Heart Attack Prediction Model

Springboard Data Science Intensive Capstone Project

Prepared by Andrew Henry

Table Of Contents

Problem Overview

- *It has been stated that misdiagnosis is more common than drug errors or wrong-site surgery. The problem at hand is the classification of patients with conditions (chest pains, high blood pressure, abnormal resting heart rate etc) likelihood/risk of suffering a heart attack or other heart disease like conditions. The incorrect assessment of a patient that is not likely to suffer some heart disease like conditions can be time consuming and costly for patients, medical units, and insurance companies and the incorrect assessment of a patient that is likely to have some serious heart complications can be life threatening. When it comes to situations that involve potentially life threatening conditions it would be in one's best interest to be as accurate as possible.*

A research study done on patients in England and Wales states that almost a third of the people are given the wrong initial diagnosis. It is suggested that if patients were given the correct diagnosis initially that over a 10 year period it is possible to prevent 250 deaths per year.

Problem Statement/ Causes For Concern

- According to an article by the Harvard Health Publishing Harvard Medical School and the CDC 200, 000 heart disease related deaths a year are preventable
- Statistics show that within one year of treatment, the risk for a heart attack can be significantly reduced
- 610, 000 people die of heart disease in the United States each year, accounting for 1 in every 4 deaths
- The average cost for treating a patient admitted to the hostpital with a heart attack is $18, 200

Goals-Objectives-Stakeholders

- The goal is to build a classification model that can accurately predict patients that are at high risk of suffering a heart attack.

- The objective is to assist physicians with the efficient prevention and health monitoring of patients that show signs of high risk of serious heart conditions

- Stakeholders – physicians, cardiac units and hospitals, medical insurance companies, patients and family members

Overview of the head of the dataset

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
0	28	1	2	130	132	0	2	185	0	0.0	?	?	?	0
1	29	1	2	120	243	0	0	160	0	0.0	?	?	?	0
2	29	1	2	140	?	0	0	170	0	0.0	?	?	?	0
3	30	0	1	170	237	0	1	170	0	0.0	?	?	6	0
4	31	0	2	100	219	0	1	150	0	0.0	?	?	?	0
5	32	0	2	105	198	0	0	165	0	0.0	?	?	?	0
6	32	1	2	110	225	0	0	184	0	0.0	?	?	?	0

Overview of the tail end of the dataset

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	num
287	50	1	4	140	341	0	1	125	1	2.5	2	?	?	1
288	52	1	4	140	266	0	0	134	1	2.0	2	?	?	1
289	52	1	4	160	331	0	0	94	1	2.5	?	?	?	1
290	54	0	3	130	294	0	1	100	1	0.0	2	?	?	1
291	56	1	4	155	342	1	0	150	1	3.0	2	?	?	1
292	58	0	2	180	393	0	0	110	1	1.0	2	?	7	1
293	65	1	4	130	275	0	1	115	1	1.0	2	?	?	1

Dataset overview/wrangling & exploratory data analysis

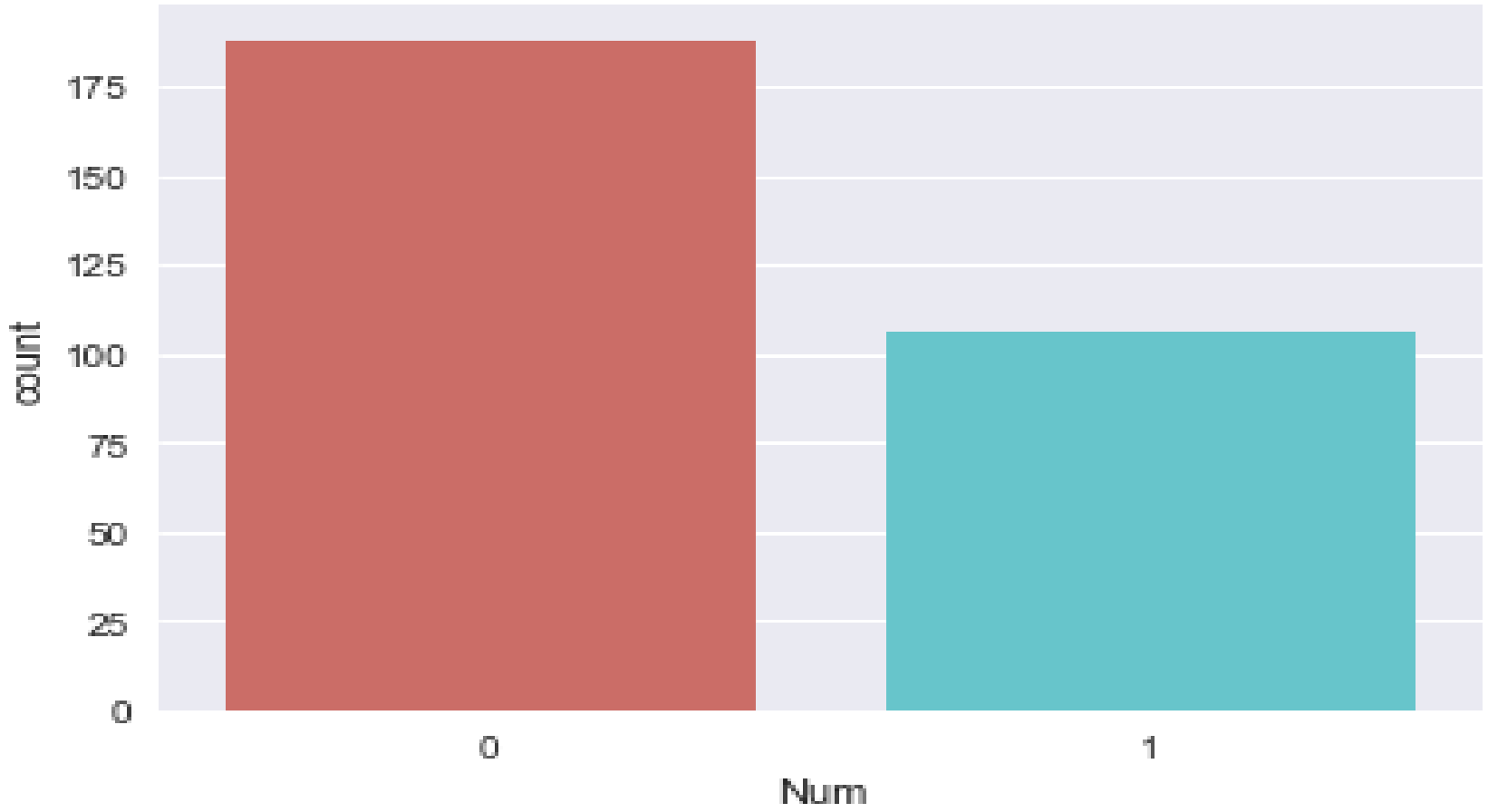
- The dataset can be found at https://www.kaggle.com/imnikhilanand/heart-attack-prediction/home
- The dataset is a combination of information from four databases concerning heart disease diagnosis
- There are 14 columns / 294 rows
- Class imbalance in the target variable
- Categorical features of sex and chest pain
- Significant amount of missing values

Summary Statistics on dataset

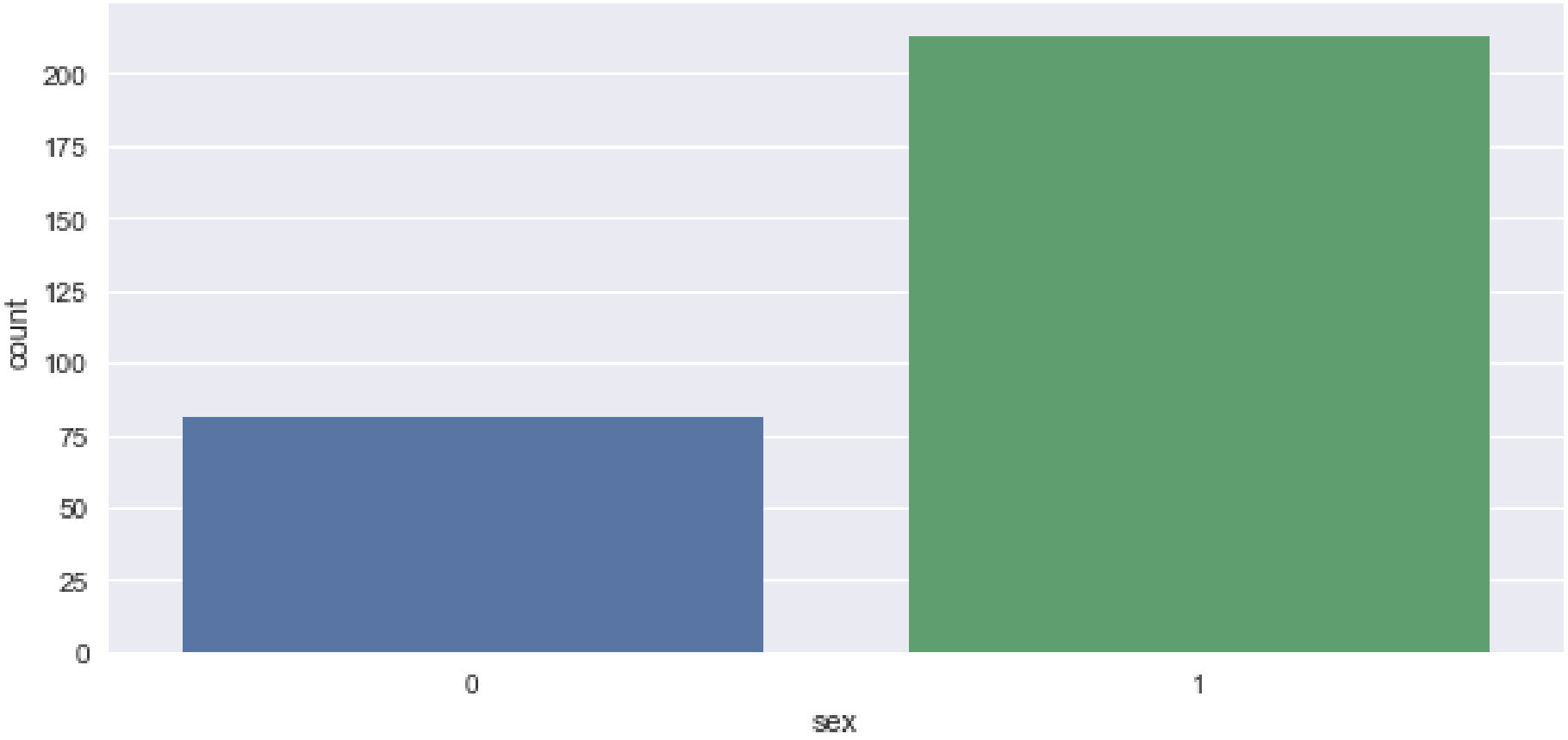
	count	mean	std	min	25%	50%	75%	max
age	294.0	47.826531	7.811812	28.0	42.00	49.0	54.0	66.0
sex	294.0	0.724490	0.447533	0.0	0.00	1.0	1.0	1.0
cp	294.0	2.982993	0.965117	1.0	2.00	3.0	4.0	4.0
trestbps	293.0	132.583618	17.626568	92.0	120.00	130.0	140.0	200.0
chol	271.0	250.848708	67.657711	85.0	209.00	243.0	282.5	603.0
fbs	286.0	0.069930	0.255476	0.0	0.00	0.0	0.0	1.0
restecg	293.0	0.218430	0.460868	0.0	0.00	0.0	0.0	2.0
thalach	293.0	139.129693	23.589749	82.0	122.00	140.0	155.0	190.0
exang	293.0	0.303754	0.460665	0.0	0.00	0.0	1.0	1.0
oldpeak	294.0	0.586054	0.908648	0.0	0.00	0.0	1.0	5.0
slope	104.0	1.894231	0.338995	1.0	2.00	2.0	2.0	3.0
ca	3.0	0.000000	0.000000	0.0	0.00	0.0	0.0	0.0
thal	28.0	5.642857	1.615074	3.0	5.25	6.0	7.0	7.0
Num	294.0	0.360544	0.480977	0.0	0.00	0.0	1.0	1.0

Imbalanced data - Plot of patients that didn't suffer a heart attack vs patients that did (target variable). No resampling of the data for the initial model tests.

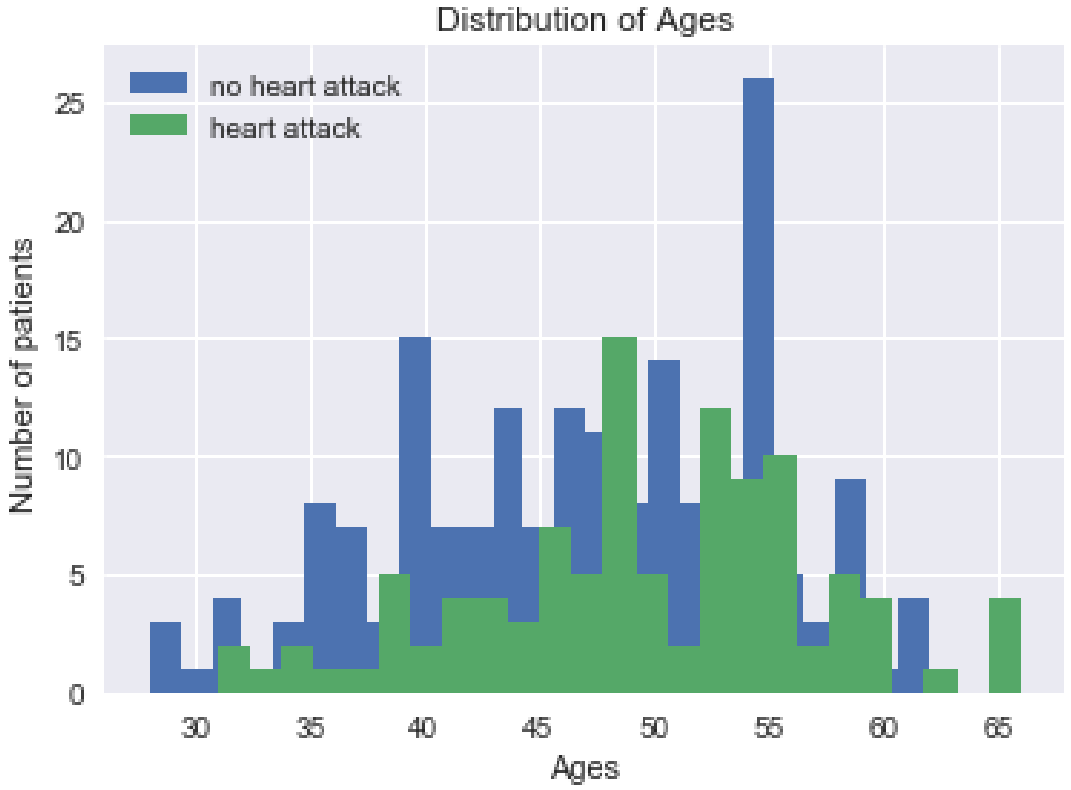
Disadvantages: random over-sampling may cause overfitting and random under-sampling may cause loss of information.

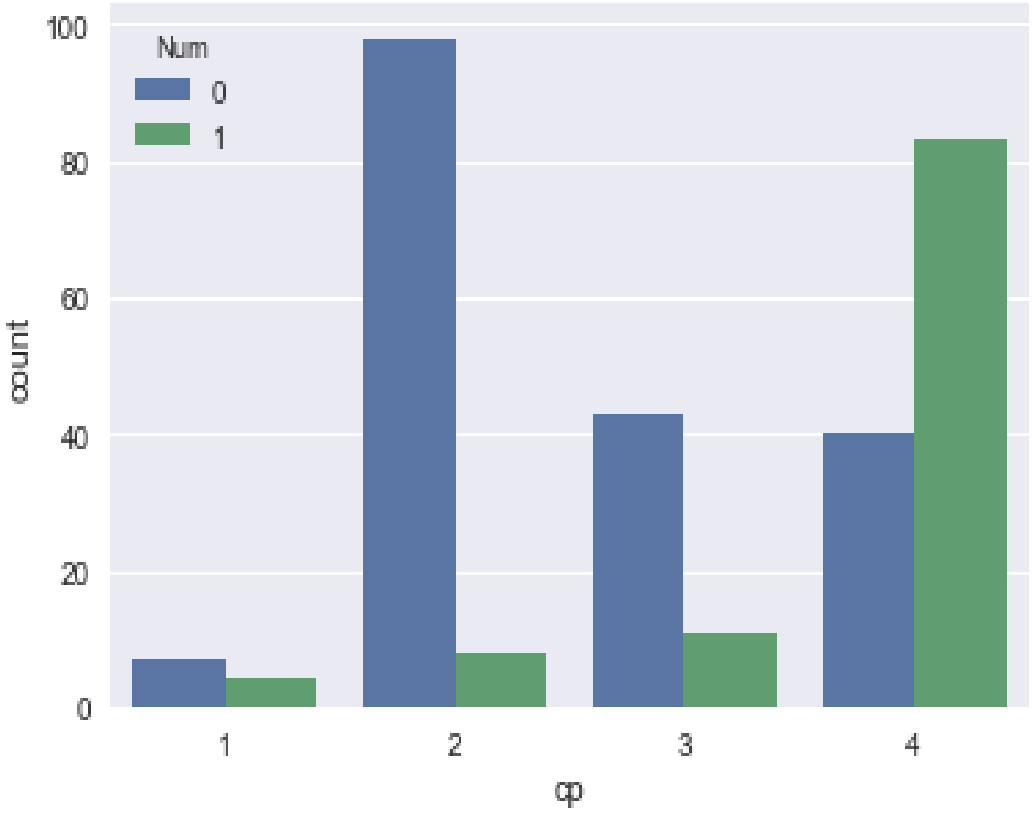


Plot of the women vs men in the data

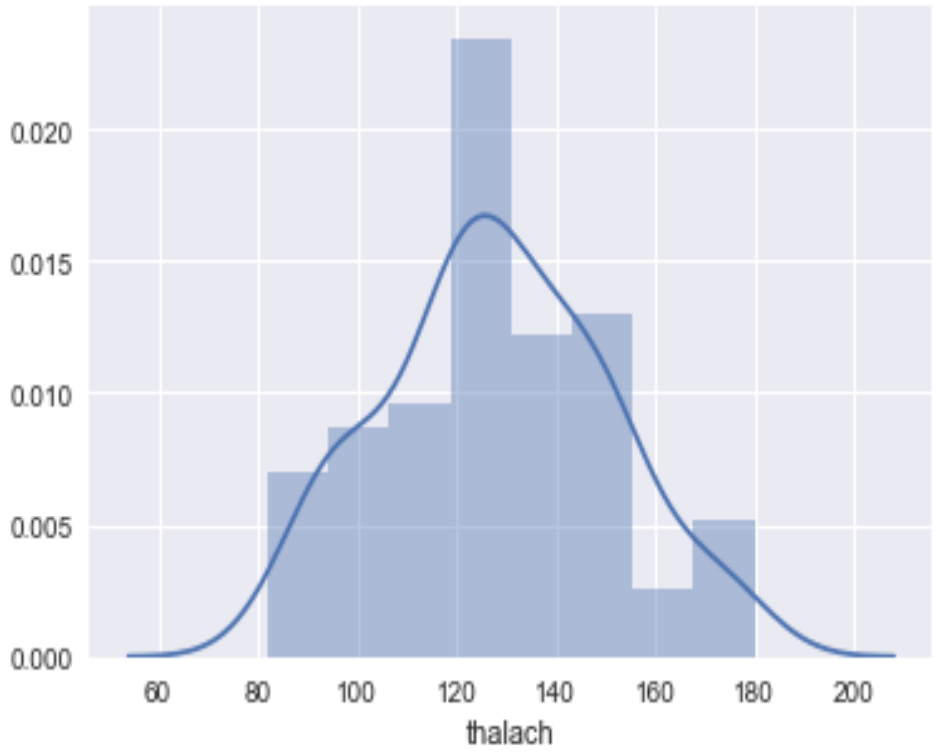


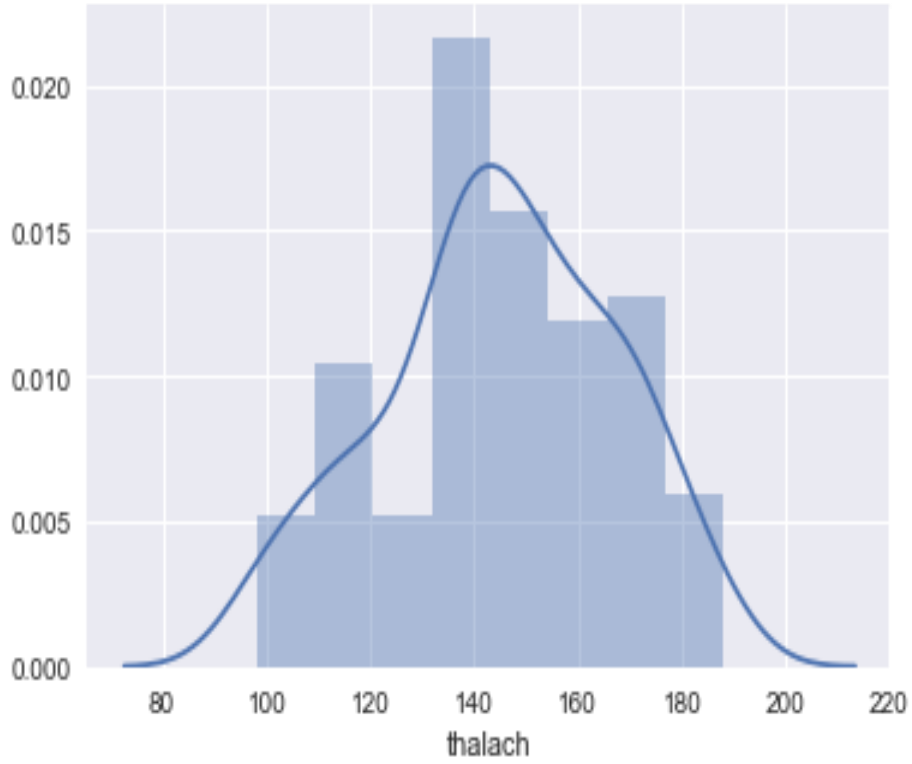
Visual overview of the data between the classes of heart attack patients and non heart attack patients



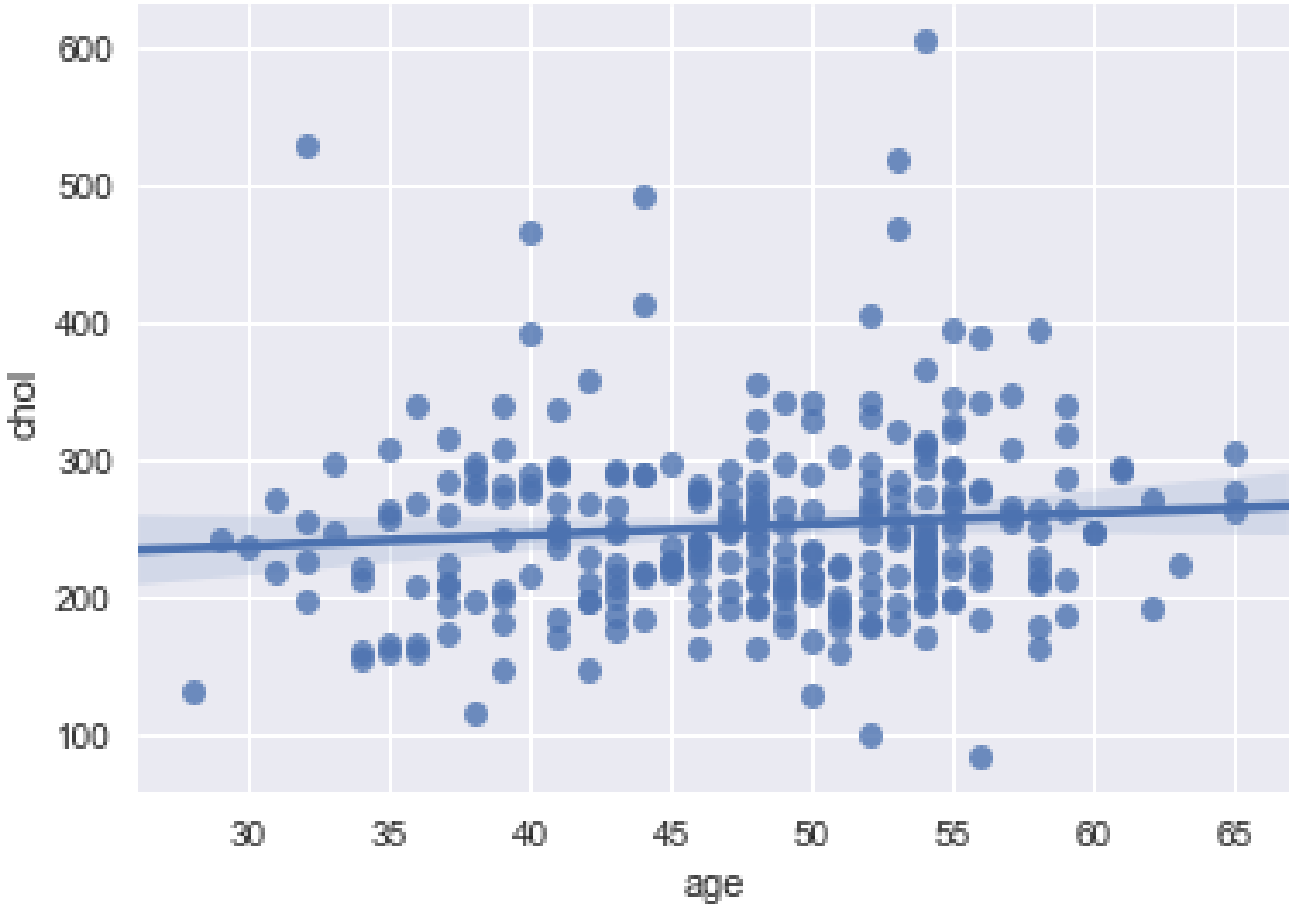


Maximum heart rate achieved of male patients that suffered a heart attack vs male patients who did not

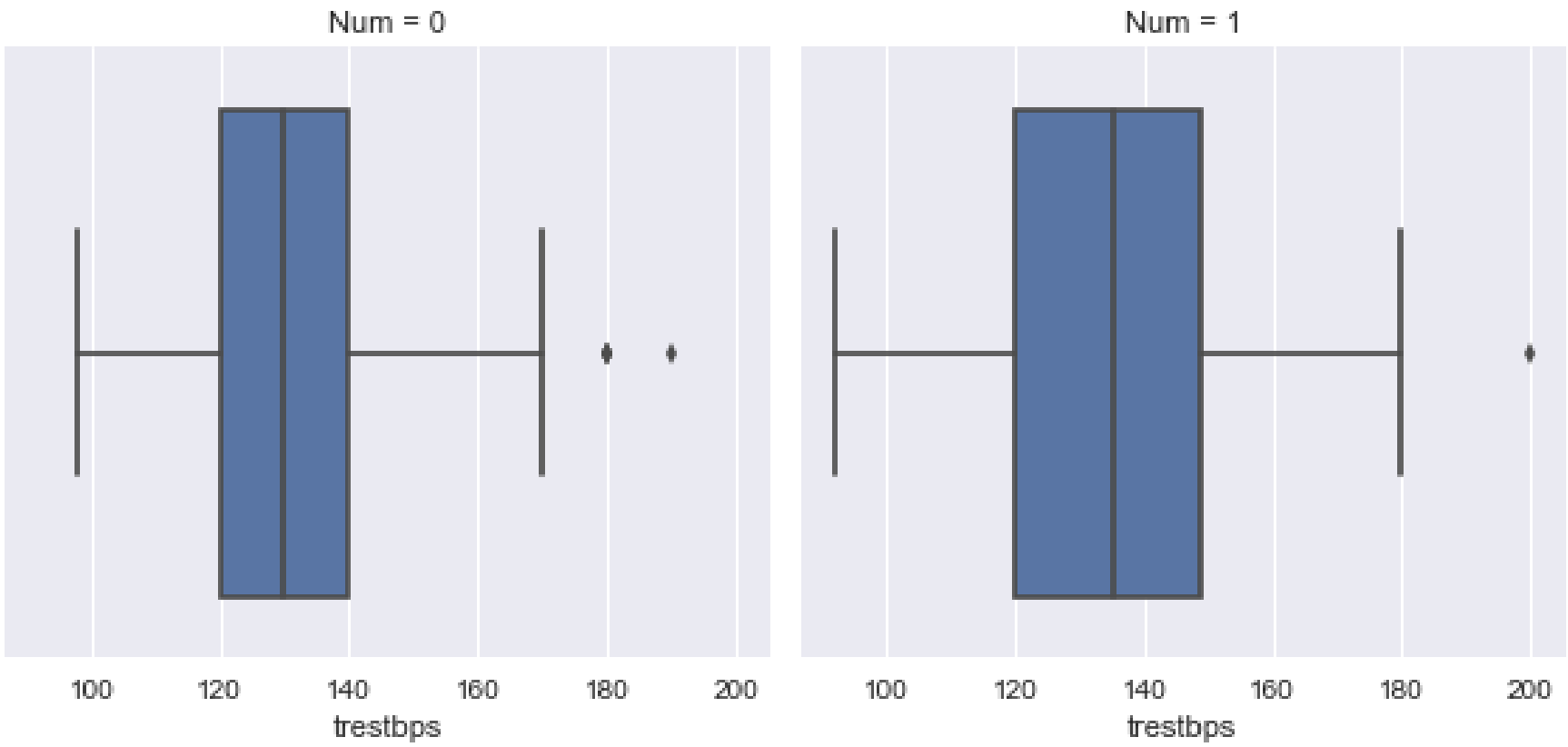




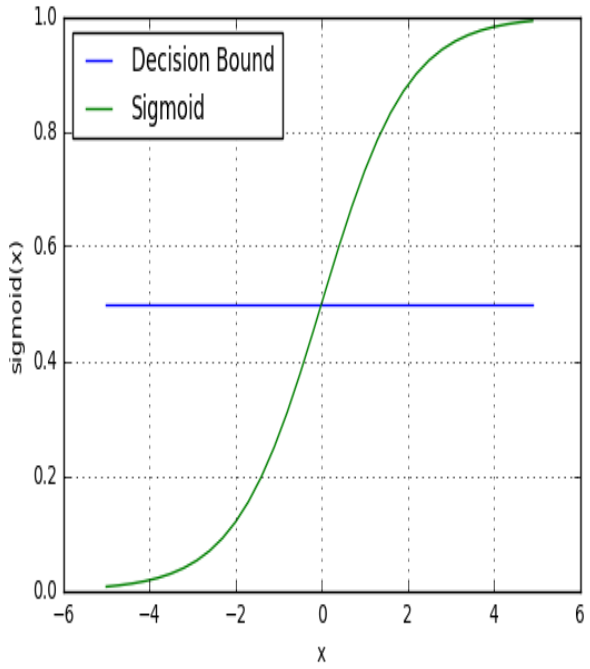
Linear regression of patients ages/cholesterol



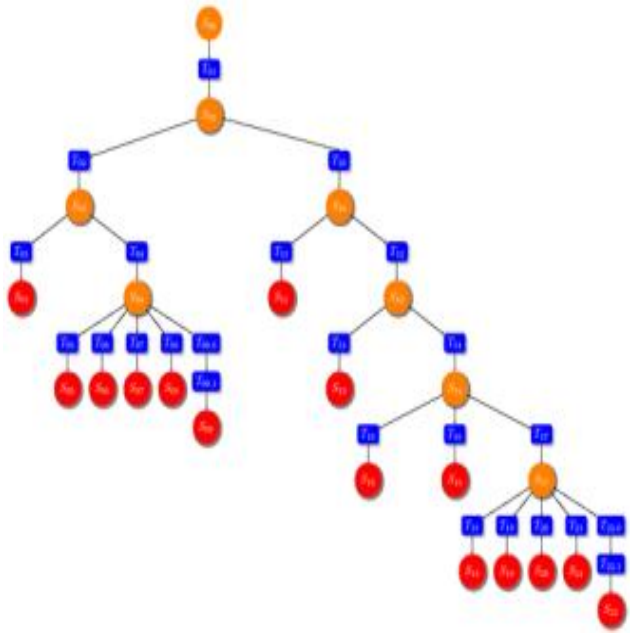
- Resting blood pressure of patients who did not suffer a heart attack vs patients who did



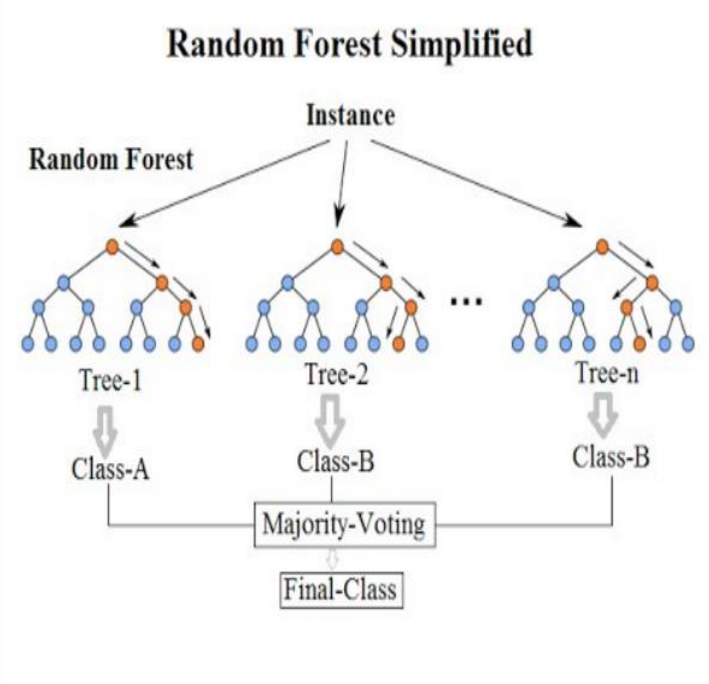
Model selections



Uses Sigmoid logistic function for binary classification



Splits data on features



A method of ensemble learning

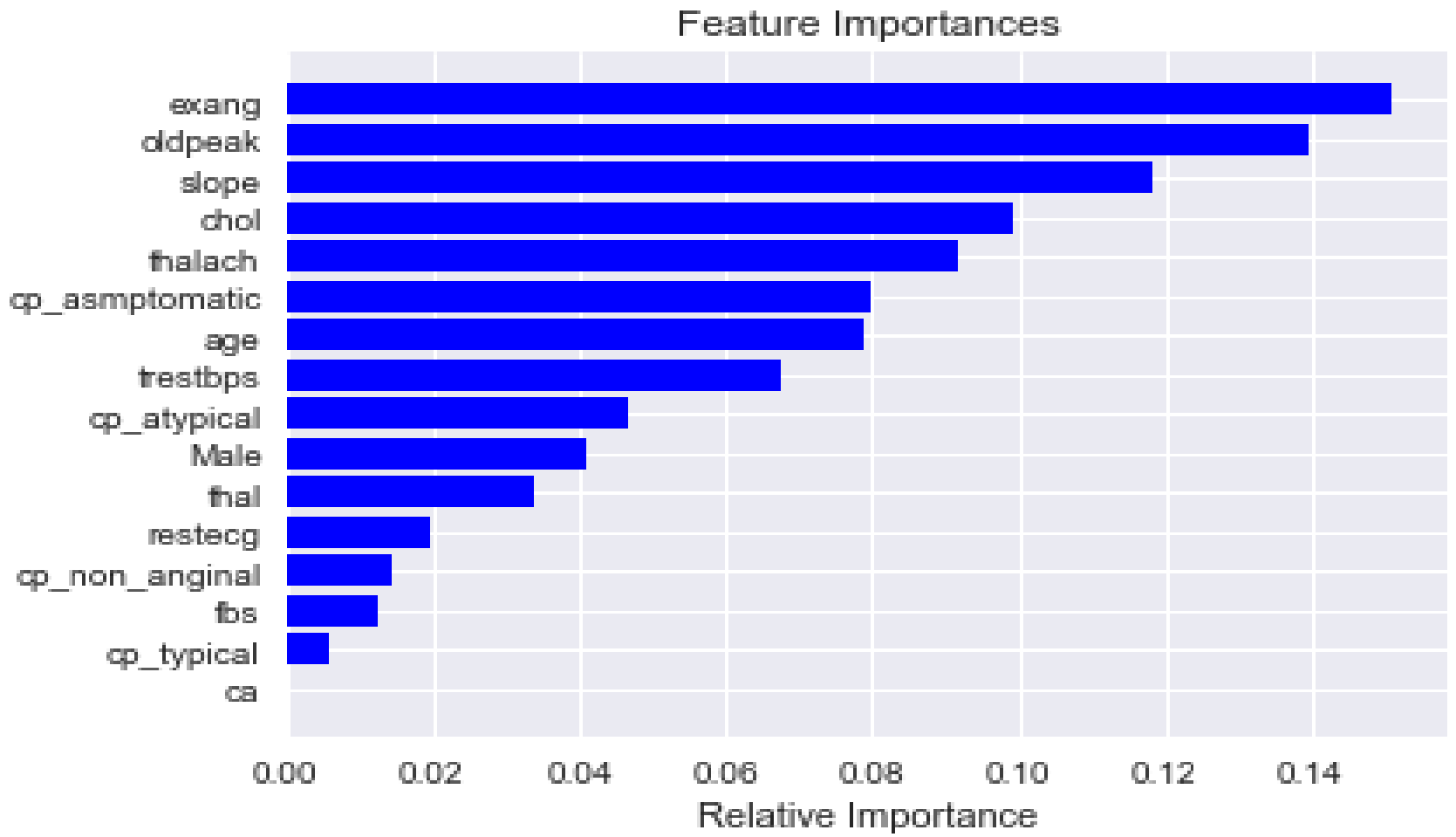
Feature Selection

- The wikipedia definition of feature selection - is the process of selecting a subset of relevant features (variables, predictors) for use in model construction.

- The reason for feature selection with our model is to give more significance to the features in our data that will make our model more efficient and accurate. We want to know what conditions/symptoms associated with each patient is most likely to help us identify patients that are at higher risks of heart attacks.

- After some analysis, there were several specific features that were shown to have a higher level of significance on or model performance (cholesterol levels, resting blood pressure etc.).

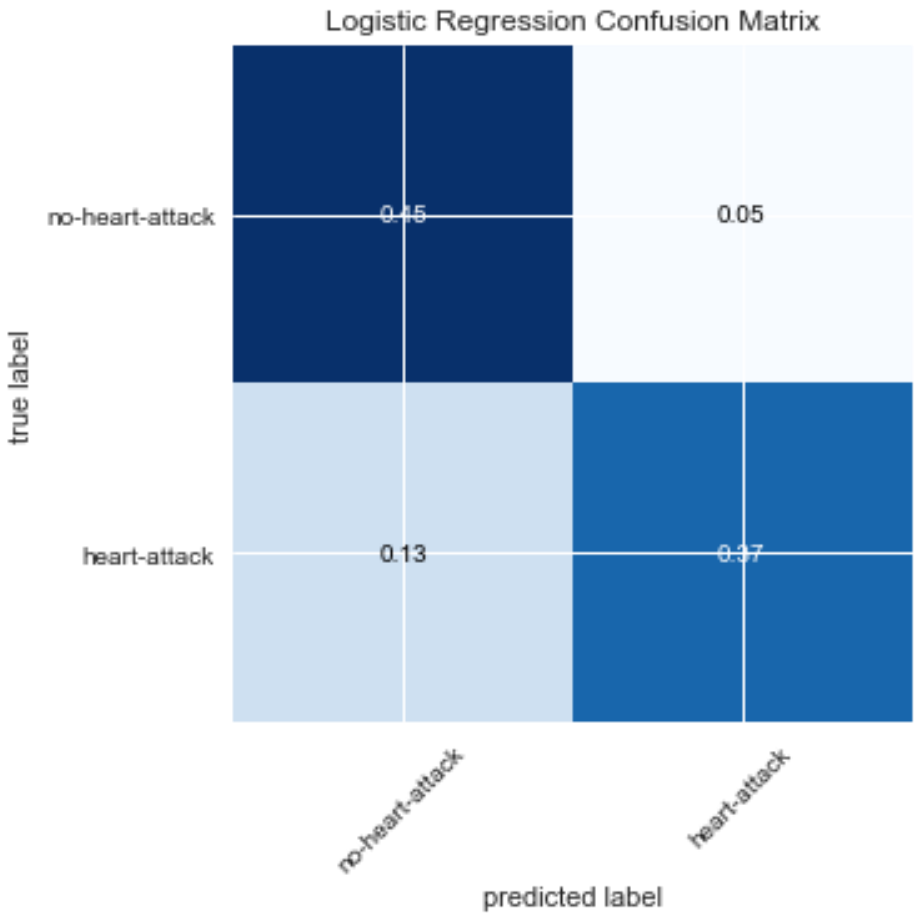
Features are graphed and ranked in order by their statistical significance to the model

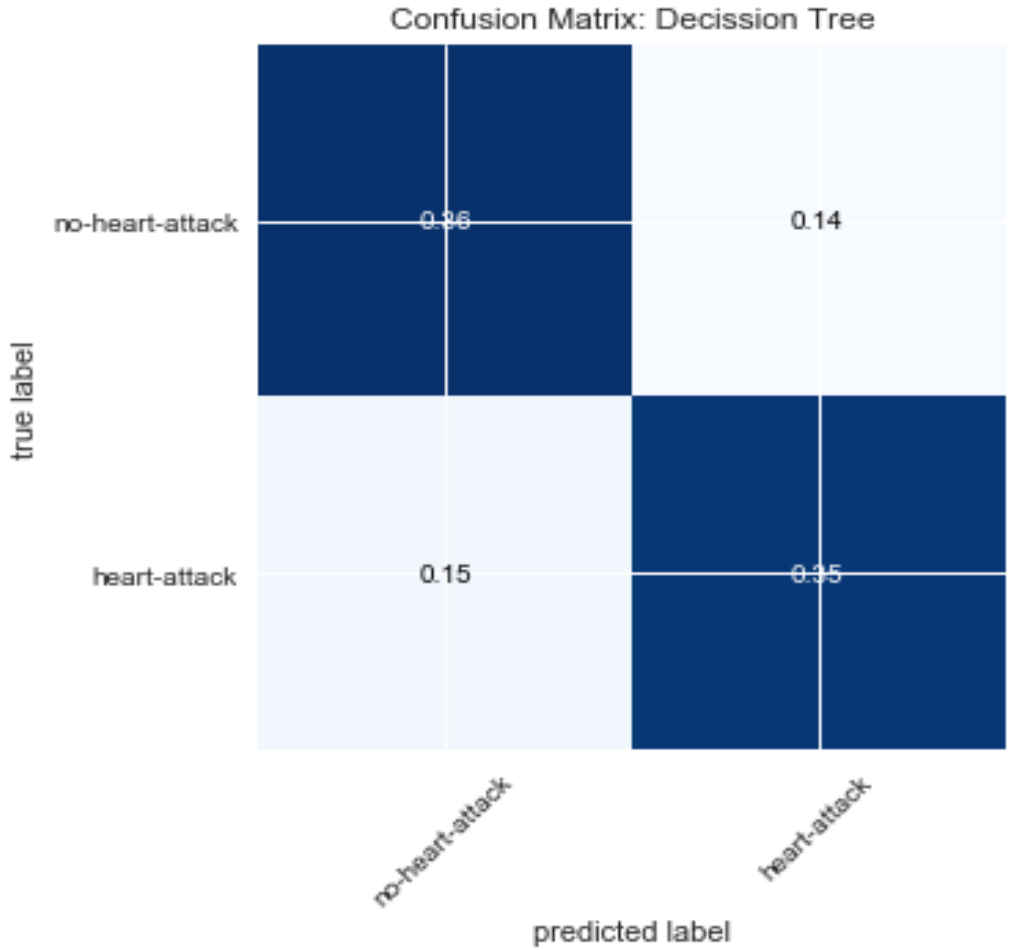


For the initial model selection I decided to go with a logistic regression model. Once I had a baseline for performace and accuracy I could try to outperform my baseline results by implementing more sophisticated decision tree and random forest models.

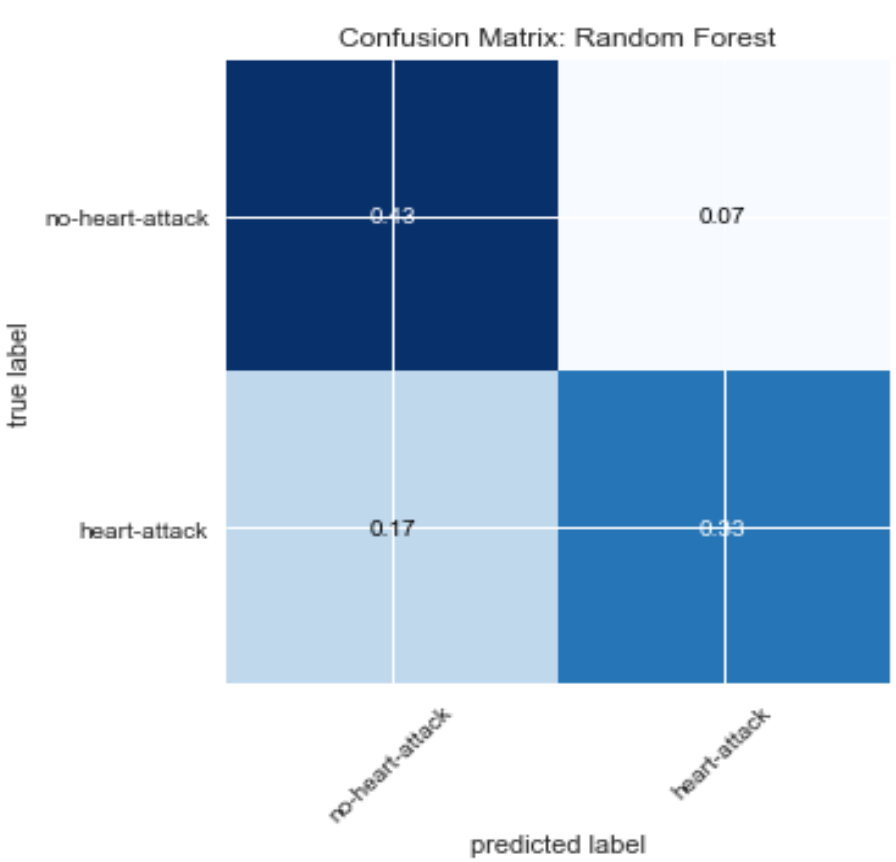
- Logistic Regression
- K-Fold Cross Validation
- Decision Tree
- Random Forest
- Random Search
- Grid Search

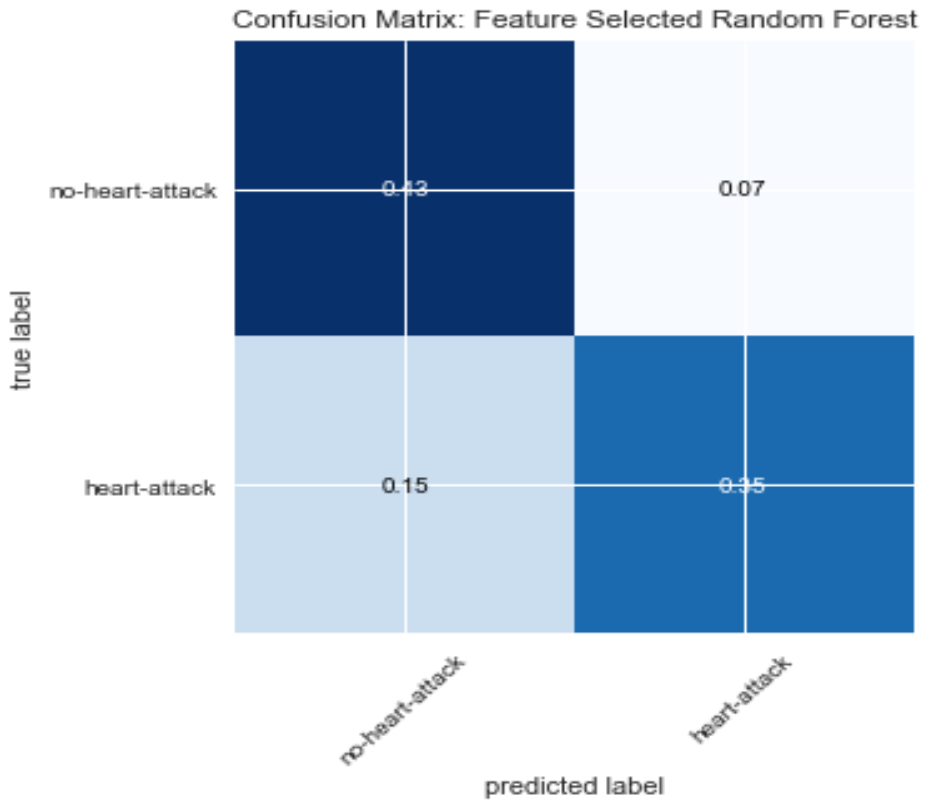
Confusion Matrix Model Comparisons



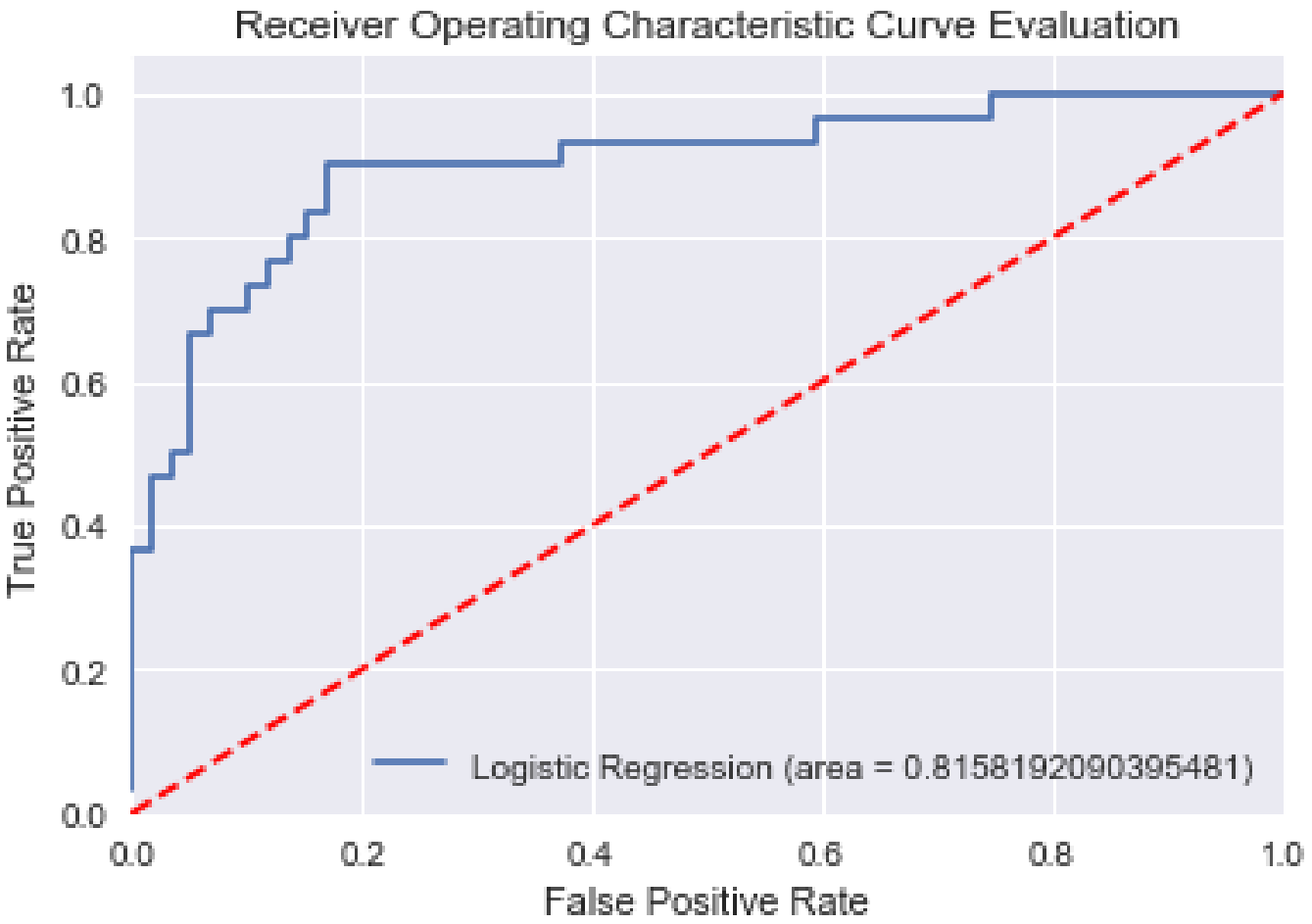


Confusion Matrix Model Comparisons





Logistic Regression Model Area Under Curve Evaluation



Metric Comparison of Models

- Accuracy of Logistic Regression is 0.84
- Precision of Logistic Regression is 0.79
- Recall of Logistic Regression is 0.73
- ROC score of Logistic Regression is 0.82

- Accuracy of Decision Tree is 0.71
- Precision of Decision Tree is 0.55
- Recall of Decision Tree is 0.70
- ROC score of Decision Tree is 0.71

- Accuracy of Random Forest is 0.80
- Precision of Random Forest is 0.71
- Recall of Random Forest is 0.67
- ROC score of Random Forest is 0.77

- Accuracy of Feature Selected RandomForest is 0.81
- Precision of Feature Selected Random Forest is 0.72
- Recall of Feature Selected Random Forest is 0.70
- ROC score of Feature Selected Random Forest is 0.78

RISKS/LIMITATIONS

- *After several model approaches we have a logistic regression model that correctly predicts 84% of the time.*

- *The logistic model incorrectly predicts more than 10% of the time.*

- The risks are the misclassification of the model especially in regards to the type 2 errors. When we take into consideration that this is in regards to patients health and can lead to a potential life threatening condition, there should be more accuracy and efficiency.

Future Work/Recommendations

- It is recommended for future work that a collection of more meaningful data and additional features are added to the model as well as improvements to the algorithm.
- Use resampling approaches to enhance the data we have for algorithms (Tomek links, Cluster centroid under-sampling, SMOTE etc)
- Collect data that involves other heart disease conditions such as arrhythmia, stroke, congestive heart failure etc. Look for potential correlation.
- Experimenting on model implementation, example - see if accurate models can be built without using some of the features that have higher statistical significance

Resampling Data Techniques For Future Work

- Random under-sampling target variable

- Random over-sampling target variable