what is the intuition and purpose for AUC.

Let's say you have built a binary classifier using Logistic Regression. Now in logistic regression as you, it gives you probability for both the class labels instead of giving you the actual class label. So now you need to decide the threshold value above which all the values would be class 1 and below which all the points would be class 0. But what would be the value of that threshold be? Let's say you decide it to be 0.5. But needless to say that it is problem specific. In some cases I have seen the right value is around 0.3 and in some other cases like financial domain the right value is 0.65, bit higher. So you need to know the best value of the threshold. AUC just checks the perfomance(TPR) of the model of different threshold values and gives you a socre. So if you can see the AUC curve with its corresponding value of threshold, then you can decide that for your business what threshold works better. Mostly AUC is not interpretable more than this. It is better to fine tune a binary classifier based on AUC instead of using it to report the model performance.

**How do we get ROC and AUC:**

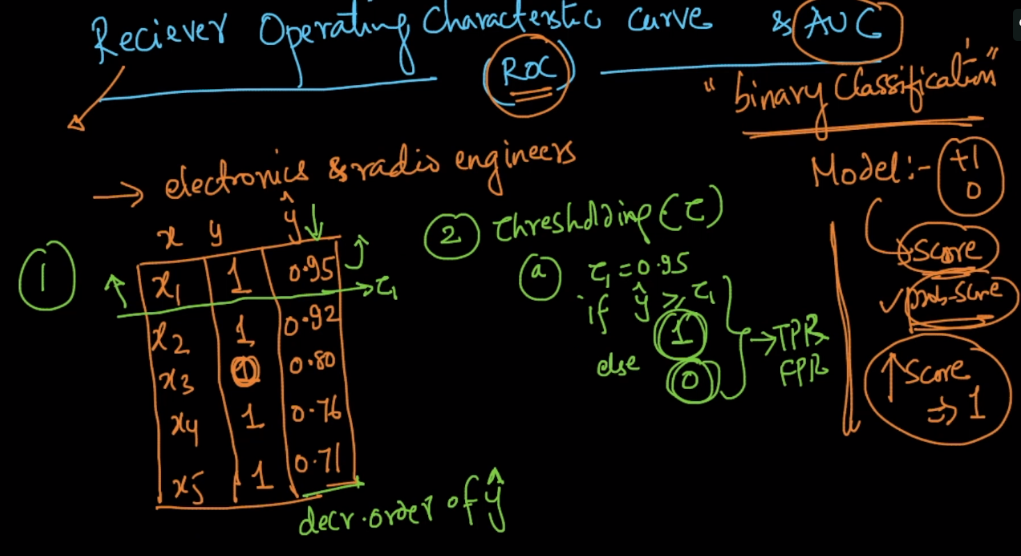
It works only for **binary classification**. Let say our model gives the probability of getting the output label as 1.

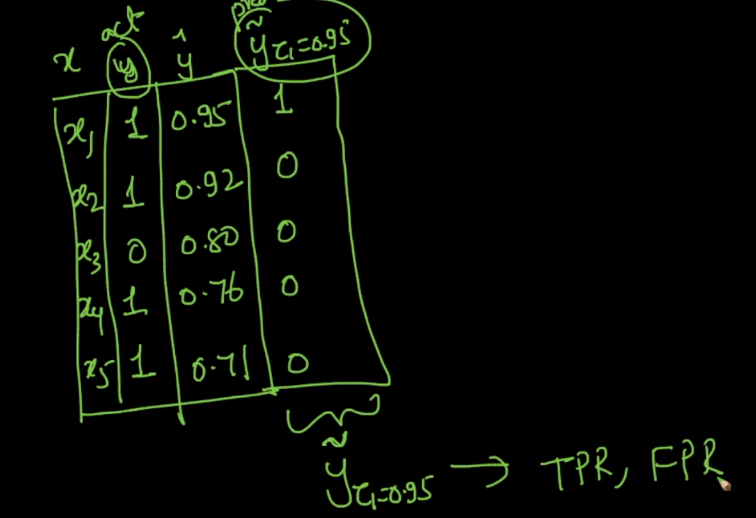
Now we sort datapoints according to probability output in decreasing order.

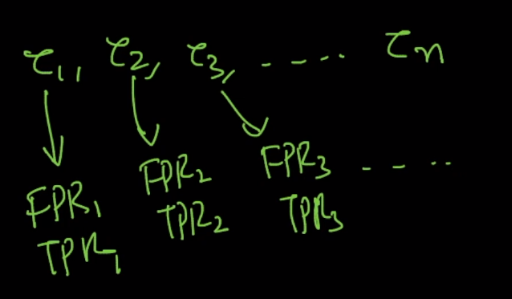
Now let’s pick one by one the threshold value of prob to decide whether for output prob the point is 1 or 0 from top to last point.

That means for given example we first pick the threshold as 0.95, now using that threshold we generate output as 1 or 0. After generating output we calculate TPR and FPR with threshold=0.95.

Similarly we do this for all output probability of ‘n’ points, so in the end we would have **n, TPR and FPR.**

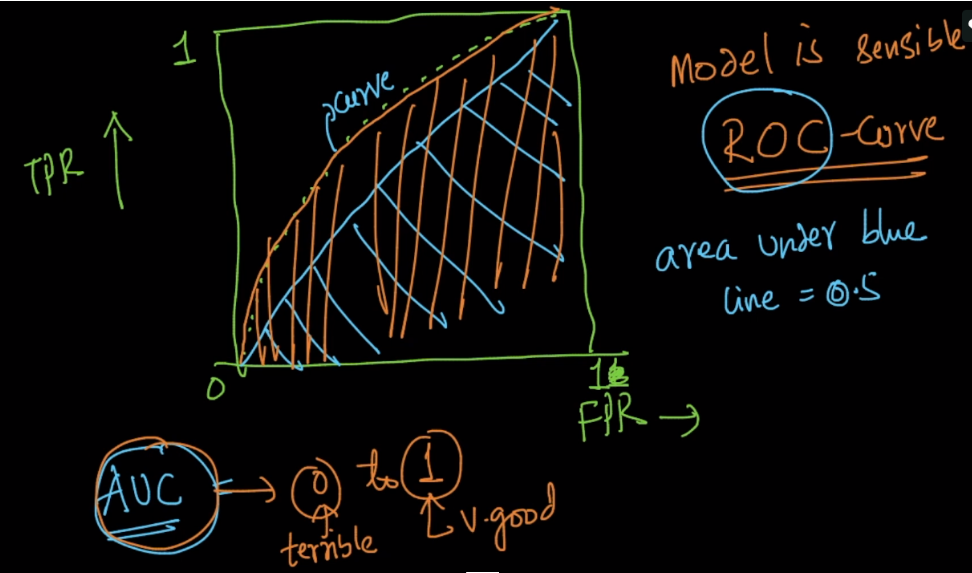






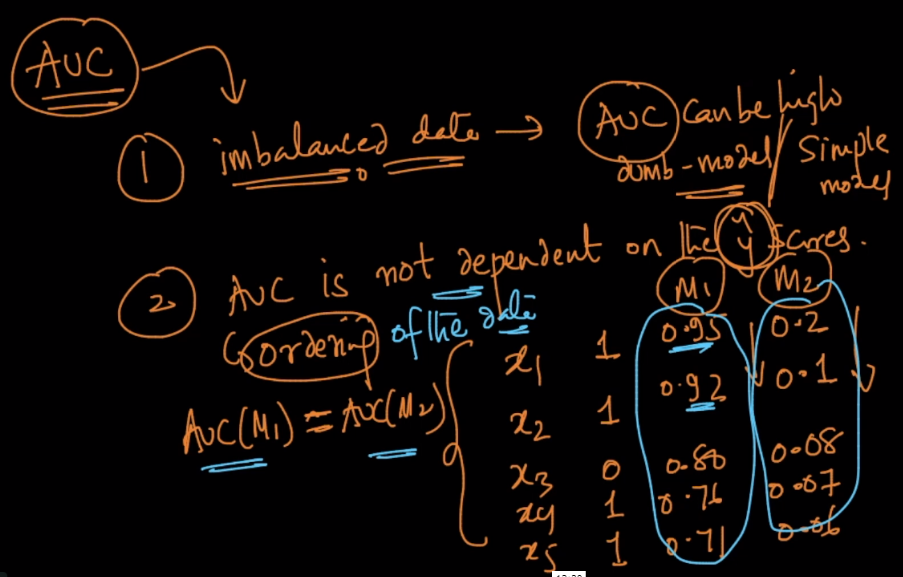
After getting n no of TPR and FPR we will draw the curve with FPR on x-axis and TPR on y-axis. The area covered under this curve is called AUC. AUC will be between 0 to 1.

As we can see in below figure area under blue line is 0.5



**Drawback of AUC:**

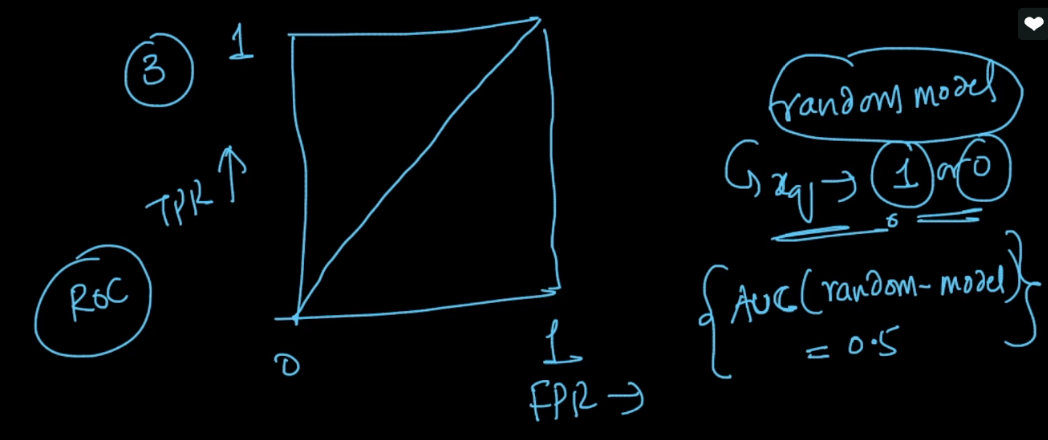
1. AUC can be high for dumb or simple model
2. AUC is dependent on the ordering of data and not on the output probability score.

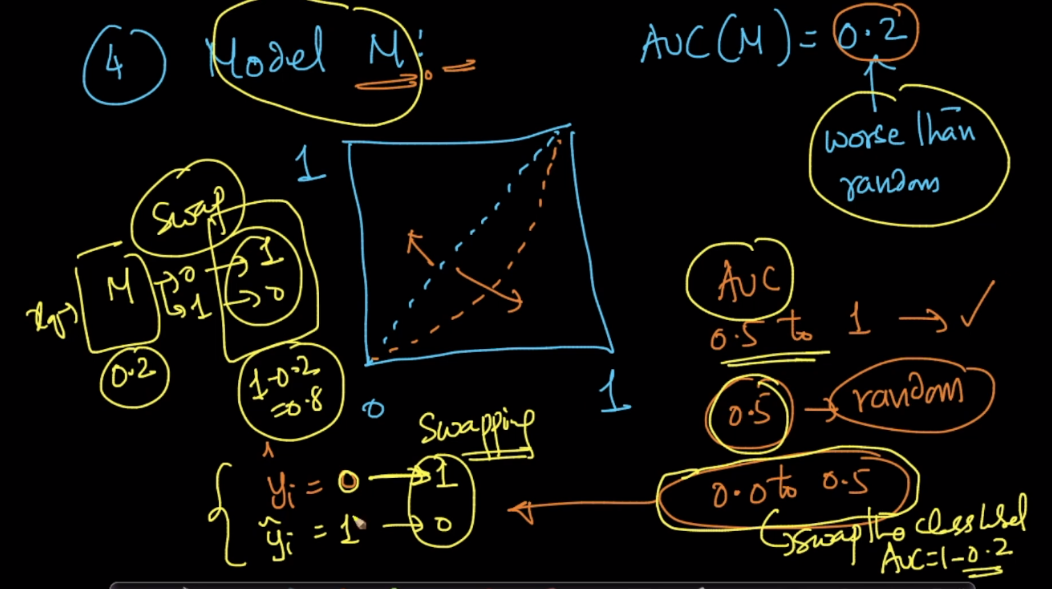


One important thing about AUC and ROC is that for random model (model which predicts output randomly) have AUC of 0.5, whose ROC line is seem as diagonal of the below figure.

So from this we can make conclusion that any model which have AUC of between 0.5-1 is considerd to be sensible model.

And if model have AUC less than 0.5 will be considered worse, because they are generating value lower than even random model.





So what we can do if model is generating AUC value lower than 0.5, so then for some cases we can swap the value that model is generating from 0 to 1, and 1 to 0. So by swapping the output value its AUC will become 1 – old\_AUC

**Must read link:** <https://medium.com/greyatom/lets-learn-about-auc-roc-curve-4a94b4d88152>

**By google ML glossary:**

**The Area under the ROC curve is the probability that a classifier will be more confident that a randomly chosen positive example is actually positive than a randomly chosen negative example is positive.**