Capstone Project

What factors are most important in determining whether a patient is likely to suffer from a heart attack?

Omar Imambaccus
August, 2018

Outline

- Problem Identification
- Understanding the dataset
- Data Wrangling Process
- What story does the data tell us?
- Inferential Statistics
- Supervised Machine Learning
- Limitations and Recommendations

Problem Identification

Problem Identification

Problem: Misallocation of resources in the Canadian health sector. Not being able to cater to cardiac patients in need of urgent care.

Client: Medical professionals in the health sector

Task: Identifying what health related factors are the strongest predictors of whether a patient is likely to have a heart attack

Data: UCI Machine Learning Repository/Heart Disease Dataset

Understanding the Dataset

A look at the dataset

Independent Variables

Age of the patient

Gender of the patient

Chest Pain Type

Resting Blood Pressure

Serum Cholesterol Levels

Fasting Blood Sugar

Resting Electrocardiographic Results

Maximum Heartrate Achieved (under stress)

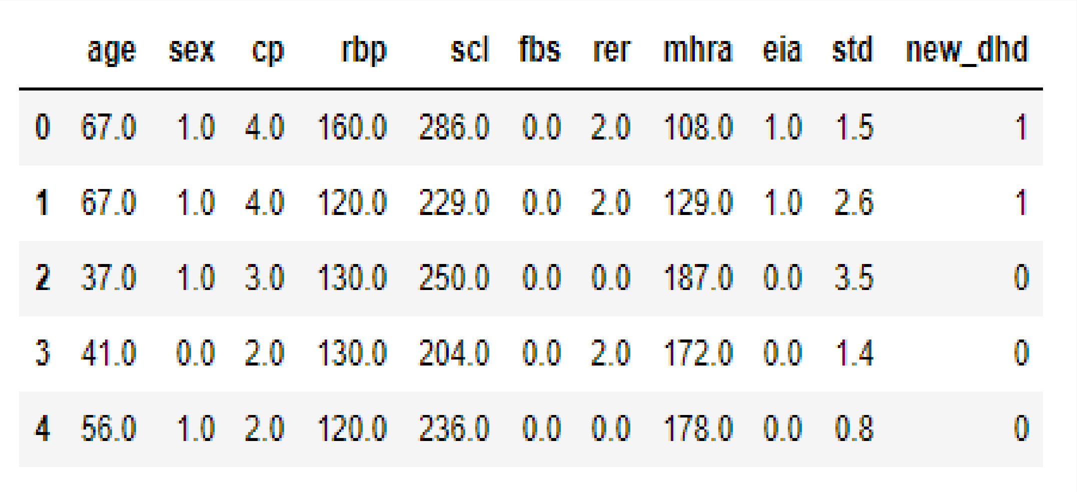
Exercise Induced Angina

ST Depression

Target (Dependent) Variable

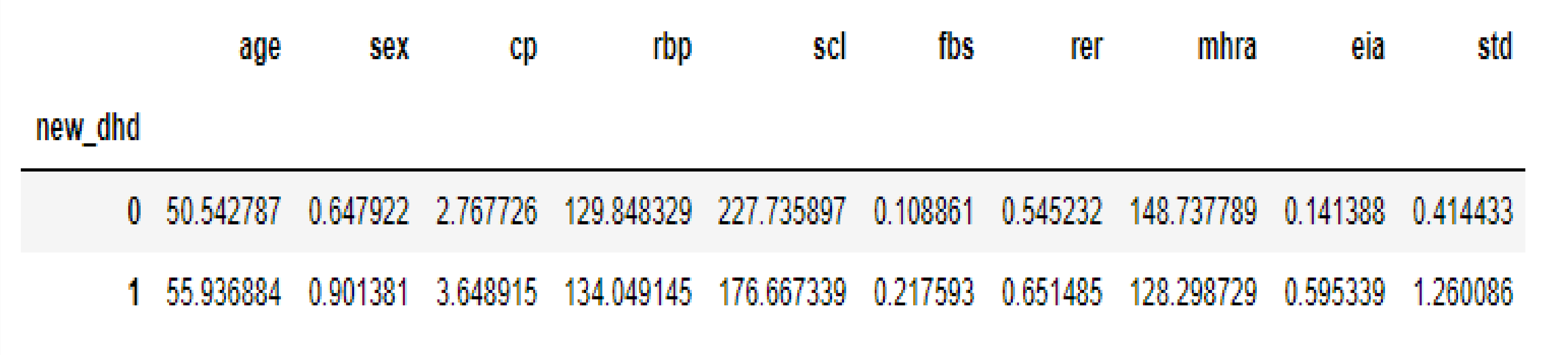
Greater or less than 50% of major vessel blocked [0,1]

Greater than 50% means patient is more likely to suffer from a heart attack.



	age	sex	cp	rbp	scl	fbs	rer	mhra	eia	std	new_dhd
0	67.0	1.0	4.0	160.0	286.0	0.0	2.0	108.0	1.0	1.5	1
1	67.0	1.0	4.0	120.0	229.0	0.0	2.0	129.0	1.0	2.6	1
2	37.0	1.0	3.0	130.0	250.0	0.0	0.0	187.0	0.0	3.5	0
3	41.0	0.0	2.0	130.0	204.0	0.0	2.0	172.0	0.0	1.4	0
4	56.0	1.0	2.0	120.0	236.0	0.0	0.0	178.0	0.0	0.8	0

Grouping the dataset by target variable



	age	sex	cp	rbp	scl	fbs	rer	mhra	eia	std
new_dhd										
0	50.542787	0.647922	2.767726	129.848329	227.735897	0.108861	0.545232	148.737789	0.141388	0.414433
1	55.936884	0.901381	3.648915	134.049145	176.667339	0.217593	0.651485	128.298729	0.595339	1.260086

In the table above, we published the mean of every variable after grouping the dataset by the target variable.

Data Wrangling Process

Data Wrangling

Dataset Description

- 4 processed datasets from Cleveland, Long Beach, Hungary and Switzerland
- Dataset is cross-sectional from 1988
- 15 explanatory variables with 916 observations in total

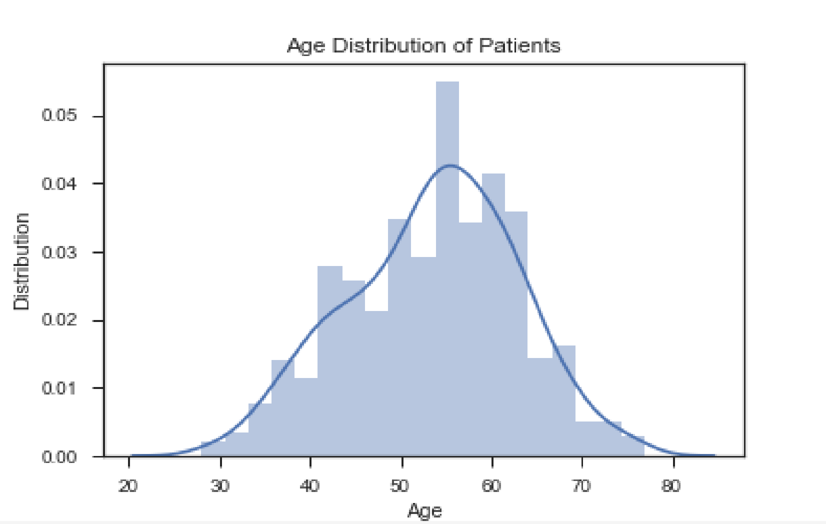
Dealing with missing (NaN) values

- Missing values denoted by '-9' or '?' entries
- Did not delete rows with NaN values initially for exploratory analysis
- Will need to delete rows for supervised learning analysis

Dropping Variables

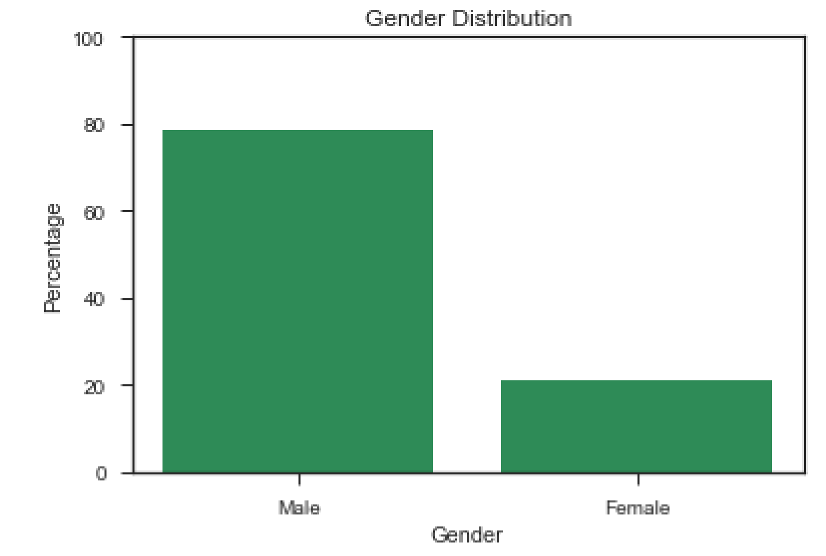
- Variables that had 50% or more of the data missing were dropped
- These variables do not add value to analysis
- Dataset now has 11 explanatory variables

What story does the data tell us?



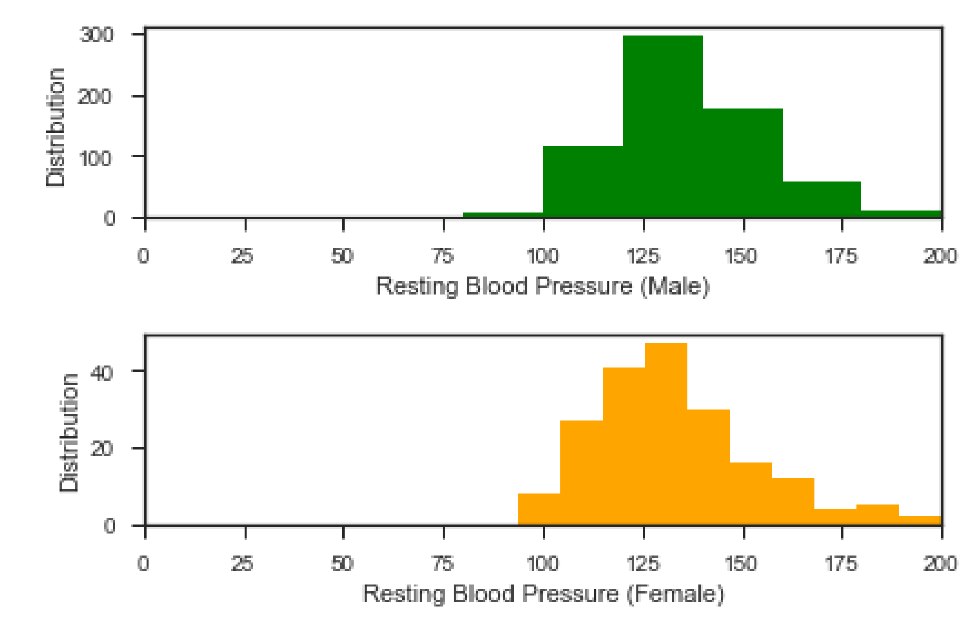
Patient Age Distribution

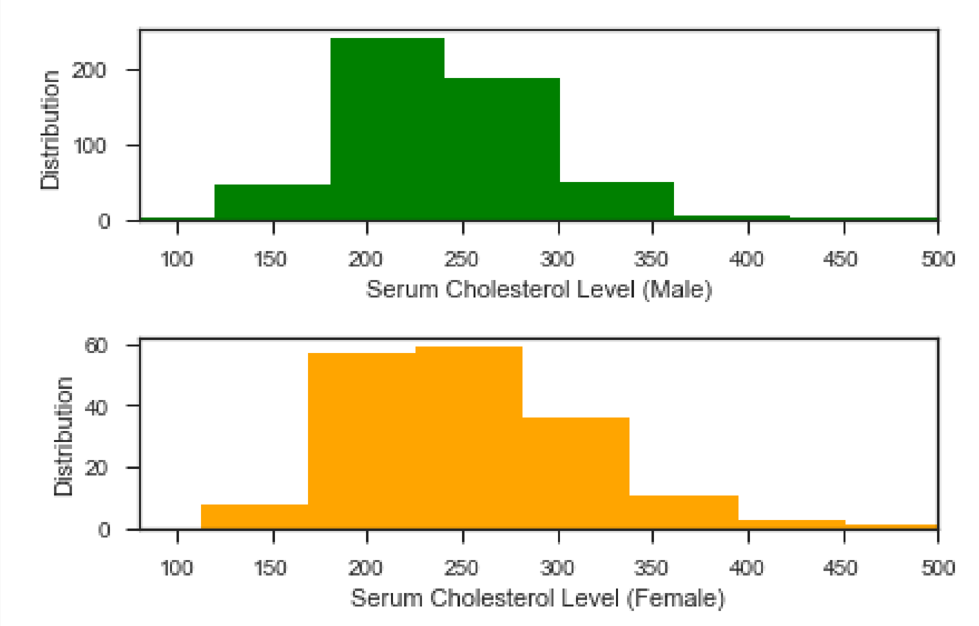
Most of patients are between the ages of 55-60 years old.



Gender Distribution

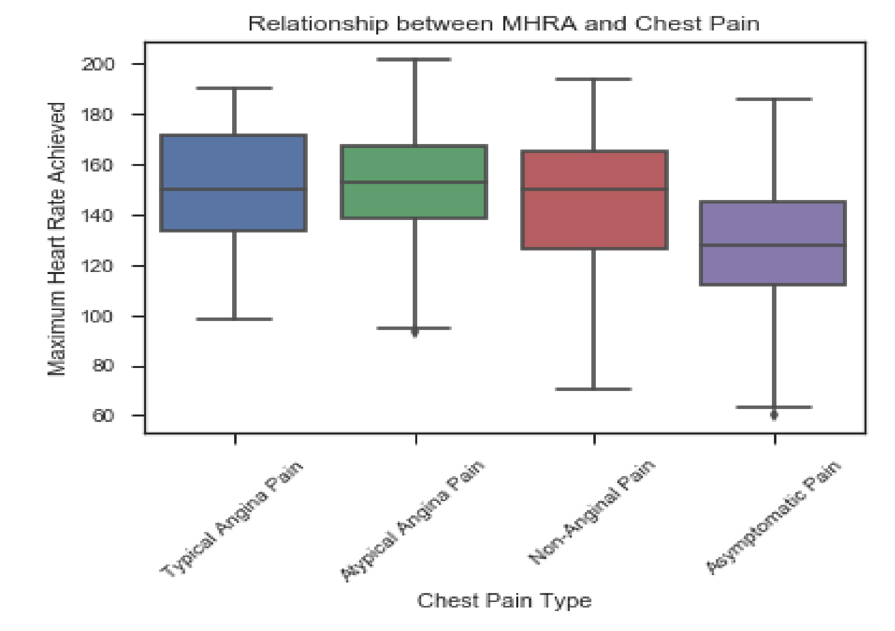
80% of patients in our analysis are males.

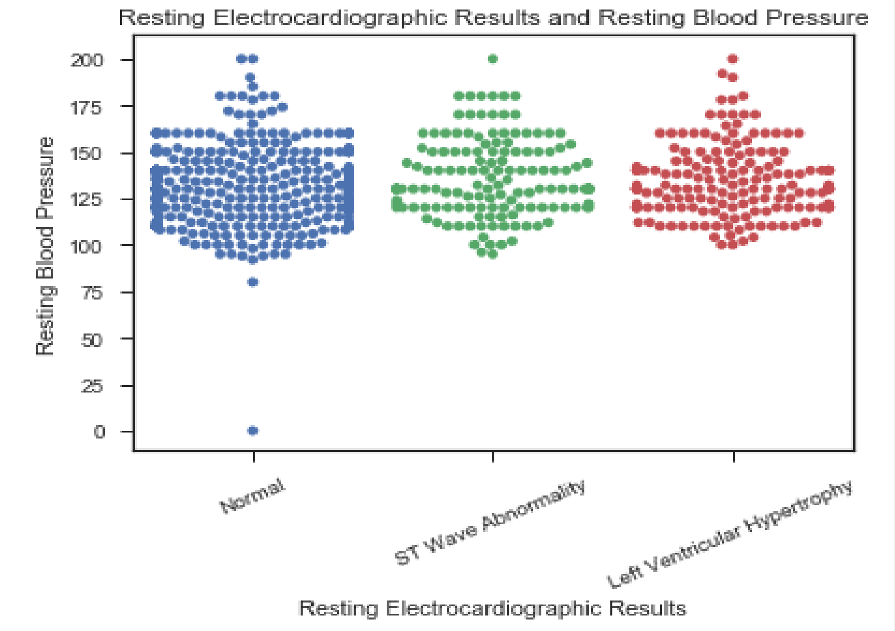




Serum Cholesterol Levels and Resting Blood Pressure | Gender Analysis

On average, male and female patients have the same average cholesterol levels and resting blood pressure.



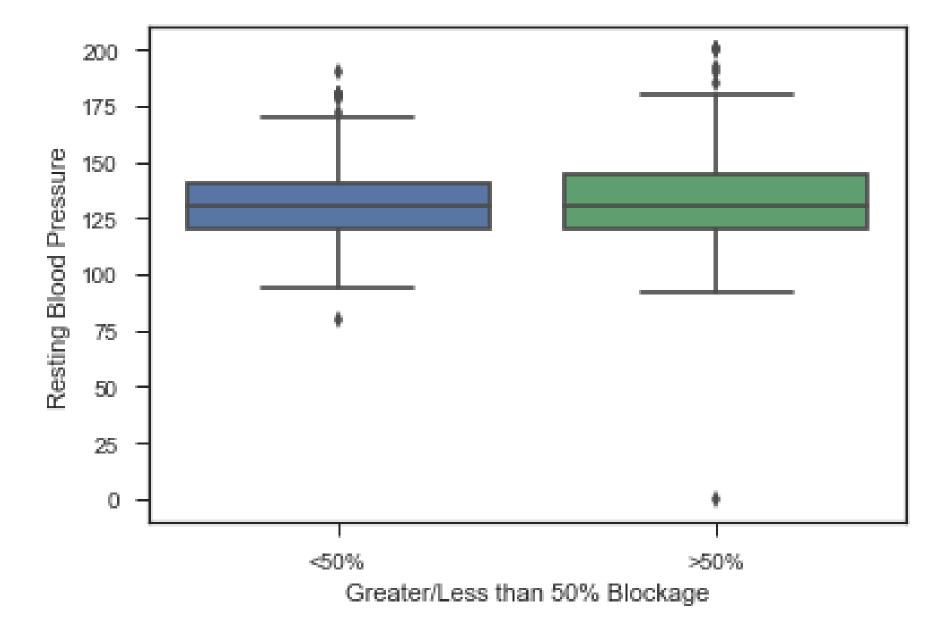


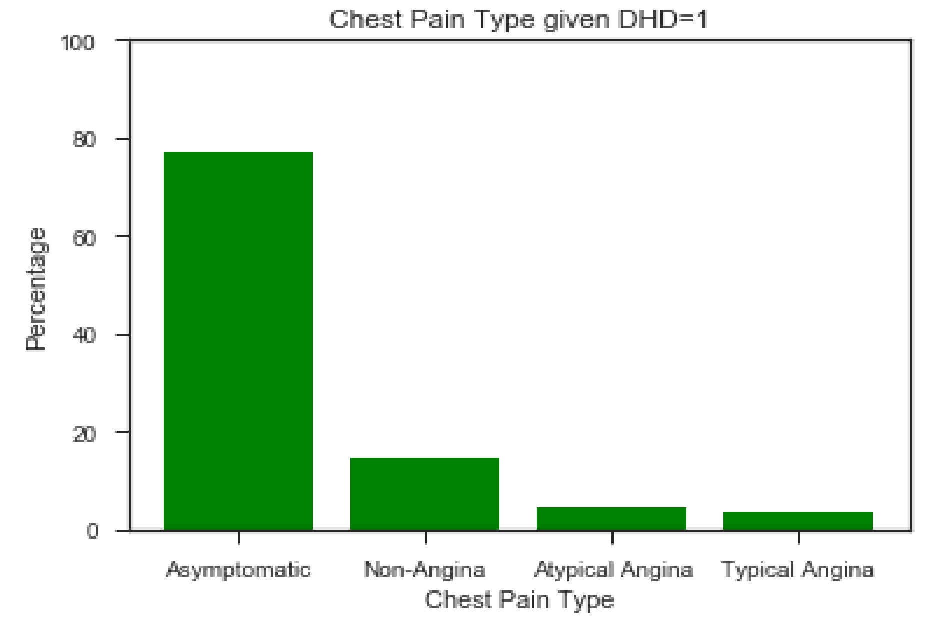
Chest Pain and Maximum Heart Rate

Patients that exhibited asymptomatic chest pains were not able to reach high maximum heart rate levels when put under stress.

RER and Resting Blood Pressure

Most patients had normal RER readings, while the height of each plot shows that the range of resting blood pressure is roughly the same across all patients.

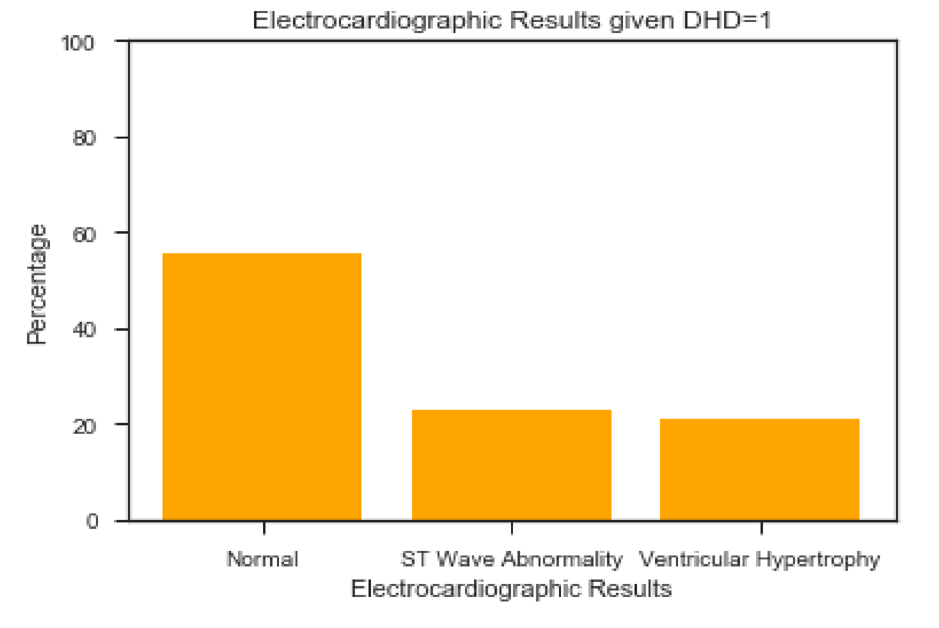


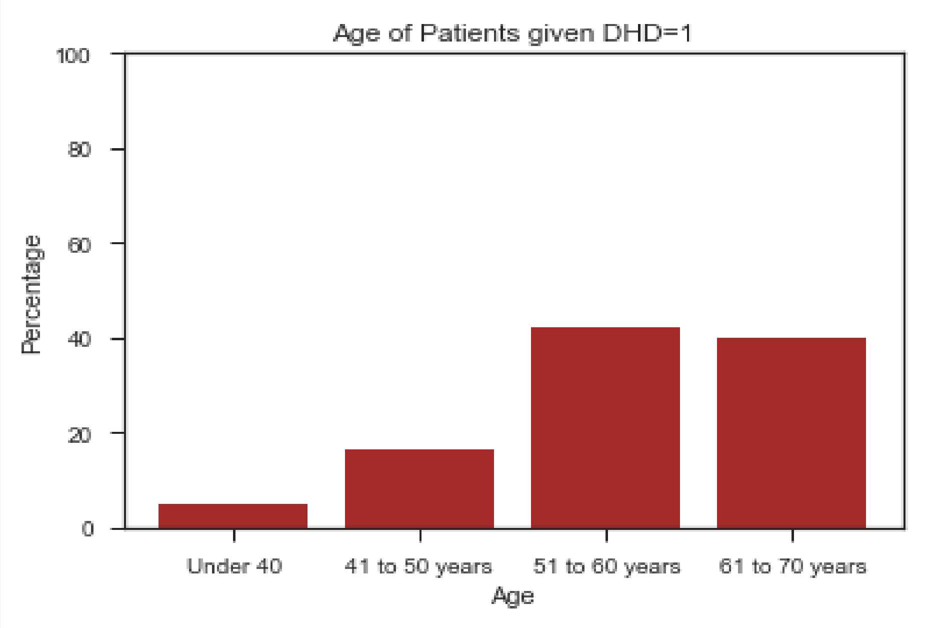


Analysis of Target Variable: Predicting whether a patient will have a heart attack (DHD)

Patients with more than 50% or less than 50% of their major vessels blocked had the same average resting blood pressure.

Patients that had more than 50% of their vessels blocked mostly suffered from asymptomatic chest pains.





Most patients with major blockage in their vessels had normal electrocardiographic readings.

Patients that had more than 50% of their vessels blocked were mostly between the ages of 51-70 years old.

Inferential Statistics

- In this section, we tested the null hypothesis that the mean of two different sample populations were the same.
- We did this to identify whether certain variables add value to our analysis and draw conclusions about their respective means. Below are some examples:

High Resting Blood Pressure v Low Resting Blood Pressure

- Firstly, we separated the dataset into two sample groups, those with high and low blood pressures.
- We tested whether the null hypothesis was true: the mean that they would have a major blockage/minor blockage was the same across both samples
- We reject this null hypothesis – the means of both samples are different

test statistic	p-value
3.43	0.0006

Normal v Abnormal Resting Electrocardiographic Results

- Firstly, we separated the dataset into two sample groups, those with normal and abnormal RER readings
- We tested whether the null hypothesis was true: the mean that they would have a major blockage/minor blockage was the same across both samples
- We reject this null hypothesis – the means of both samples are different

test statistic	p-value
-2.98	0.002

Supervised Machine Learning

Model: Logistic Regression Model

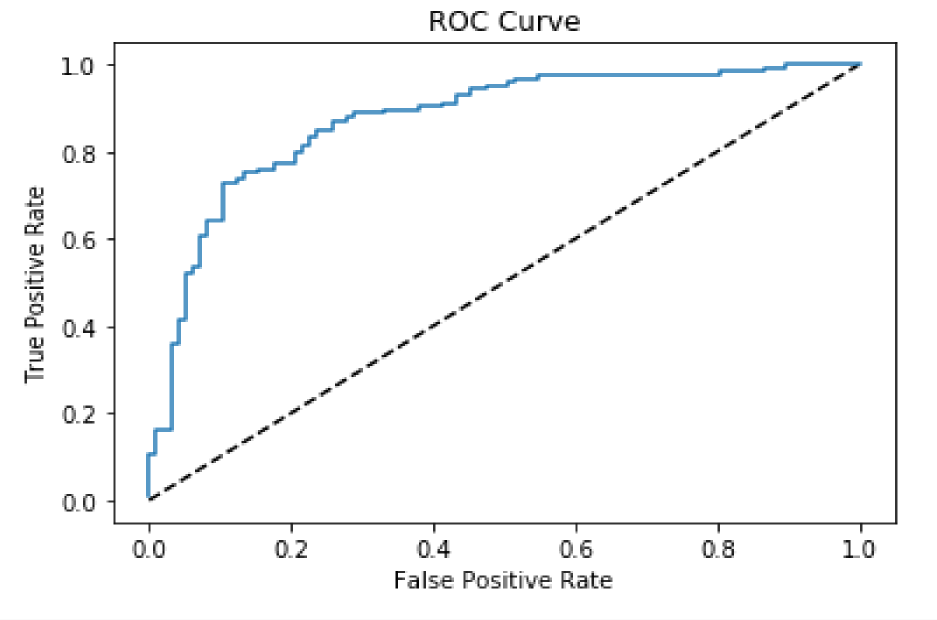
We began our analysis by using a logistic regression model to predict our results. Below are the coefficients of our analysis:

Variable	Coefficient	Standard Error	Odds Ratio
Age	0.003	9.394	1.002
Sex	1.108	0.425	3.023
Chest Pain	0.542	0.935	1.720
Resting BP	0.003	18.597	1.003
Serum Cholesterol	-0.002	93.694	0.998
Fasting Blood Sugar	0.404	0.356	1.499
Resting Electrocardiographic Results	0.277	0.839	1.32
Maximum Heart Rate	-0.02	25.830	0.981
Exercise Induced Angina	1.2	0.49	3.312
ST Depression	0.701	1.08	2.018

- Coefficients of a logistic regression are hard to interpret
- For interpretation we transform coefficients into odds ratio
- Features with the most impact are ones with the **highest** odds ratio and **smallest** standard error
- In this case, we see the standout features are gender, chest pain, fasting blood sugar, resting electrocardiographic results and exercise induced angina

Model: Logistic Regression Model Evaluation

We evaluate the competency of our model using several approaches. For this presentation, we will focus on a confusion matrix and an ROC curve.



True Positive

77	20
26	**99**

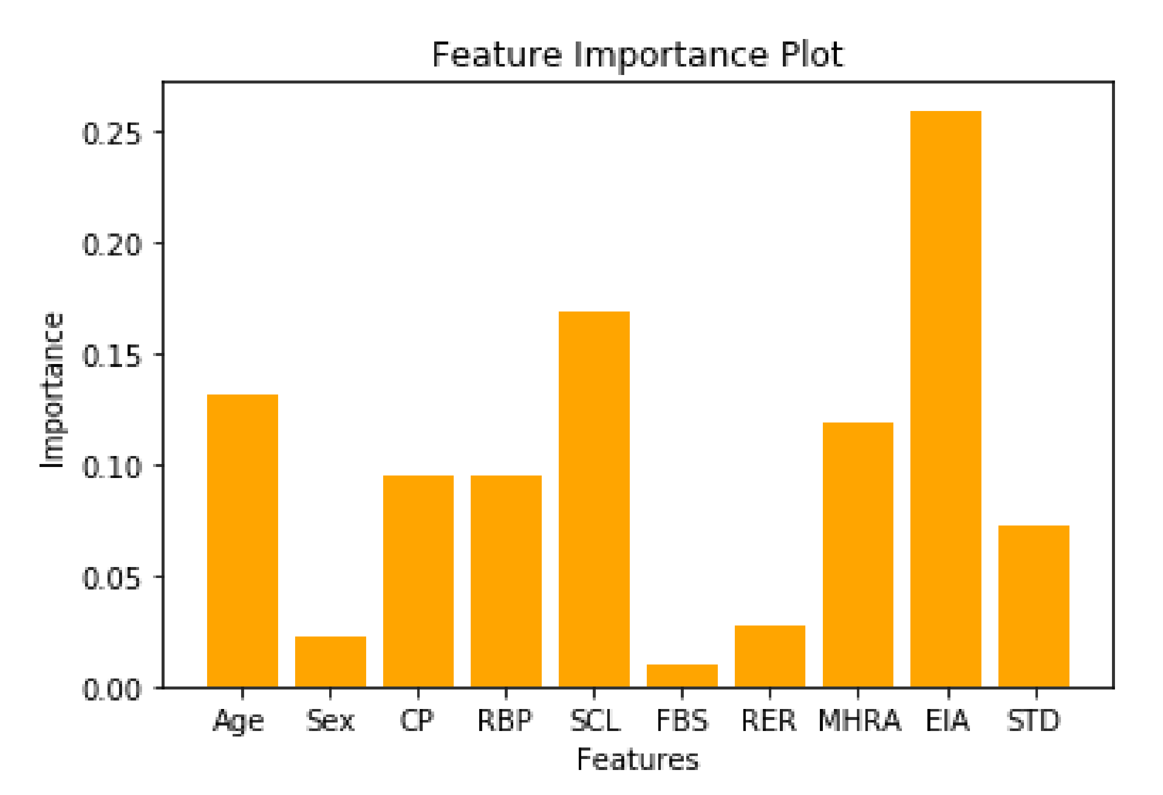
True Negative

- The greater the area under our ROC curve, the more favourable is our model.
- Based on the latter, our model faired well.

- In a confusion matrix, we want to focus on the True Positive Rate and True Negative Rate (the whole numbers are in bold above)
- We see that the TPR and TNR are 79%, which is favourable

Model: Random Forest Model

To compare our analysis, we used a stronger and more stable model. The coefficients, in terms of feature importance, from this model are below:



- Unlike the previous model, a random forest is able to compute each individual feature importance
- From the plot, we see that there are three major stand out features
- These features are age, serum cholesterol levels and exercise induced angina

Model: Random Forest Model Evaluation

We evaluate the competency of our random forest model by studying the **mean squared error (MSE)** and the **R Score** on the training and test sets.

R Score on Training Set	R Score on Test Set
0.88	0.3

- Our model performed well on the training set but lacked quality in predicting the test set

MSE on the Training Set	MSE on the Test Set
0.172	0.03

- The MSE was low on both training and test set
- MSE decreases on the test set
- Our predictions are more stable

Limitations and Recommendations

Limitations

- Lack of observations means a lack of variety in analysis and therefore lower confidence in predictions
- Many missing values from the dataset
- Dataset is 30 years old, lifestyle choices, nutrition and the medical sector has changed since then
- Male dominant dataset – analysis might not be applicable to an average female patient

Recommendations

- Add more relevant features (variables) to the model
- Add more observations and complete patient rows to the analysis
- Transform study into a longitudinal one to capture trends and changes over time
- Collect more recent data for the analysis