Heart Attack Analysis

DSC-680 Project 1 Draft

Isabella Sturm

2021-12-18

Business Problem

Heart Disease is the leading cause of death in the United States, killing nearly 700,000 people each year. Heart attacks are a form of heart disease, affecting about 805,000 people every year. Detecting a heart attack early is important in reducing damage to the heart, so beyond knowing what to look out for if a person is experiencing one, knowing whether a person is at high risk for heart disease and heart attacks is also important.

The goal of this project is to determine key features that determine who is at high risk for heart attacks.

Background

As the leading cause of death in the United States, being able to predict whether someone is at risk of heart disease or a heart attack is important in order to save lives. By catching the risk of heart disease before major symptoms, early intervention can take place to work on changing habits and starting treatment if necessary to reduce factors (such as blood pressure and cholesterol) that can lead to serious complications of heart disease, including heart attack and death.

The data used in this project has a binary target meaning that we will use classification to group people into the two categories: True (1, at risk of heart attack) or False (not a risk of heart attack). Performing some statistical analysis and doing some exploratory analysis will help us determine factors that most increase chances of heart disease. Then modeling the data will allows us to make predictions based on the several features affecting the risk of heart attack. The models I have chosen to start with are Logistic Regression, K-Nearest Neighbors, Decision Tree and Random Forest.

Data Explanation

The dataset I am using for this project is a Heart Disease Dataset (Comprehensive) from IEEE.org. It is a combined dataset of five other heart disease datasets: Cleveland, Hungarian, Switzerland, Long Beach VA, and Statlog (Heart) Data Set:

https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive

The dataset has 1190 rows (1180 after removing outliers) with 11 features detailed in the following table:

Feature	**Description**
Age	Age of the patient
Sex	Sex of the patient (1=M, 0=F)
Chest pain type	Chest pain types: 1 = typical angina 2 = atypical angina 3 = non-anginal pain 3 = asymptomatic
Resting bp s	Resting blood pressure (mm Hg)
Cholesterol	Cholesterol (gm/dl) from BMI sensor
Fasting blood sugar	1 = true (fasting blood pressure > 120 mg/dl), 0 = false
Resting ecg	Resting electrocardiographic results: 0 = normal 1= having ST-T wave abnormality (T wave inversion and/or ST elevation or depression of > 0.05 mV) 2 = showing probably or definite left ventricle hypertrophy by Estes' criteria
Max heart rate	Maximum heart rate achieved
Exercise angina	Exercise induced angina (0 = false, 1 = true)
Oldpeak	ST depression
ST slope	The slope of the peak exercise ST segment: 1 = upsloping 2 = flat 3 = downsloping
Target	Class: 0 = normal 1 = heart disease

Methods

For this project, I decided to test out a few different models good for classifying binary target variables. The models I chose include Logistic Regression, K-Nearest Neighbors, Decision Tree and Random Forest. I chose these four because they are good for classification problems and in this case the classification is either True (have or at risk for heart disease/attack) or False (no heart disease/risk of heart attack).

Analysis

To analyzing the data, I began with some exploratory analysis. I created histograms and bar charts of all 11 features as well as the target to get an idea of how the data was distributed amongst the features. Especially when it came to the target variable, it was split nearly 50/50 which was good because I was worried a more drastic split would be the data would be biased and therefore make less accurate predictions.

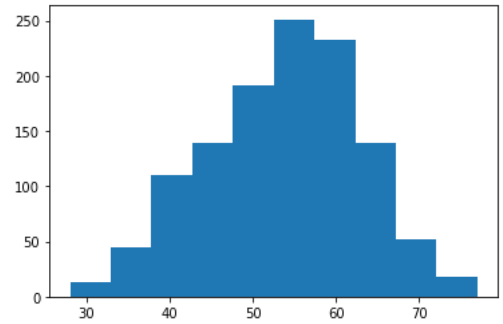


Figure 1 Histogram of Age Feature

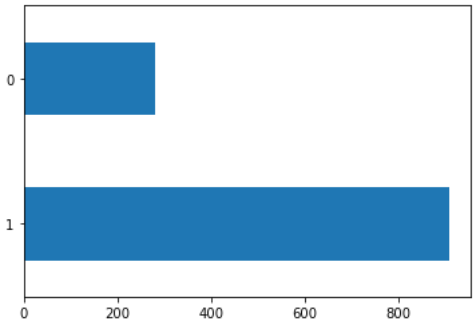


Figure 2 Bar Chart of Sex Feature

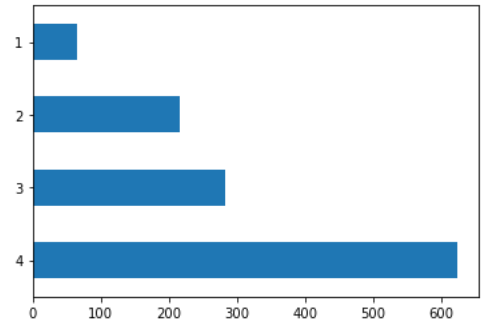


Figure 3 Bar Chart of Chest Pain Type

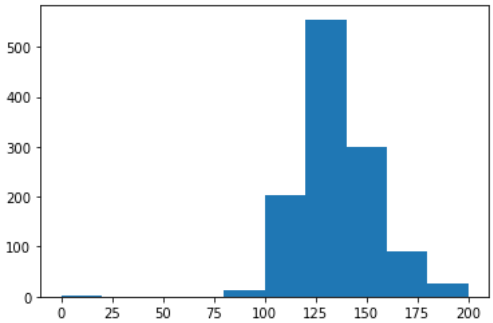


Figure 4 Histogram of Resting Blood Pressure Feature

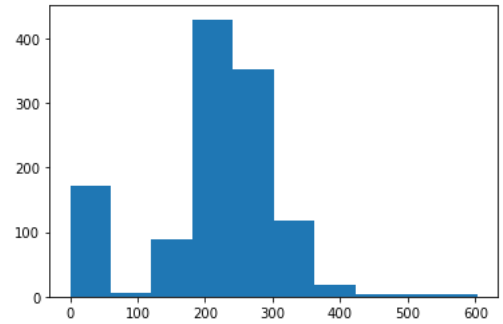


Figure 5 Histogram of Cholesterol Feature

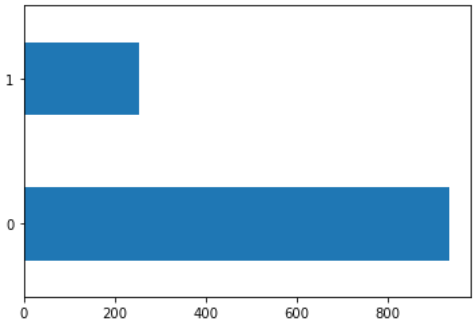


Figure 6 Bar Chart of Fasting Blood Sugar Feature

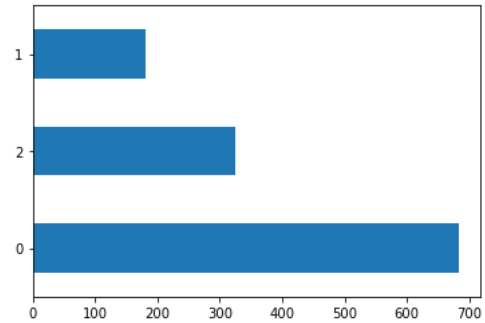


Figure 7 Bar Chart of Resting ECG Feature

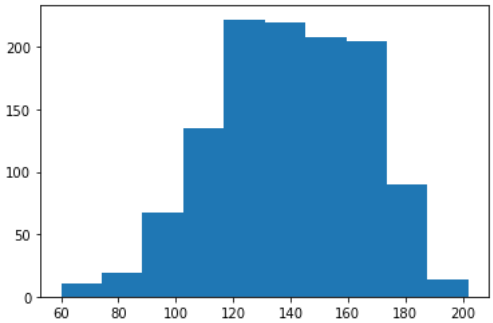


Figure 8 Histogram of Max Heart Rate Feature

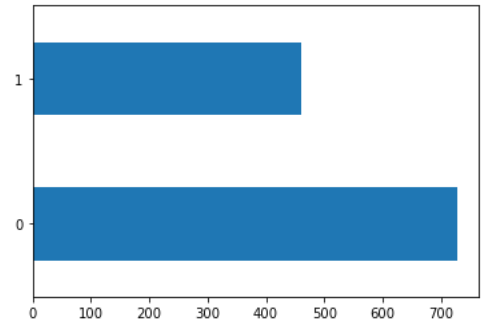


Figure 9 Bar Chart of Exercise Angina Feature

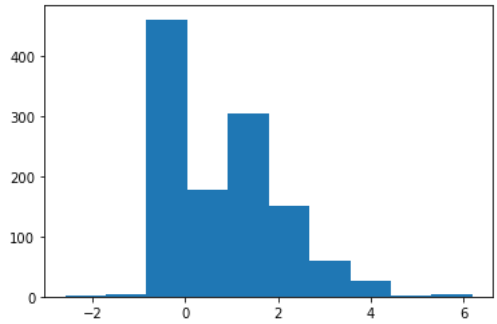


Figure 10 Histogram of Old Peak Feature

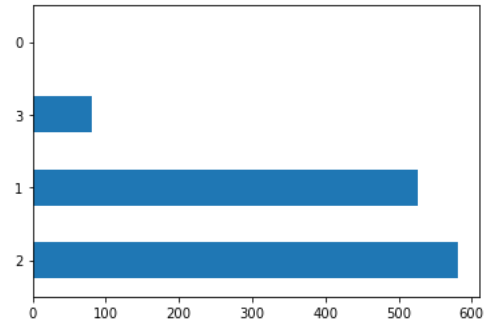


Figure 11 Bar Chart of ST Slope Feature

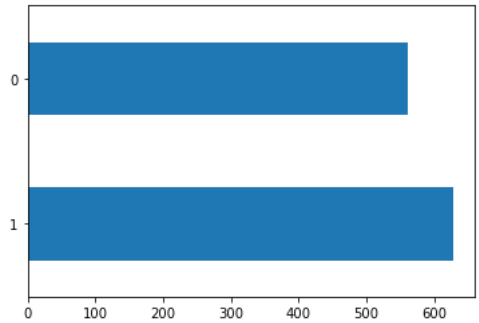
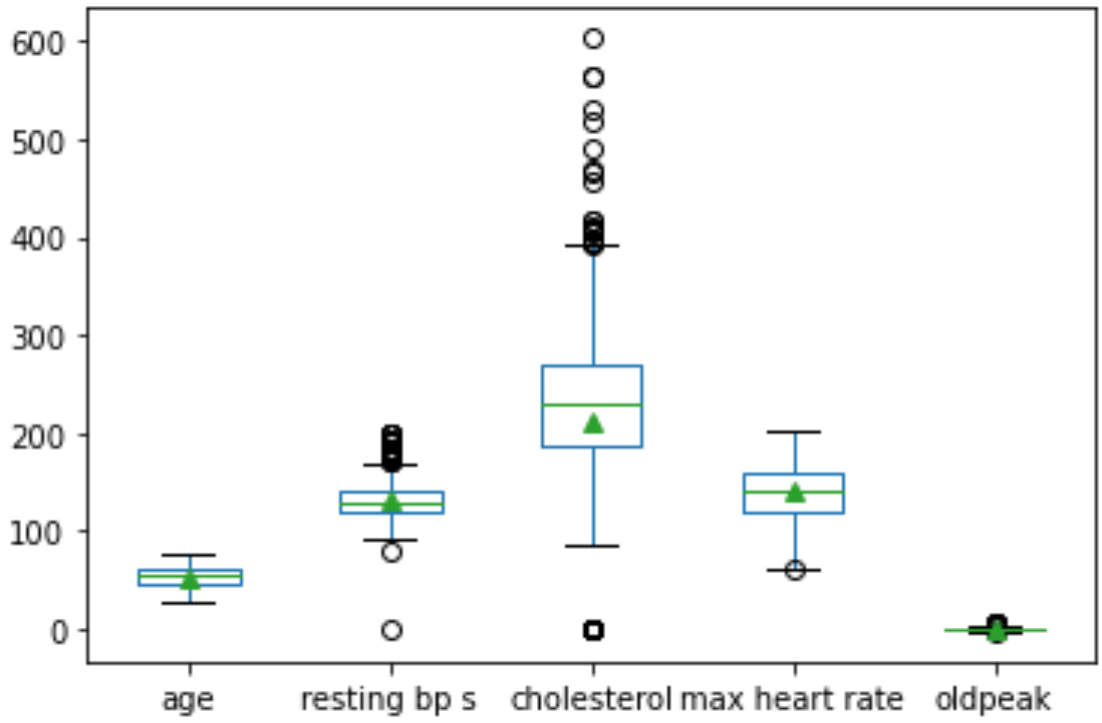


Figure 12 Bar Chart of the Target

Then I created a confusion matrix of all the features with the target included. The features that have the highest correlation to each other and to the target are Max Heart Rate, Testing ECG, Exercise Angina, Old Peak and ST Slope. Max Heart Rate has a negative correlation between 0.35 and 0.41 (except it is only -0.187 to Old Peak) for the correlated features. The other three features are positively correlated roughly between 0.4 and 0.5. Most the other features have relatively low correlations to each other and to the target.

I also created a plot with a box plot for each continuous numeric field to see if there are any outliers, and I printed out the outliers to see how many there were for each variable. There ended up being a total of only 10 records with outlier values, so I removed them.



After cleaning and exploring the data, I started modeling the data First I split the data into training and test datasets. I ended up doing a data split of 75% training and 25% testing. I also tested with 67% to 33% split and 80% to 20% but the 75% to 25% seemed to have better accuracies with the models – the sweet spot.

First, I trained the Logistic Regression Model and got an 85% accuracy. Then I trained the data using the K-Neighbors Classifier and got a 76% accuracy. Next, I trained a Decision Tree Classifier and got an 88% accuracy. Finally, I trained the Random Forrest Classifier and got a 93% accuracy. Below are the confusion matrices for each model.

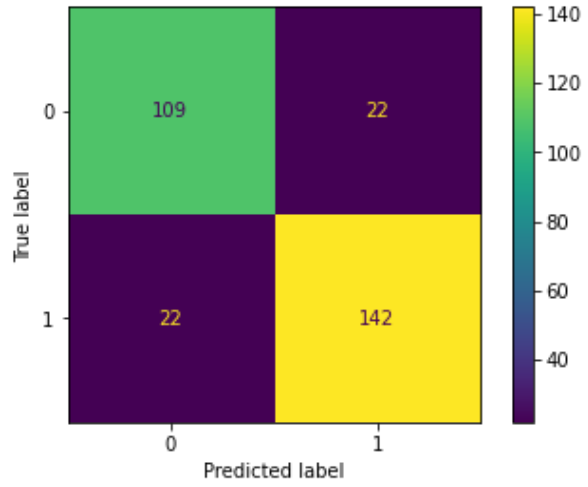


Figure 13 Logistic Regression Classifier: 85.085% accuracy

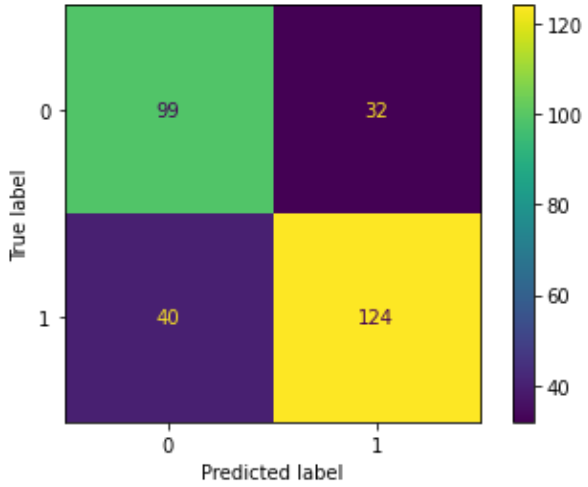


Figure 14 K-Neighbors Classifier: 75.593% accuracy

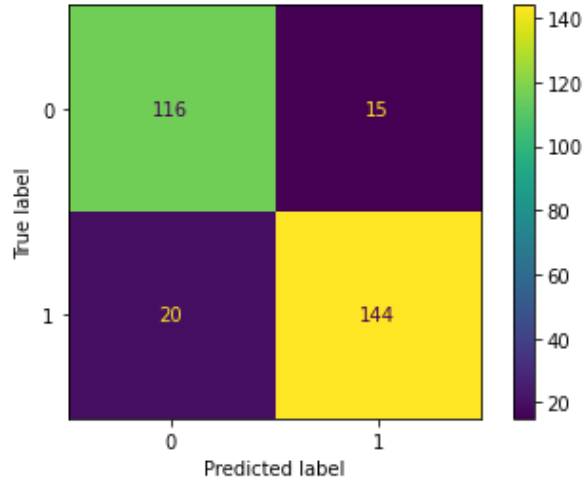


Figure 15 Decision Tree Classifier: 88.136% accuracy

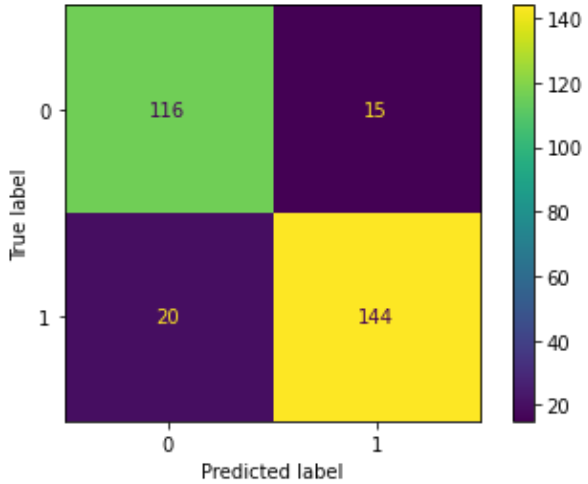


Figure 16 Random Forest Classifier: 93.220% accuracy

Conclusion/Recommendations

With a 93% accuracy, the Random Forest Classifier has the highest accuracy of the models I have selected thus far. While the other three models resulted in fair models (75-88% accuracy), the Random Forest Classifier would be the best model to deploy with the analysis I have done up to this point.

With more data available, the results of the analysis may have turned out differently. In a future iteration of this project, finding more data or perhaps scaling the dataset would allow for more accurate predictions or better confidence in the results.

Assumptions

At this point, the major assumption I have made is that the limited dataset and the varying distribution of data in each feature is still making accurate enough predictions to be useful. Because the target data is at least evenly distributed, I did not think it was necessary to go through and try to balance the rest of the data, especially since in the end I should limit my analysis to the features that most affect the target.

Limitations/Challenges

The primary limitation for this project is the data size. There are less than 1200 records to work with for analysis and modelling. While this doesn't mean the predictions will inherently be bad or wrong, it is not really a lot of data especially considering how prolific heart disease and heart attacks are in America.

Aside from the dataset limitation, I have not come across other challenges. I would like to try new models I have not used before and do some tweaking to the features (perhaps only using the most correlated features for the modelling), but I have not tried this yet. I still do not foresee any major challenges in doing this though.

Future Uses/Additional Applications

This type of medical data could be used to predict the age someone who is at risk of a heart attack or heart disease might start experiencing more severe symptoms of heart disease. I

think this would require more progression/time data of patients (joined by a non-identifying patient ID) so you can analyze how the different features change over time. This type of application would help with diagnosing early signs of heart disease and perhaps encouraging positive change or treatment before anything serious and life-altering occurs (such as a heart attack or death).

Implementation Plan

In the real world, it would be best to be able to connect live data to the model and continuously train and tweak the deployed model.

As is this model could still be used to predict if someone is at risk of heart disease if there was an interface to interact with that allowed inputs to the model and the model could then return a prediction. The risk here, as noted above, is the limited dataset meaning predictions, while more useful than no predictions could still be wrong and in the case of falsely predicting False could be just as harmful if not more than not knowing at all, where in the case of falsely predicting True, positive lifestyle adjustments would not result in a more negative effect than predicted.

Ethical Assessment

With medical related data, an obvious ethical concern is patient privacy. In this case, all identifying information is already removed. However, in a scenario where I would be collecting data or have all the patient data, it would be a major ethical concern to use that data improperly.

References

Centers for Disease Control and Prevention. (2021, January 11). *Heart attack symptoms, risk, and recovery*. Centers for Disease Control and Prevention. Retrieved December 20, 2021, from https://www.cdc.gov/heartdisease/heart_attack.htm

Centers for Disease Control and Prevention. (2021, September 27). *Heart disease facts*. Centers for Disease Control and Prevention. Retrieved December 20, 2021, from https://www.cdc.gov/heartdisease/facts.htm

Siddhartha, M. (2020, November 6). *Heart disease dataset (comprehensive)*. IEEE DataPort. Retrieved December 20, 2021, from https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive