Trey Hibbard

Write Up – Cardiovascular Heart Disease Classifier Models

Variables and Explanation:

The dataset includes one target and twelve predictor variables. They are the following.

- “target”
  - Binary
  - 0 = No Heart Disease
  - 1 = Has Heart Disease

1. “age”
   - Numeric
2. “gender”
   - Binary
   - 0 = Female
   - 1 = Male
3. “chestpain”
   - Nominal
   - 0 = Typical angina
   - 1 = Atypical Angina
   - 2 = Non-anginal Pain
   - 3 = Asymptomaic
4. “restingBP”
   - Numeric
   - Resting Blood Pressure
5. “serumcholestrol”
   - Numeric
6. “fastingbloodsugar”
   - Binary
   - 0 = < 120 mg/dl
   - 1 = > 120 mg/dl
7. “restingrelectro”
   - Nominal
   - 0 = Normal
   - 1 = Having ST-T wave abnormality
   - 2 = Showing probable or definite left ventricular hypertrophy
8. “maxheartrate”
   - Numeric

9. “exerciseangia”
    - Binary
    - Exercise Induced Angia
    - 0 = No
    - 1 = Yes
10. “oldpeak”
    - Numeric
    - Numeric
11. “slope”
    - Nominal
    - 0 = Unknown
    - 1 = Upsloping
    - 2 = Flat
    - 3 = Downsloping
12. “noofmajorvessels”
    - Ordinal
    - Number of existing major blood vessels affected
    - 0,1,2, or 3

Overview and Methodology:

The preference for most projects is to take the simpler model if it performs competitively, if not better, than a more complicated black box model. For this research, I wanted to investigate to scenarios.

1. Every variable is available to you.

2. Only variables measurable at home are available.

Home measurable variables include.

- “age”
- “gender:
- “restingBP”
- “chestpain”
- “exerciseangia”

Age and gender are clearly measurable, but the other options have some room for error. To measure resting blood pressure, someone could acquire their own hardware in the form arm pressure monitor. Chest pain angina would need to be properly classified by the person at home and exercise angina would be done similarly, just after having done a workout.

Room for error is very present, but the idea of an at home test with fair predictive accuracy (90%) is impressive.

Data Cleaning and Preparation:

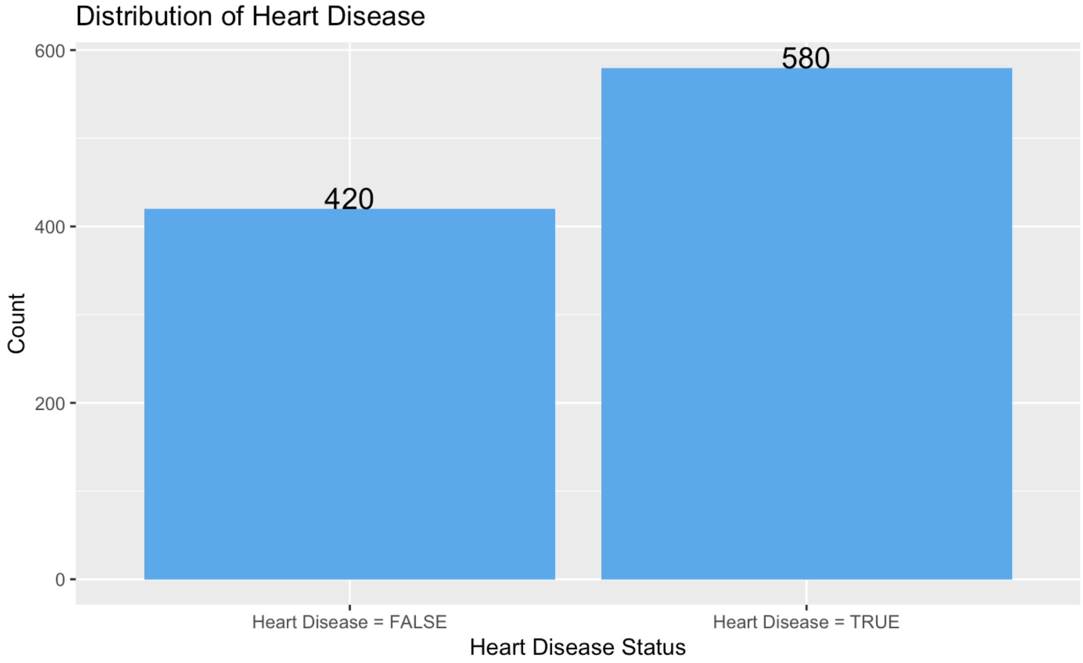
After importing the dataset, its quickly seen that the dataset is very clean, but there is one issue regarding the "slope". 180 values are classified as "0", which is not described in the pdf. I chose to label these as "unknown" and move on. Further on in analyzing the data, it seems that every "unknown" value is directly correlated to no heart disease.

No data is missing from the dataset.

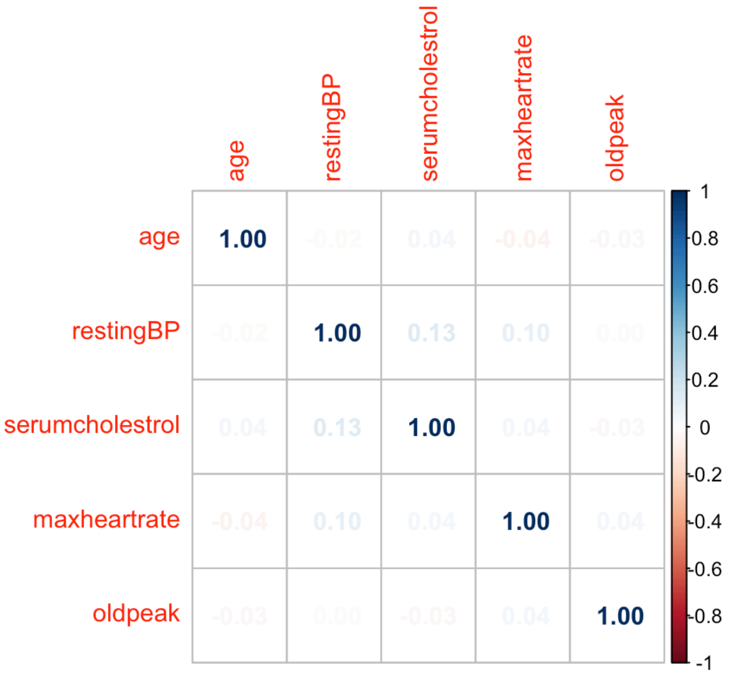
Exploratory Data Analysis:

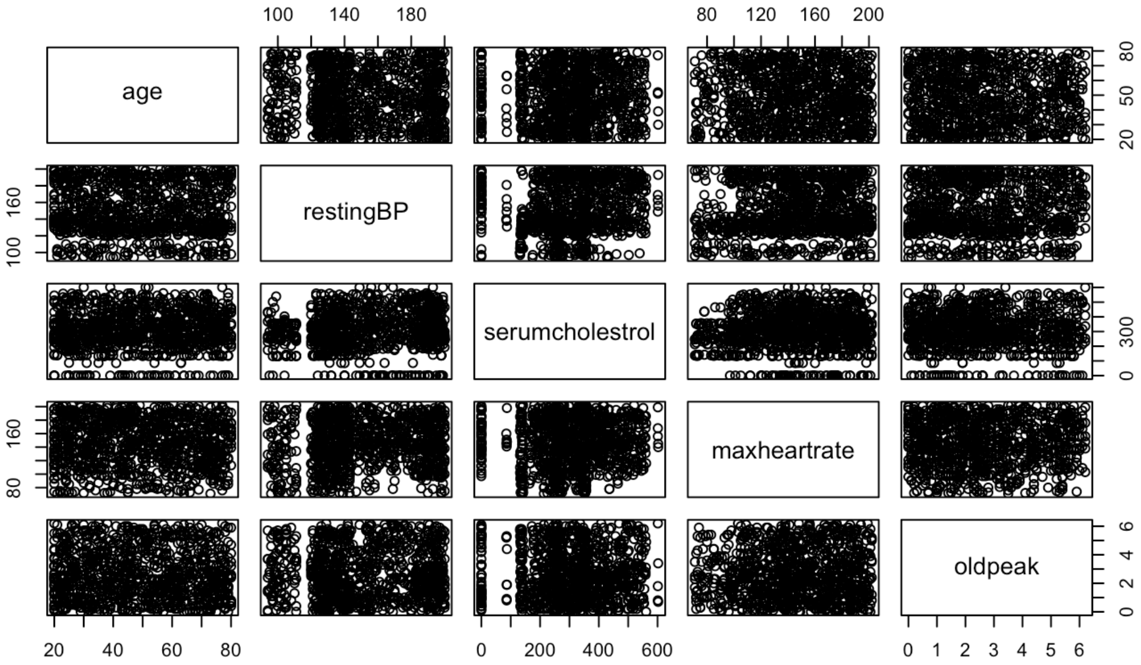
As part of early data analysis for the classification problem, I organized my data in terms of "numeric", "nominal", and "binary" data. These are the results from these observations.

1. Numeric

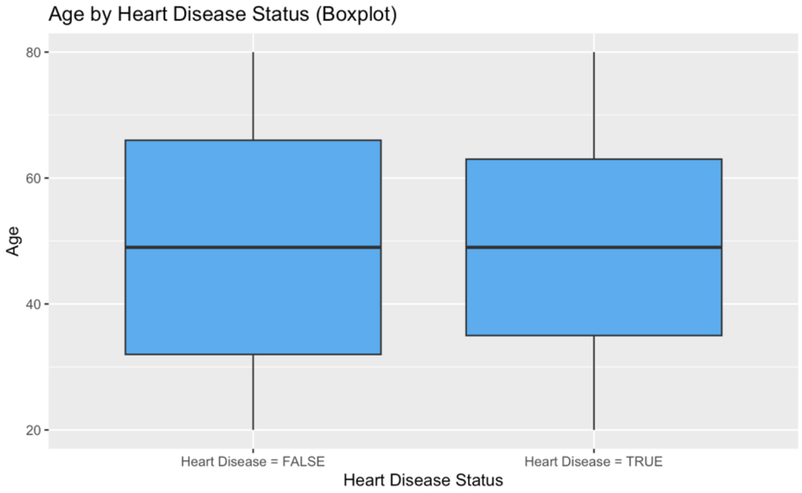


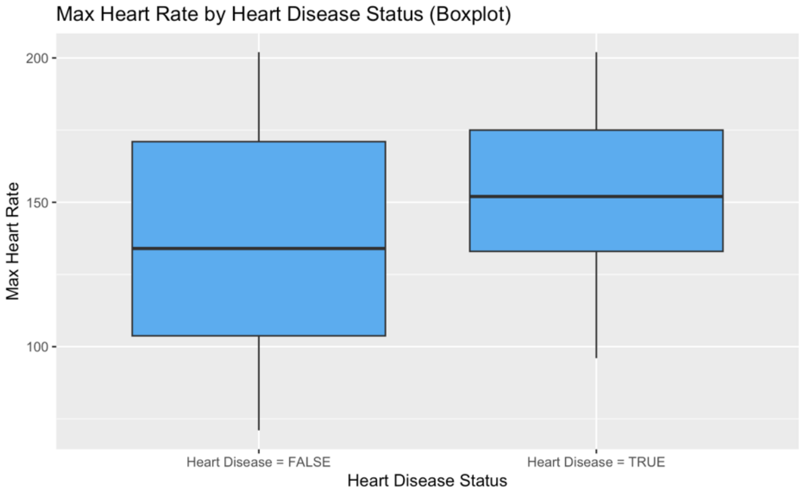
The distribution of heart disease is near evenly split, with a slight lean towards patients with heart disease present.

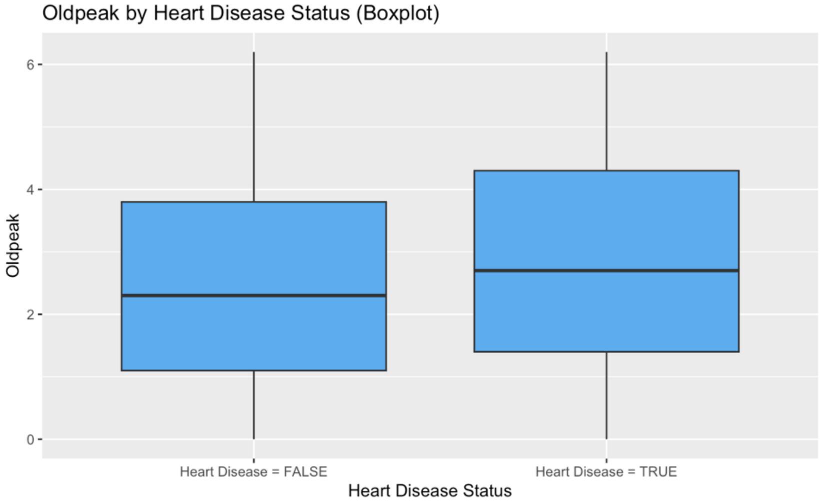




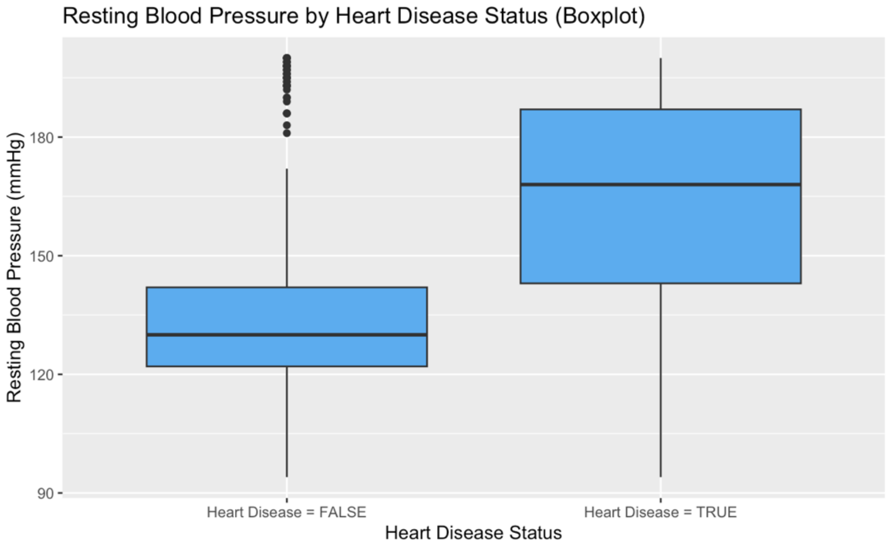
These two plots indicate a lack of correlation between our numeric variables. This gives us confidence that each variable (if important) tells us something different about our target variable.





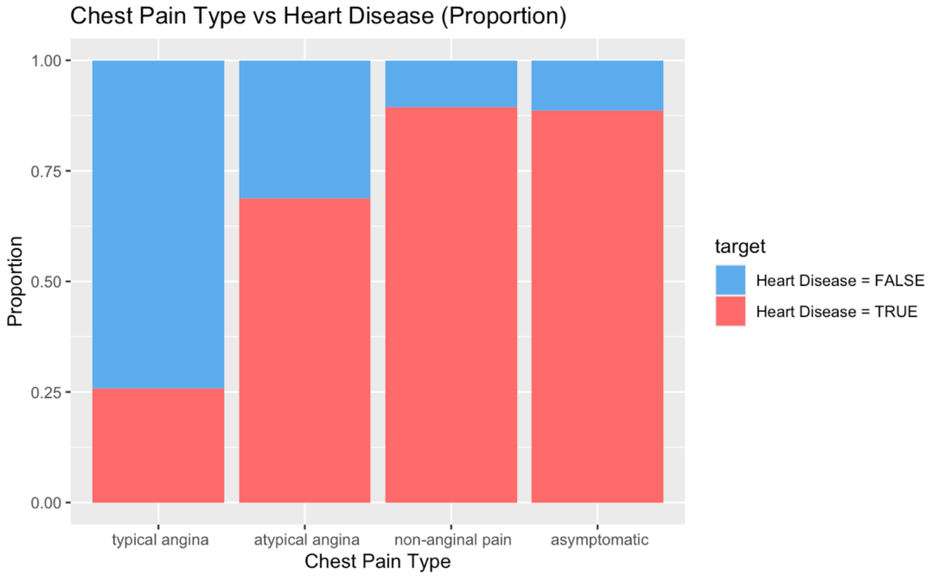


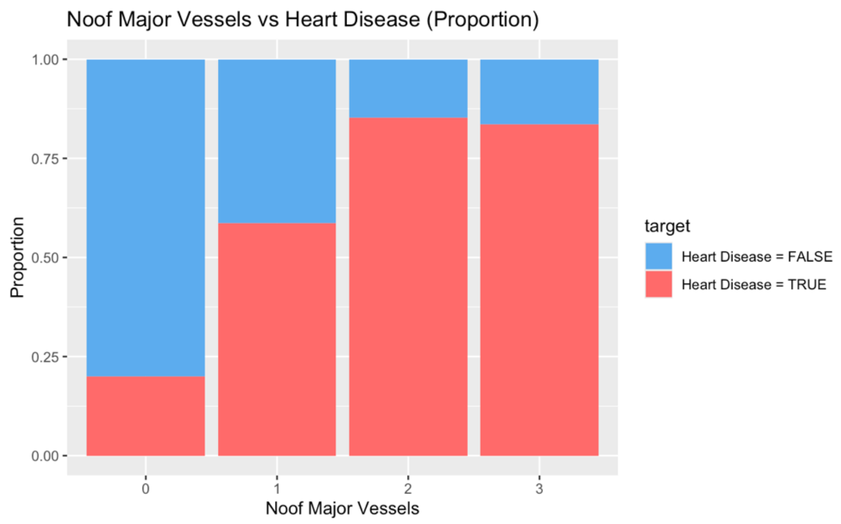
Aside from maybe max heart rate, it appears that age, max heart rate, and old peak show very little importance in predicting heart disease. If these centered around difference means, we may have considered some statistical significance.

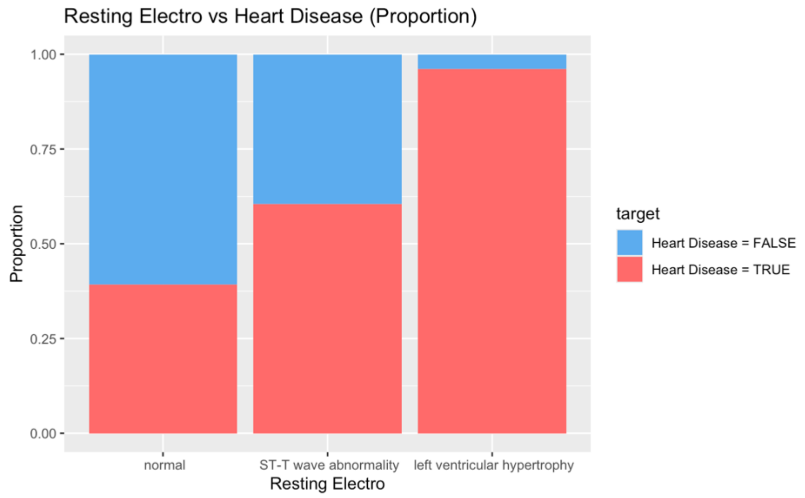


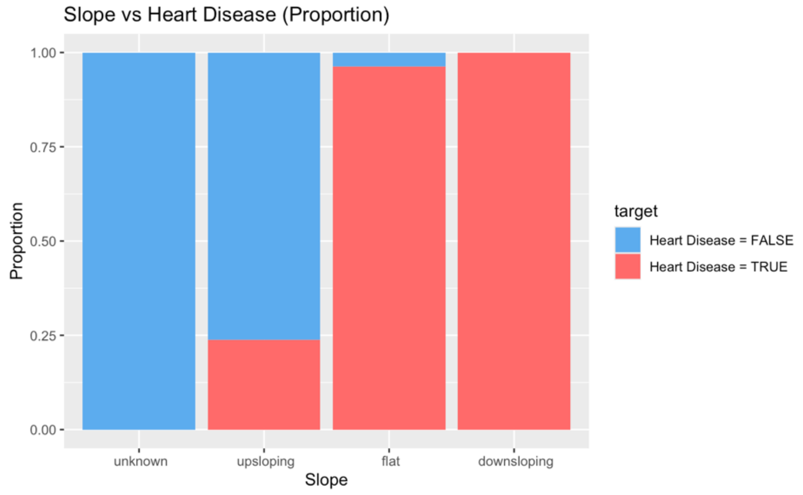
We can see that resting blood pressure appears very differently with respect to the target variable. A higher resting blood pressure often indicates heart disease present, but clearly by the outliers present when heart disease is not present, we can't always assume this to be true. There are too many outliers to consider them something to remove.

2. Nominal Data





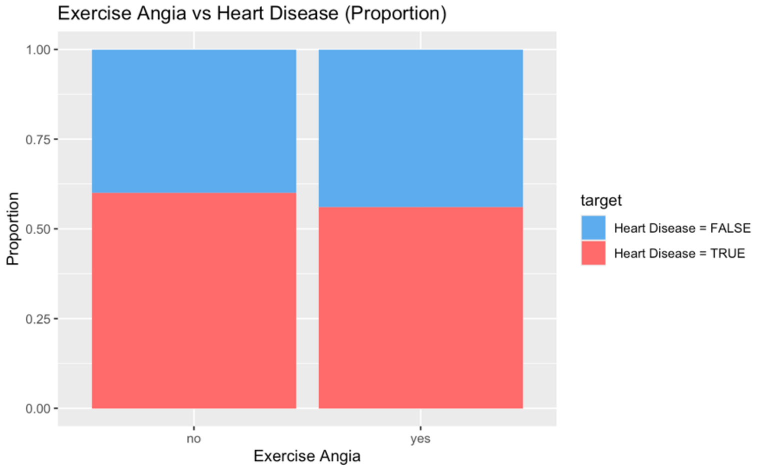


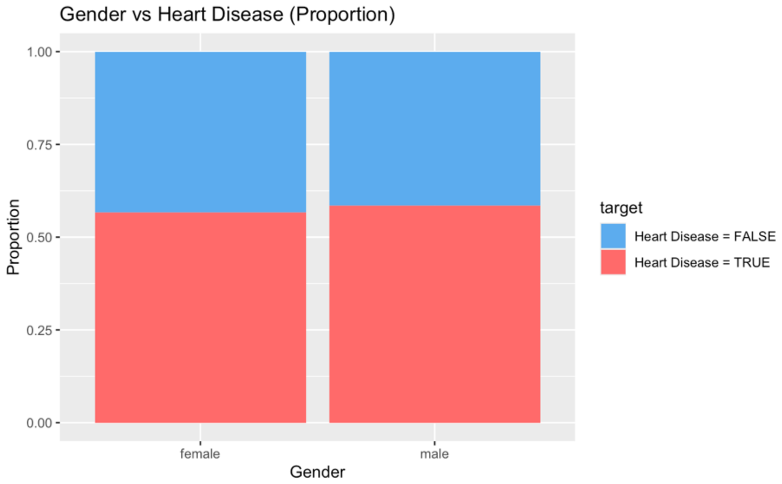


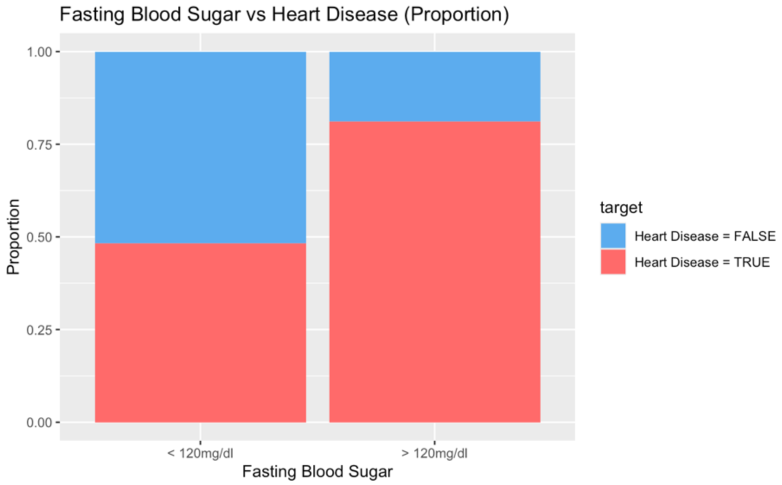
As indicated by the proportion of each of these nominal indicators, there is always at least one class which gives strong predictive accuracy. Recall that chest pain is something we can analyze at home, so the strong predictive accuracy of that variable is very insightful to us.

Additionally, down sloping can near perfectly predict heart disease in a patient. This is something measurable at an office, not at home. A larger dataset should be investigated to see how true this is, but the value of this variable cannot be understated.

3. Binary







a

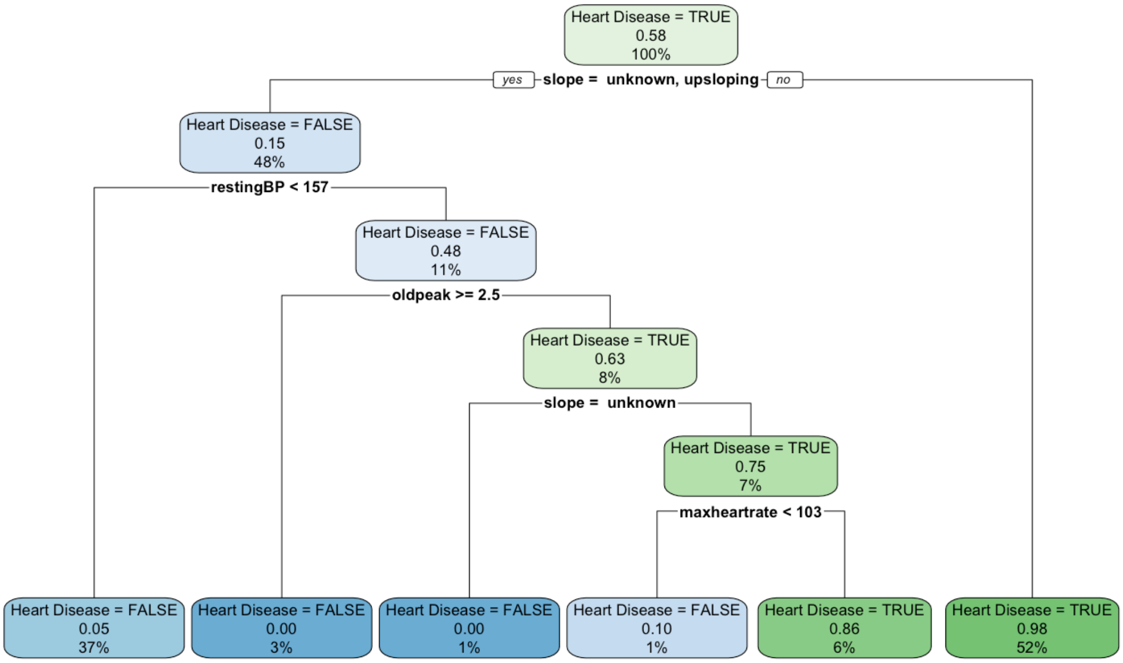
The only binary predictor we should expect to have much predictive accuracy is fasting blood sugar. However, our models show that there is not much strong value from this predictor.

Models for Medical Professional Use:

Two models were tested for medical professional use

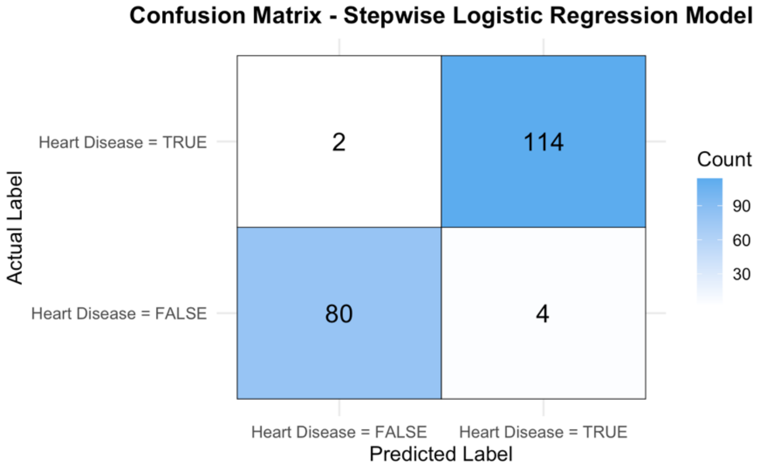
1. Logistic Regression
2. Random Forest

Additionally, a basic tree model was created to gather a basic understanding and more insight of our data.



The tree tells us that slope is by and far our most important predictor, only failing 2% of the time. Second to slope is resting blood pressure, one of the variables we can predict at home.

1. The Logistic Regression Model



Forward stepwise selection based on AIC was used to create a logistic regression model with a 97% accuracy. The model uses the following formula.

target ~ slope + oldpeak + restingBP + gender + chestpain + restingrelectro +
fastingbloodsugar + maxheartrate

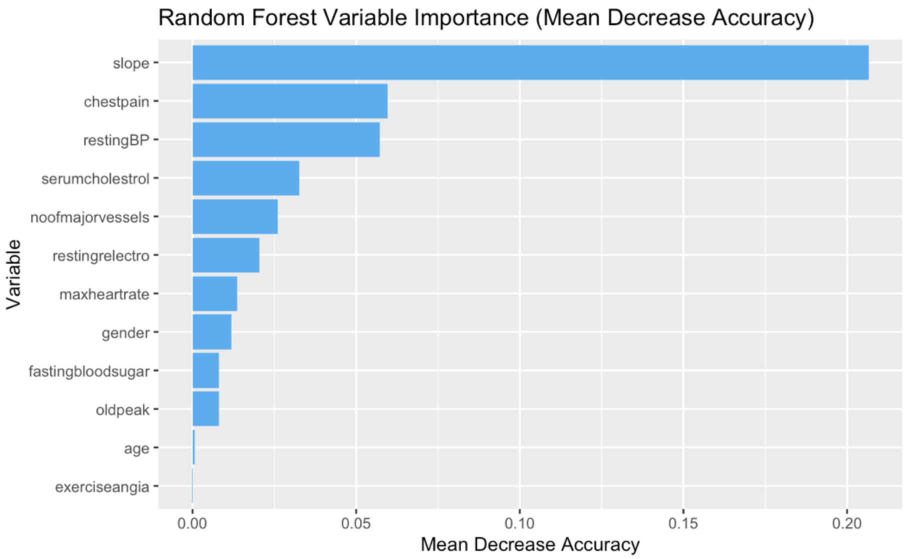
The model has a confidence interval of (0.9358, 9889) and provides us with a very simple understandable model with high predictive accuracy. Type one errors are more common than type two errors, from the few errors at all.
Through variable elimination and statistical analysis, the most important variables in this model turn out to be “slope”, “chestpain”, and “restingBP”.

The random forest model supports these results.

2. The Random Forest Model

   The random forest model also had a 97% accuracy. Since it is a more complicated model however, I opted not to use it as my final model.

   Using mean decrease accuracy, we can get a sense of variable importance. The results align with what we saw in logistic regression, which not only gives us confidence that our random forest model is good, but that we can still rely on our simpler logistic regression model.

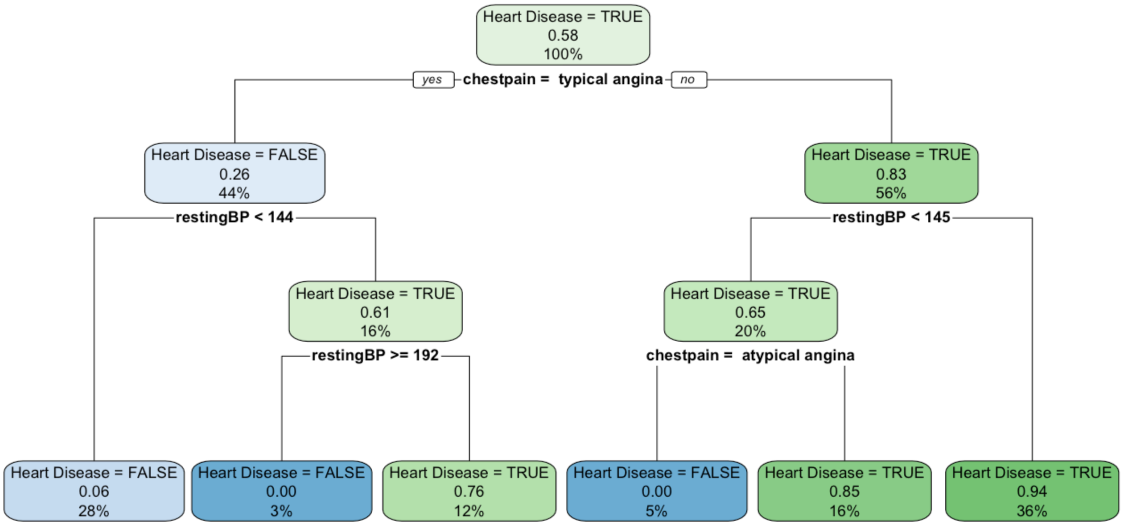


Models for Personal Use:

The following Logistic regression and random forest model are limited to the following at home measurable variables as mentioned above.

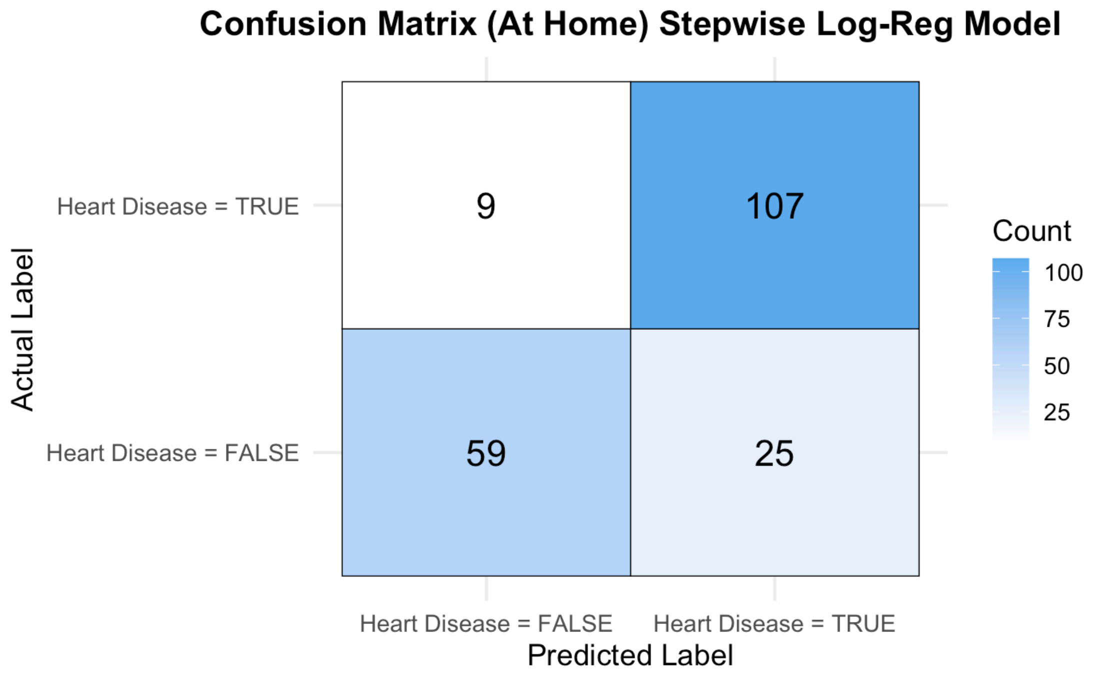
- “age”
- “gender:
- “restingBP”
- “chestpain”
- “exerciseangia”

We find that both models rely heavily on “restingBP” and “chestpain”.



This basic decision tree highlights this pattern, with chest pain being the most important for each.

1. Logistic Regression


Confusion Matrix (At Home) Stepwise Log-Reg Model

The logistic regression model for at home analysis loses a lot of the predictive accuracy we had with medical measurements with an accuracy of only 83%. Using stepwise forward selection based on AIC, we return a formula that.

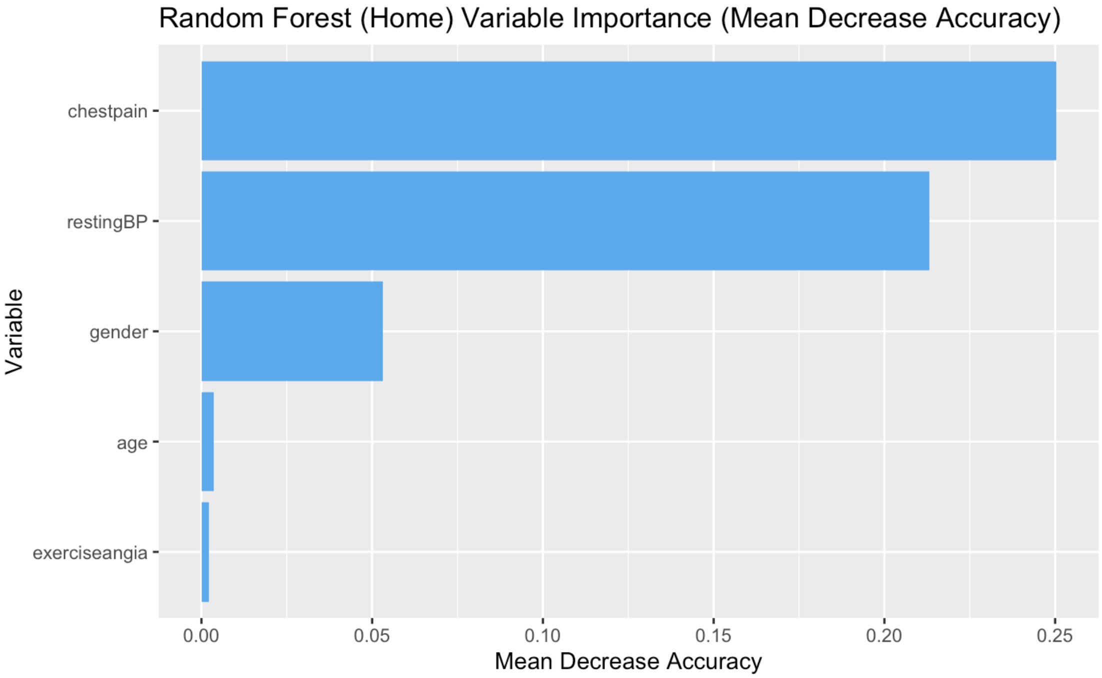
target ~ chestpain + restingBP

Considering our exploratory data analysis from the beginning, this aligns with our predictions about the predictive accuracy of these features.

2. Random Forest

The random forest model for at home measurement performs surprisingly well, with 90% accuracy. Though a more complicated model, if you were unable to find medical assistance for a period of time, this black box random forest model could provide insight into your potential for health disease.

The mean decrease accuracy measurements highlight variable importance supported in the basic decision tree.



Takeaways from this analysis:

Provided you have access to all of the variables, it would obviously be a much better decision to go with a full model, however, a limited model for personal use can still prove beneficial if you are willing to accept the lacking interpretability of the random forest model.

For the full model however, both logistic regression and random forest achieve the same accuracy of 97%, so choosing the simpler logistic regression model would be my personal choice as the best model option for predictive heart failure.