Simplifying the ROC and AUC metrics.

Taking the confusion out of classification metrics.

Parul Pandey Follow
Mar 3, 2019 · 8 min read

Photo by Daniele Levis Pelusi on Unsplash

> “The definition of genius is taking the complex and making it simple.”– Albert Einstein

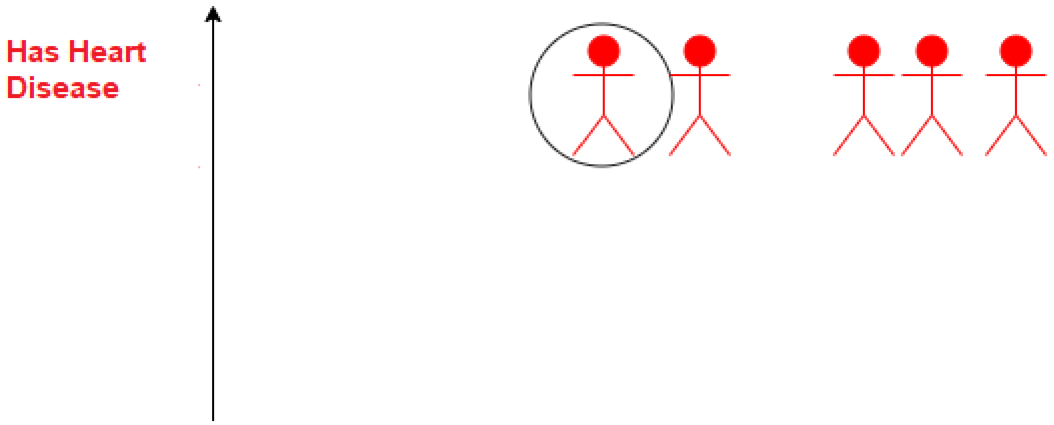
ROC and AUC curves are important evaluation metrics for calculating the performance of any classification model. These definitions and jargons are pretty common in the Machine learning community and are encountered by each one of us when we start to learn about classification models. However, most of the times they are not completely understood or rather misunderstood and their real essence cannot be utilized. Under the hood, these are very simple calculation parameters which just needs a little demystification.

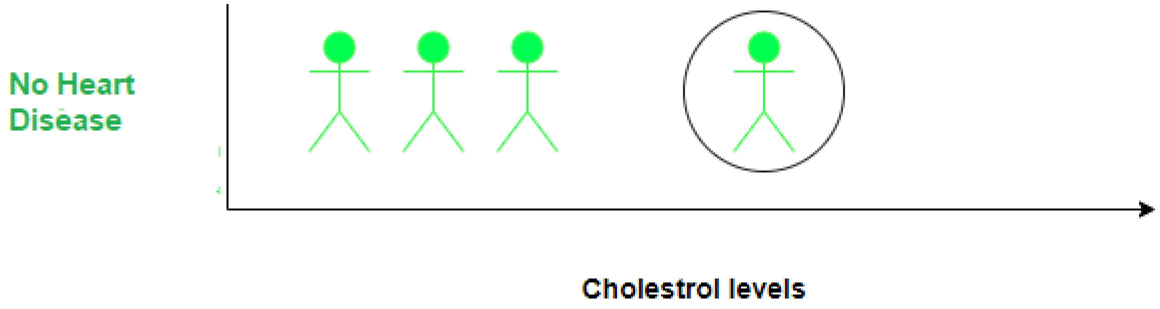
. . .

> *The concept of ROC and AUC builds upon the knowledge of Confusion Matrix, Specificity and Sensitivity. Also, the example that I will use in this article is based on Logisitic Regression algorithm, however, it is important to keep in mind that the concept of ROC and AUC can apply to more than just Logistic Regression.*

. . .

Consider a **hypothetical example** containing a group of people. The y-axis has two categories i.e `Has Heart Disease` represented by red people and `does not have Heart Disease` represented by green circles. Along the x-axis, we have **cholesterol** levels and the classifier tries to classify people into two categories depending upon their cholesterol levels.





Something to notice:

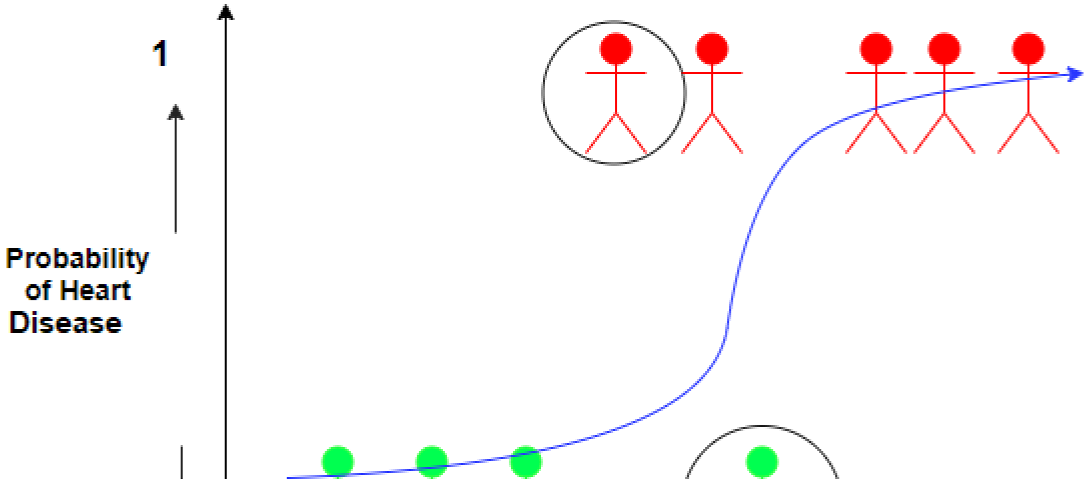
- Circled Green person has a high level of cholesterol but does not have heart disease. This may be due to the reason that now the person is observing a better lifestyle and exercising regularly.
- Circled Red person has low cholesterol levels still had a heart attack. This may be due to the reason that he has other heart-related issues.

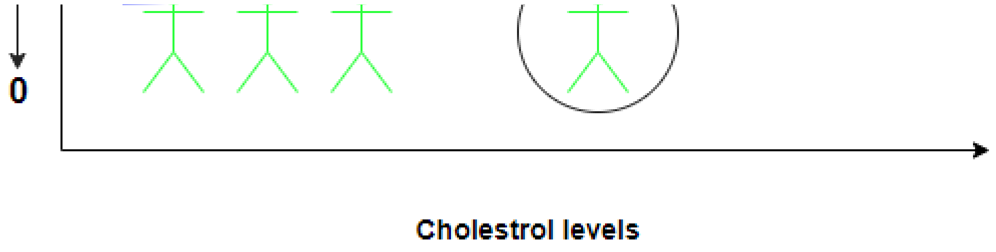
This is a hypothetical example so the reasons are also hypothetical 😃

. . .

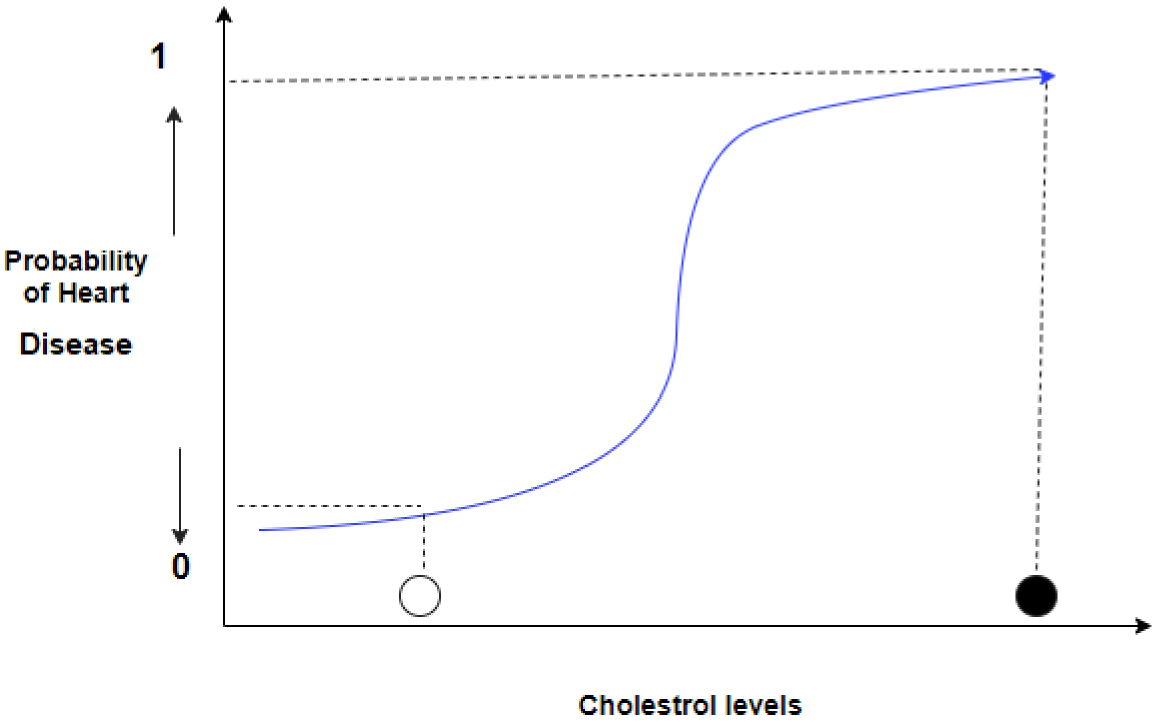
Logistic Regression

Now if we fit a Logistic Regression curve to the data, the Y-axis will be converted to the **Probability** of a person having a heart disease based on the Cholesterol levels.

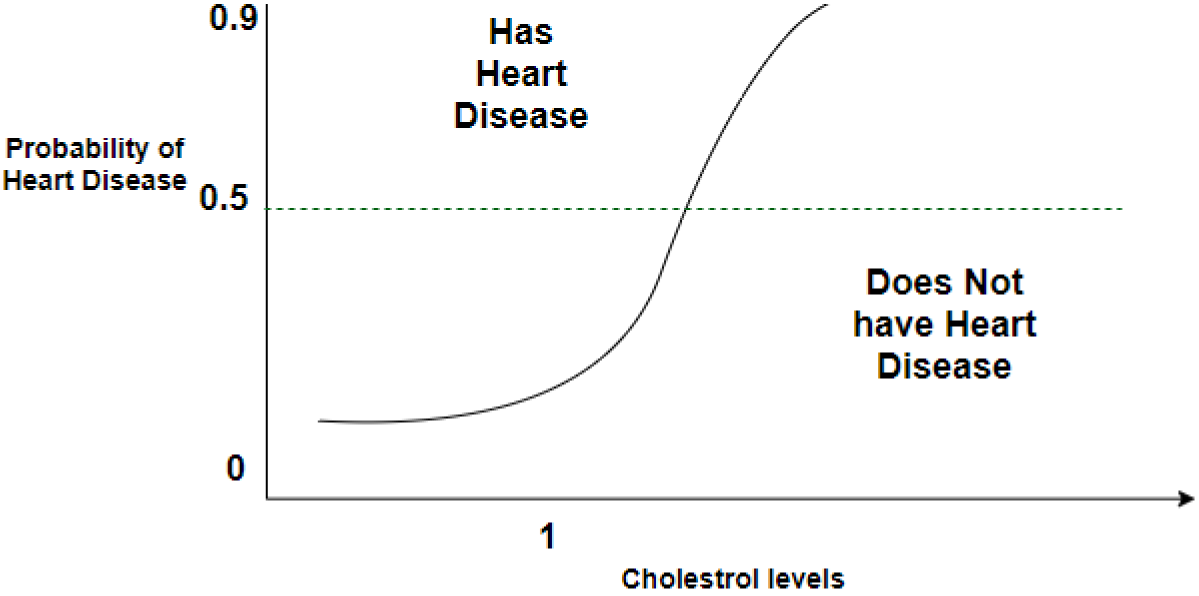




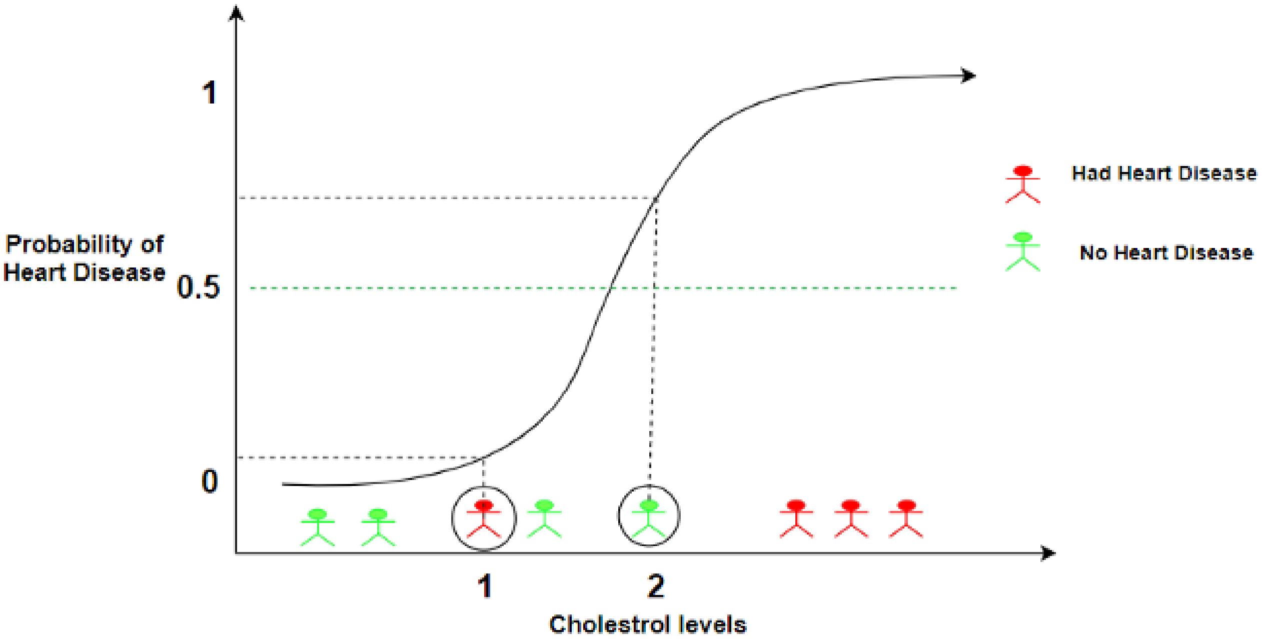
The white dot represents a person having a lower heart disease probability than the person represented by the black dot.



However, if we want to classify the people in the two categories, we need a way to turn probabilities into classifications. One way is to set a threshold at 0.5. Next, classify the people who have a probability of heart disease > 0.5 as "**having a heart disease**" and classify the people who have a probability of heart disease < 0.5 as " **not having a heart disease**".



Let us now evaluate the effectiveness of this logistic regression with the classification threshold set to 0.5, with some new people about whom we already know if they have heart disease or not.



Our Logistic Regression model correctly classifies all people except the persons 1 and 2.

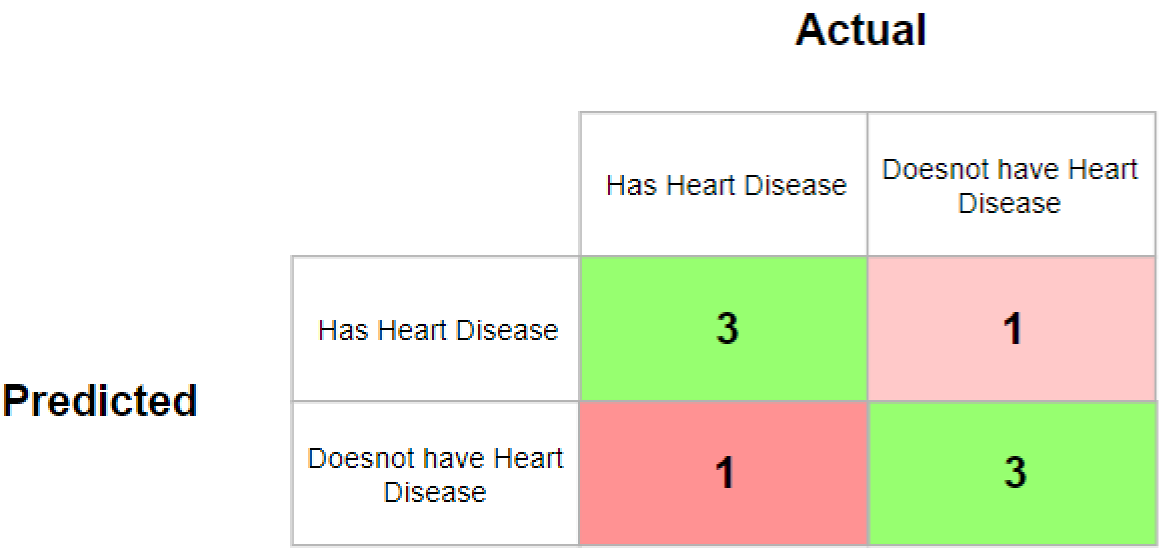
- We know Person 1 has heart disease but our model classifies it as otherwise.

- We also know person 2 doesn't have heart disease but again our model classifies it incorrectly.

. . .

Confusion Matrix

Let's create a Confusion Matrix to summarize the classifications.



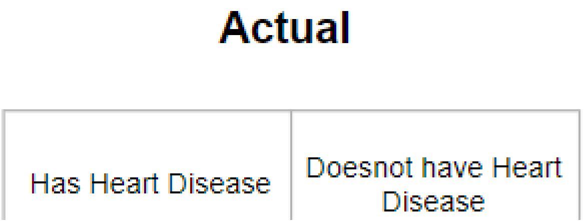
		Actual	
		Has Heart Disease	Doesnot have Heart Disease
Predicted	Has Heart Disease	3	1
	Doesnot have Heart Disease	1	3

Once the confusion matrix is filled in, we can calculate the **Sensitivity** and the **Specificity** to evaluate this logistic regression at 0.5 threshold.

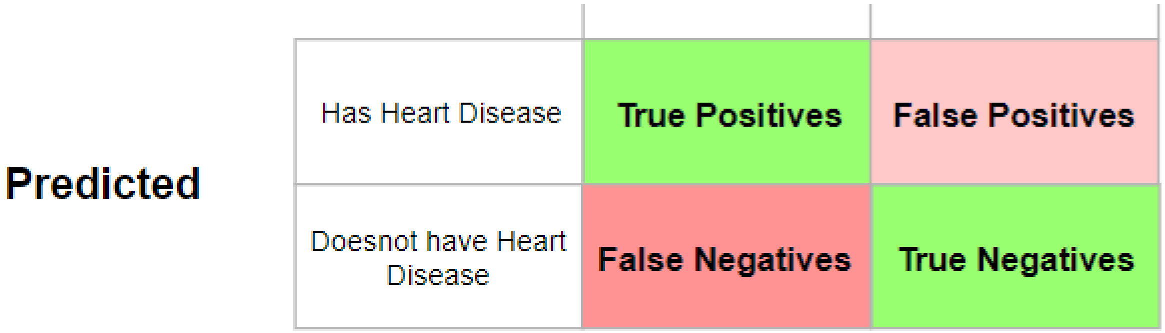
. . .

Specificity and Sensitivity

In the above confusion matrix, let's replace the numbers with what they actually represent.



Actual	
Has Heart Disease	Doesnot have Heart Disease



Predicted	Has Heart Disease	True Positives	False Positives
	Doesnot have Heart Disease	False Negatives	True Negatives

- **True Positives (TP):** People who *had heart disease* and were also predicted to have heart disease.
- **True negatives (TN):** People who *did not have heart disease* and were also predicted to not have heart disease.
- **False negatives (FN):** People who have heart disease but the prediction says they don't.
- **False positives (FP):** People who *did not have heart disease* but the prediction says they do.

We can now calculate two useful metrics based upon the confusion matrix:

Sensitivity

Sensitivity tells us what percentage of people *with* heart disease were actually correctly identified.

$$\text{sensitivity} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}}$$

This turns out to be: 3/3+1 = 0.75

This tells us that **75%** of people with heart disease were correctly identified by our model.

Specificity

Specificity tells us what percentage of people *without* heart disease were actually correctly identified.

$$\text{specificity} = \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}}$$

This turns out to be: 3/3+1 = 0.75

This tells us that again **75%** of people *without heart disease* were correctly identified by our model.

If correctly identifying positives is important for us, then we should choose a model with higher Sensitivity. However, if correctly identifying negatives is more important, then we should choose specificity as the measurement metric.

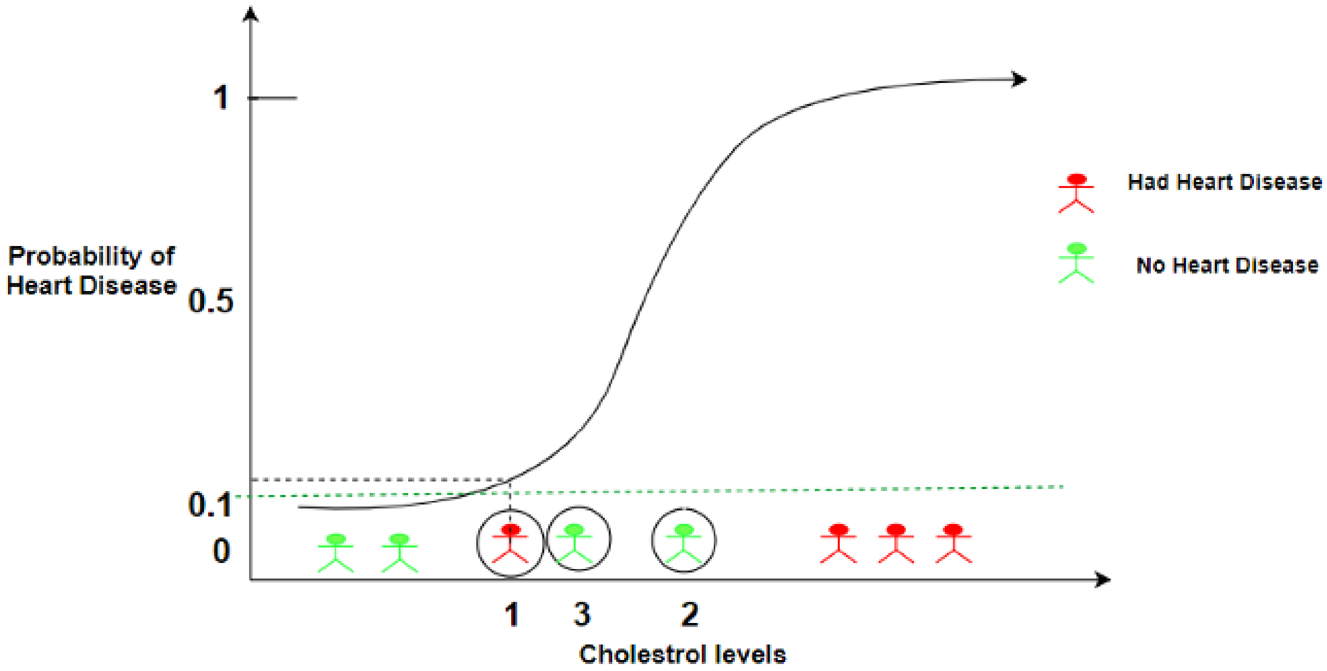
. . .

Identifying the Correct Thresholds

Now, let's talk about what happens when we use a different threshold for deciding if a person has heart disease or not.

- **Setting the Threshold to 0.1**

This would correctly identify all people who have heart disease. The person labeled 1 is also correctly classified to be a heart patient.

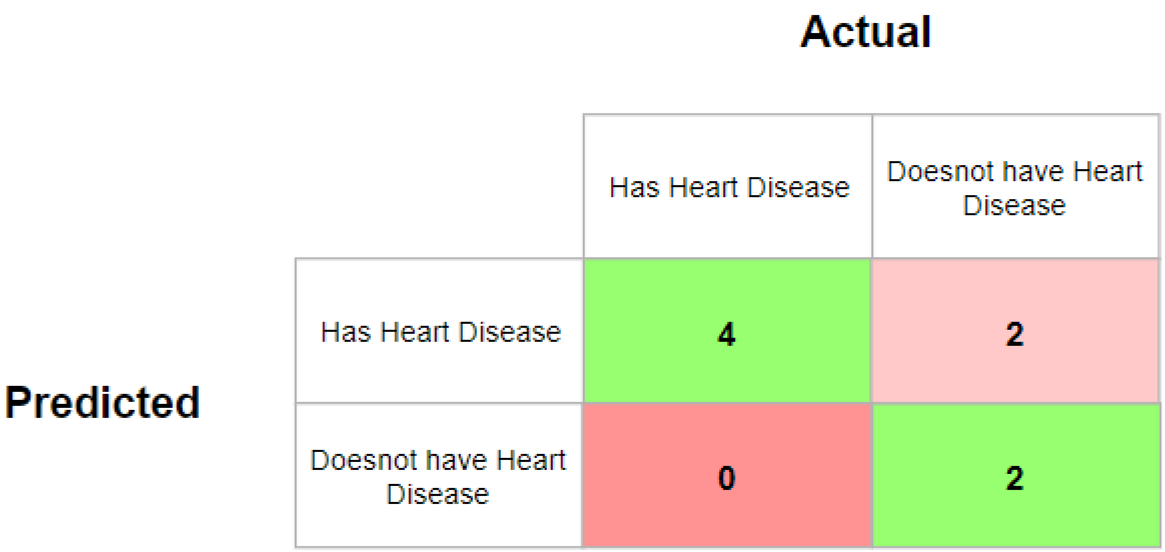


However, it would also increase the number of False Positives since now person 2 and 3 will be wrongly classified as having heart disease.

Therefore a lower threshold:

- Increases the number of False Positives
- Decreases the number of False Negatives.

Recalculating the confusion matrix :

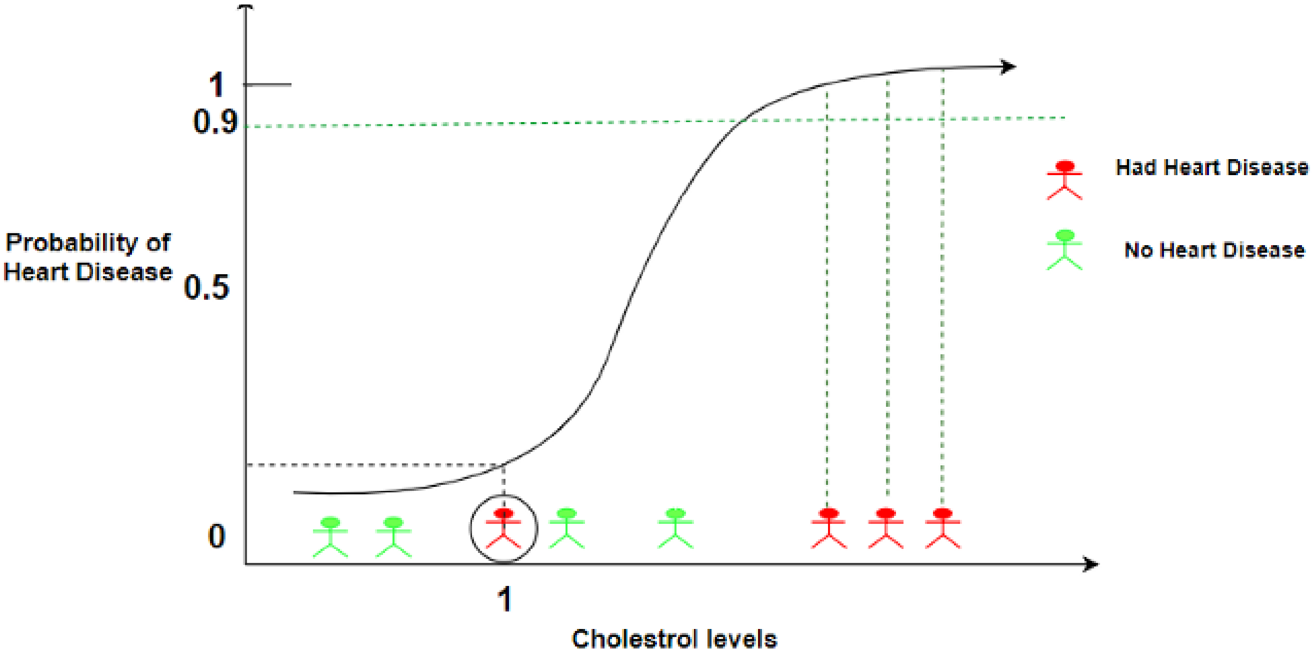


		Actual	
		Has Heart Disease	Doesnot have Heart Disease
Predicted	Has Heart Disease	4	2
	Doesnot have Heart Disease	0	2

In this case, it becomes important to identify people having a heart disease correctly so that the corrective measures can be taken else heart disease can lead to serious complications. This means lowering the threshold is a good idea even if it results in more False Positive cases.

- **Setting the Threshold to 0.9**

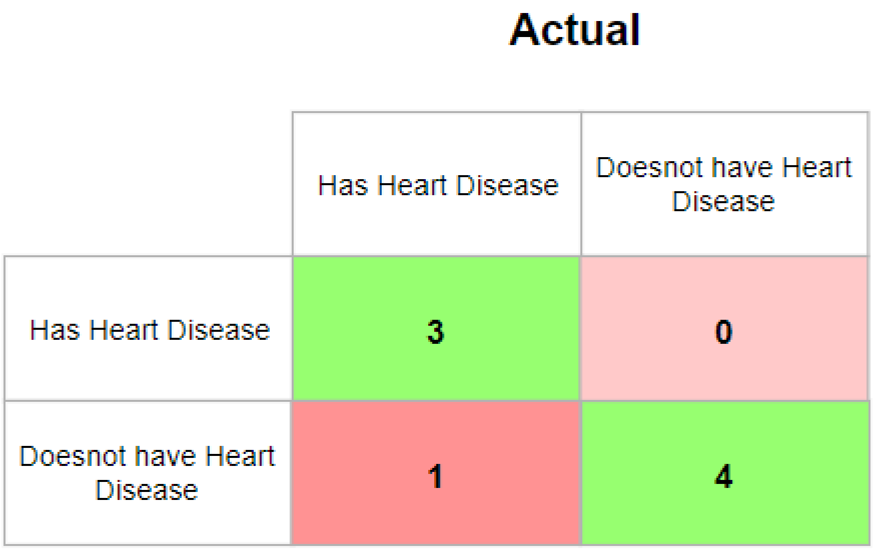
This would now correctly identify all people who do not have heart disease. The person labeled, however, person 1, would be incorrectly classified having no heart disease.



Therefore a lower threshold:

- Decreases the number of False Positives
- Increases the number of False Negatives.

Recalculating the confusion matrix :



Actual

	Has Heart Disease	Doesnot have Heart Disease
Has Heart Disease	3	0
Doesnot have Heart Disease	1	4

The threshold could be set to any value between 0 and 1. So how do we determine which threshold is the best? Do we need to experiment with all the threshold values? Every threshold results in a different confusion matrix and a number of thresholds will result in a large number of confusion matrices which is not the best way to work.

. . .

ROC Graphs

ROC(Receiver Operator Characteristic Curve) can help in deciding the best threshold value. It is generated by plotting the True Positive Rate (y-axis) against the False Positive Rate (x-axis).

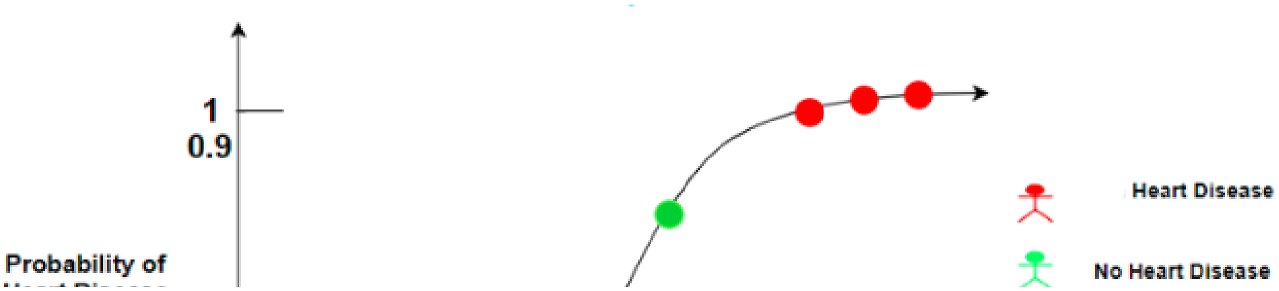
$$\text{True Positive Rate} = \text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

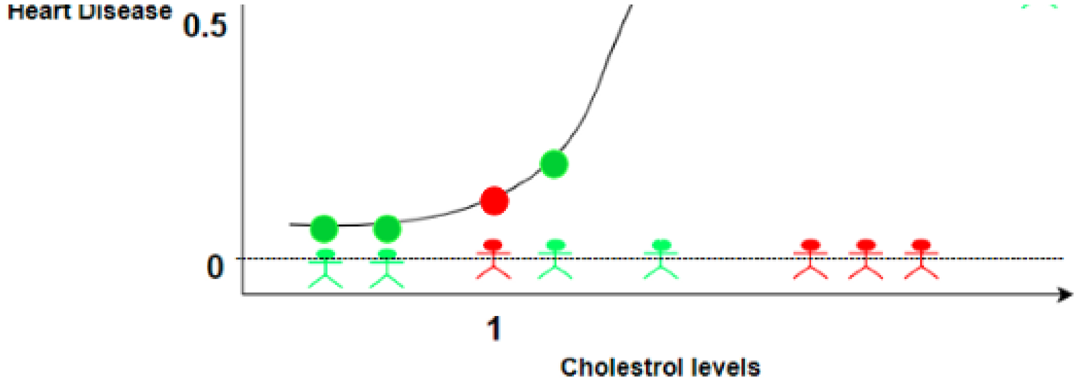
$$\text{False Positive Rate} = (1 - \text{Specificity}) = \frac{\text{False Positives}}{\text{False Positives} + \text{True Negatives}}$$

> *True Positive Rate indicates what proportion of people '**with heart disease**' were correctly classified.*
>
> *False Positive Rate indicates the proportion of people classified as '**not having heart disease**', that are False Positives.*

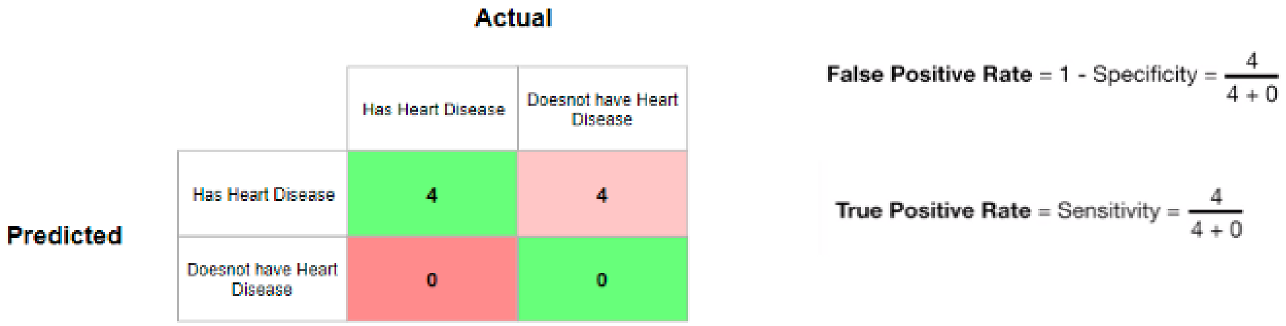
To get to know the ROC better, let's draw one from scratch.

- **Threshold classifying all people as having heart disease.**





The confusion matrix will be:



Actual

		Has Heart Disease	Doesnot have Heart Disease
Predicted	Has Heart Disease	4	4
	Doesnot have Heart Disease	0	0

False Positive Rate = 1 - Specificity = $\frac{4}{4+0}$

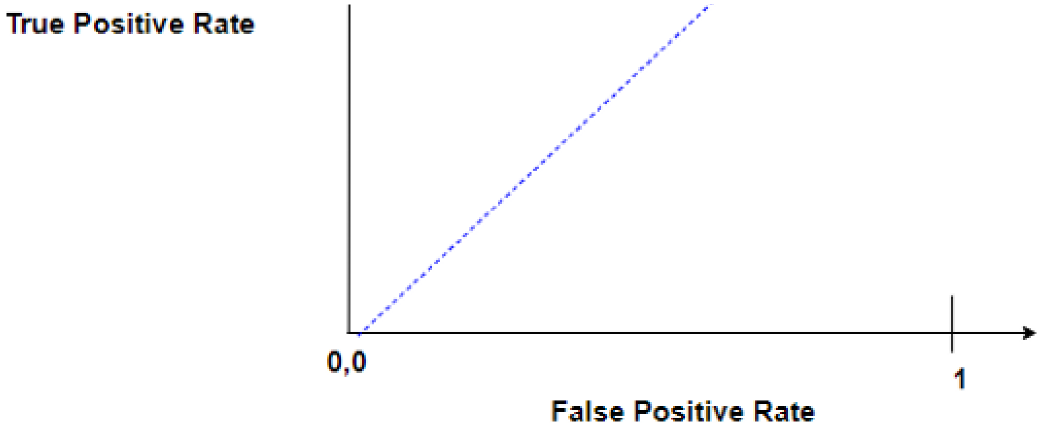
True Positive Rate = Sensitivity = $\frac{4}{4+0}$

This means the True Positive Rate when the threshold is so low that every single person is classified as having heart disease, is 1. This means that every single person with heart disease was correctly classified.

Also, the False Positive Rate when the threshold is so low that every single person is classified as having heart disease, is also 1. This means that every single person without heart disease was wrongly classified.

Plotting this point on the ROC graph:

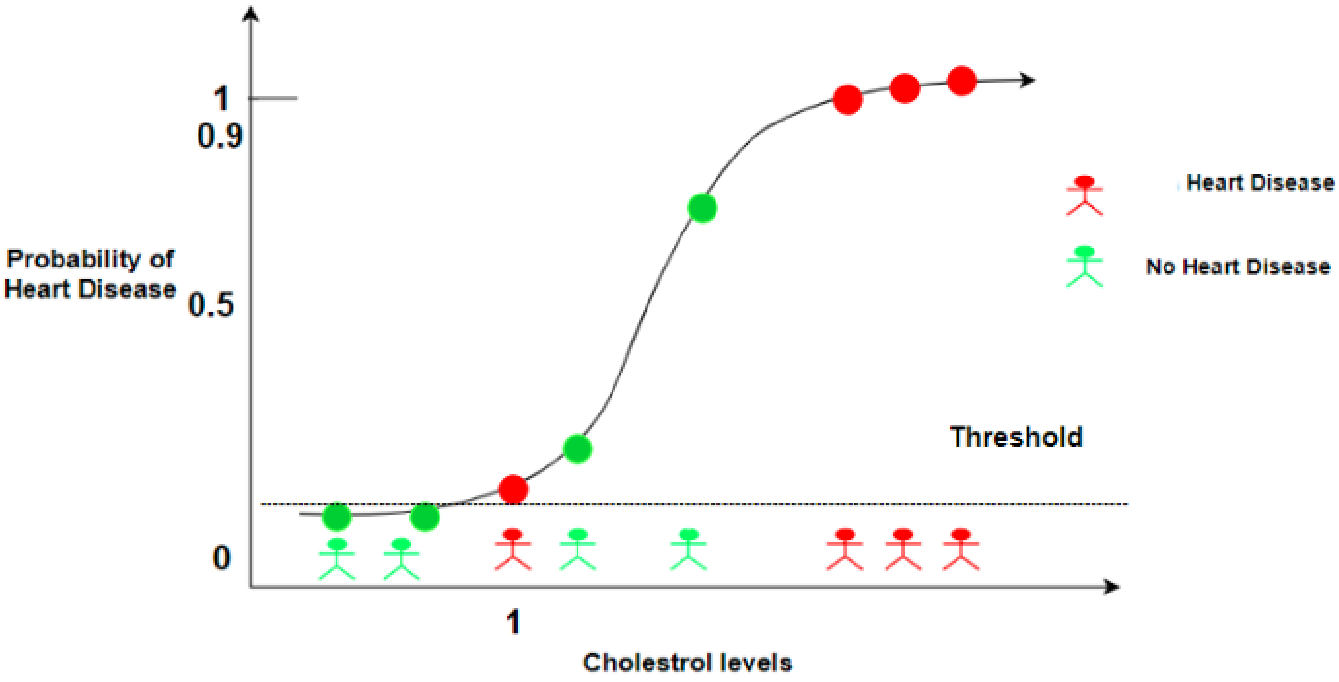
At point(1,1), we correctly classified all heart patients but also incorrectly classified healthy patients as heart patients



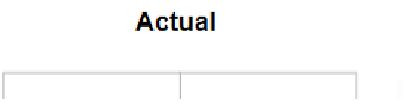
Diagonal Line :True Postive Rate = False Positive Rate

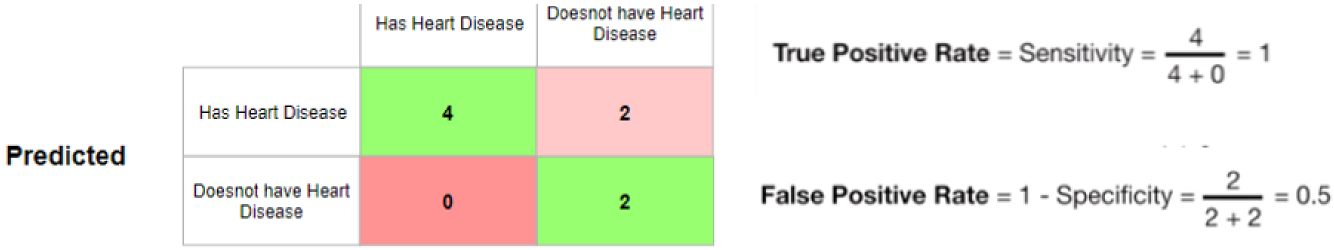
Any point on the Blue Diagonal Lines means that the proportion of correctly classified samples is equal to the proportion of incorrectly classified samples.

- **Increasing the Threshold slightly so that only the two people with the least cholesterol value are below the threshold.**



The confusion matrix will be:


Actual

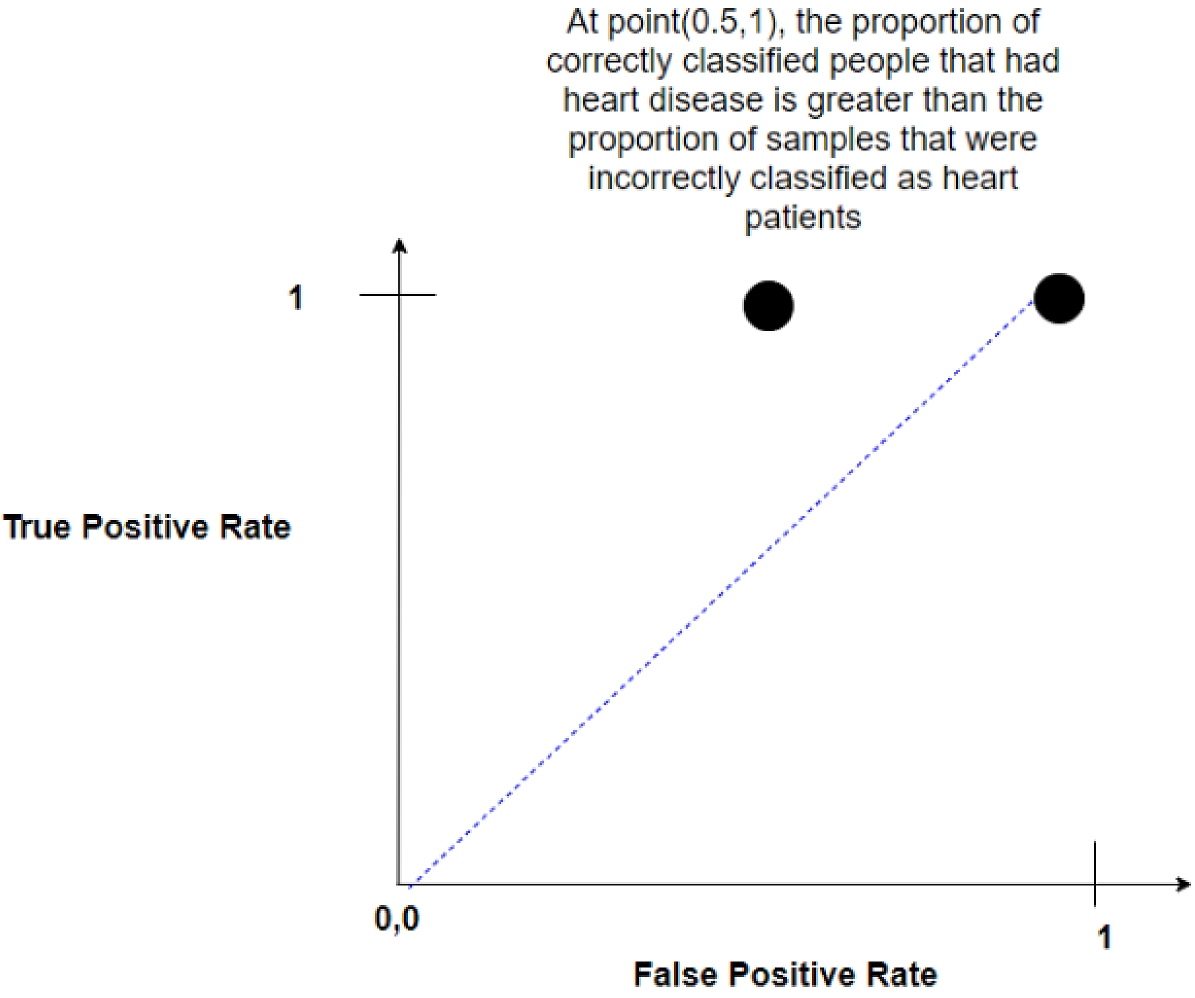


		Has Heart Disease	Doesnot have Heart Disease
Predicted	Has Heart Disease	4	2
	Doesnot have Heart Disease	0	2

$$\textbf{True Positive Rate} = \text{Sensitivity} = \frac{4}{4+0} = 1$$

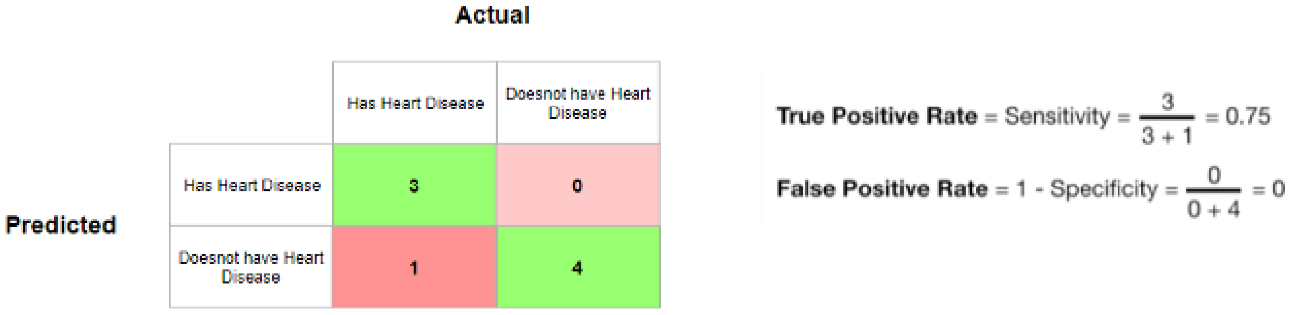
$$\textbf{False Positive Rate} = 1 - \text{Specificity} = \frac{2}{2+2} = 0.5$$

Let’s plot this point (0.5,1) on the ROC graph.



This means this threshold is better than the previous one.

- **Now if go on increasing the threshold values, and reach a point where we get the following confusion matrix:**

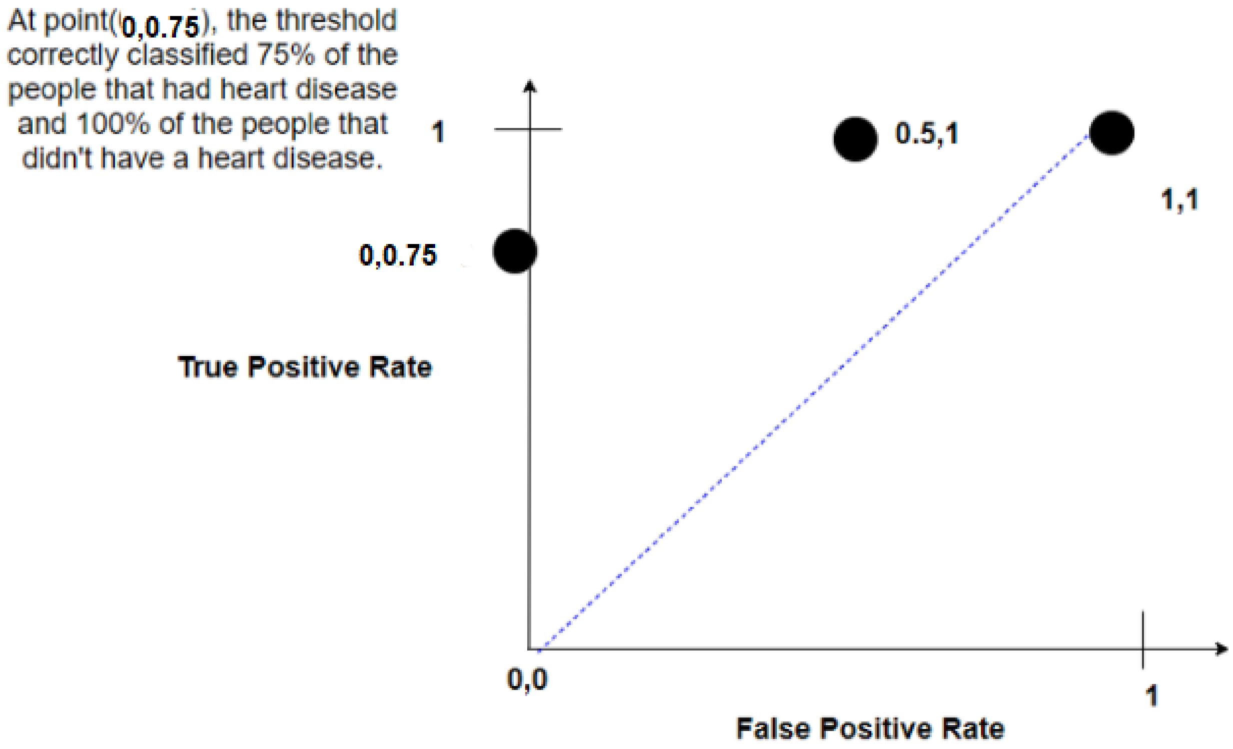


		Actual	
		Has Heart Disease	Doesnot have Heart Disease
Predicted	Has Heart Disease	3	0
	Doesnot have Heart Disease	1	4

True Positive Rate = Sensitivity = $\frac{3}{3+1}$ = 0.75

False Positive Rate = 1 - Specificity = $\frac{0}{0+4}$ = 0

Let's plot this point (0,0.75) on the ROC graph.

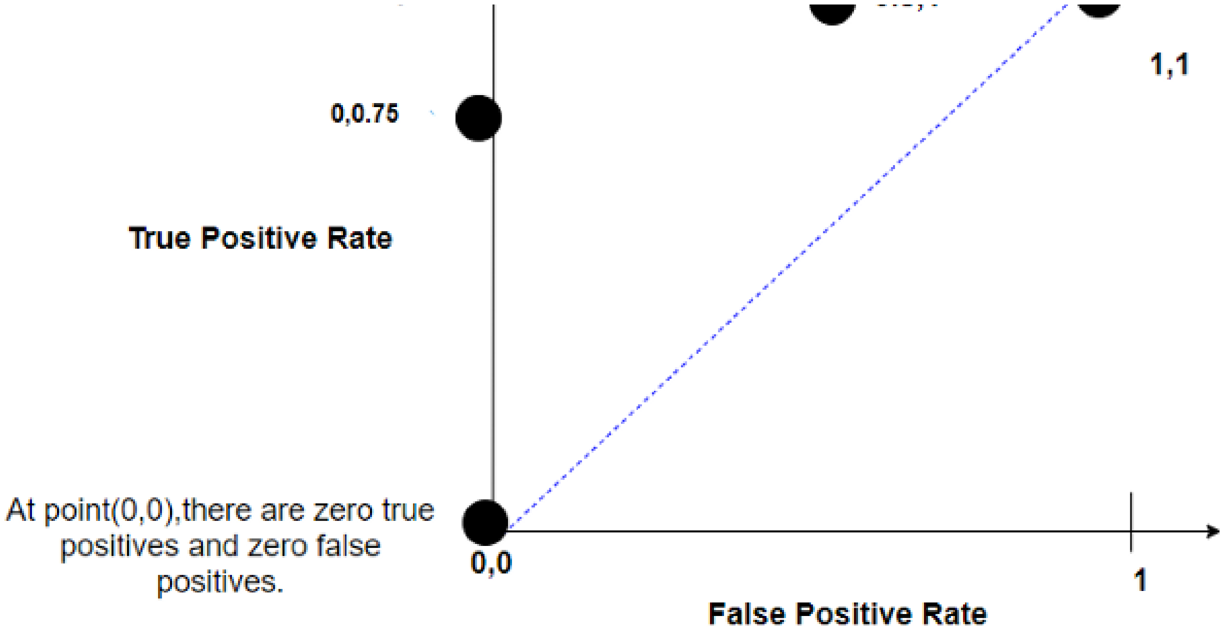


At point(0,0.75), the threshold correctly classified 75% of the people that had heart disease and 100% of the people that didn't have a heart disease.

By far this is the best threshold that we have got since it predicted no false positives.

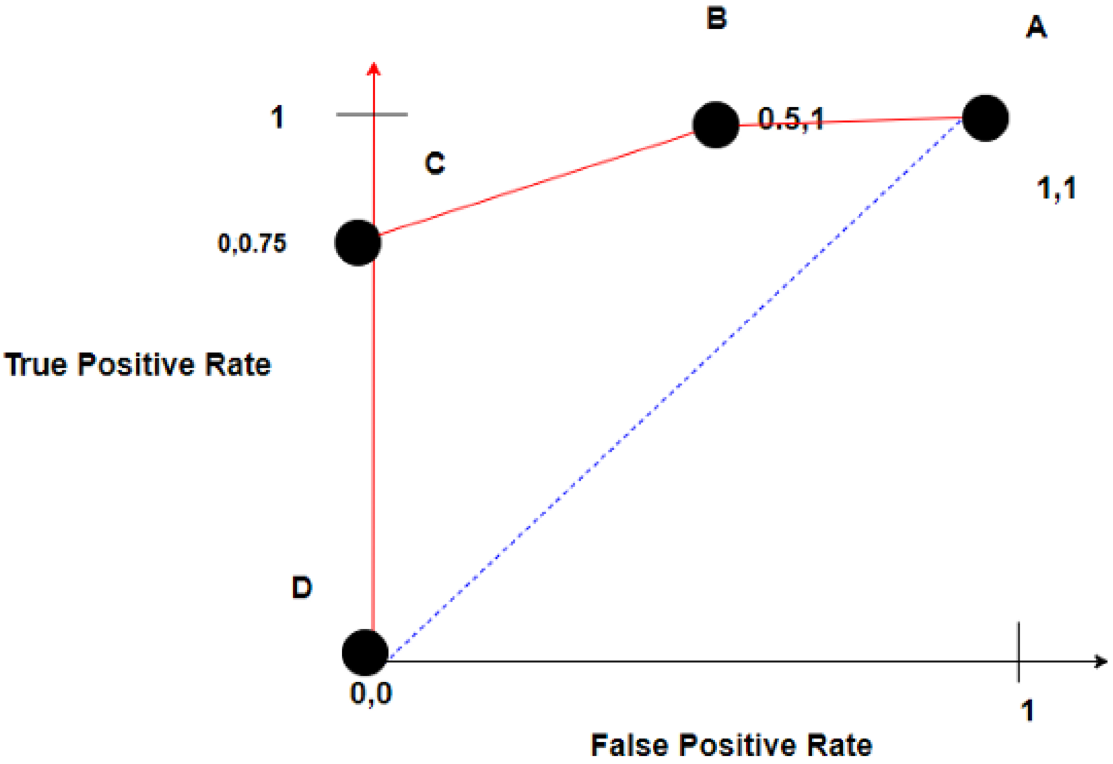
- **Lastly, we choose a threshold where we classify all people as not having a heart disease i.e Threshold of 1.**

The graph, in this case, would be at (0,0):



At point(0,0),there are zero true positives and zero false positives.

We can then connect the dots which gives us a ROC graph. The ROC graph summarises the confusion matrices produced for each threshold without having to actually calculate them.

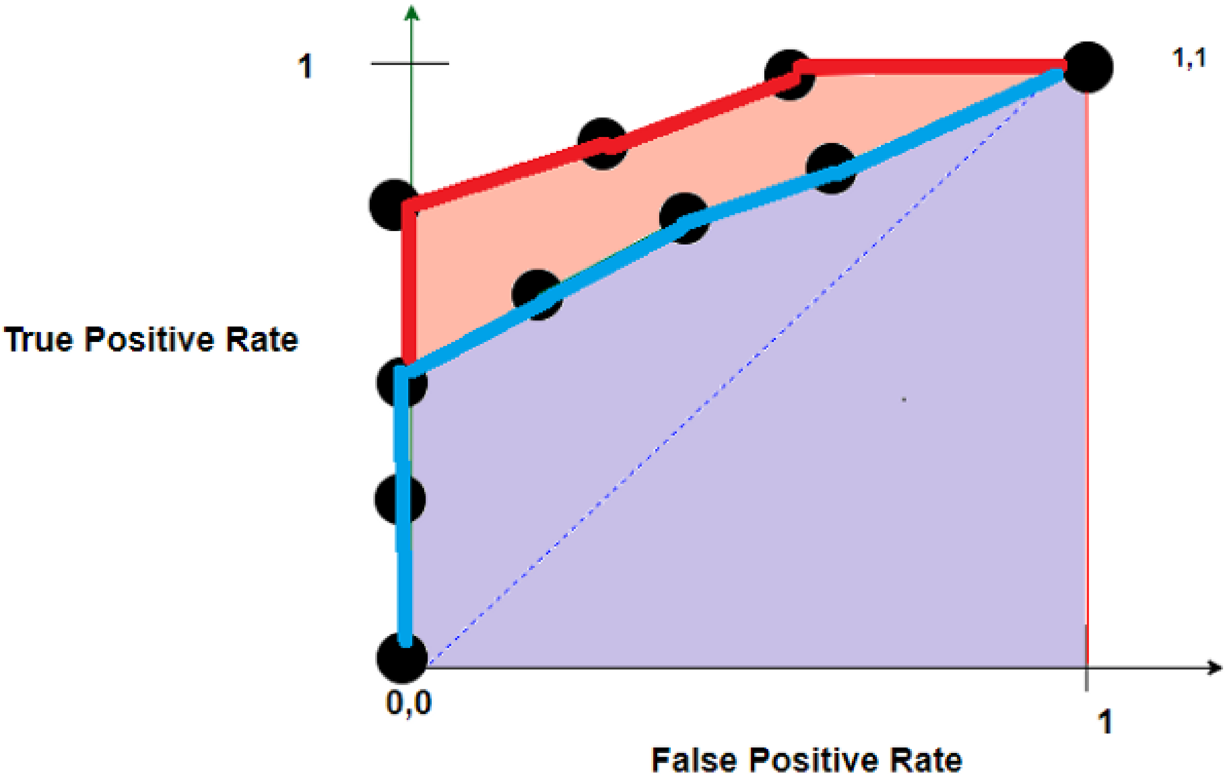


Just by glancing over the graph, we can conclude that threshold C is better than threshold B and depending on how many False Positives that we are willing to accept, we can choose the optimal threshold.

. . .

AUC

AUC stands for **Area under the curve**. AUC gives the rate of successful classification by the logistic model. The AUC makes it easy to compare the ROC curve of one model to another.



The **AUC** for the red **ROC** curve is greater than the **AUC** for the blue **ROC** curve. This means that the Red curve is better. If the Red ROC curve was generated by say, a Random Forest and the Blue ROC by Logistic Regression we could conclude that the Random classifier did a better job in classifying the patients.

. . .

Conclusion

AUC and ROC are important evaluation metrics for calculating the performance of any classification model's performance. Therefore getting to know how they are calculated is as essential as using them. Hopefully, next time when you encounter these terms, you will be able to explain them easily in the context of your problem.

. . .

Reference

The article is an adaptation of this excellent video by Josh Starmer on ROC and AUC. I'll recommend you to watch this video for more clarity and many other videos on various statistics and ML for more clarity.

Machine Learning Data Science Statistics Classification Metrics