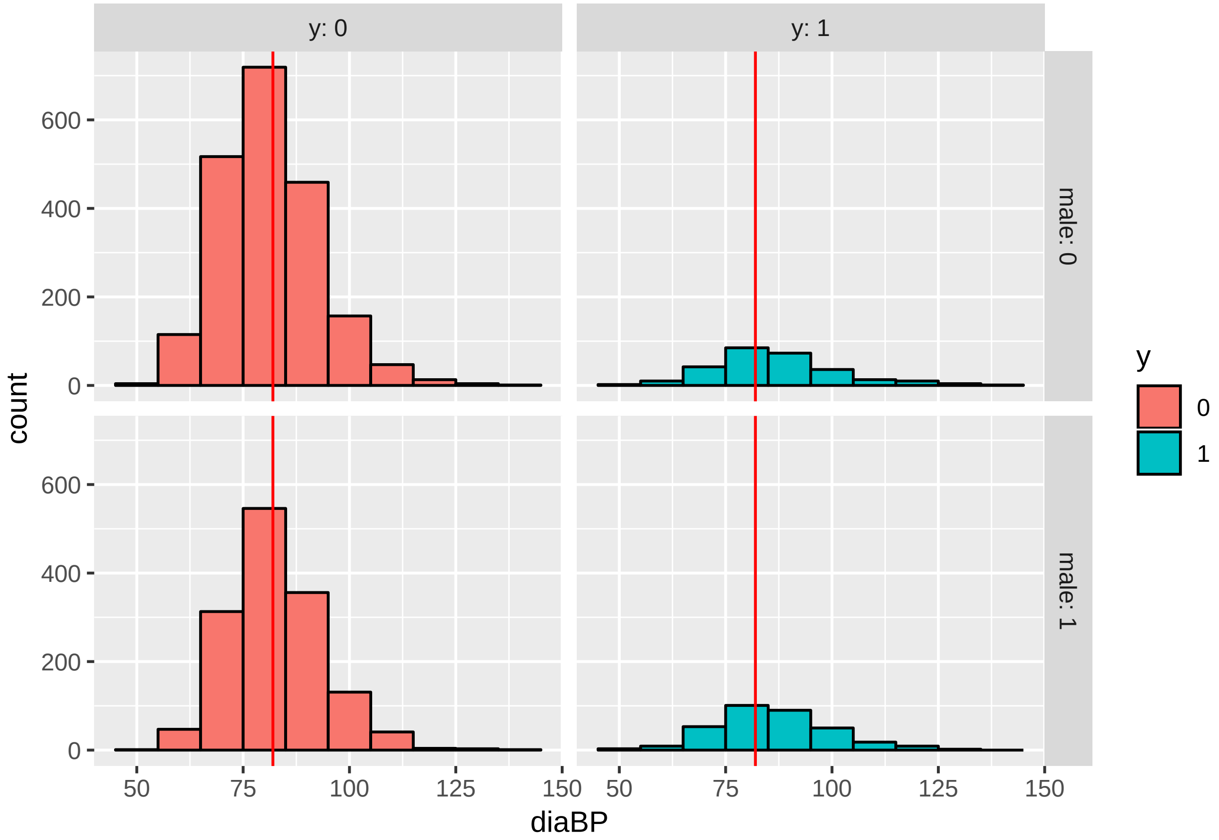

y: 0
y: 1
male: 0
male: 1
count
2000
1500
1000
500
0
−0.5
0.0
0.5
1.0
1.5
diabetes
y
0
1



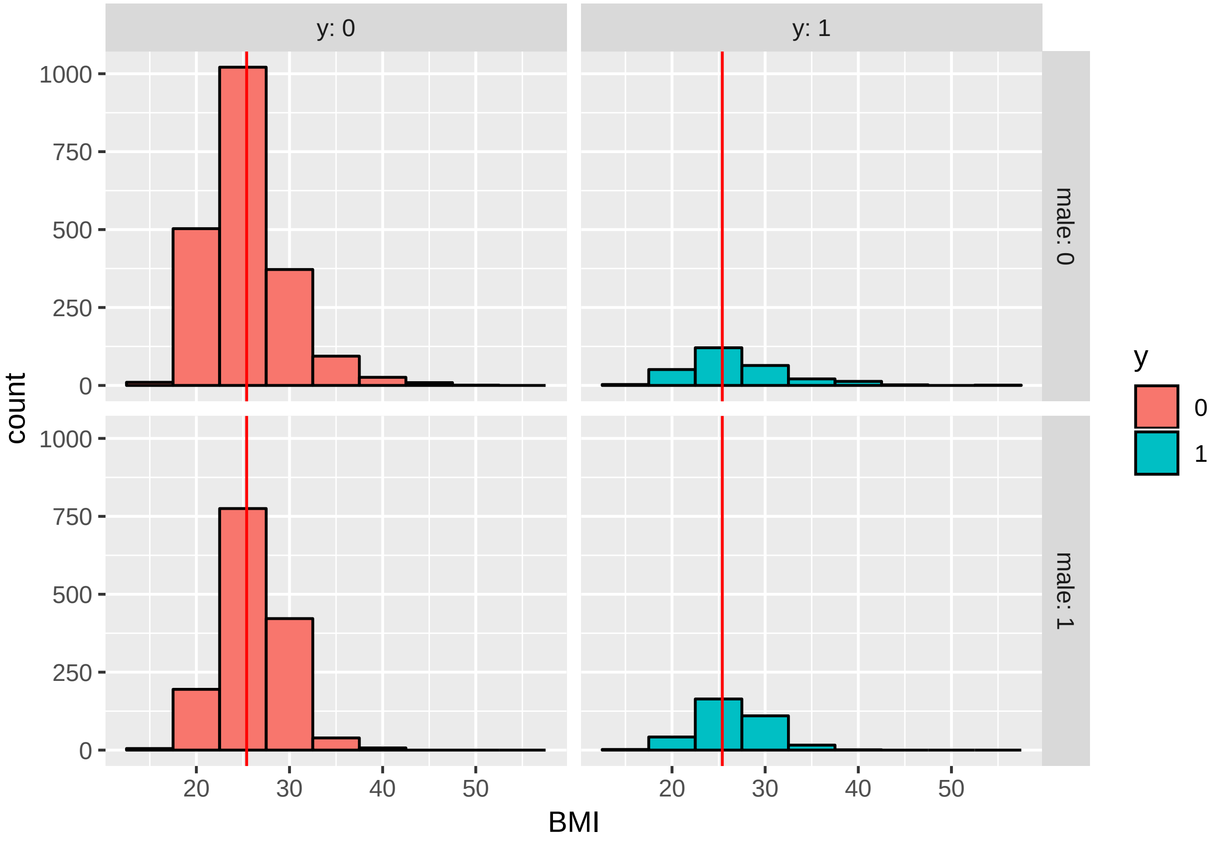
```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    113.0   206.0   234.0   236.7   263.0   696.0
```



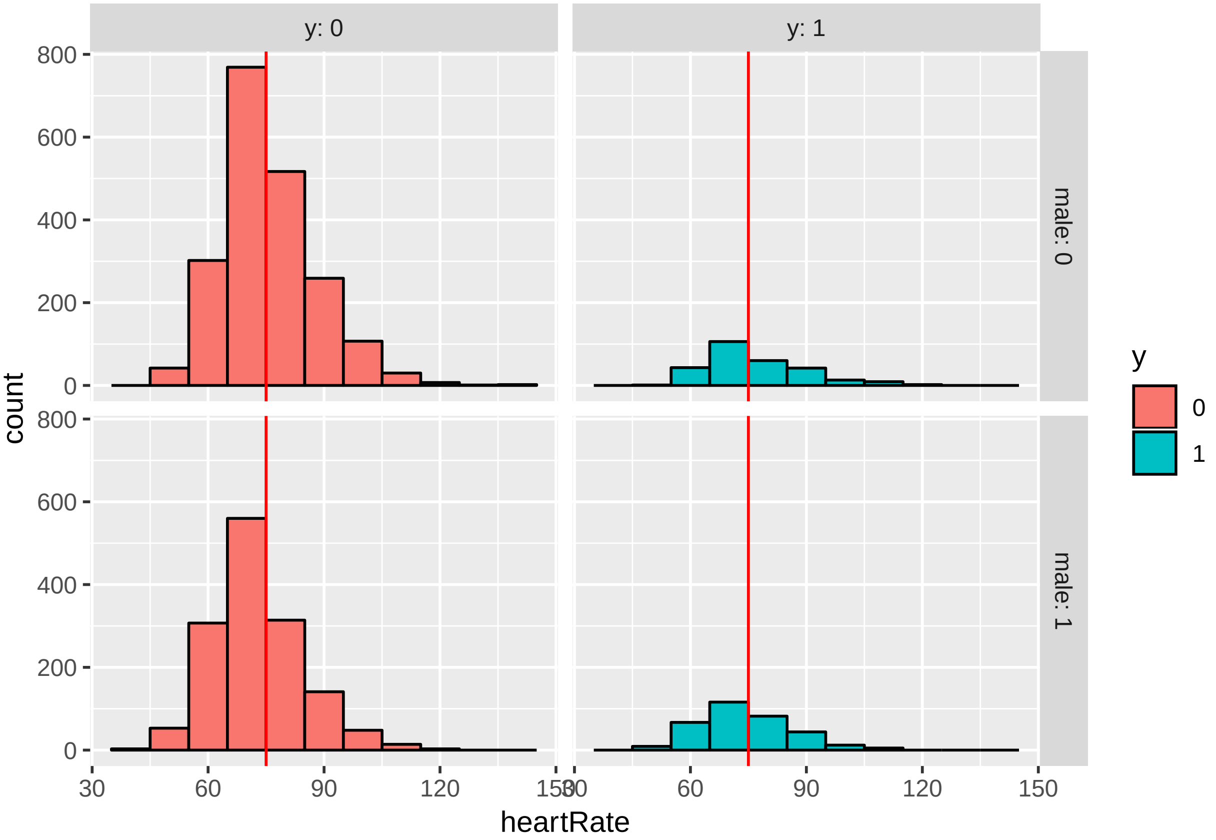
```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     83.5   117.0   128.0   132.2   143.5   295.0
```



```
##     Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    48.00   75.00   82.00   82.89   89.50  142.50
```



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.54   23.07   25.40   25.80   28.04   56.80
```



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   44.00   68.00   75.00   75.84   83.00  143.00
```

Also, prevalent stroke does not seem to be present much (only 22 times, and only slightly favors a diseased outcome). But it has a slight tilt towards increased risk.

```
## # A tibble: 4 x 3
## # Groups:   prevalentStroke [2]
##   prevalentStroke y         n
##             <dbl> <fct> <int>
## 1               0 0      3465
## 2               0 1       603
## 3               1 0        14
## 4               1 1         8
```

Modeling approach

I use the caret package to Split the data into training and test sets.

Since I have significant amount of data (over 4000 rows), I allocate 20% of data for test and 80% for training, still giving me sufficient data to train

Since this is a classification problem with binary outcomes, precision and specificity (and related complex metrics like F score) will be used for analyzing different models and their permutations. Specificity will be valued much higher because false negatives have high consequences: when an at-risk patient drops off from life-saving treatment options.

The goal of this analysis is to train the most appropriate model for this problem. In addition to specificity and precision of prediction, emphasis is also given to performance and simplicity of the model.

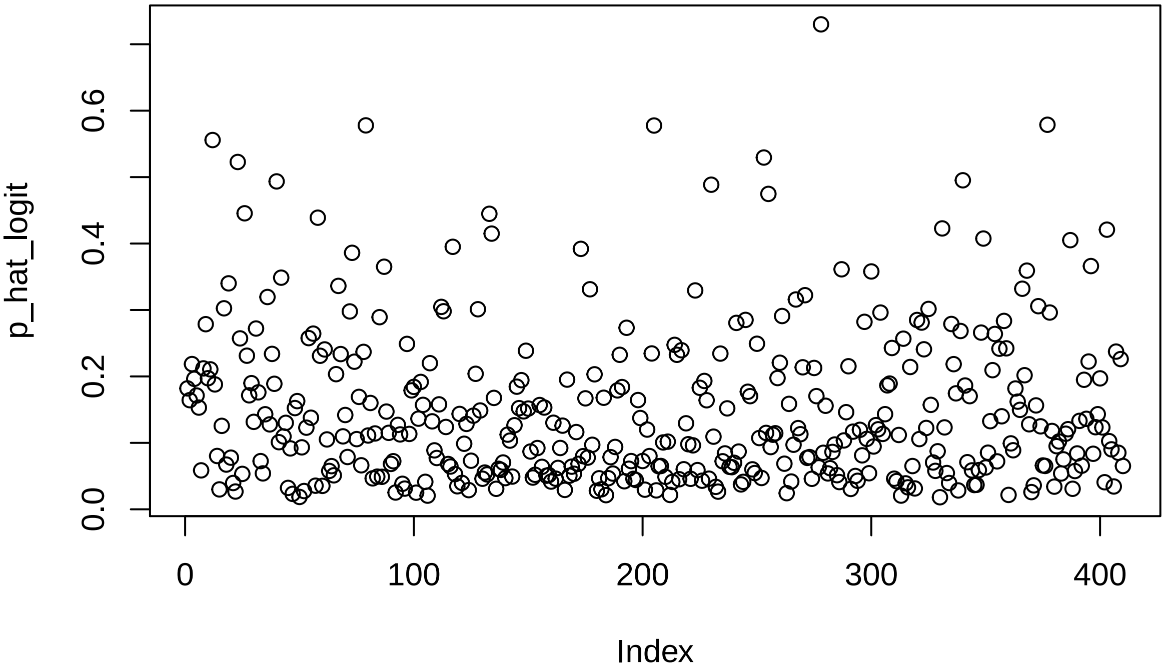
I repeat the model for the UCI/Cleveland Heart dataset and gain some insights around the subtle difference between predicting what has happened vs. the chance of something happening (predicting the prediction, so to speak).

Results and Outcomes

Train the models using the Framingham dataset

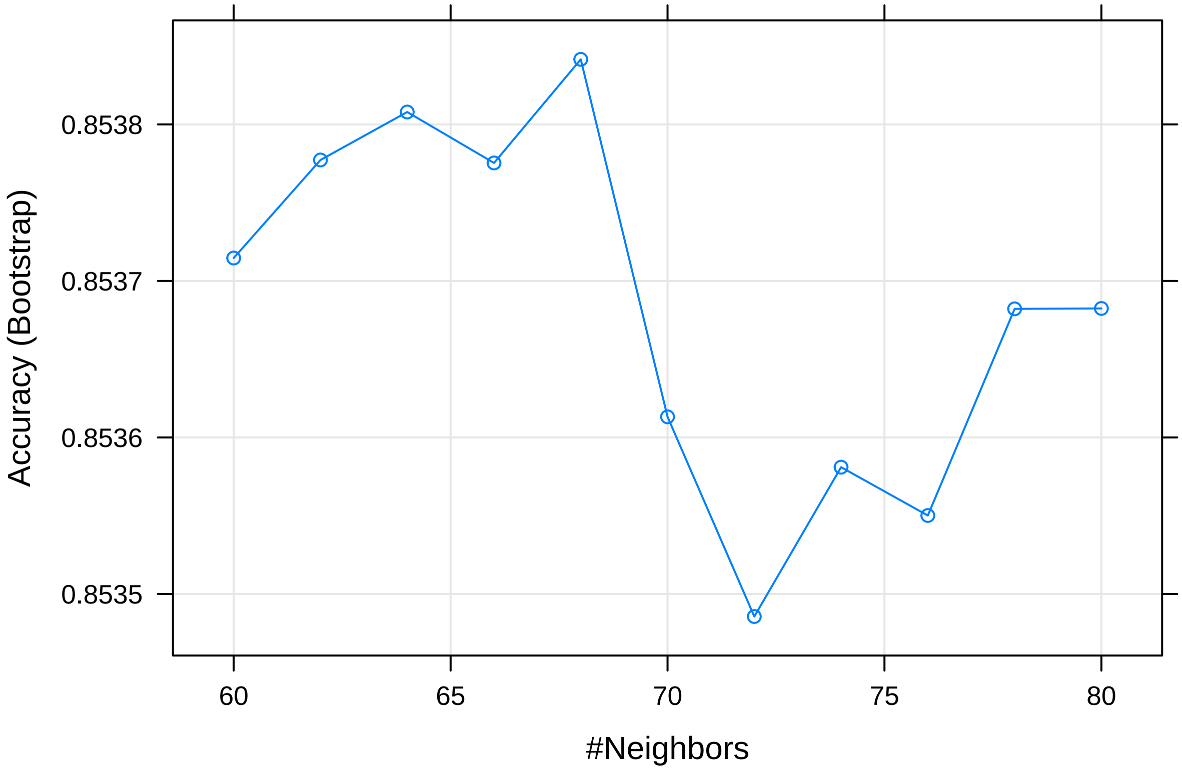
Train six different models and collect the results . . .

Model 1: Logistic Regression - with all predictors



Model 2: K-Nearest Neighbors (KNN)

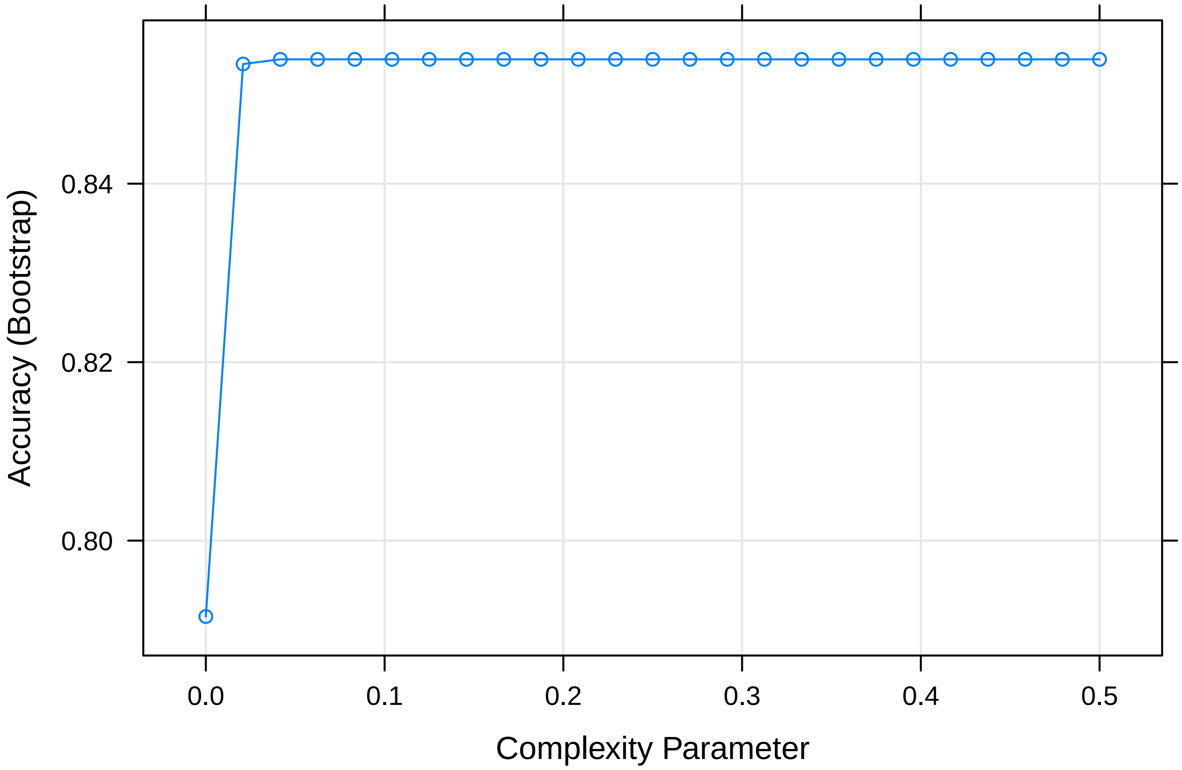
```
##    k
## 5 68
```



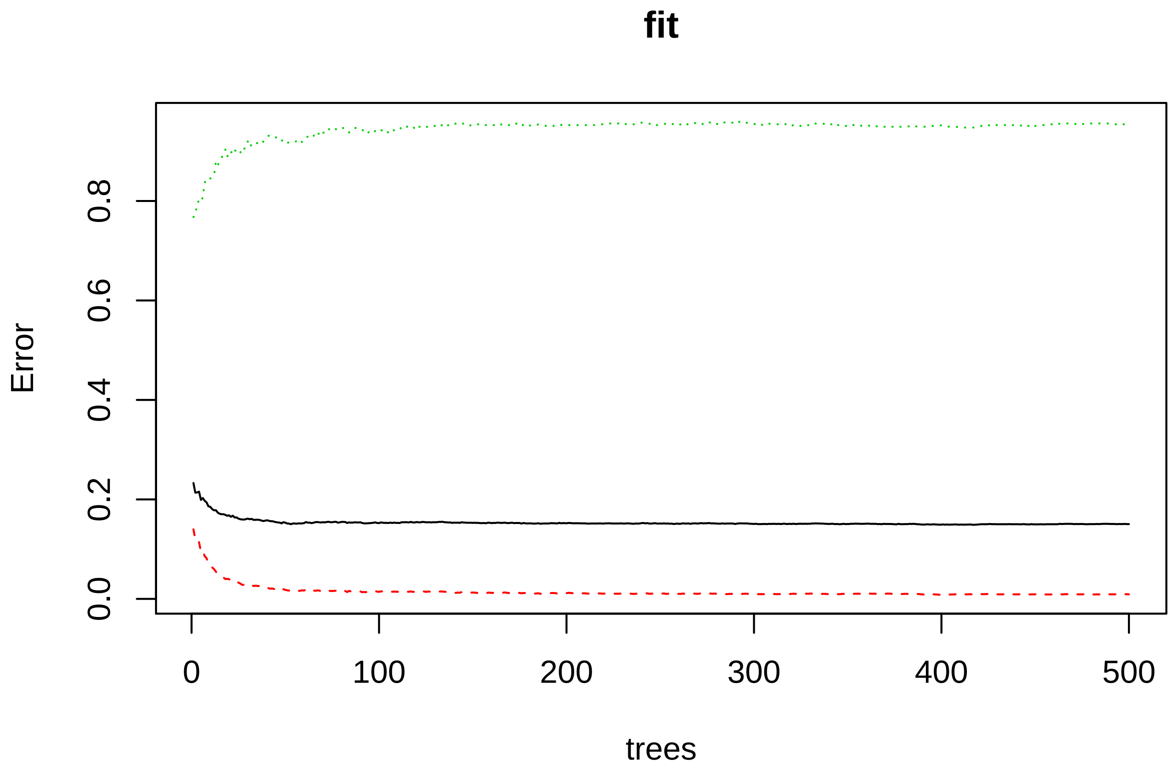
Model 3: Quadratic Discriminant Analysis (QDA)

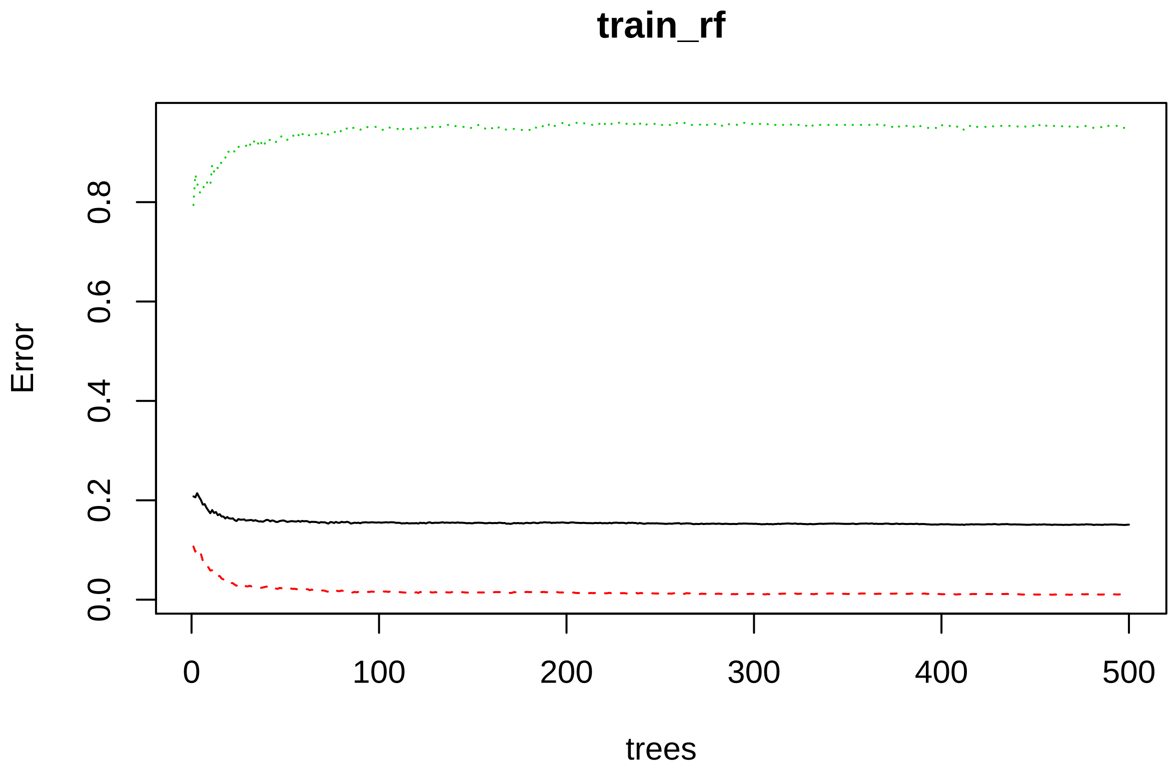
Model 4: Linear Discriminant Analysis (LDA)

Model 5: Decision Tree - using rpart



Model 6: Random Forest


fit
Error
0.0
0.2
0.4
0.6
0.8
0
100
200
300
400
500
trees



The following table summarizes the results obtained for Accuracy and Specificity, respectively for each trained model.

Accuracy:

```
## # A tibble: 6 x 2
##   method                                accuracy
##   <chr>                                    <dbl>
## 1 logistic regression - all params         0.851
## 2 knn caret                                0.849
## 3 quadratic discriminant analysis (QDA)    0.834
## 4 linear discriminant analysis (LDA)       0.854
## 5 decision tree (CART)                     0.849
## 6 random forest                            0.856
```

Specificity:

```
## # A tibble: 6 x 2
##   method                                specificity
##   <chr>                                       <dbl>
## 1 logistic regression - all params           0.0645
## 2 knn caret                                  0
## 3 quadratic discriminant analysis (QDA)      0.194
## 4 linear discriminant analysis (LDA)         0.0968
## 5 decision tree (CART)                       0
## 6 random forest                              0.0968
```

Final Model Selection

The best performing are those that have high accuracy and specificity. These include * logistic regression * QDA * LDA * Random forest

QDA has best specificity at the slight relative loss of accuracy Random forest has the best accuracy Logistic regression also shows are reasonable combination of accuracy and prediction while at the same time being simpler to use with a much better performance compared to some of the others.

Of course, we can play with hyper-parameters of each model and further pre-process some of the inputs to further tweak the numbers up and down.

I pick logistic regression as it is a simpler model and there is no need to use a more complex classification model if logistic regression does the trick.

Furthermore, since specificity is more important than precision, I tune the model to give more weight to specificity rather than precision when training.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 168  12
##          1 180  50
##
##                Accuracy : 0.5317
##                  95% CI : (0.4821, 0.5808)
##     No Information Rate : 0.8488
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1368
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.4828
##             Specificity : 0.8065
##          Pos Pred Value : 0.9333
##          Neg Pred Value : 0.2174
##              Prevalence : 0.8488
##          Detection Rate : 0.4098
##    Detection Prevalence : 0.4390
##       Balanced Accuracy : 0.6446
##
##        'Positive' Class : 0
##
```

```
##
## Call:
## glm(formula = y ~ male + age + currentSmoker + cigsPerDay + BPMeds +
##     prevalentStroke + prevalentHyp + diabetes + totChol + sysBP +
##     diaBP + BMI + heartRate, family = "binomial", data = .)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4114  -0.5837  -0.4253  -0.2860   2.8446
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -8.1151719  0.7151322 -11.348  < 2e-16 ***
## male             0.5128679  0.1154236   4.443 8.86e-06 ***
## age              0.0634810  0.0070716   8.977  < 2e-16 ***
## currentSmoker    0.0140050  0.1640713   0.085  0.93198
## cigsPerDay       0.0205325  0.0064303   3.193  0.00141 **
## BPMeds           0.2804471  0.2446094   1.147  0.25158
## prevalentStroke  0.7038507  0.5474883   1.286  0.19858
## prevalentHyp     0.2132526  0.1464048   1.457  0.14523
## diabetes         0.6308278  0.2504774   2.519  0.01179 *
## totChol          0.0023780  0.0011482   2.071  0.03835 *
## sysBP            0.0162007  0.0040625   3.988 6.67e-05 ***
## diaBP           -0.0031051  0.0067922  -0.457  0.64756
## BMI              0.0011676  0.0134907   0.087  0.93103
## heartRate       -0.0003753  0.0043580  -0.086  0.93137
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2790.0  on 3310  degrees of freedom
## Residual deviance: 2473.9  on 3297  degrees of freedom
## AIC: 2501.9
##
## Number of Fisher Scoring iterations: 5
```

I see that male, age and sysB are the most contributing inputs towards the training of this model, while cigPerDay, diabetes and totlChol also have some contribution. The other inputs do not show any significant contribution towards predicting a patient's 10-year chance of getter a heart stroke.

Let's drop the unused parameters, retrain the model, and check the performance again.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 167  13
##          1 181  49
##
##                Accuracy : 0.5268
##                  95% CI : (0.4772, 0.576)
##     No Information Rate : 0.8488
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.1279
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.4799
##             Specificity : 0.7903
##          Pos Pred Value : 0.9278
##          Neg Pred Value : 0.2130
##              Prevalence : 0.8488
##          Detection Rate : 0.4073
##    Detection Prevalence : 0.4390
##       Balanced Accuracy : 0.6351
##
```

```
##        'Positive' Class : 0
##

##
## Call:
## glm(formula = y ~ male + age + cigsPerDay + diabetes + totChol +
##     sysBP, family = "binomial", data = .)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4485  -0.5889  -0.4289  -0.2875   2.8737
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.674609   0.483285 -17.949  < 2e-16 ***
## male         0.512416   0.113078   4.532 5.86e-06 ***
## age          0.064578   0.006830   9.456  < 2e-16 ***
## cigsPerDay   0.020642   0.004360   4.735 2.20e-06 ***
## diabetes     0.644413   0.248364   2.595  0.00947 **
## totChol      0.002457   0.001144   2.149  0.03167 *
## sysBP        0.018604   0.002297   8.100 5.51e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2790.0  on 3310  degrees of freedom
## Residual deviance: 2479.8  on 3304  degrees of freedom
## AIC: 2493.8
##
## Number of Fisher Scoring iterations: 5
```

There is only a slight but not appreciable decrease. We can train with either option, but it’s good to know where is the most buck for the benefit, in case we have to train on large amount of data or if we are investing a significant amount of money on collecting the inputs that are not useful.

Comparing Framingham Data Set with UCI Data Set

The following table summarizes the results obtained for Accuracy and Specificity, respectively for each trained model for the UCI dataset.

UCI/Cleveland Accuracy:

```
## # A tibble: 6 x 2
##   method                                accuracy
##   <chr>                                    <dbl>
## 1 logistic regression - all params         0.885
## 2 knn caret                                0.885
## 3 quadratic discriminant analysis (QDA)    0.820
## 4 linear discriminant analysis (LDA)       0.918
## 5 decision tree (CART)                     0.918
## 6 random forest                            0.869
```

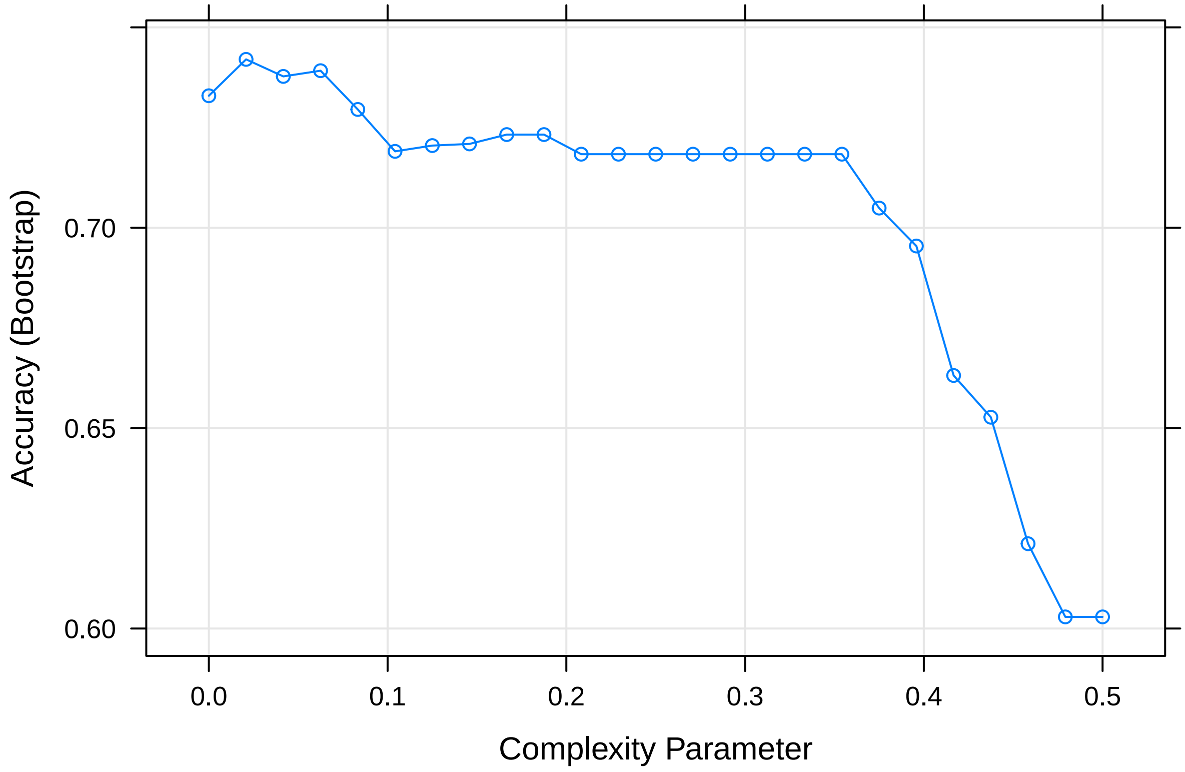
UCI/Cleveland Specificity:

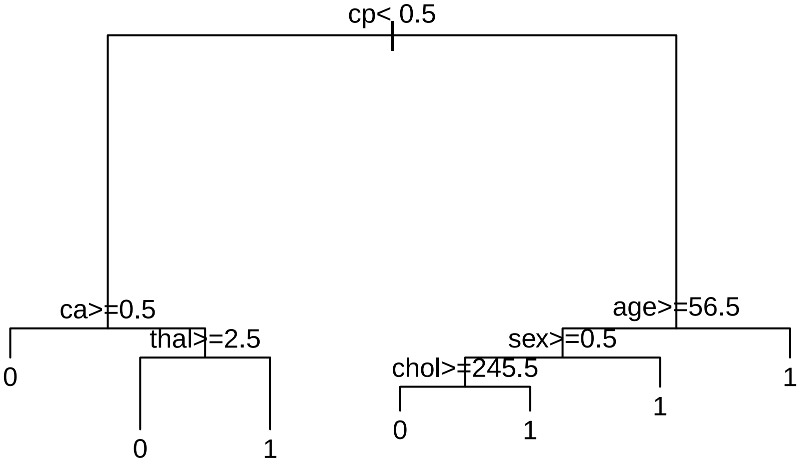
```
## # A tibble: 6 x 2
##   method                                specificity
##   <chr>                                      <dbl>
## 1 logistic regression - all params            0.909
## 2 knn caret                                   0.909
## 3 quadratic discriminant analysis (QDA)       0.788
## 4 linear discriminant analysis (LDA)          0.939
## 5 decision tree (CART)                        0.939
## 6 random forest                               0.879
```

In this case I obtain much better numbers and both good precision and specificity. This is because we are predicting an existing thing, and not future probability which is much more dependent on future events and lifestyle changes during the next ten years.

LDA and decision tree show the best overall results with both high accuracies and specificities.

I pick decision tree for this problem as it also provides a much better visual and logical output that healthcare professional can interactively look at as a comprehensive diagnostic approach, either for the purpose of improving their own diagnostics, or improving the machine learning model based on their own subject matter expertise around diagnosing heart disease. The following decision tree output demonstrates this:





Conclusion

This became a bigger undertaking than what I started out with (and what was required from me). In addition to finding the best model and predicting it for diagnostic purposes, I wanted to see why the given dataset (Framingham, in this case) was not getting very good performance despite doing all the necessary preparation. And I needed run the model on another dataset and compare the results.

What I found through data exploration and modelling, is that the Framingham dataset predicts the risk, while UCI predicts the actual disease. This is one reason UCI is more accurate even with small data to train with. This is an important consideration in problem formulation: there are problems that predict what happened or is happening (this has the uncertainty of prediction); and there are problems that predict something that may happen (i.e. with double uncertainty: that of prediction and that of the chance of it happening in real life).

For training Framingham Study dataset, Logistic Regression was selected and was further tuned to favor high specificity with significant but limited loss (by less than 50%) loss of precision. There is a good amount of data to train the model quite well.

For training UCI Cleveland Study dataset Decision Tree was selected as the preferred and better-performing model. This dataset, in contrast to Framingham, shows much improved results even with a very small training set (for the reason described above).

However, one can notice that the UCI model somehow followed the common human bias around gender and age when it comes to heart treatment: the tree graph for decision tree model (in the results section) shows that women aged 56 or less will never be treated for heart disease if their cholesterol is low dispite having a poor HDL/LDL ratio, if the treatment decision is solely based on the decision tree trained here. Of course, more data (UCI dataset is very small) should improve such shortcomings.

On an interesting side note: our brains probably also make small-data-driven, simplistic models (as we are not suited for large number crunching at high speeds, for obvious reasons) and that may be the implicit cause of many of our human biases (like the UCI dataset possesses to a smaller degree). In machine learning it is important to not further feed our human biases in the model, either during training or during data gathering; and this is the reason why I removed education column from the dataset before training when I didn't see any significant correlation with the output during data exploration.

With age, there is probably another confounder, the data collection process itself: people who are young and healthy probably won't seek heart healthcare (not seeking heart care thus being a confounder to being young) and hence will appear at a lower proportion (of their demographic) in training data as healthy.

Recommendations

Improved data collection campaigns, especially incorporating genomic data (both population/stratified and individual), health metrics from wearable devices, moving forward.

Shortcomings and Areas of Improvement

The models provided useful but less-than-perfect predictions. There are still several false negatives, emphasizing that the data can only pinpoint general improvement (in saving lives and cutting the cost of many unnecessary treatments). At the population level, this can improve guidelines and help triage more effectively; on individual level, this is not a sure-shot predictor of disease (at least not yet, until we embark on more ML-targeted data collection campaigns, and develop improved modeling with more health science, biochemical, genetic, and biometric insights, the disease accurately). The more data, and the more relevant data, the more such models can pinpoint outcomes better.

References

- Introduction to Data Science - Dr. Rafael A. Irizarry: https://rafalab.github.io/dsbook/
- Centers for Disease Control (CDC) - Heart Disease Facts: https://www.cdc.gov/heartdisease/facts.htm
- Framingham Heart study dataset: https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset
- Heart Disease UCI: https://www.kaggle.com/ronitf/heart-disease-uci
- GitHub repository for this project: https://github.com/samseatt/fram-heart