David Lattimer

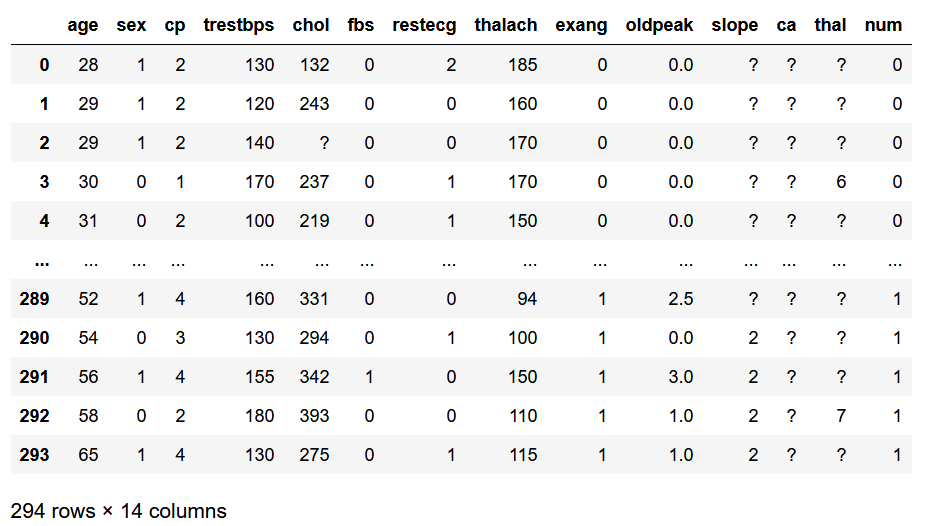
2/19/2021

DSC 550

Heart Disease Case Study

In this case study, I wanted to look into the dataset found here: <https://www.kaggle.com/imnikhilanand/heart-attack-prediction> to help us predict people that might have heart disease. By looking at this case study, many of the attributes in the original set of data are parts of normal doctors tests and check ups, so if a patient’s measurables get put into our model and come out as having or at risk for heart disease, we can test for it and treat it before it becomes a problem that we have can’t fix. Being able to measure just a few things and knowing that the patient may have heart disease could save lives, so it is important to go through the data and make the model as accurate as possible. So first of all, let’s look through the data and see what we have and see if we can get any preliminary clues as to what might contribute to heart disease.

So let’s look at the data and work on making it all usable and clean it up. Our original data looks like this:

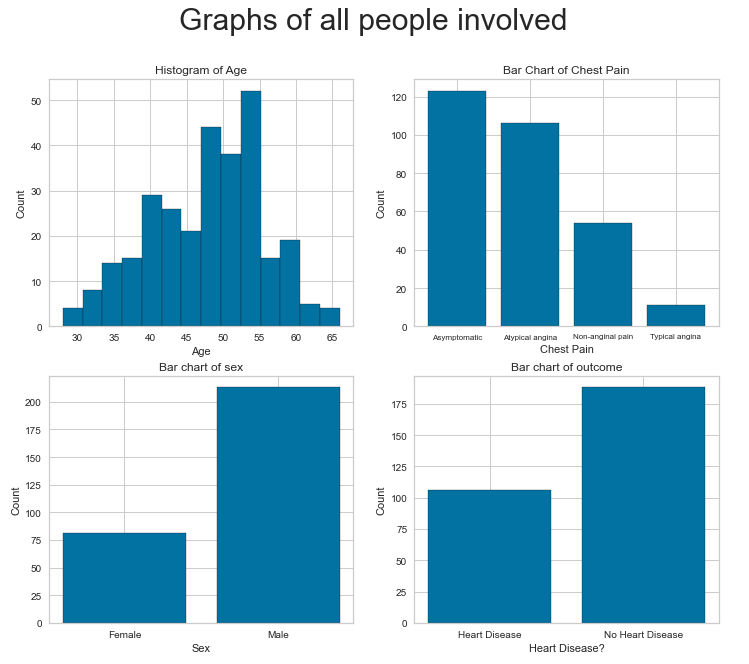


Which has 294 rows of data and 14 columns. The columns are coded based on the study that took place in 1988, so we can go through them now.

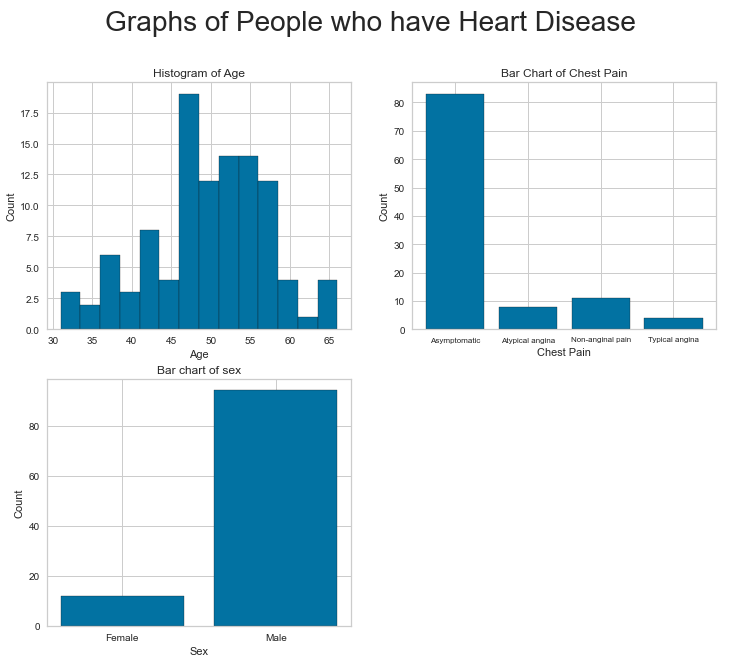
* Age: Age of the patient. It is estimated that chances of heart disease and heart attacks triples every decade someone is alive, so this is a huge factor
* Sex: Sex of the patient. Men tend to have a higher chance of heart disease than women, especially before menopause. After, it is predicted the likelihood about evens out.
  + 1: Male
  + 0: Female
* CP: Chest Pain. This category is classified into 4 groups to show the chest pain someone is experiencing in the dataset.
  + 1: Typical Angina
  + 2: Atypical Angina
  + 3: Non-anginal pain
  + 4: Asymptomatic.
* Trestbps: Resting Blood Pressure. This is one of our only numerical columns, many of the others are coded as numbers but are actually categorical.
* Chol: Cholesterol. Another numerical column, the reason why we won’t see a hugely impactful correlation on heart disease with cholesterol is because we have bad cholesterol (LDL) which will narrow arteries, and good cholesterol (HDL) which will lower the risk of heart disease.
* FBS: Fasting Blood Sugar. The higher the resting blood sugar, the less insulin your body is making and the more likely you are to have heart problems. This is also the main factor in having both types of Diabetes.
  + 0: Blood sugar is less than 120 mg/dl
  + 1: Blood sugar is greater than 120 mg/dl
* Restecg: Resting Electrocardiographic Results. This is sorted into 3 categories:
  + 0: normal
  + 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
  + 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
* Thalach: maximum heart rate achieved. It is estimated that a higher maximum heart rate leads to increased risk of heart disease. It is guessed that an increase in 10 beats per minute increases the chance of heart disease by more than 20%.
* Exang: Exercise Induced Angina. Since our chest pain column showed us who had the anginal pain, this column expands on it.
  + 0: No exercise induced angina
  + 1: Exercise induced angina
* Oldpeak: ST depression induced by exercise relative to rest. This column is complicated, but most people surveyed will have a 0 shown here, and increasing numbers are a strong indicator of heart disease.
* Slope: the slope of the peak exercise ST segment. The three options are:
  + 1: upsloping
  + 2: flat
  + 3: downsloping
* CA: number of major vessels (0-3) colored by fluoroscopy. This entire column is left missing in the dataset and won’t be included in the model.
* Thal
  + 3 = normal
  + 6 = fixed defect
  + 7 = reversible defect
* Num: Having heart disease. This is our target column where:
  + 0 = No Heart Disease
  + 1 = Heart Disease

Now that we know what we are working with and know we only have a handful of actual numerical columns, we can take a look at what we are working with. When cleaning up the dataframe, we find out the entire CA column is missing, so we can remove that, and then the missing values are coded as ‘?’s. We can change the ‘?’s into NaN’s and then from there we are going to take the median values for missing values. We don’t want to use means because even though the dataframe is coded as numbers, fractional numbers wouldn’t work for columns like chest pain, where each number is its own specific category. So we use median to make sure we get whole numbers.

Now let’s explore our data a little bit more to see some trends before building the model. Here is a histogram of the full dataset and some of the important categories involved:

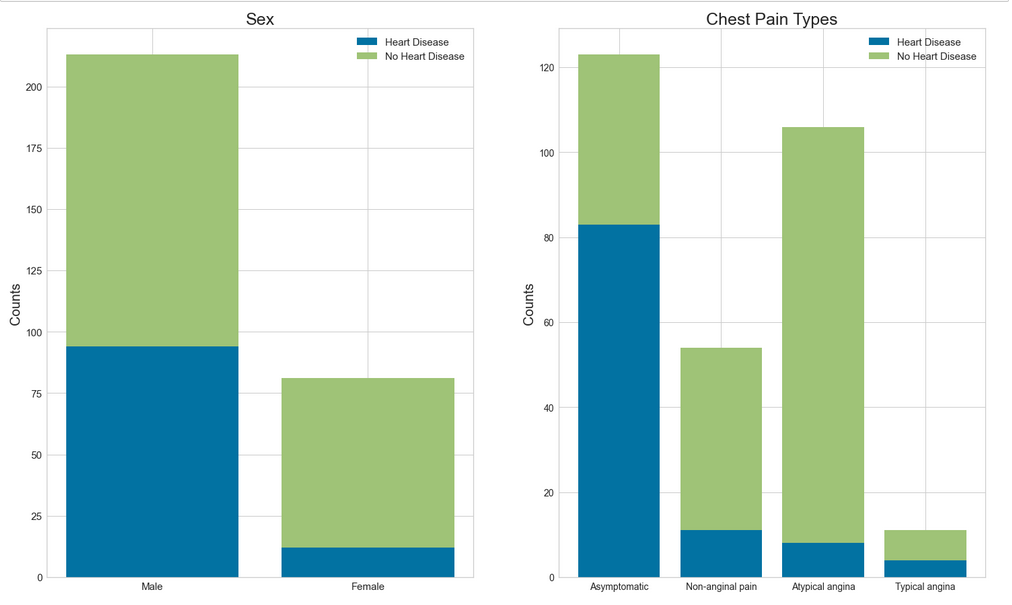


We can see from the histograms and bar charts what we are working with. We have an approximately normal shaped histogram of ages, the dispersion of chest pain people were feeling, the male vs. female count, and of course the amount of people that had heart disease vs did not. There is nothing too out of the ordinary here, but a few things to note, many of the people did not have chest pain, we have a lot more males than females and we have only about a third of our patients having heart disease. This is useful in some ways, but if we are looking for indicators of what might lead to heart disease, it would be helpful to see these same graphs but only including people with heart disease (except the outcome of course). They end up looking like:

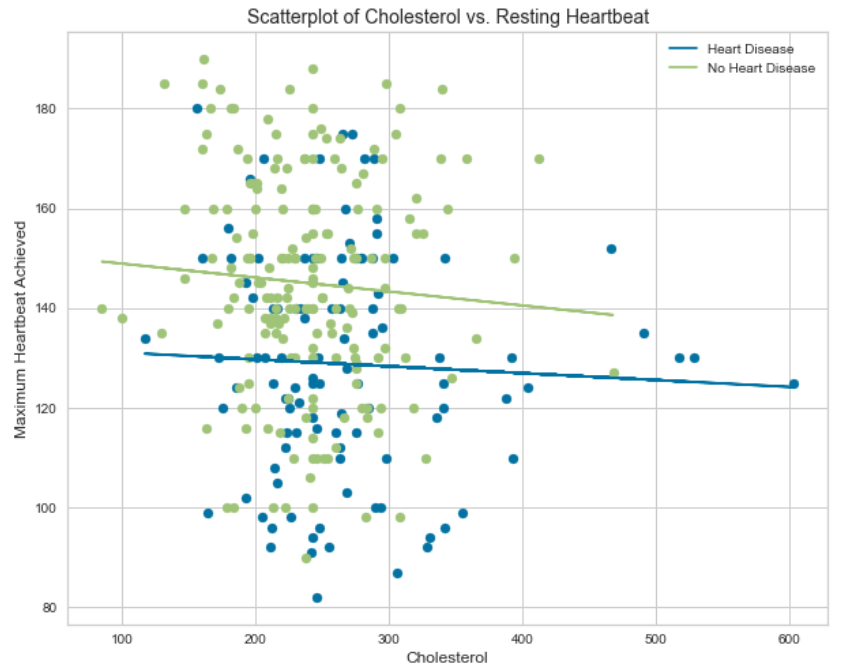


This shows up a lot different than the histograms and bar charts we graphed earlier and it is important to notice this distinction. Age looks like it skews a lot more toward the older people having heart disease, which is something we would expect, and there is a lot to unpack from the chest pain bar chart. For one, most of the people who ended up with heart disease had no chest pain symptoms. Also it seems that atypical angina is a bad indicator of heart disease since we had a large number of cases, but very few had heart disease. And lastly, even though we had a large gap in the entire dataset between males and females, the gap of people with heart disease widened even more. From looking at this, age and sex are good indicators of heart disease, and as strange as it sounds, at least from this set of individuals, people that were asymptomatic ended up having heart disease more often than people with chest pain.

We can also look at two of these graphs laid over each other just to make the comparisons easier. This can be found here:

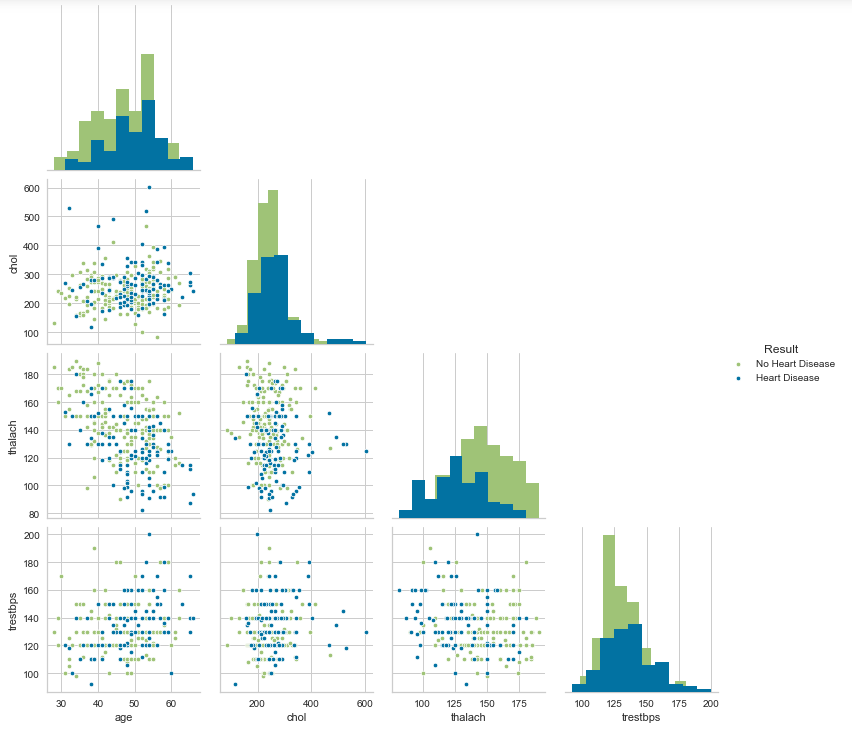


This doesn’t add a ton more except to condense two graphs into one and help show the scale of these two graphs on top of each other. It helps show the portions of the entire survey population that have and don’t have heart disease.

Next let’s look at two of our numerical values and see if we can find any trends in what they mean and see how strong the correlations are. 

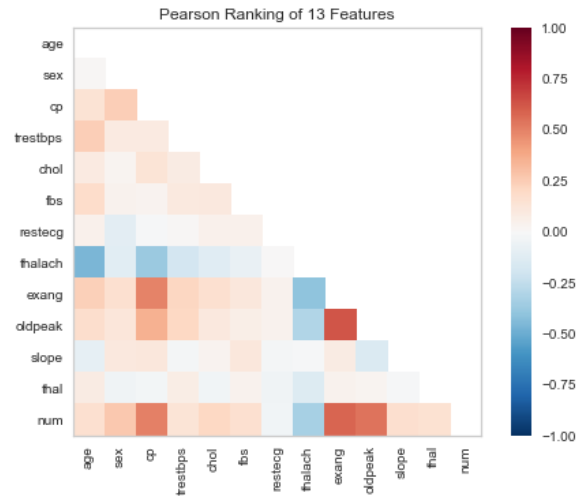
From here, we can see that the best fit lines help us sum up what we are seeing in the scatterplots. From here, we can see that along the x-axis is cholesterol, something we touched on earlier that may not give off the result we would anticipate due to both good and bad forms of cholesterol. We can see that the people that have heart disease tend to skew more to the right and have higher cholesterol, but the difference between our people with and without heart disease isn’t terribly drastic. We will see that again in our Pearson’s correlation chart which should have a positive correlation, but not terribly strong, much like we see here. We can also see that our people with heart disease tended to have a lower maximum heart rate, which is the opposite of what we would have expected. This could be just the data points we had or it could be a sign that we were off on what we thought would happen. This will also show in the Pearson’s correlation chart as a negative relationship since having heart disease has a lower maximum heart rate.

Now let’s take a look at all of our numerical values and see the relationships between them all. We can do that easily by using the pairplot function. Our numerical values that we have are age, cholesterol, maximum heart rate achieved and resting blood pressure. We can create scatterplots of all of these, including histograms of them all diagonally across the chart. We also separate the people with and without heart disease so we can see if one group of people are skewed left, right, up or down on the scatter of the two variables. Here is the chart:



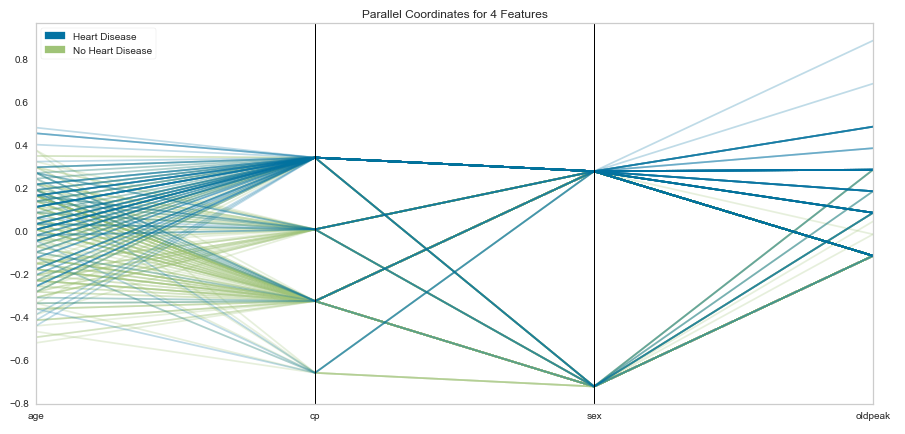
This includes the graph we made up above a little earlier, but also shows us all the relationships relatively easily so we don’t have to code each one and choose what to include. So now we can see them all. This helps us look at all of the variables in one quick graph and the trends we can see are age skewing right, indicating that older people are more susceptible to heart disease as well as the relationships between all of our numerical values.

Speaking of the Pearson’s correlation chart, we can graph all of our variables and look at the strength of their correlations and the relationships the variables have to each other. We graph the Pearson’s correlation chart and it looks like:

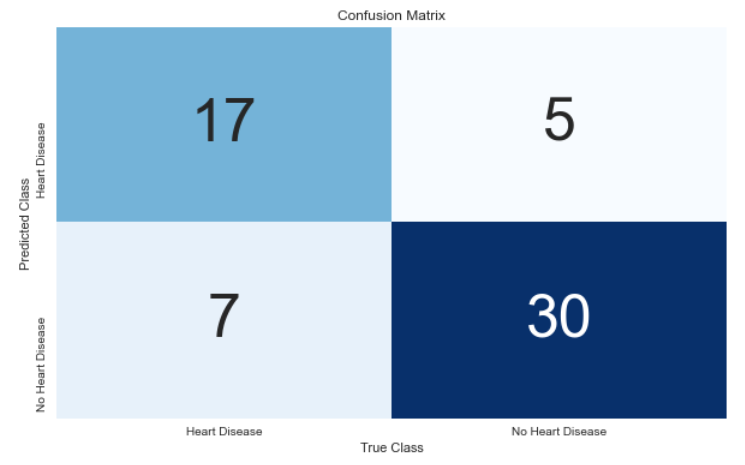


This chart shows us a lot of what we have already discovered, mostly all positive relationships, with varying degrees of strength. The most important row of these is looking at the correlations between the target column (num, aka heart disease) and the other variables in the dataset. Strong correlations between exercise induced anginas (exang), ST depression induced by exercise relative to rest (oldpeak) and chest pain (CP). We also only have one negative correlation, between maximum heart rate measured and having heart disease, which is decently strong as well. These strengths and correlations are also shown in our graphs above and reconfirmed here.

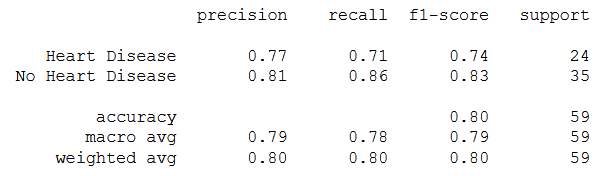
Another graph we can make is a parallel coordinates graph, which really shows us each case and shows where they end up in specific stopping points and shows patterns as well, like this:



There are some pretty obvious things shown here, such as the tendency of people with heart disease to be older, and then how the chest pains break down. It is nice to be able to look at how the chest pains split and are distributed. One of which, CP=1 which is typical angina, the split from there from male to female shows that all females with this chest pain do not have heart disease, whereas most males end up with heart disease. It also shows that a higher oldpeak has increased likelihood of heart disease. Although we have looked at a lot of similar graphs and how they affect heart disease, this breaks up each individual case and shows how hard it may be to predict each case and why we will get errors because there is no perfectly split up category. But knowing the tendencies let’s us make a best guess and find the probability of heart disease, which can lead to extra screenings and catching things before they are a problem.

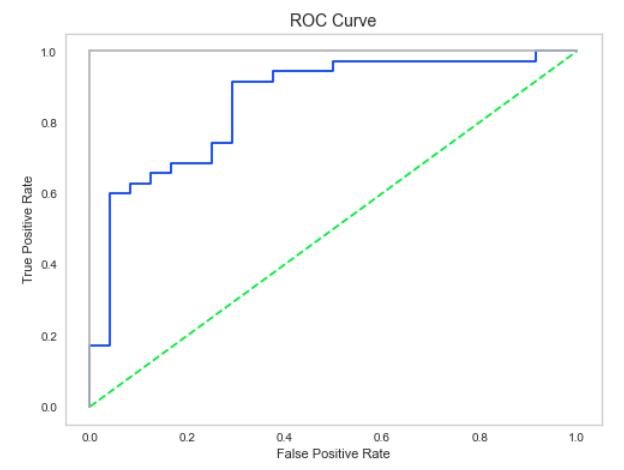
And now for the model making, which means splitting up the data, teaching our model with 80% of the cases, and then testing that model on the remaining 20% of the data. Using the columns we have been looking at and figuring out if the person does or doesn’t have heart disease through our logistic regression model. Of course, we can test to make sure our model is good, especially using a confusion matrix, which helps us see our predicted vs actual results. Our confusion matrix looks like: 

Which shows that our model is relatively efficient but still contains errors. There is a chance that if we were in a medical field, we would want to avoid one of these errors over the other, whether it be predicting they have heart disease when they don’t, which could lead to stress and taking medication that is not needed. Or maybe we don’t want errors where we predict they do not have heart disease when they do, and could be treated and helped before it is a problem. If we were to create a model, we would probably lean more toward avoiding getting false negatives because if we say someone doesn’t have heart disease and they do, they will miss getting treatment and know the ways they can work on being healthier and try to avoid heart attacks and heart disease in the future. For our model, we can see how accurate and how often our model is right by running a classification report,



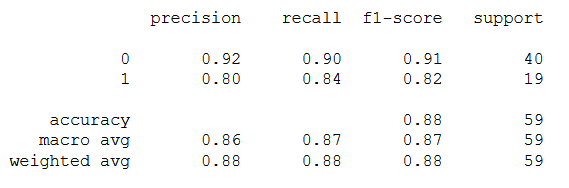
Which shows that we are hovering right around the 80% mark for how our model performed, which isn’t too bad. Not amazing, but being able to be right 80% of the time on the condition of the heart with just 13 measurable categories really helps detect these issues early and help people before it’s too late.

Lastly we will look at an ROC curve that shows the trade-off between collecting all the true positives and the false positives that come with that. Our curve should be between the gray lines on the axis that would be perfect and the dotted line that shows what would happen if we were to essentially guess each target value without a model. As you can see:

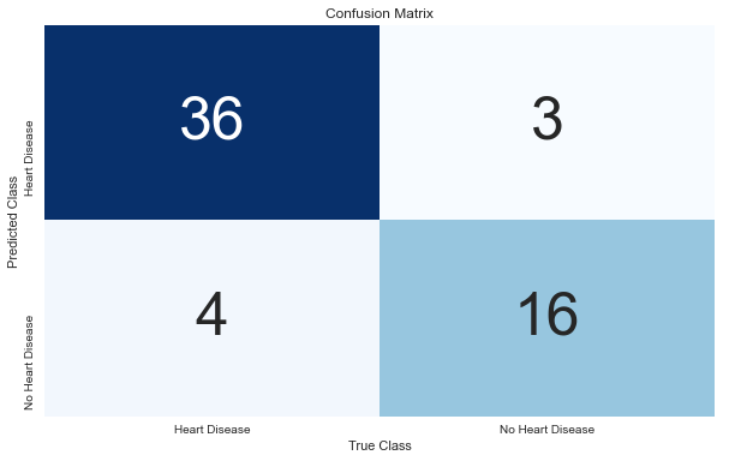


Our curve is between these two which is what we would expect and want.

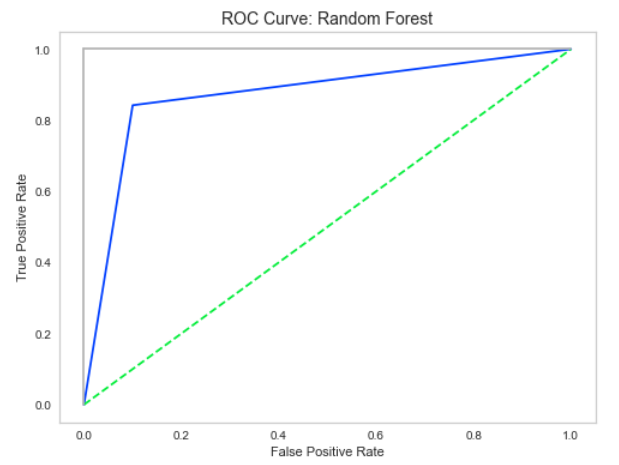
Let’s do a different model now, just to see if we can predict who has heart disease a little better. We are now going to try a random forest model. We can set this up and we get a result that is better than the logistic regression model we just set up. Random forest sets up decision trees to break down the data and find the most likely outcome based on breaking it all down and combining the results of each tree. When we build the model, we end up getting results that look better than our first:



As we can see, our model performed very well when run and will likely average a bit lower when run repeatedly. Our confusion matrix looks like:



Which is great, and our ROC curve looks like:



And that looks fine. This ended up being a much better model for our data.

This case study helped us explore some of the more readily available health information we could have on a patient and are able to reasonably estimate that a patient has or doesn’t have heart disease. At the very least, we could take these into account and hopefully catch the risk before heart issues become a problem. By taking these into account we can estimate if someone has heart disease and build a model that could find if someone has or doesn’t have heart disease at approximately 87% accuracy, precision and recall. Overall our model is good and should help with predictions and catching these problems before they get worse.