R.O.C!

- Why ROC curve?

- You have built your machine learning model you need to evaluate and validate how good or bad it is, so you can decide whether to implement it. That's where ~~so~~ AUC-ROC Curve comes in -

- It's just saying that we are calculating AUC (Area Under Curve) for ROC (Receiver Operating Characteristic).

- Although, it works only for binary classification problem.

What is ROC Curve?

→ A ROC curve is like a graph that shows how well a classification model performs.

→ It helps us ~~to~~ see how the model makes decision at different levels of certainty.

The curve has two lines -

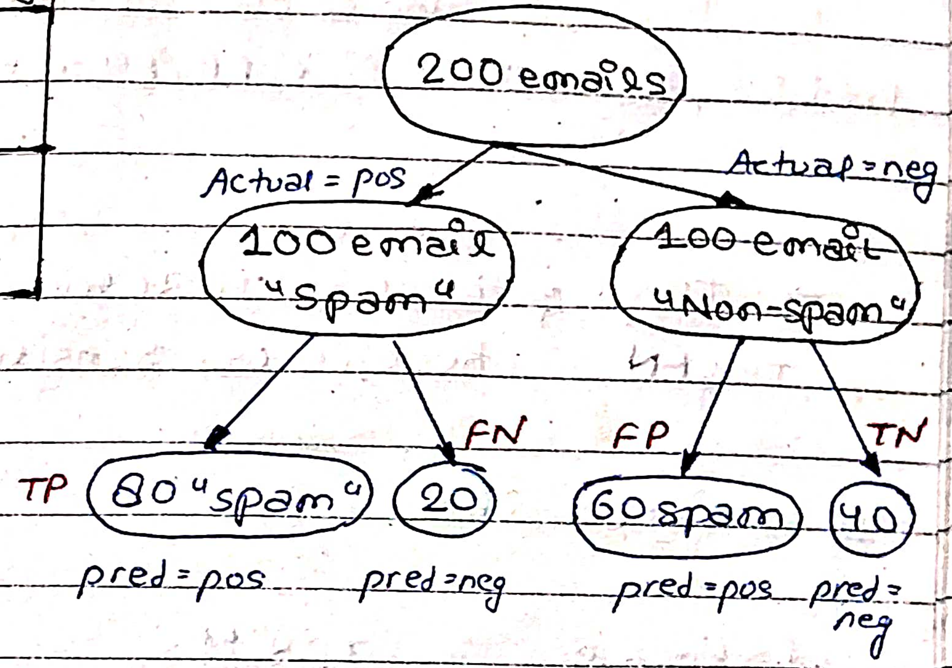
- One for how often the model correctly identifies positive cases (true-positives).

- and another for how often it mistakenly identifies negative cases as positive (false-positives).

~~confusing~~ confusion matrix :

Actual		class = pos	class = neg
	class = pos	TP	FN
	class = neg	FP	TN

TP = 80	FN = 20
FP = 60	TN = 40



- Accuracy ⇒ $\frac{TP + TN}{TP + TN + FP + FN}$; $\frac{\text{correctly predicted observo}}{\text{total nos of observo}}$

$$\frac{80 + 40}{60 + 40 + 80 + 20} \Rightarrow \frac{120}{200} \Rightarrow 0.666$$

- precision ⇒ $\frac{TP}{TP + FP}$; $\frac{\text{correctly predicted positive observo}}{\text{total predicted pos observo}}$

$$\frac{80}{80 + 60} \Rightarrow \frac{80}{140} \Rightarrow 0.571$$

- Recall ⇒ $\frac{TP}{TP + FN}$; $\frac{\text{correctly predicted poso obso}}{\text{all observo in actual class}}$

$$= \frac{80}{80 + 20} \Rightarrow \frac{80}{100} \Rightarrow 0.80$$

- F1 score ⇒ $\frac{2(\text{Recall} * \text{Precision})}{\text{Recall} + \text{Precision}} \Rightarrow \frac{2(0.80 * 0.57)}{0.80 + 0.57}$

$$\Rightarrow \frac{0.912}{1.37} \Rightarrow 0.665$$

F1 score $\xrightarrow{\text{work best}}$ FP & FN have similar cost, uneven class distribution.

Precision, Recall $\xrightarrow{\text{work best}}$ FP, FN cost different

→ $\frac{TP}{TP+FN}$: recall or TPR (true positive rate) or Benefit or sensitivity

ie email-classifier

best-case

=80	=20
=20	=80

$\frac{80}{80+20} \Rightarrow 80\%$

TPR ↑↑ (good-thing)

worst-case

=20	=80
=20	=80

$\frac{20}{80+20} \Rightarrow 20\%$

→ $\frac{FP}{FP+TN}$: FPR (false positive rate) or "1-specificity" Cost

best-case

0	100

$\frac{0}{0+100} \Rightarrow \boxed{0} = FPR$

FPR ↑↑ (bad-thing)

worst-case

100	0

$\frac{100}{100+0} \Rightarrow \boxed{1} = FPR$

Specificity / TNR (True Neg rate) TN / TN + FP
what proportion of neg class got correctly classified.

=~~S~~

ROC curve,

data

cgpa	iq	placement(yes/no)

train

⇓

model
(logistic reg)

test

threshold

0.5, 0.6, 0.7, 0.8, ...

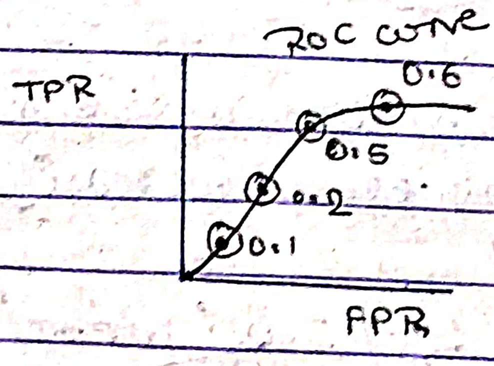
↓

Confusion matrix

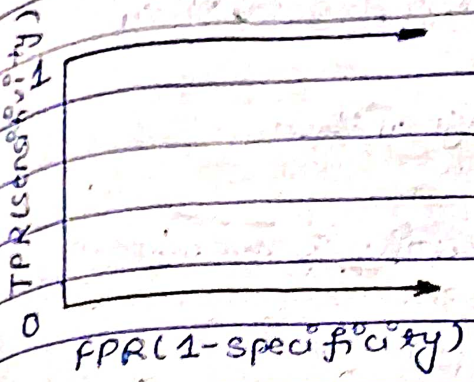
↓

TPR, FPR calculate

(plot)

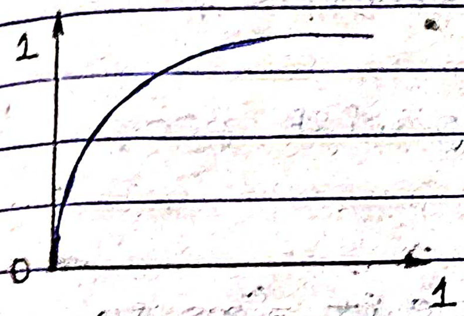


* Different cases —



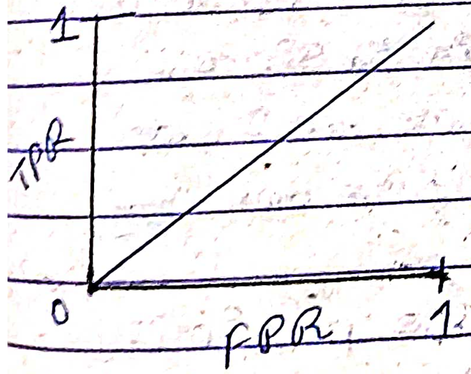
when AUC = 1, the classifier can correctly distinguish b/w all the positive and negative class points.

If, however the AUC had been 0, then classifier would predict all negatives ~~and~~ as positives and all positives as negatives.



when AUC (0.5 < AUC < 1), there is high chance that the classifier will be able to distinguish the positive class values from the negative ones.

This is so because the classifier is able to detect more numbers of True positives and True neg. than false negatives & false positives.



when AUC = 0.5, then classifier is not able to distinguish b/w positive & negative class points.

Meaning, that the classifier either predicts a random class or a constant class for all the data points.