HEART DISEASE CLASSIFIER

Trey Hibbard

DATASET, DETAILS, AND EXPLANATION

- Data for this research comes from data.medeley.com
- Target variable: Cardiovascular Heart Disease

DATASET, DETAILS, AND EXPLANATION

- Numeric Variables:
  - Age
  - Resting Blood Pressure
  - Serum Cholesterol
  - Max Heart Rate
  - Old Peak

DATASET, DETAILS, AND EXPLANATION

- Nominal Variables:
  - Chest Pain
  - Resting electrocardiogram results
  - Slope
  - Number of Major Vessels

DATASET, DETAILS, AND EXPLANATION

- Binary Variables:
  - Gender
  - Fasting Blood Sugar
  - Exercise Angia

GOALS

- Identify a best model using all the available information.
- Cross validate the results of two separate models and consider the best option.

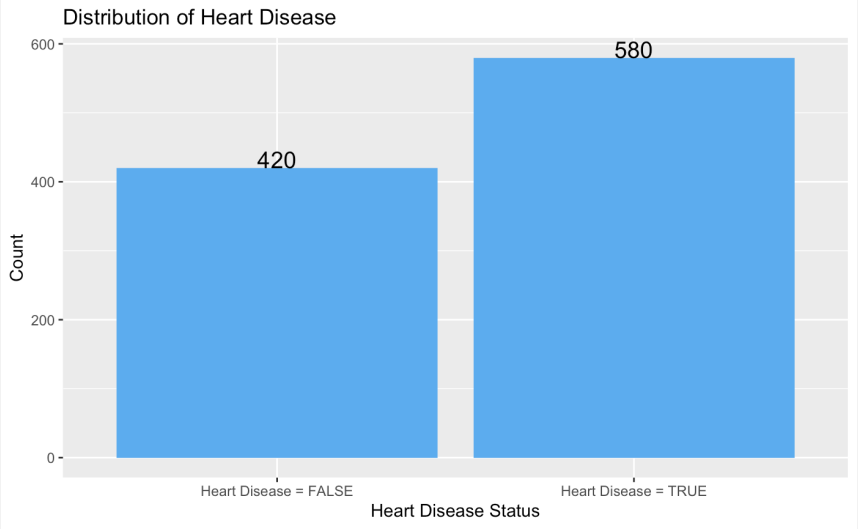
- Identify a simpler model with variables that can be investigated at home, using the same cross validation technique.

VARIABLES FROM HOME

- At home variables:
  - Age
  - Gender
  - Resting Blood Pressure
  - Chest Pain
  - Exercise Angia

EXPLORATORY DATA ANALYSIS

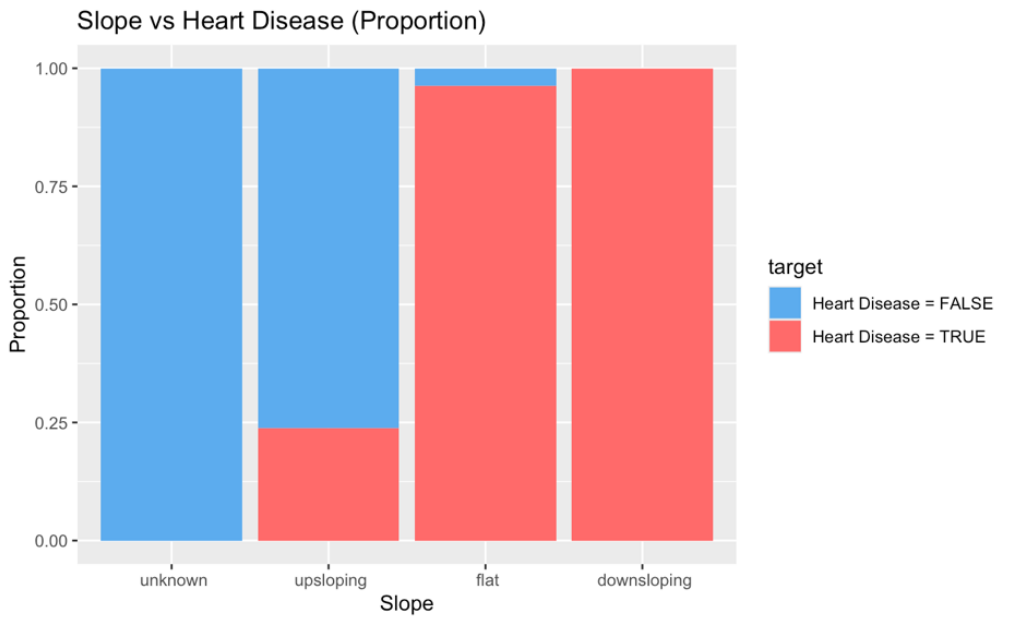
- There is a near even split of the target data



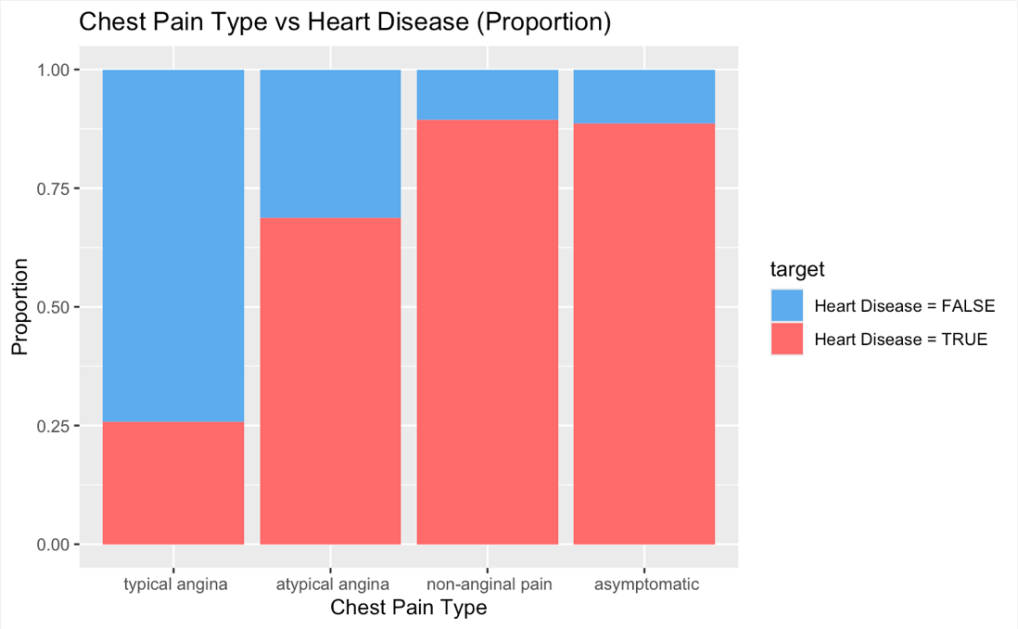
INITIAL FINDINGS / THINGS TO KEEP IN MIND

- Documentation fails to clarify slope = 0, which is present for 180/1000 patients
- The dataset is very clean and isn't missing any rows
- The numeric data has negligible correlation
- The most important features tend to be
  - Slope
  - Chest pain
  - Resting Blood Pressure

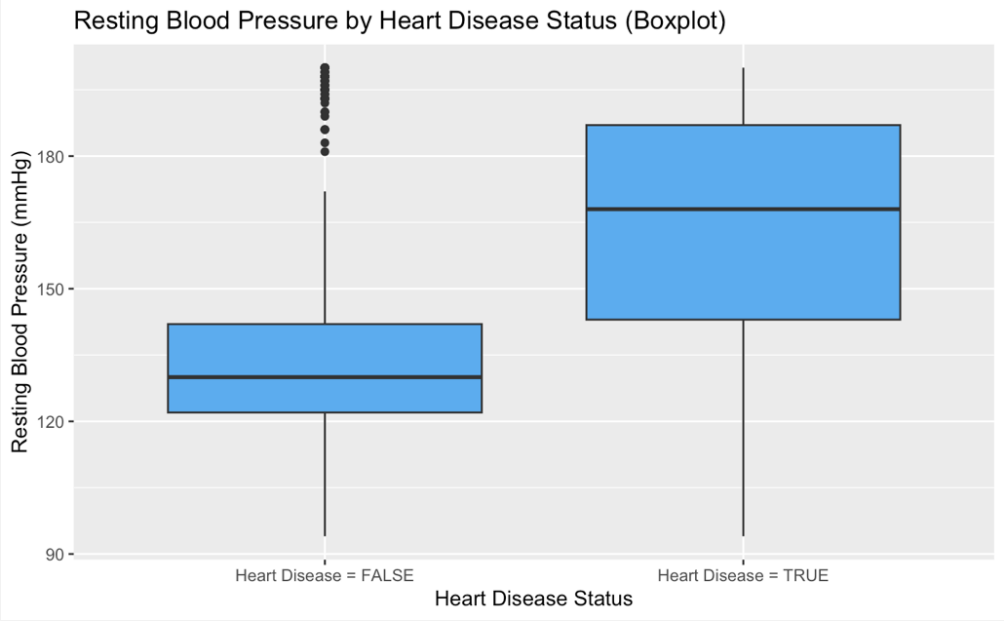
INITIAL FINDINGS / THINGS TO KEEP IN MIND



INITIAL FINDINGS / THINGS TO KEEP IN MIND

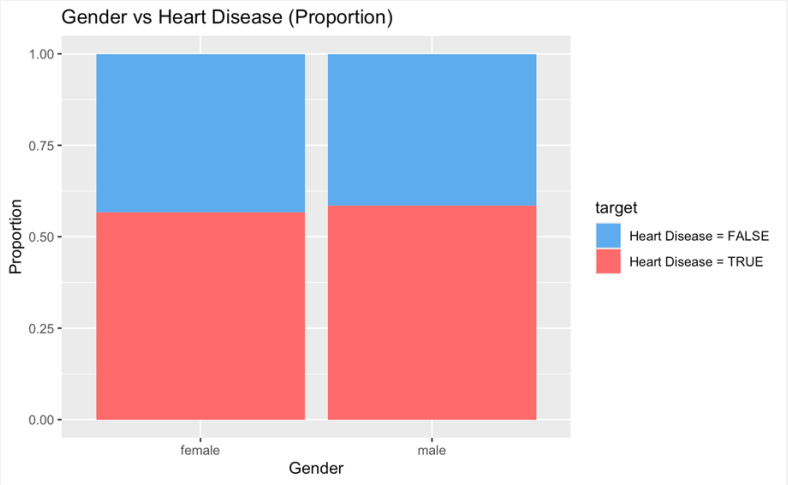


INITIAL FINDINGS / THINGS TO KEEP IN MIND

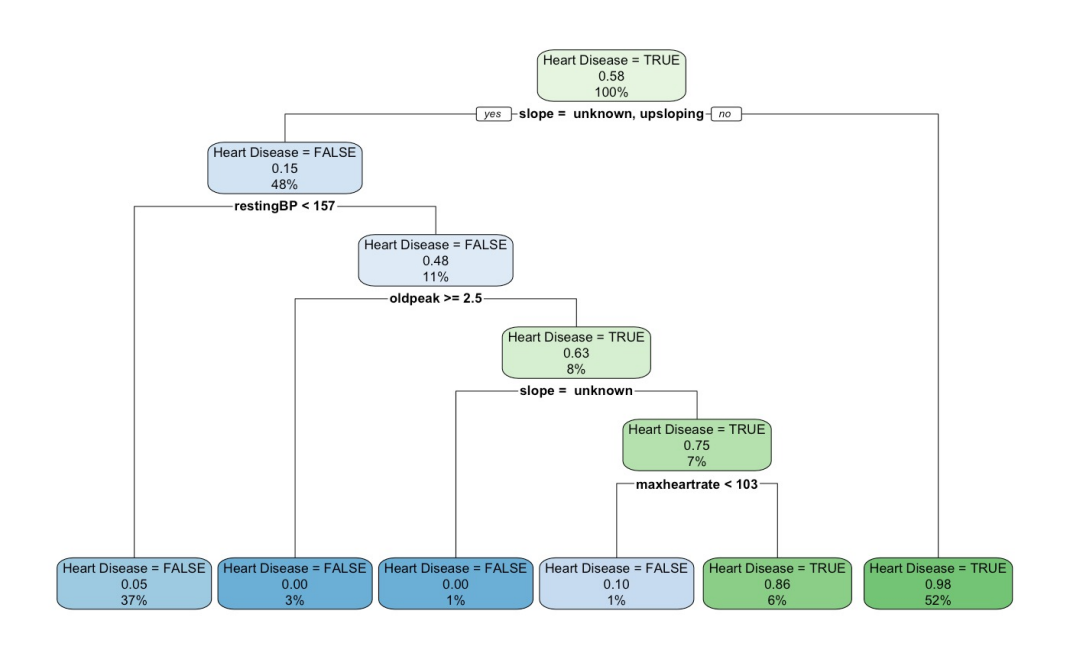


EXPLORATORY DATA ANALYSIS

- Some features such as gender do not give valuable insight to the target



BASIC DECISION TREE

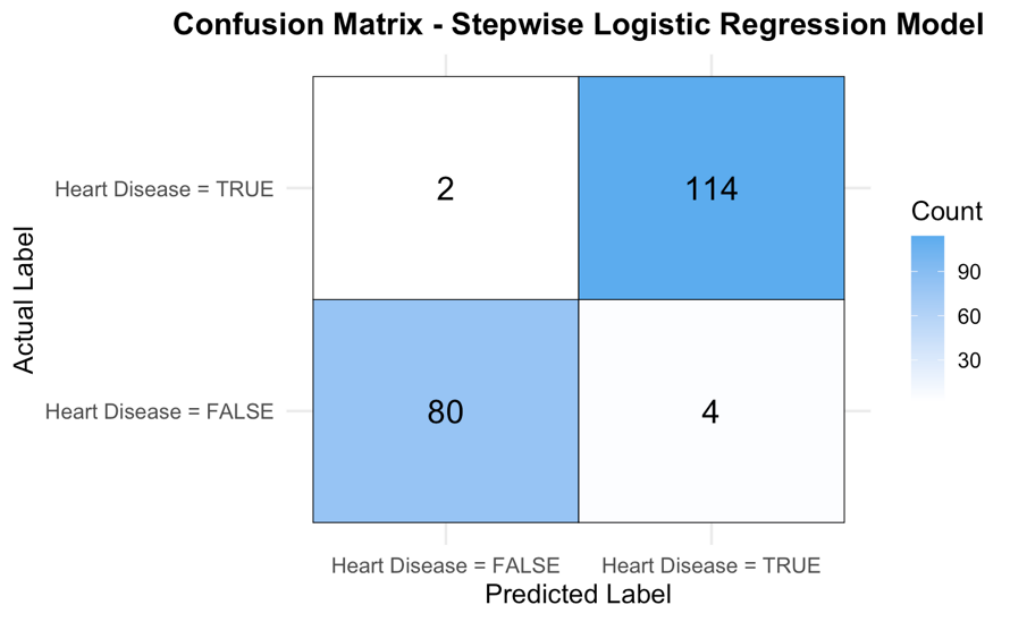


MODELS WITH ALL VARIABLES

- The best logistic regression model using AIC and forward stepwise selection
- target ~ Slope + OldPeak + RestingBP + Gender + ChestPain + RestingRElectro + FastingBloodSugar + MaxHeartRate

- Emphasis in the model relies on slope, chest pain, and resting blood pressure.
- The Logistic Regression model returned a 97% accuracy

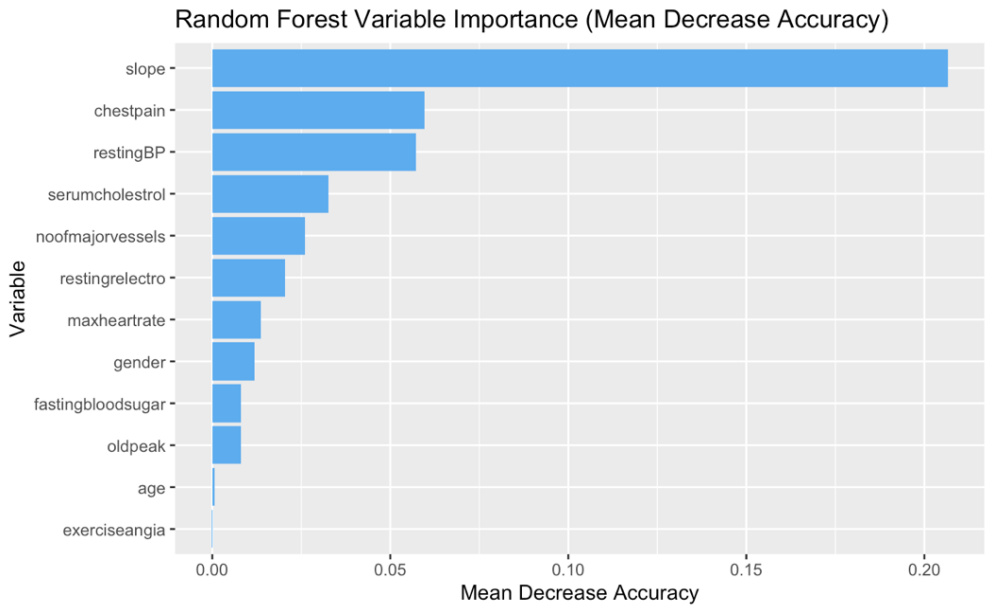
MODELS WITH ALL VARIABLES



MODELS WITH ALL VARIABLES

- Random Forest Model
- Received a similar accuracy of 97%
- Provided insights into the reliability of our logistic regression model

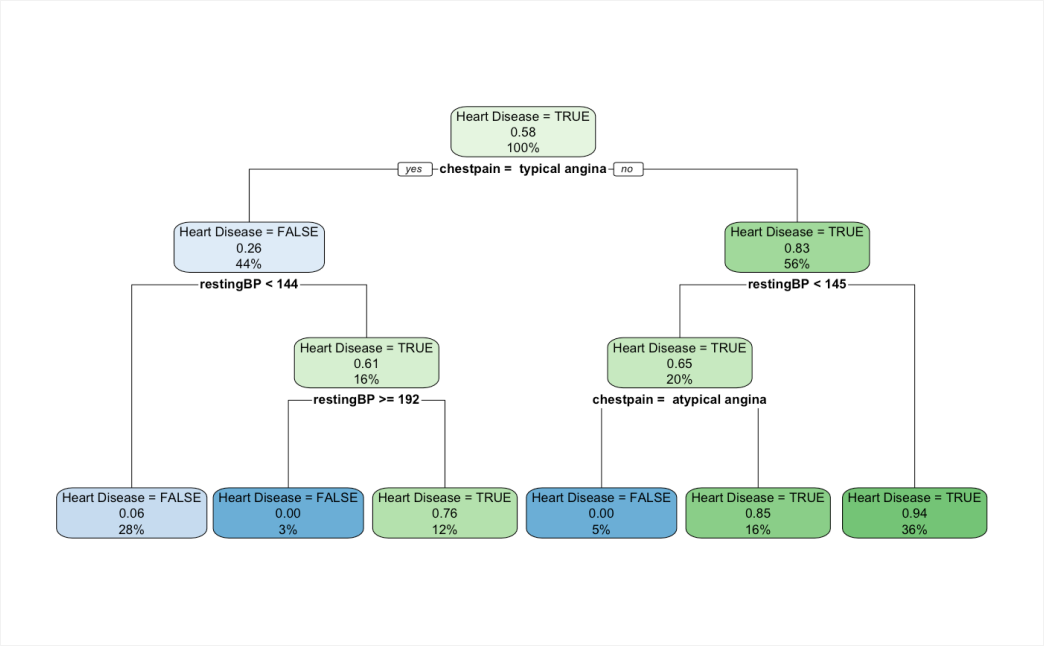
MODES WITH ALL VARIABLES



WHAT ABOUT A SIMPLER AT HOME MODEL?

- An interesting look at the affordable accuracy you can have just using measurements from home.
- Recall the available features
  - Age
  - Gender
  - Resting Blood Pressure
  - Chest Pain
  - Exercise Angia

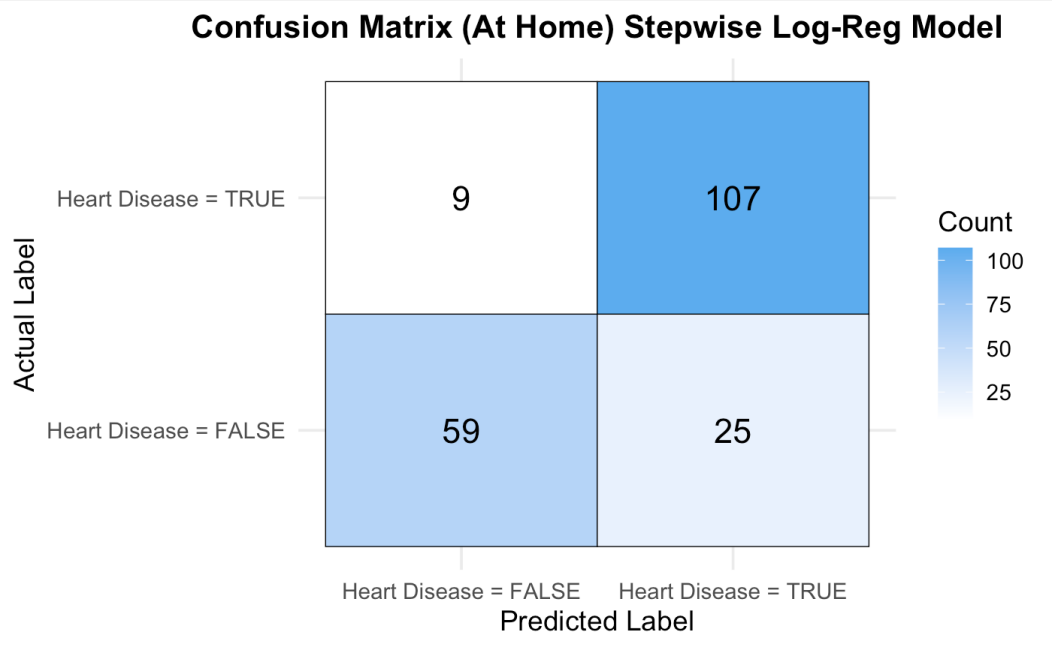
BASIC DECISION TREE (HOME)



MODELS WITH HOME VARIABLES

- The best logistic regression model using AIC and forward stepwise selection
- target ~ ChestPain + restingBP
- Predictive accuracy of only 83%.

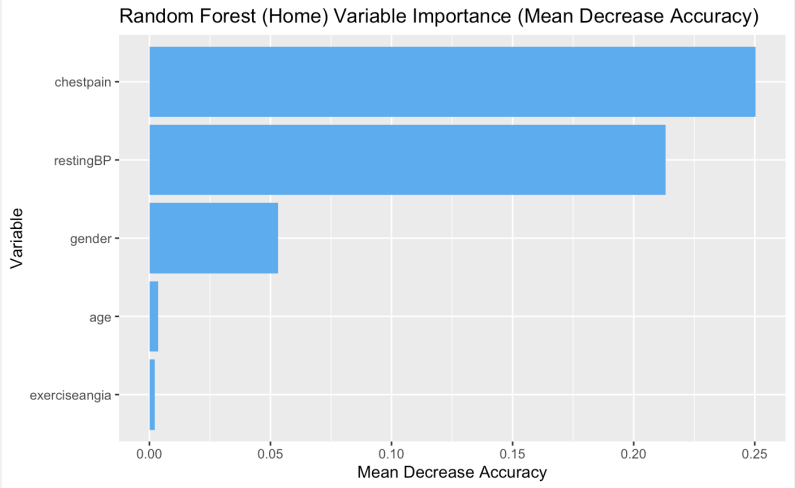
MODELS WITH HOME VARIABLES



MODELS WITH HOME VARIABLES

- Random Forest Model
- Accuracy of 90% !!

MODELS WITH HOME VARIABLES



WHAT THIS TELLS US

- We are able to achieve very high predictive accuracy using all of the info available to us.
- Minimizes type two error
- An at home model is strong, though not reliable enough

HOW WE CAN USE THIS INFORMATION

- Affordable self assessment
- Powerful medical classification tool

THANK YOU