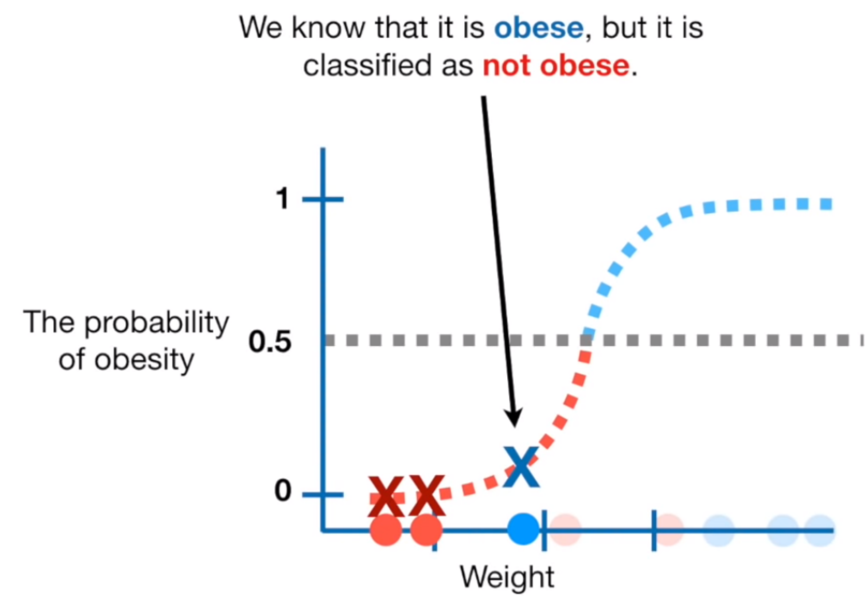
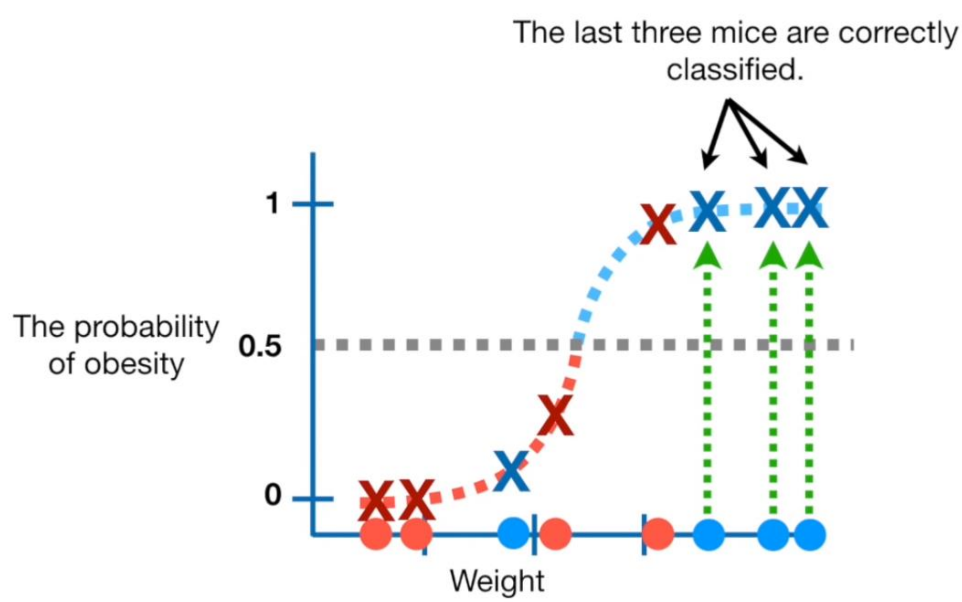
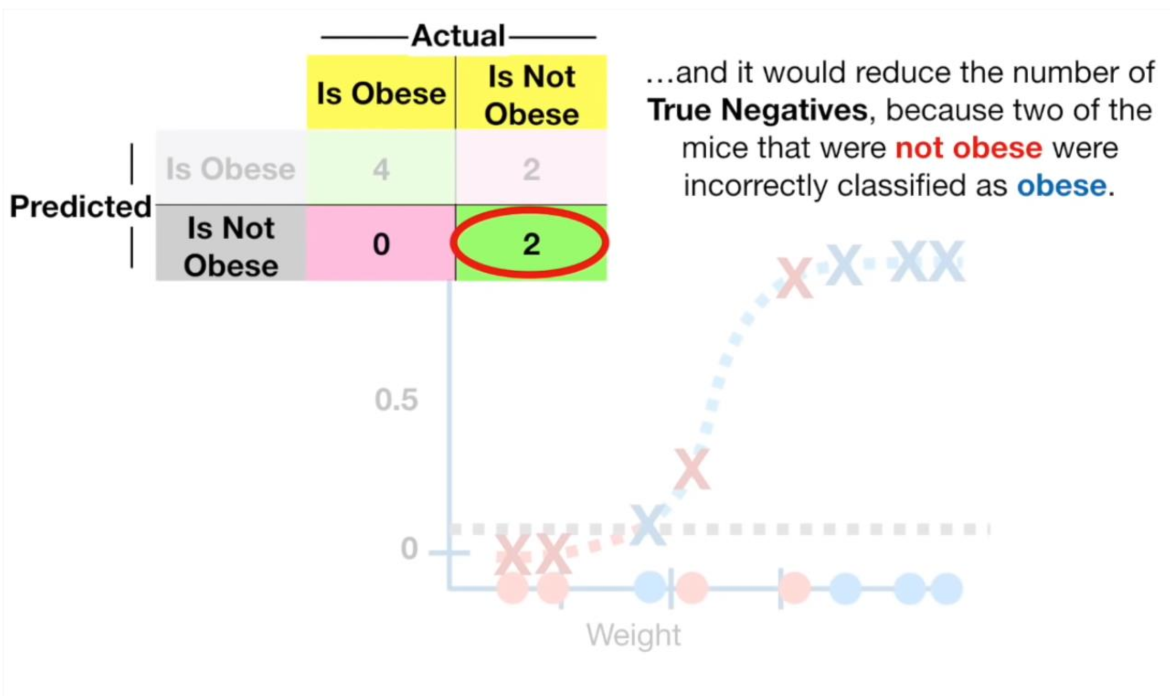
AUC AND ROC

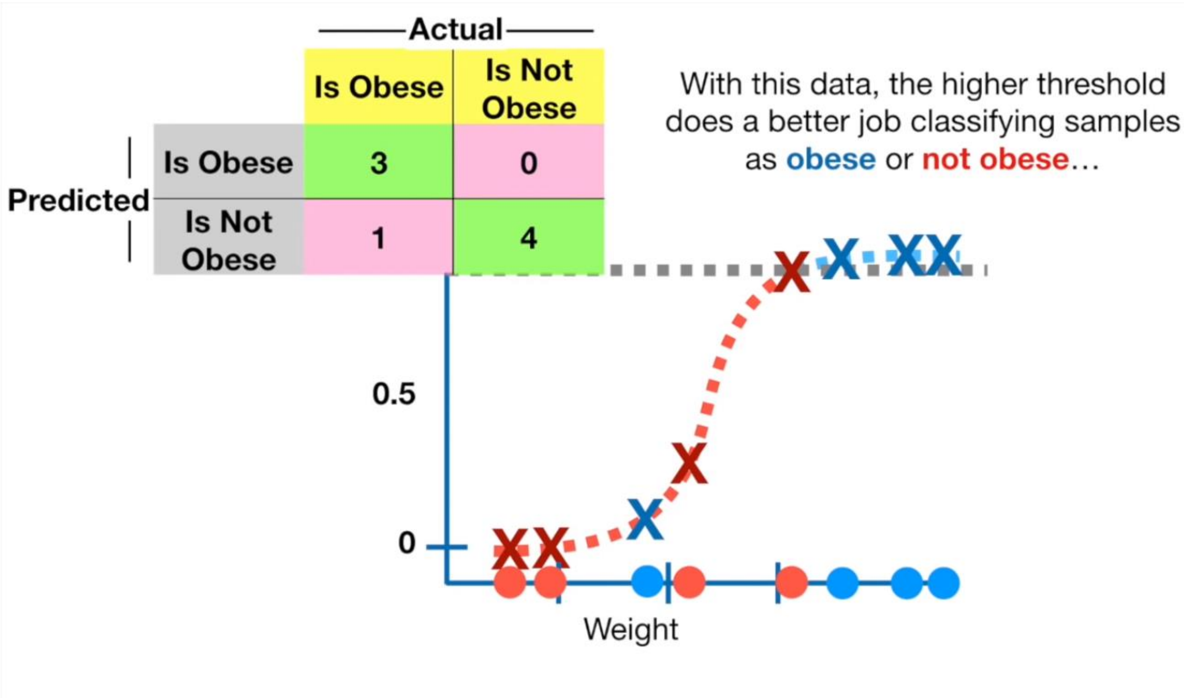
A complete explanation with plotting





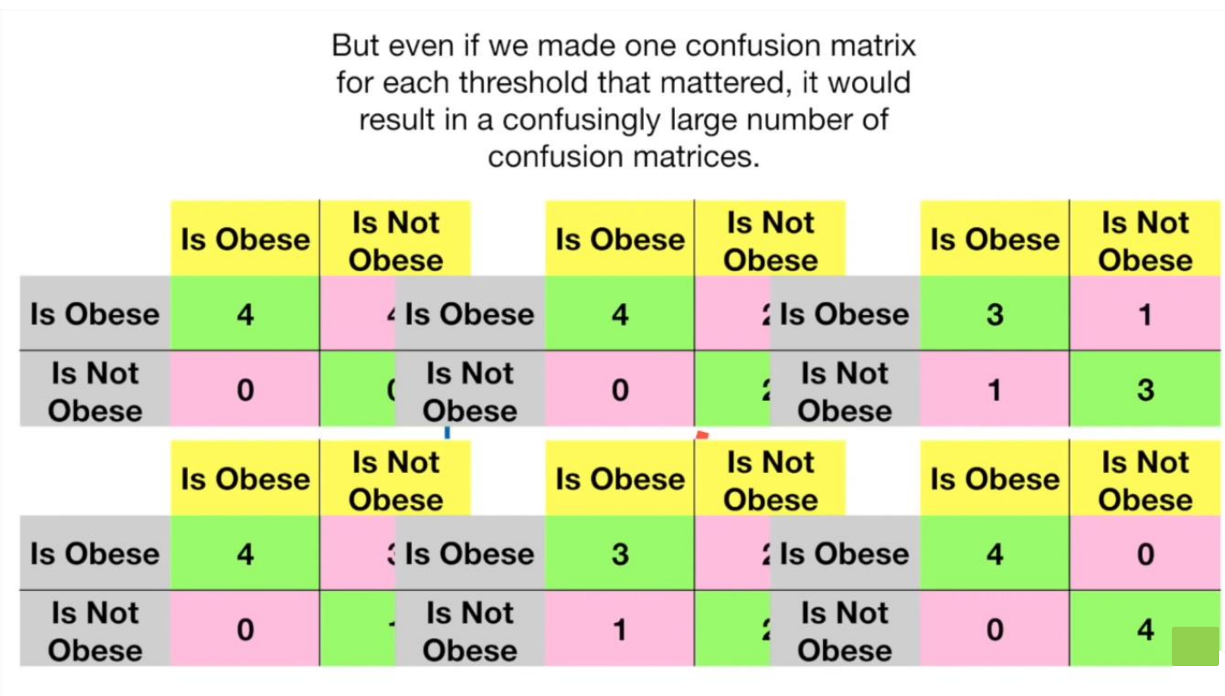
If we change the threshold frequency:





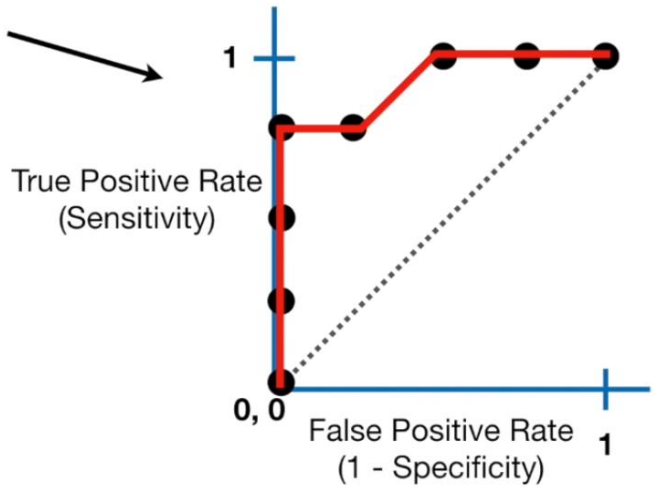
Our main aim is to show the best threshold so that, we would don’t get Type errors.

So making confusion matrix for each threshold is not valid like below:



Rather than doing all the calculations, we for a technique called **ROC**, **AUC**.

So instead of being overwhelmed with confusion matrices, **Receiver Operator Characteristic** (**ROC**) graphs provide a simple way to summarize all of the information.



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

The **True Positive Rate**, when the threshold is so low that every single sample is classified as **obese**, is **1**.

		Actual: Is Obese	Actual: Is Not Obese
Predicted	Is Obese	4	4
	Is Not Obese	0	0

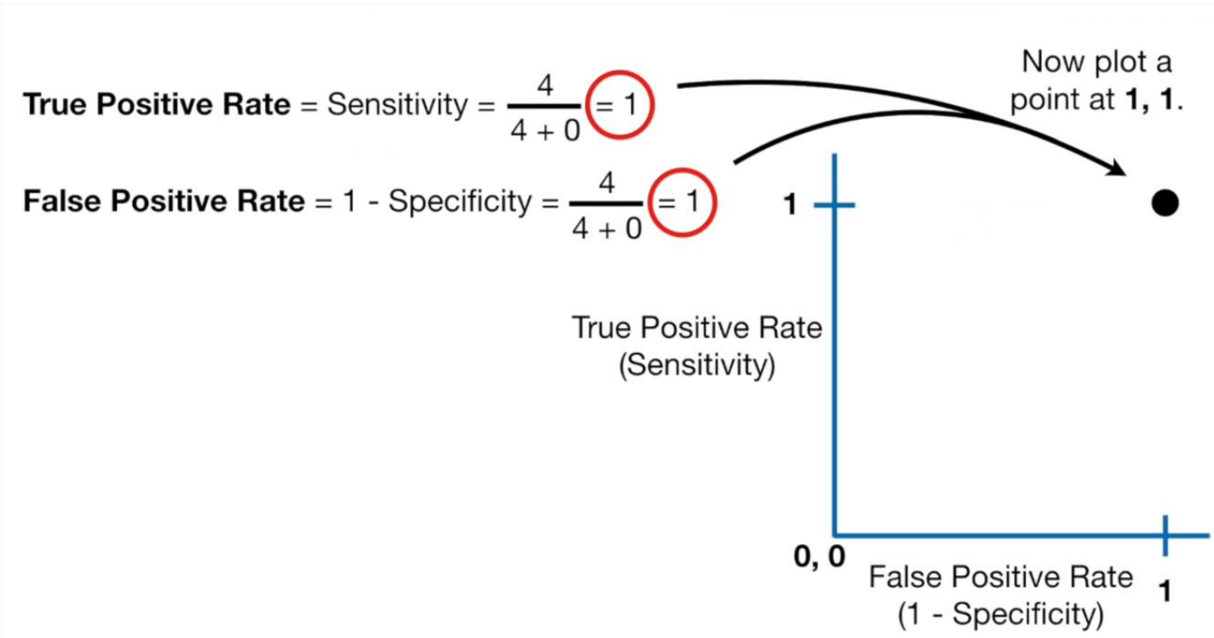
True Positive Rate = Sensitivity = $\frac{4}{4+0} = 1$

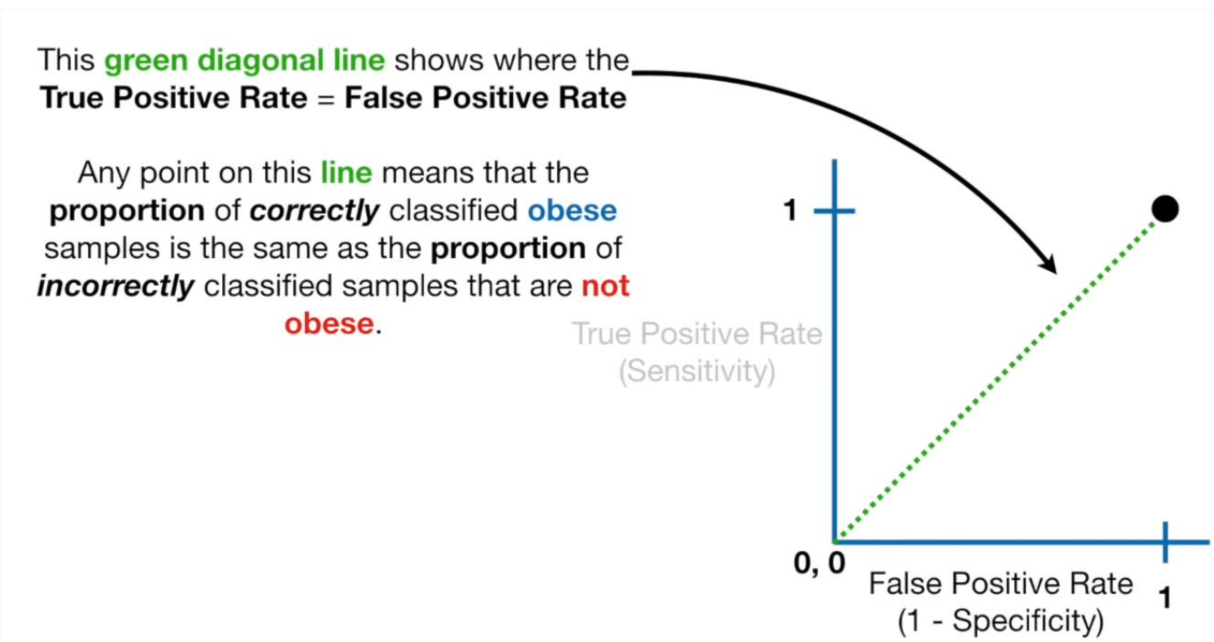
False Positive Rate = 1 - Specificity = $\frac{4}{4+0} = 1$

The **False Positive Rate**, when the threshold is so low that every single sample is classified as **obese**, is also **1**.

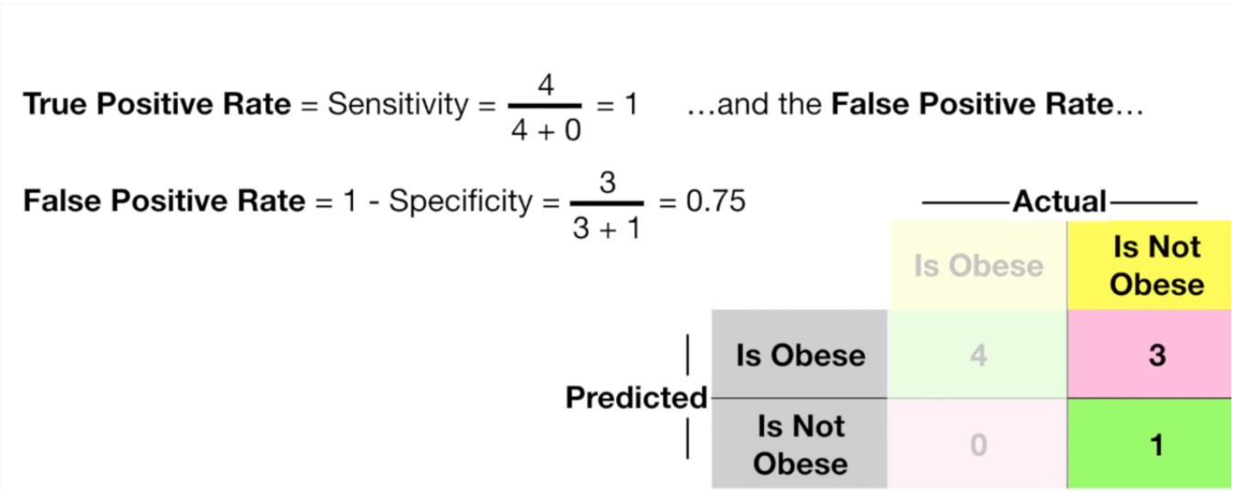
		Actual: Is Obese	Actual: Is Not Obese
Predicted	Is Obese	4	4
	Is Not Obese	0	0

Now let's go to plot the point (1, 1)

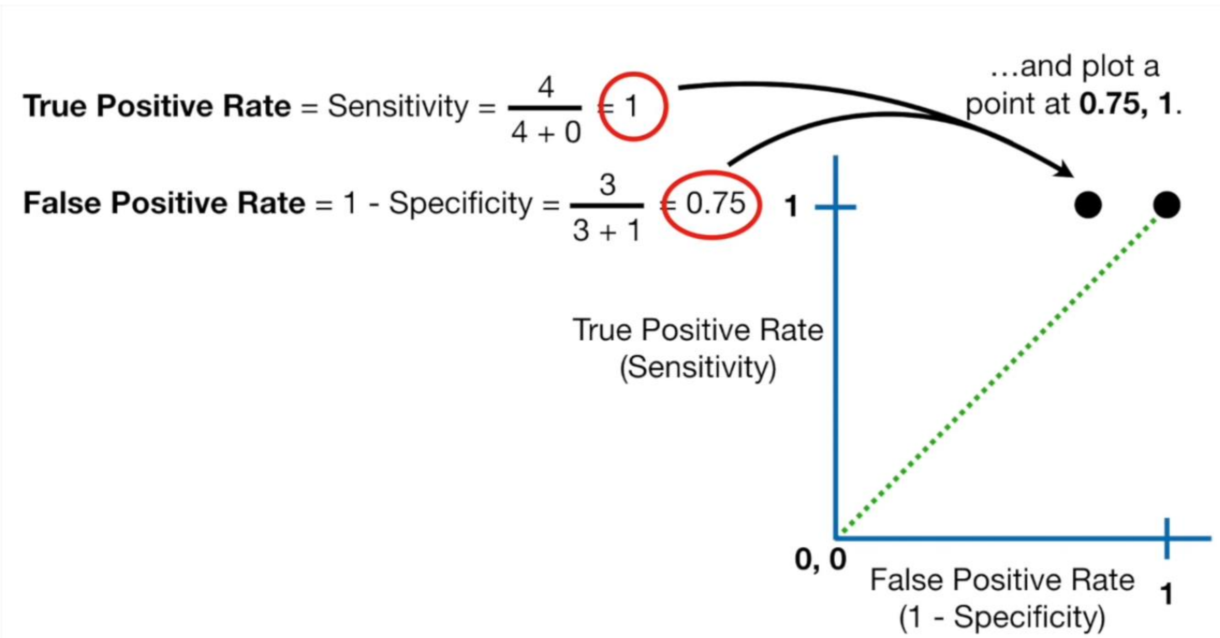




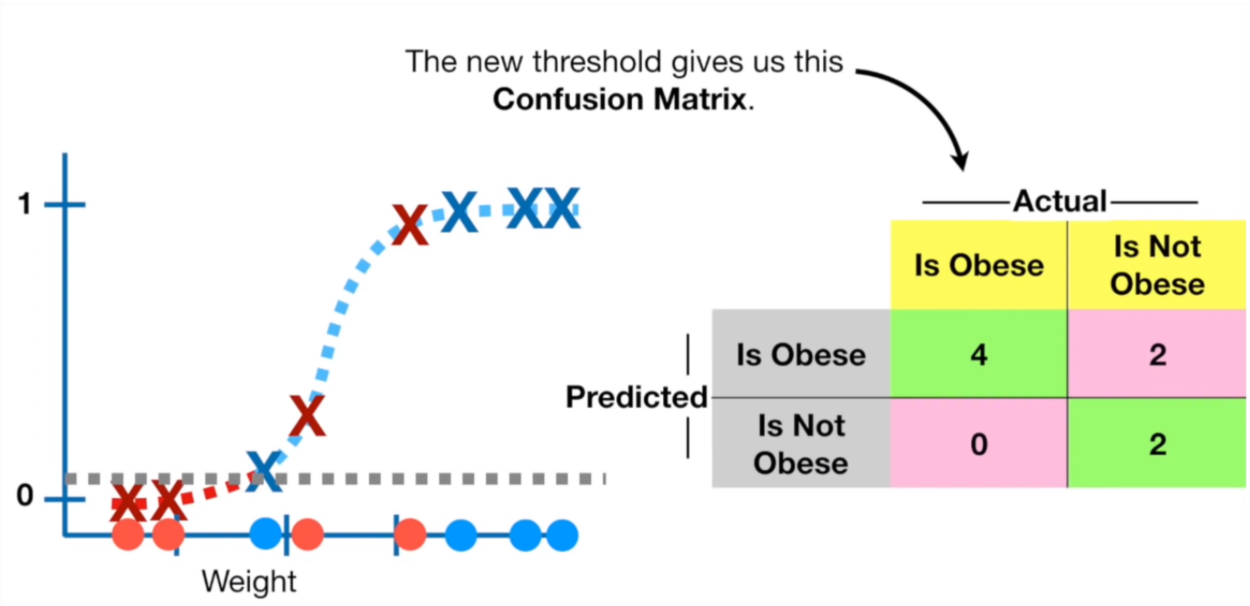
For Another confusion matrix of different threshold:

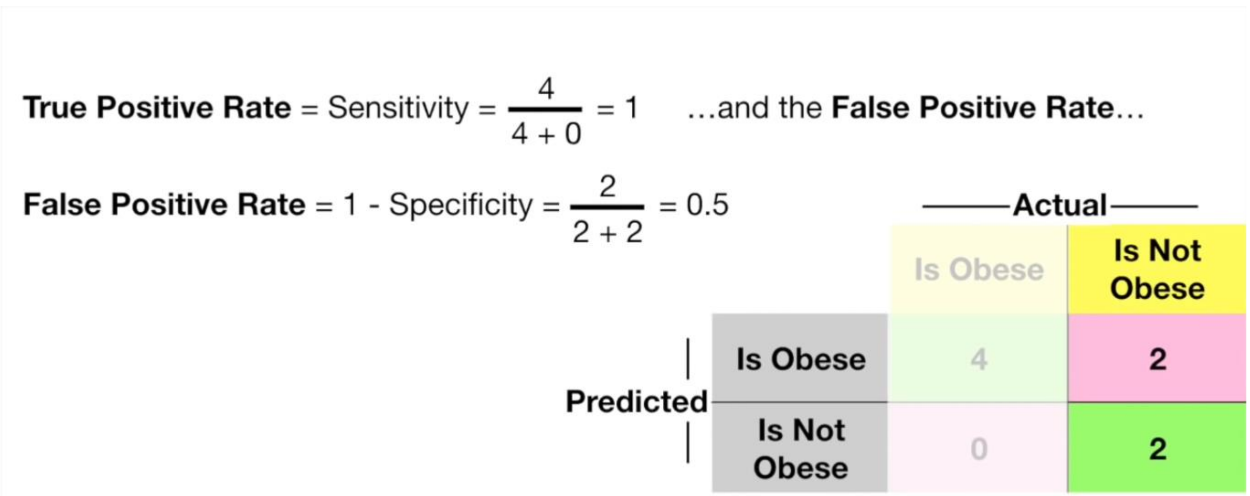


Let’s again plot it on the graph (0.75, 1):

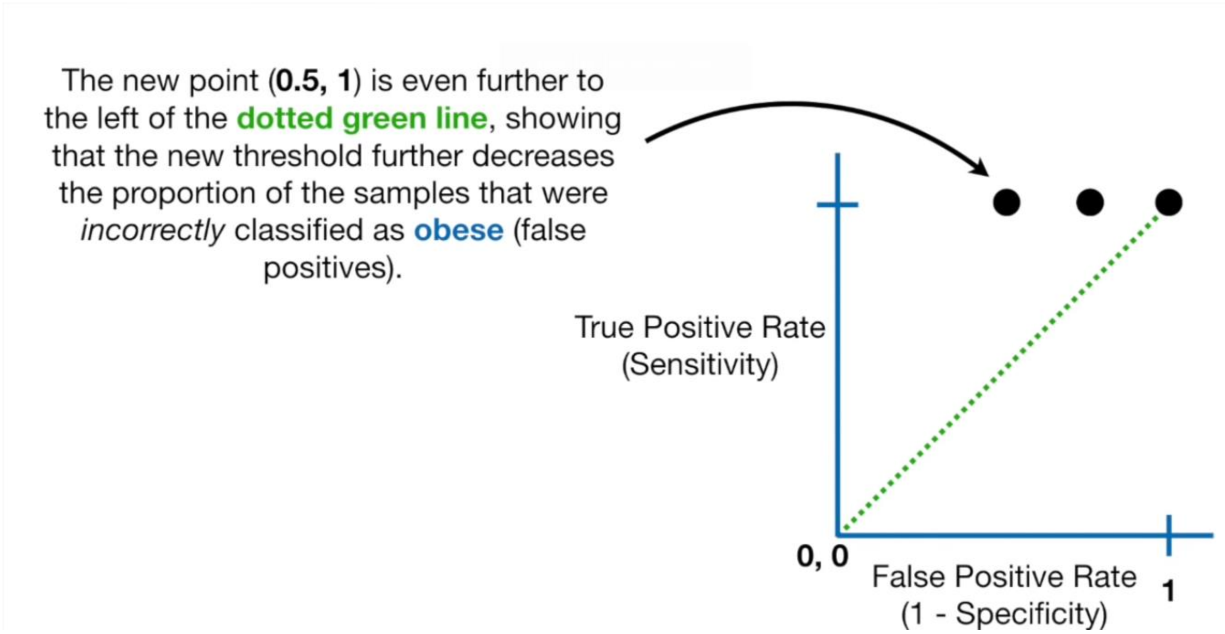


Again let’s set new threshold matrix and see the confusion matrix:

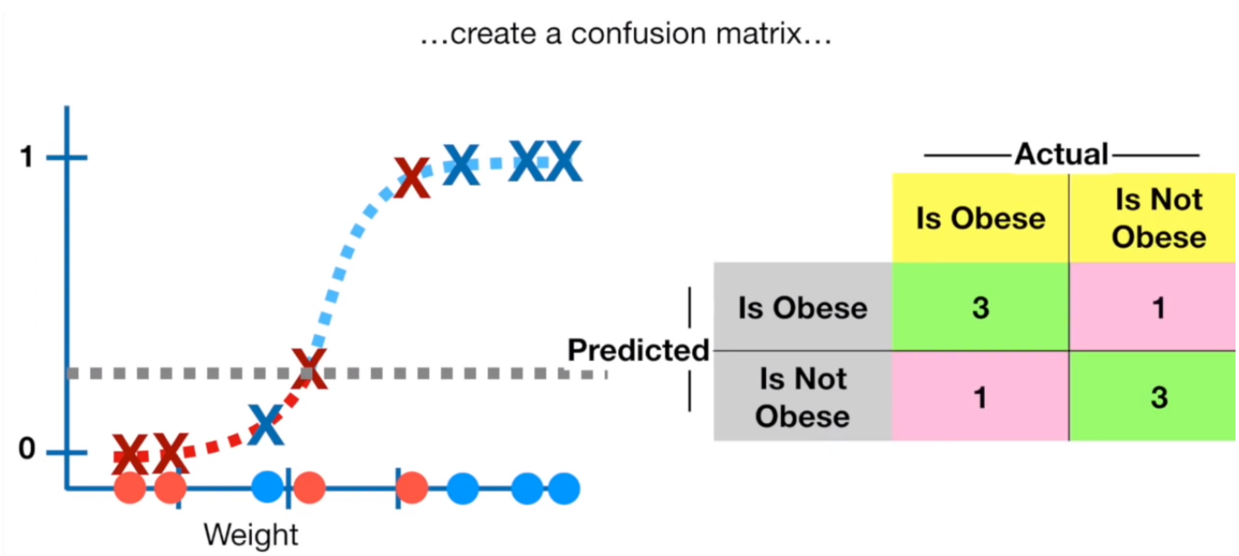


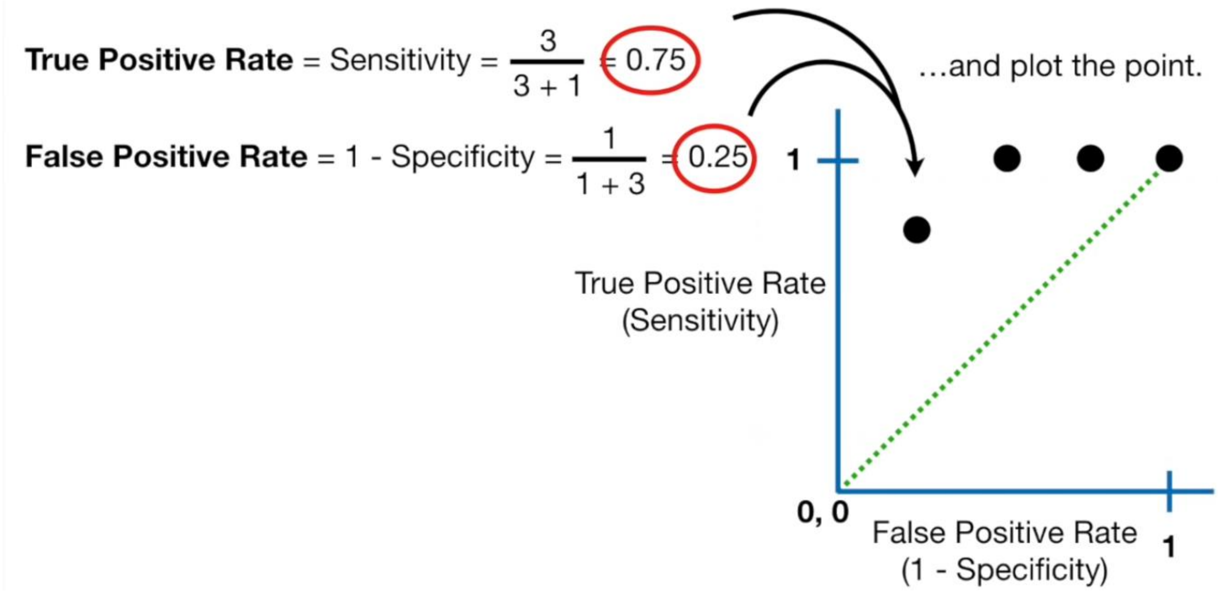


Let's again plot points on (0.5, 1):

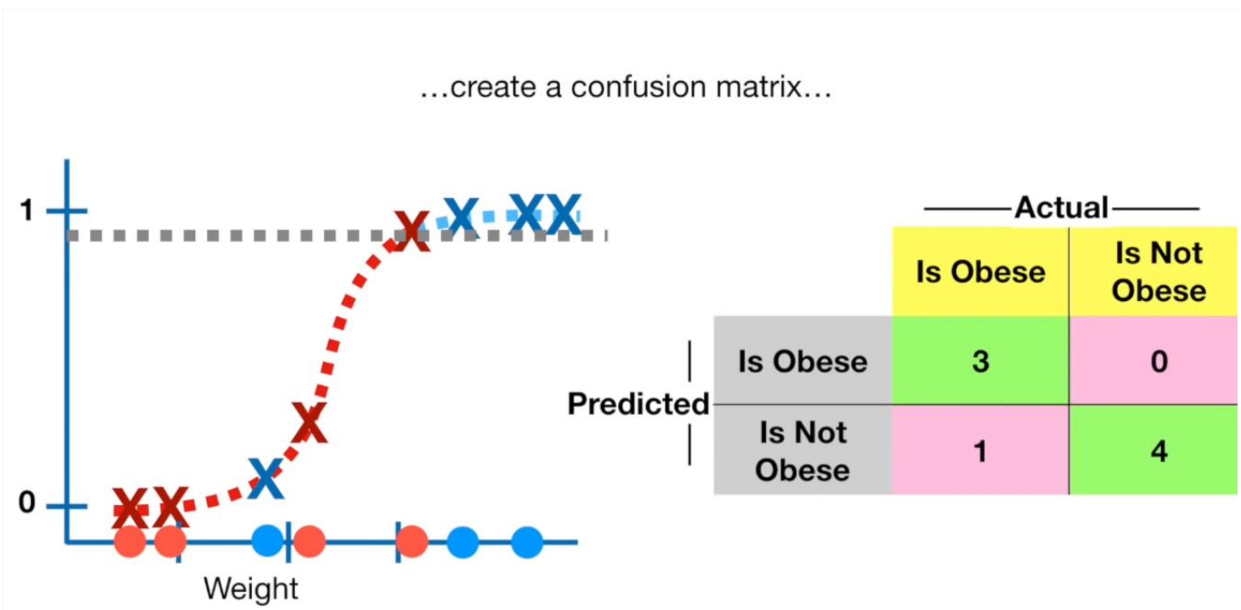


Again we change the threshold and calculate the confusion matrix:





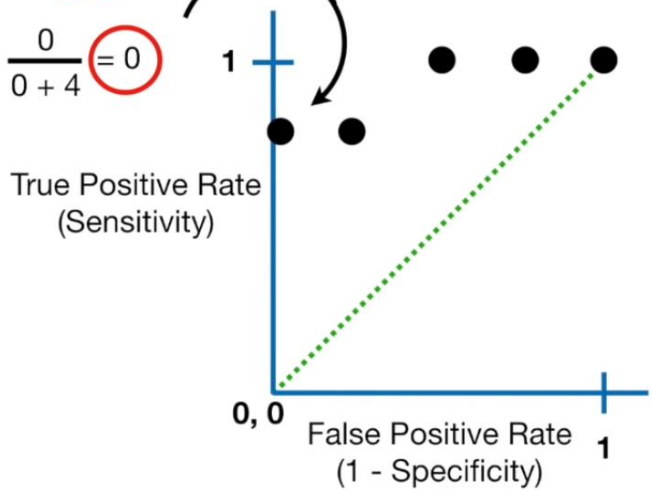
Again we change the threshold with calculations:



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

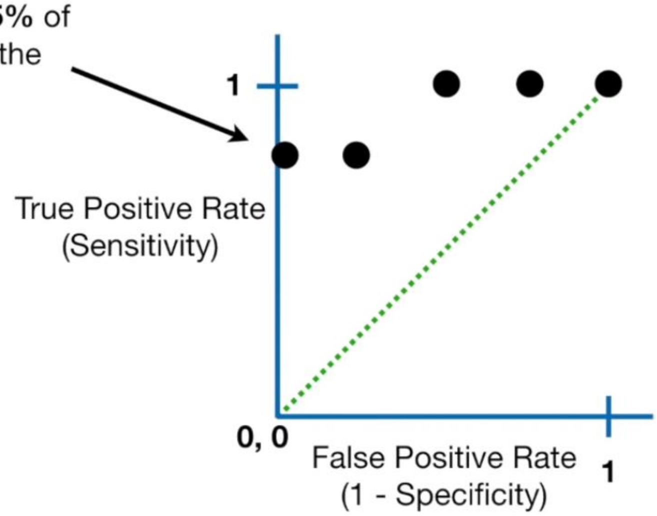
False Positive Rate = 1 - Specificity = $\frac{0}{0+4} = 0$

…and plot the point.

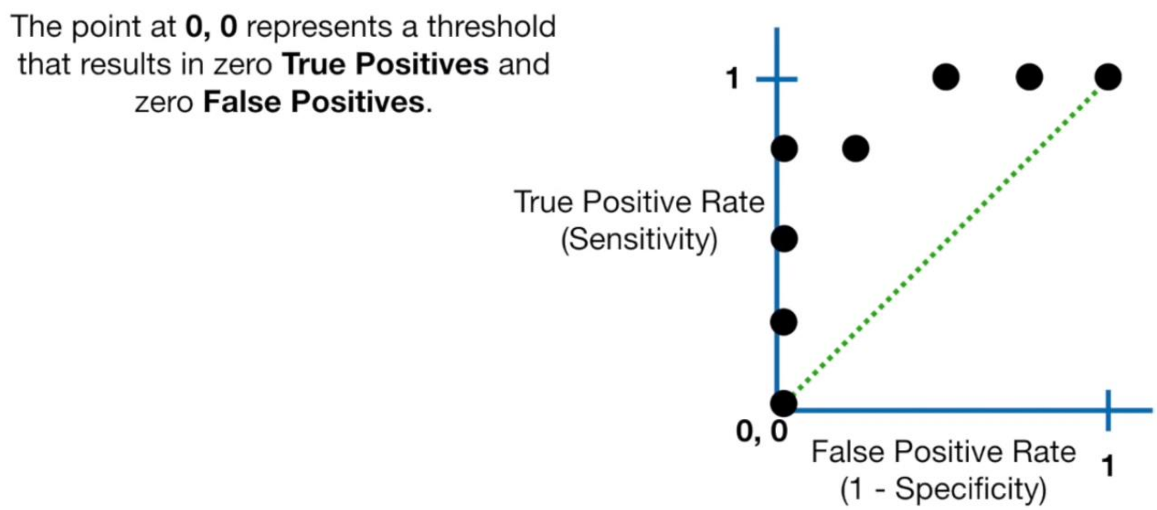


The threshold represented by the new point (**0, 0.75**) correctly classified **75%** of the **obese** samples and **100%** of the samples that were **not obese**.

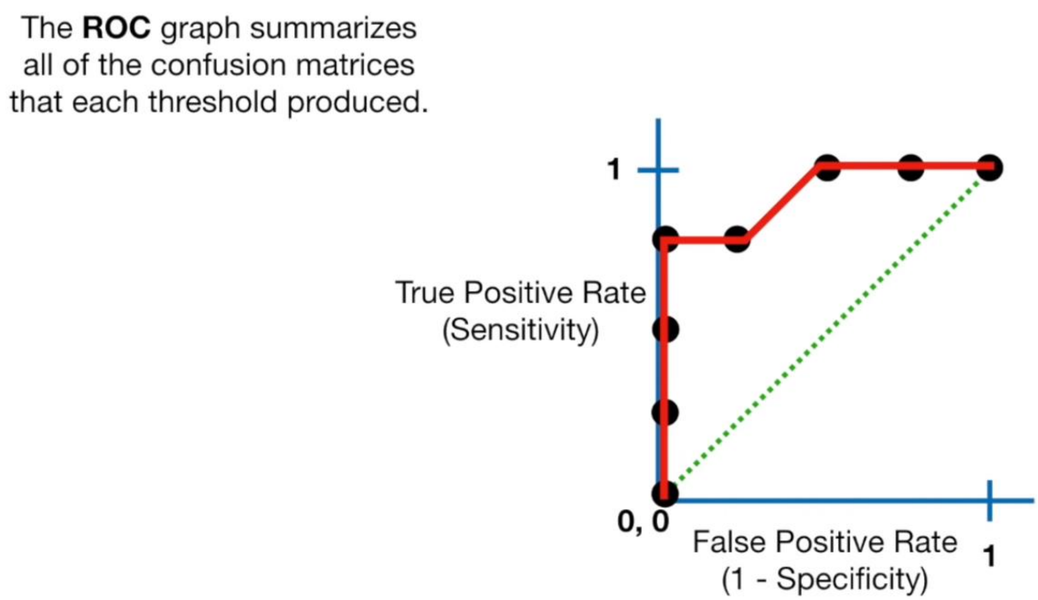
In other words, this threshold resulted in no **False Positives**.



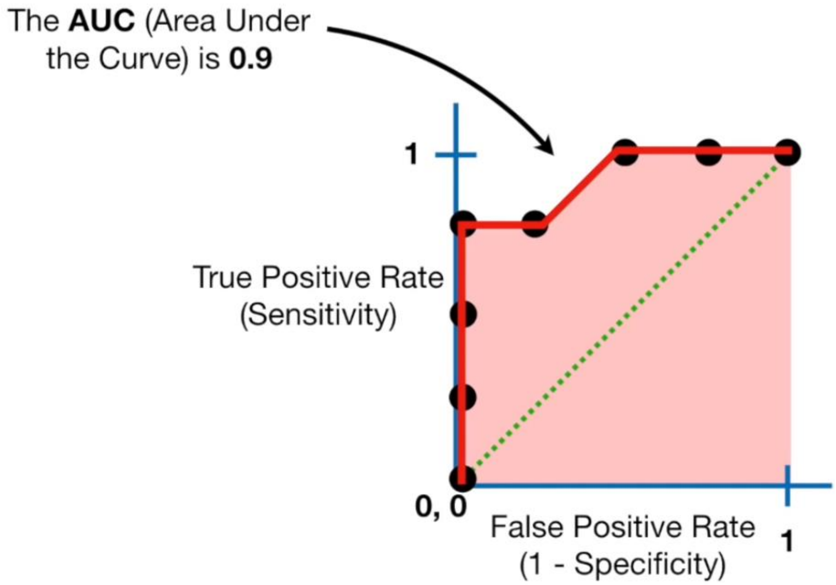
If we get all False-Positive (Error) = 0, we got touched to Sensitivity. But if we increase the threshold again going to decrease our accuracy. And we touch the origin.

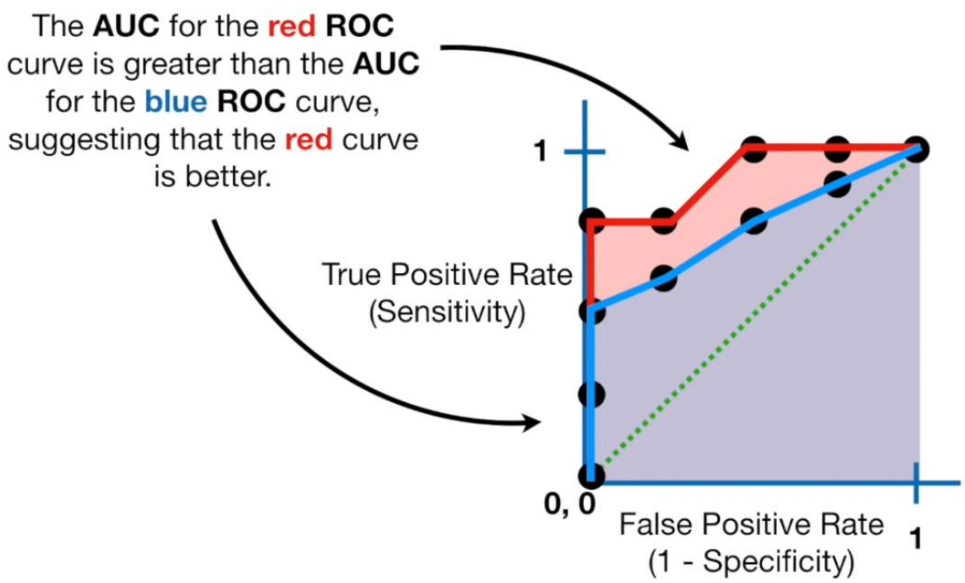


If we connect these lines we get ROC (Receiver Operator Characteristics) graph.

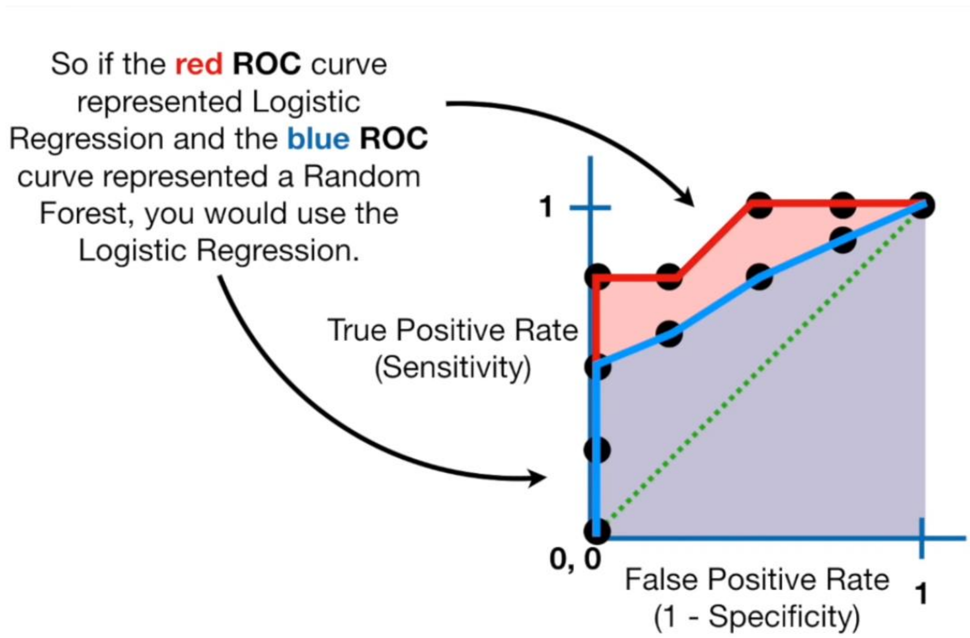


Now what is AUC (Area Under the Curve)?


The AUC (Area Under the Curve) is 0.9
1
True Positive Rate (Sensitivity)
0, 0
False Positive Rate (1 - Specificity)
1


The AUC for the red ROC curve is greater than the AUC for the blue ROC curve, suggesting that the red curve is better.
1
True Positive Rate (Sensitivity)
0, 0
False Positive Rate (1 - Specificity)
1

Conclusion:


So if the red ROC curve represented Logistic Regression and the blue ROC curve represented a Random Forest, you would use the Logistic Regression.
1
True Positive Rate
(Sensitivity)
0, 0
False Positive Rate
(1 - Specificity)
1