Dorwin Liang, David Huang

Heart Disease Prediction Analysis

Agenda

1. Introduction
2. Data Pre-processing
3. Data Mining Methods
4. Model Performances
5. Findings & Actionable Steps

1 Introduction

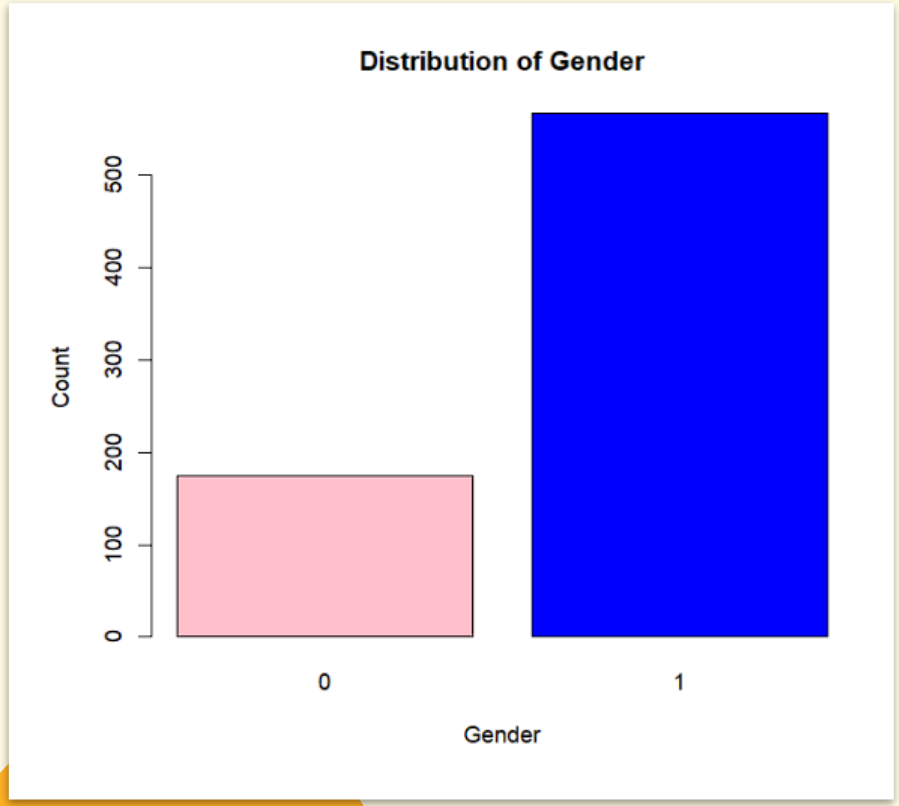
Dataset : UCI Heart Disease Dataset

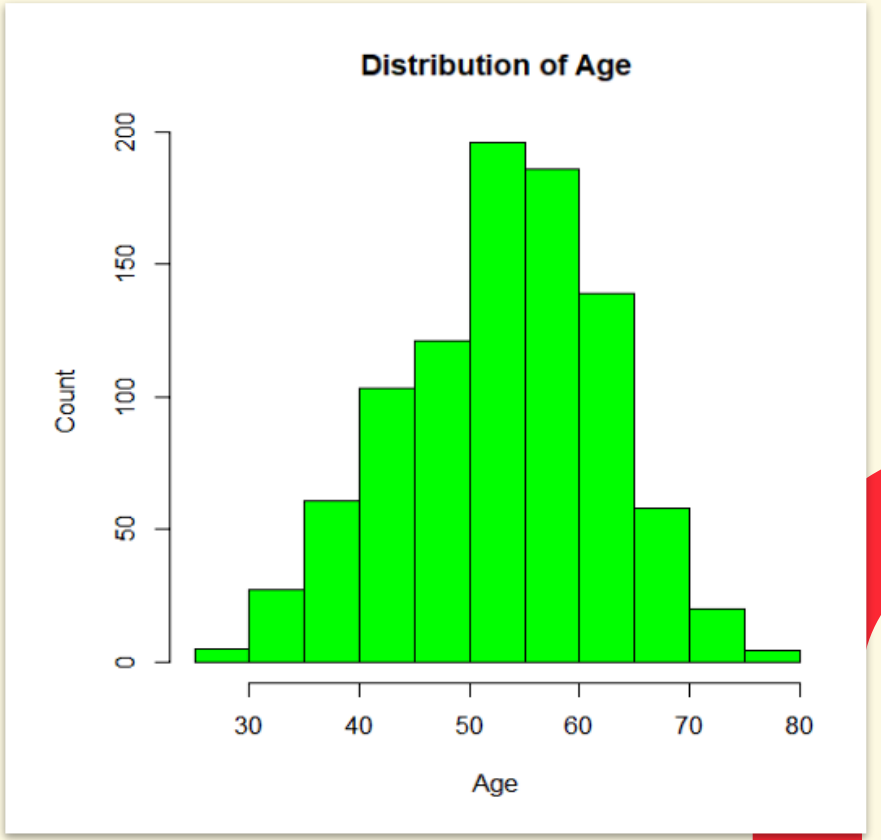
Business Objective : Provide insights that help public health officials decide how to improve health disease education

Question to answer : Which health metrics and tests correlate the most to heart disease?

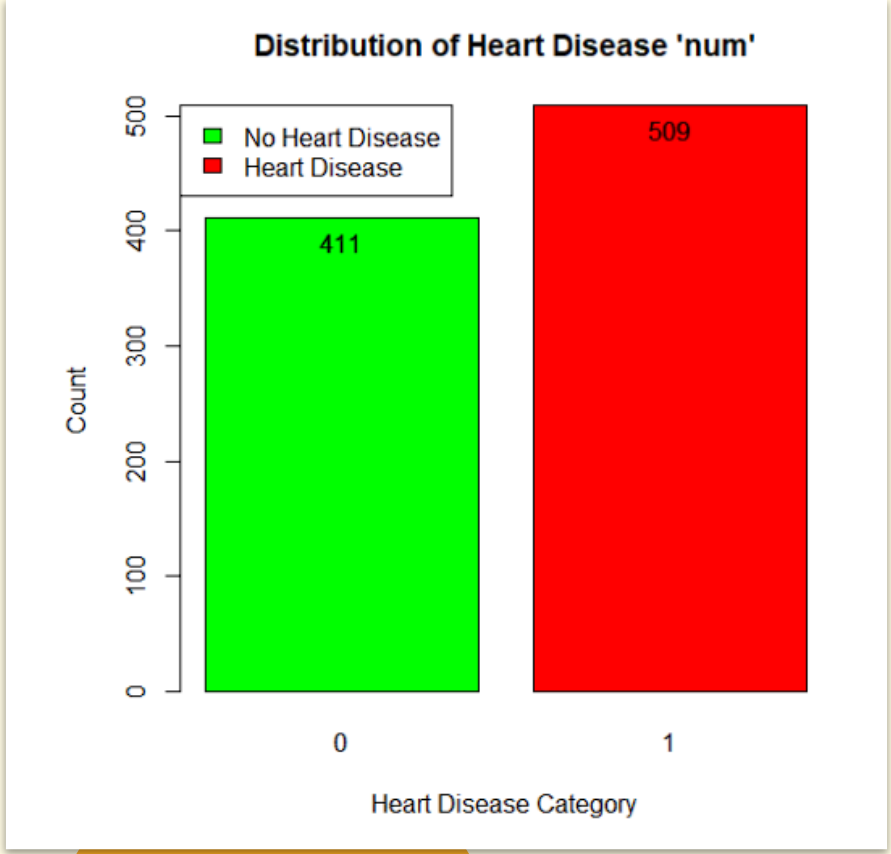
Target audience : Public Health Officials

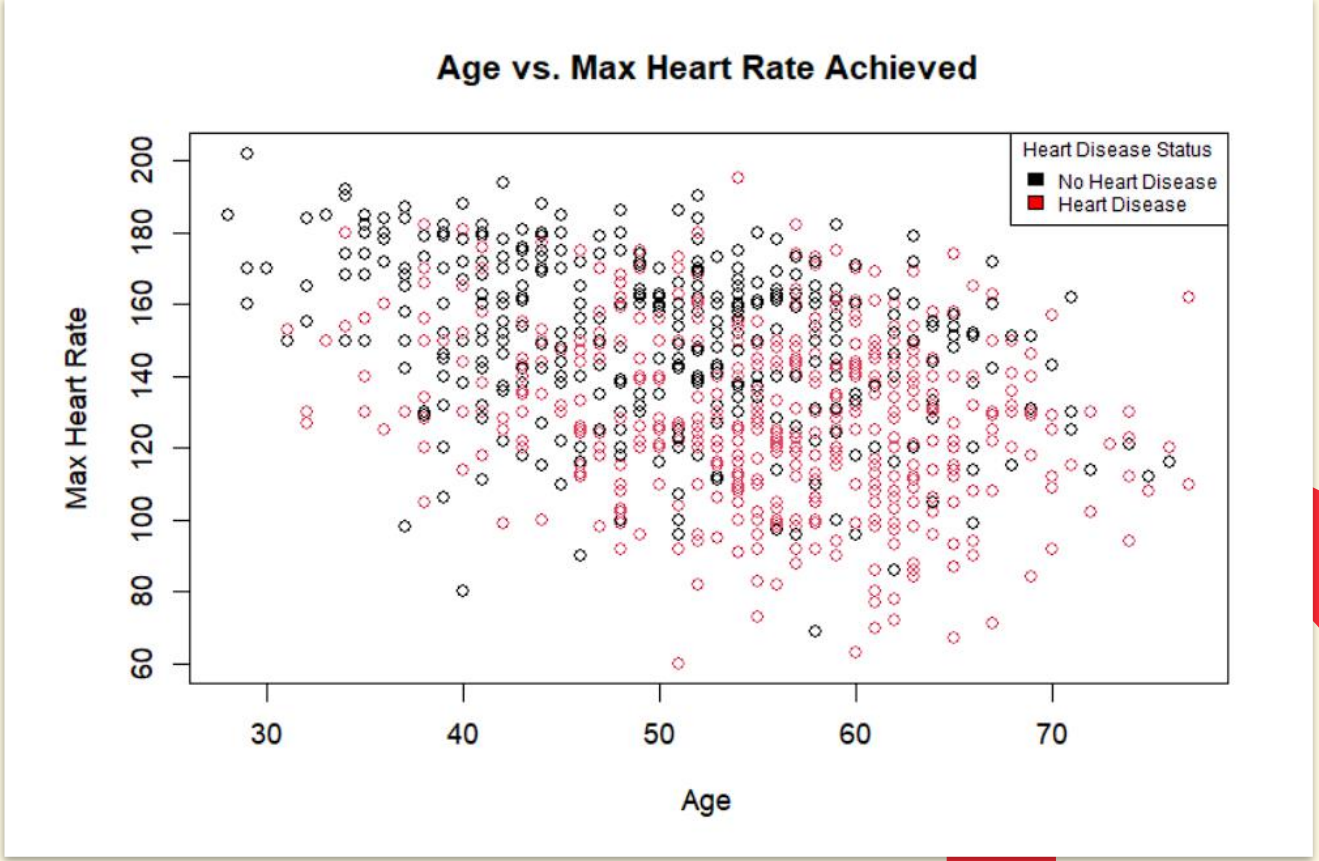
1 Data Exploration



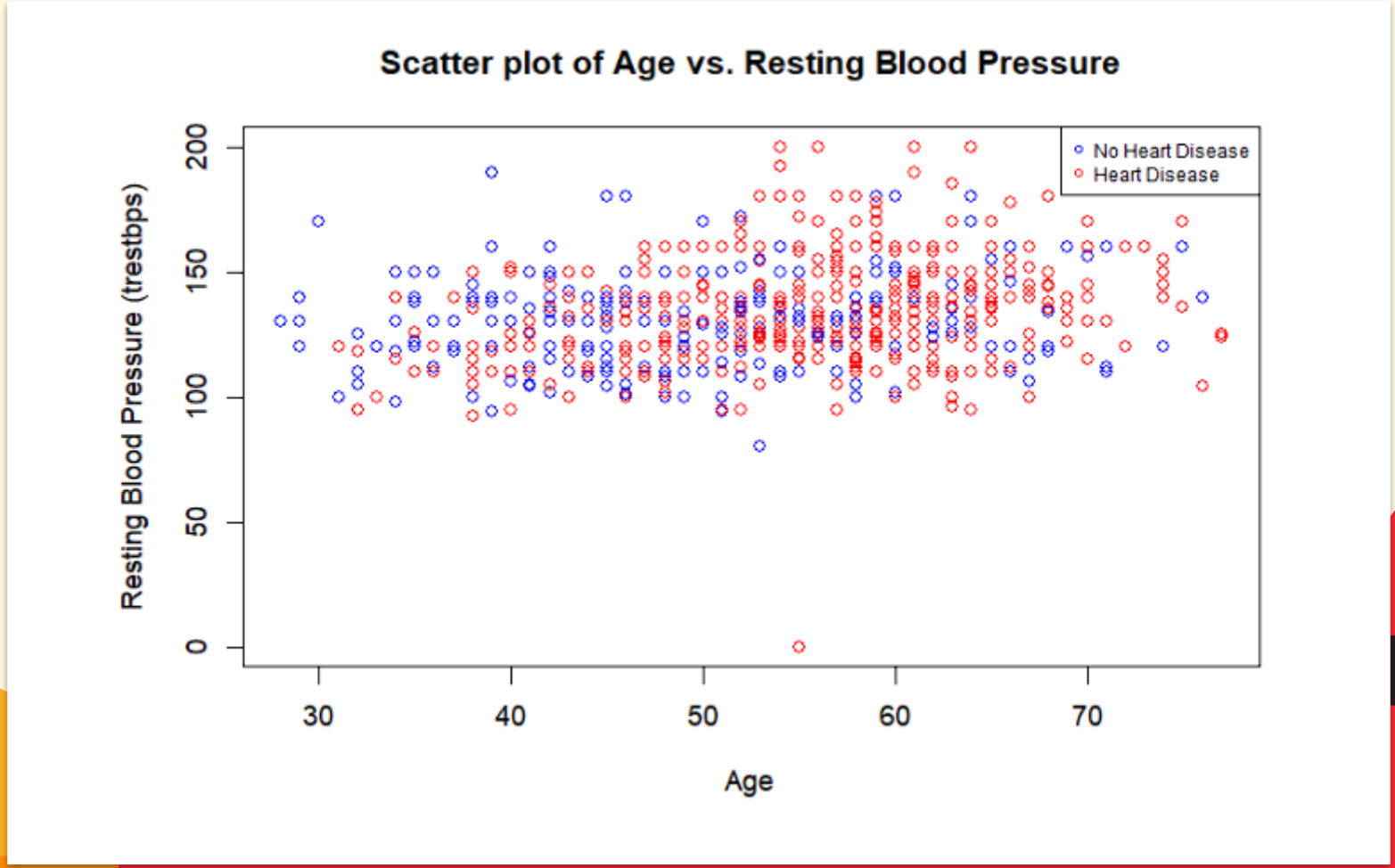


1 Data Exploration





1 Data Exploration



2 Data Pre-processing

- In the beginning, we converted our target variable into a binary column for logistic regression

- To prepare for our modeling, we initially excluded three columns due to an excessive of missing data
  - ‘ca’ ➔ number of major vessels
  - ‘thal’ ➔ heart defects
  - ‘slope’ ➔ slope of the peak exercise ST segment

- Removed rows with missing values

2 Data Pre-processing

- Converted categorical variables into separate binary columns
  - 0 or 1 to indicate its presence or not
  - Utilized a R library ➔ fastDummies

Cp
Cp_asymptomatic
cp_atypical angina
cp_non-anginal
Cp_asymptomatic
Cp_non-anginal

→

Cp_asymptomatic	cp_atypical angina	cp_non-anginal
0	0	1
1	0	0
0	1	0
1	0	0
0	0	1

2 Data Pre-processing

- Got rid of multicollinearity issue
  - This is when multiple variables are extremely correlated with one another, affecting the accuracy of the models.

- Converted the rest of the categorical variables into binary format for the models to interpret.
  - ‘sex’ , ‘fbs’, ‘exang’

- Last step was to scale the numerical variables to standardize them.
  - Uses the built-in Scale() function

3 Data Mining Methods

Logistic Regression Model

- **Why we chose this** : Directly correlated with the question we wanted to answer.
- This model would help us find which variables greatly contribute to the presence of heart disease.
- Its interpretability allows for our target audience, public health officials, to understand the impact of each variable.
- **Pre-requisite steps** : one-hot encoding

3 Data Mining Methods

KNN Model

- **Why we chose this** : Wanted a model that had a classification approach and dealt with the data's proximity to one another
- Specializes in predicting patterns and clusters in the dataset
- We wanted to have a model with a classification approach
- **Pre-requisite steps** : scaling to standardize numerical columns.

3 Data Mining Methods

Naïve bayes Model

- **Why we chose this** : Gives a different perspective on the dataset by showing how related the variables are to one another.
- **Pre-requisite steps** : import necessary packages for Naïve Bayes

4 Model Performance

	Average	Last Fold		
Model Type	5-fold Cross-validation Accuracy	Recall	Precision	Accuracy
Logistic Regression	0.8077	0.8169	0.7733	0.8101
KNN	0.9100	0.8947	0.9067	0.9051
Naïve Bayes	0.7854	0.8636	0.7600	0.8291

***Figure 1**: Table of the average 5-fold cross-validation accuracy as well as the recall, precision, and accuracy of the last fold for each of the three models.*

5 Findings & Actionable Steps

Best Model to use for our question is

Logistic Regression Model

- **Directly answers the question we are looking for.**
- **Summary allows us to identify statistically significant variables and their correlations to presence of heart disease**

5 Findings & Actionable Steps

- **Summary showed 4 statistically significant variables**

Variables	Estimate	Std. Error	Z value	Pr(>|z|)
Sex (gender)	1.3767	0.26861	5.125	2.97E-07
Exang	0.9898	0.23771	4.164	3.13E-05
Oldpeak	0.78088	0.12913	6.047	1.47E-09
Cp_asymptomatic	1.2814	0.44356	2.877	0.004012

***Figure 2**: The summary output of our logistic regression model which shows four statistically significant variables with positive correlations to presence of heart disease*

Exercise-induced chest pain

ST Depression induced by exercise

Asymptomatic Chest Pain

5 Findings & Actionable Steps

- **Summary showed 4 statistically significant variables**

Variables	Estimate	Std. Error	Z value	Pr(>|z|)
Sex (gender)	1.3767	0.26861	5.125	2.97E-07
Exang	0.9898	0.23771	4.164	3.13E-05
Oldpeak	0.78088	0.12913	6.047	1.47E-09
Cp_asymptomatic	1.2814	0.44356	2.877	0.004012

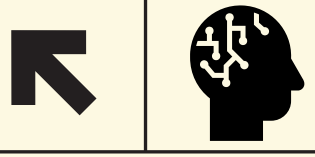
***Figure 2**: The summary output of our logistic regression model which shows four statistically significant variables with positive correlations to presence of heart disease*

5 Findings & Actionable Steps

Actionable Steps for public health officials

- ✓ Given that men are more likely to develop heart disease and exercise-induced chest pain correlates highly with the presence of heart disease…
  - ✓ **Solution** : Create targeted ad campaigns for men who have chest pain while exercising and persuade them to get a heart health screening

- ✓ Given that the presence of heart disease also prevalent in those with asymptomatic symptoms such as chest pain with no symptoms and ST depression induced by exercise, which is only discoverable with an ACG scan…
  - ✓ **Solution** : Make another ad campaign focusing on the entire population, focusing on addressing misconceptions about heart disease, emphasizing that early signs of heart disease can be asymptomatic.

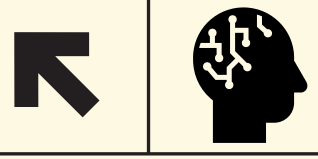


Conclusion

- Best performing model was KNN, but Logistic Regression directly gave us the answer to our question
  - Average Cross-validation Accuracy
    1. KNN ➔ 0.9100
    2. Logistic Regression ➔ 0.8077
    3. Naïve Bayes ➔ 0.7854

- Logistic Regression Model shows four statistically significant variables
  1. Gender
  2. Exercise-induced chest pain
  3. ST Depression induced by exercise
  4. Asymptomatic chest pain

- **Solution** : Public health officials should use the four variables to target and inform people of high risk of heart disease



Limitations

- The dataset we used has a limited number of variables.
  - variables that have high correlation to the presence of heart disease that were not in the dataset could exist.

- After data cleaning and processing, we removed a total of 179 rows which accounted for roughly 19.4% of the data.
  - Left with 741 rows of data
  - Not much data to work with given something as importance as presence of heart disease where precision and accuracy are important, and the size of a dataset used to train models greatly affect those measures.

Thank you!

Dorwin Liang

David Huang