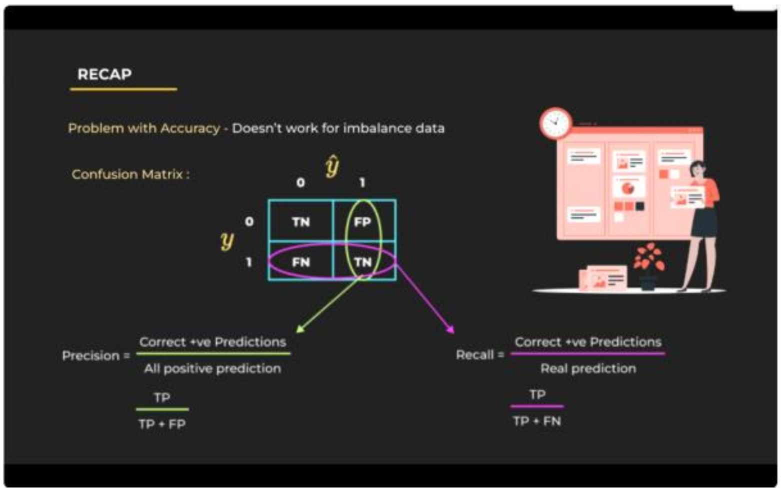
Recap

15 August 2025 18:27



Positive → 1 → Class 1

Negative → 0 → class 0

True → $\hat{y} = y_i$

False → $\hat{y} \neq y_i$

Sensitivity and Specificity

15 August 2025 18:36

Sensitivity is same as True positive rate / Recall

$$= \frac{TP}{TP + FN}$$ => Want to predict as many True positives as possible

Against false negatives

Ex:- Want to predict those who have cancer correctly as much as possible. Don't want to miss any cancer patient's diagonosis.

{ Hence **high Sensitivity** of the screening test becomes **crucial**:

- As the **consequences of failing to treat the disease worsens the patient's condition** }

Speificity → Correctly predict as many true negatives as possible.

Maximize => TN   Minimize → FP

$$\text{True Negative Rate} = \frac{TN}{TN + FP}$$

Ex:- Let us say there is a newly discovered drug, which in combination with other drugs may promote quicker relief from an illness.

However, when a healthy individual takes it, they face severe side-effects.

We want to maximize TN so that healthy patients ($y=0$) are not given the drug.

quiz

15 August 2025 18:48

What to say when screening test identifies 92 Cancer patients out of 100?

Choices

- [x] test has high sensitivity
- [ ] test has low sensitivity
- [ ] test has no sensitivity
- [ ] cannot be determined

Cancer → 1 → positive

non-cancer → 0 → negative

$$\left\{\frac{TP}{TP+FN}\right\} = 92\%$$

Sensitivity

92 True positives

out

of 100 positives

Actually neg, predicted positive

$$FPR = \frac{FP}{FP + TN} \Rightarrow$$ lot of negatives being predicted as positives.

$\left\{\right.$ Lot of non-cancer patients falsely being diagonosed with cancer. $\left.\right\}$ $\Rightarrow$ Chaos, panic, system

$$\{ FPR = 1 - TNR \} \quad \Uparrow TNR \quad \Downarrow FPR$$

TNR - Correctly classifies negatives
FPR — Incorrectly classifies negatives (as positives)

$$FNR = \frac{FN}{FN + TP}$$

Lot of cancer patients being classified as healthy.

$$FNR = 1 - TPR$$

TPR → Correctly classifies positives
FNR → Incorrectly classifies positives (as negatives)

In a credit fraud detection system, which is more important?

21 users have participated

- A Sensitivity 29%
- B Specificity 29%
- C both 43% ✓

FP => Not fraud but predicted as fraud

FN => Fraud of 5 L Rs but predicted as non-Fraud

Max - $1 - FPR = TNR$

Max - $1 - FNR = TPR$

ROC AUC

15 August 2025 19:01

Logistic Regression outputs probabilities.

{By default, $P \geq 0.5$ is considered as positive class (1)}

{$P < 0.5$ is considered as negative class (0)}

But what if we change this threshold?

Obs	y_true	y_pred_prob	y_pred_label (th=0.3)	y_pred_label (th=0.5)
1	1	0.9	1 ✓	1 ✓
2	0	0.2	0 ✓	0 ✓
3	1	0.4	1 ✓	0 ✗
4	0	0.1	0 ✓	0 ✓
5	1	0.35	1 ✓	0 ✗

$\frac{5}{5}$ $\frac{3}{5}$

Metrics

Threshold	Accuracy	TPR (Recall)	TNR (Specificity)
0.3	100% ✓	1.0 ✓	1.0 ✓
0.5	60% ✗	0.33 ✗	1.0 ✓

→ Example 1

Acceptable model or not?

Obs	y_true	y_pred_prob	y_pred_label (th=0.2)	y_pred_label (th=0.3)	y_pred_label (th=0.5)
1	1	0.4	1	1	0
2	0	0.35	1	1	0
3	1	0.3	1	1	0
4	0	0.45	1	1	0
5	1	0.25	1	0	0

Threshold	TP	TN	FP	FN	Accuracy	TPR (Recall)	TNR (Specificity)
0.2	4	0	2	0	4/6 = 66.7%	1.0	0.0
0.3	3	0	2	1	3/6 = 50%	0.75	0.0
0.5	0	1	1	4	1/6 = 16.7%	0.0	0.5

Example 2

{Acceptable model or not?}

{No!!}

{Thresholds play a huge role in determining whether a model is good or not.}

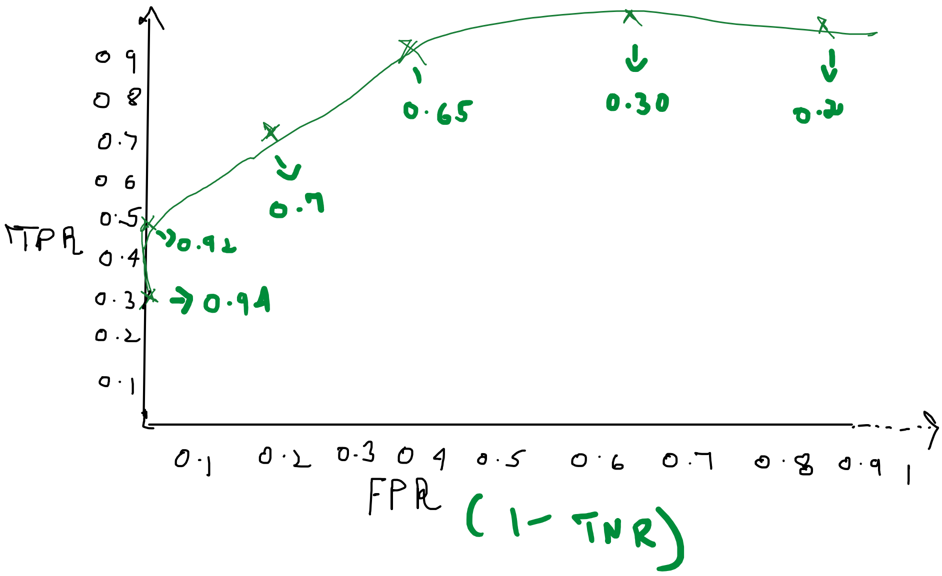
{But attempting multiple thresholds and then finding out that a model is bad will

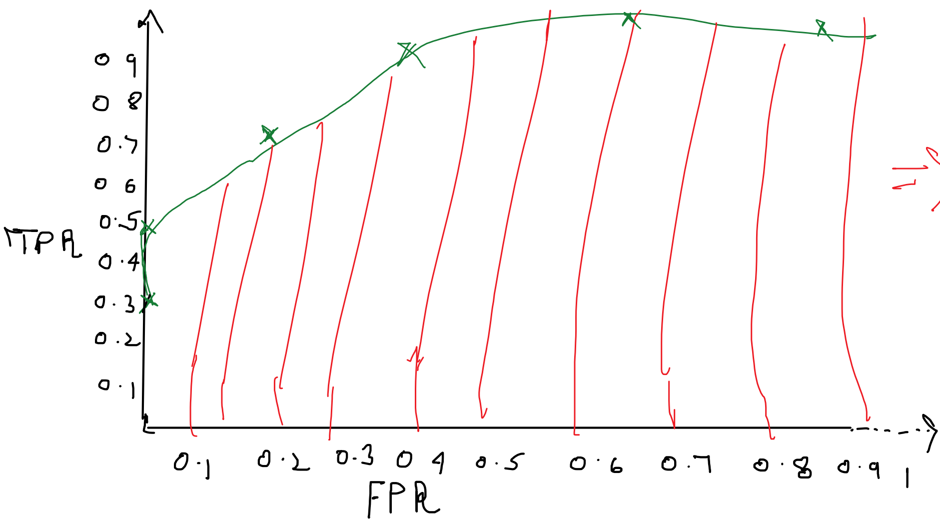
Waste us time!

Any way to know through a single metric whether a model will do well in atleast some threshold?

(Threshold)

P	TPR	FPR
0.94	0.33	0.00
0.92	0.50	0.00
0.70	0.67	0 2
0.65	0.9	0.4
0.30	1.00	0.67
0.20	1.00	1.00





{Area under the curve}

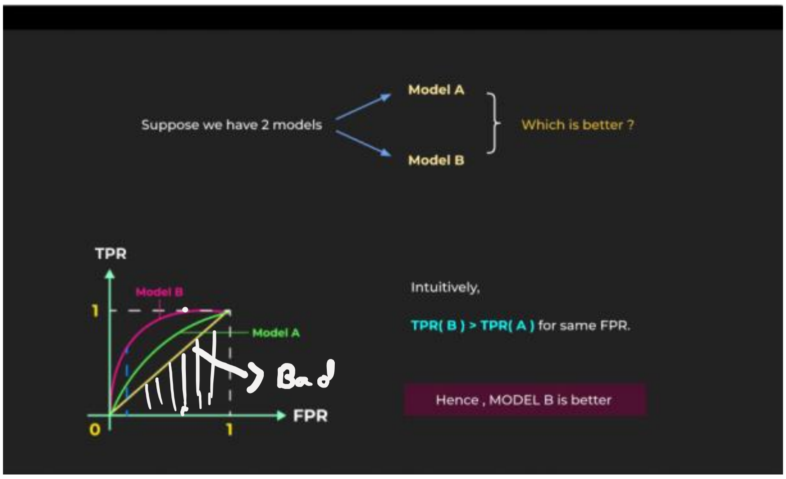
= AUC ROC

Area under Curve for Receiver Operating Characteristic

Name derived from Electronics concept

Higher the AUC score, better the model

High Roc-AUC means there exists atleast one 'threshold when the model is doing very well!



ROC AUC score around 0.4 – 0.6 represents a bad Model (Random Guessing)

0.7 – 0.9 ⇒ Decent Model ✓

0.9+ → Great Model ✓

How many points are typically used to plot an ROC curve?

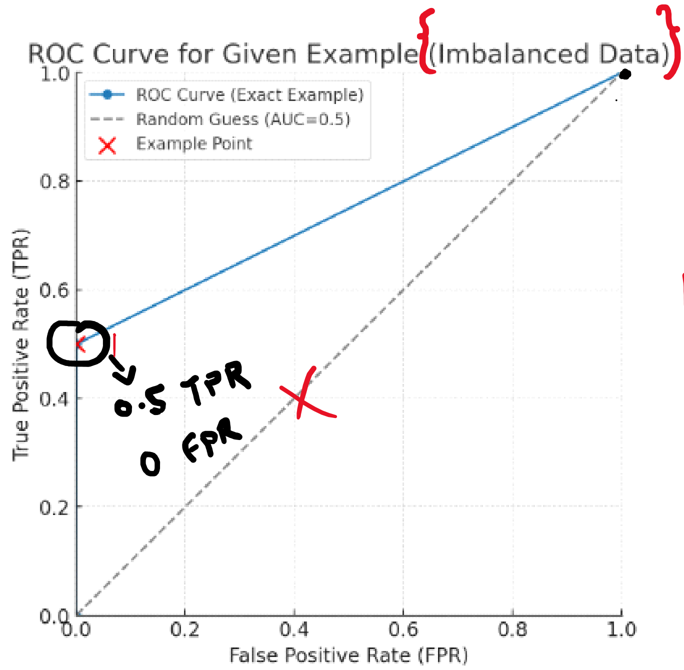
11 users have participated

- A — 2 points (0,0) and (1,1) — 27%
- B — 3 points representing the thresholds 0.25, 0.5, and 0.75 — 18%
- C — 10 points equally spaced between 0 and 1 — 0%
- D — Depends on the number of unique threshold values — 55% ✓

Problem with ROC AUC

15 August 2025 20:01

ROC-AUC can be misleadingly high when dataset is imbalanced.



TPR = 1
FPR = 1

ROC-AUC
=> 0.75

ROC-AUC will fail for imbalanced data!!

imbalanced

If data contains 50 spam and 300 non-spam samples then which is true?

0 users have participated

A	ROC may overestimate the model's performance.	0%
B	ROC may underestimate the model's performance.	0%
C	ROC does provide useful information.	0%

If data contains 50 spam and 300 non-spam samples then which is true?

0 users have participated

A	ROC may overestimate the model's performance.	0%
B	ROC may underestimate the model's performance.	0%
C	ROC does provide useful information.	0%
D	ROC cannot be created	0%

PR Curve

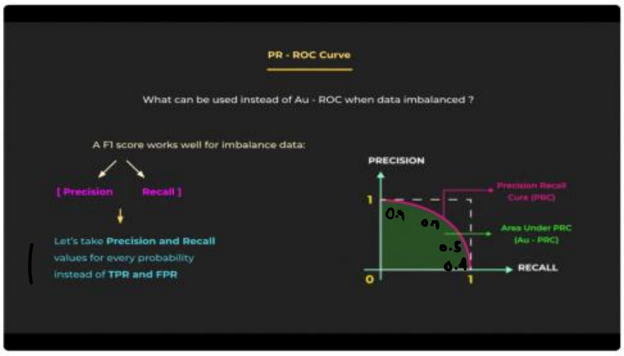
15 August 2025 21:40

F1 - Score

Will a certain threshold give me a good model?

Precision - Recall curve

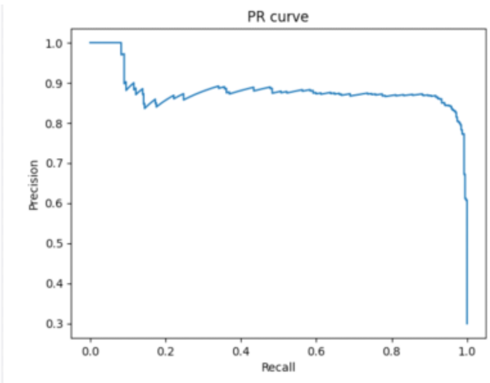
What can be used instead of AU-ROC curve, when data is imbalanced ?



Ans: Since F1 score works well for imbalance data:

Plot precision vs recall for every probability threshold.

Area Under PRC → Higher the better



Use PR - AUC score when focus is to predict minority class correctly

as in spam example very few 1s, but we want to do well on 1s.

Area under Precision vs Recall curve for different probability thresholds.

Interview ⇒ What is the whole point or indication we get by looking at ROC-AUC or PR-AUC scores?

Answer :- They tell us how well the classification boundary is able to differentiate b/w positive and negative class.

Imbalanced data

15 August 2025 21:59

80% of 1 class

20% of other class

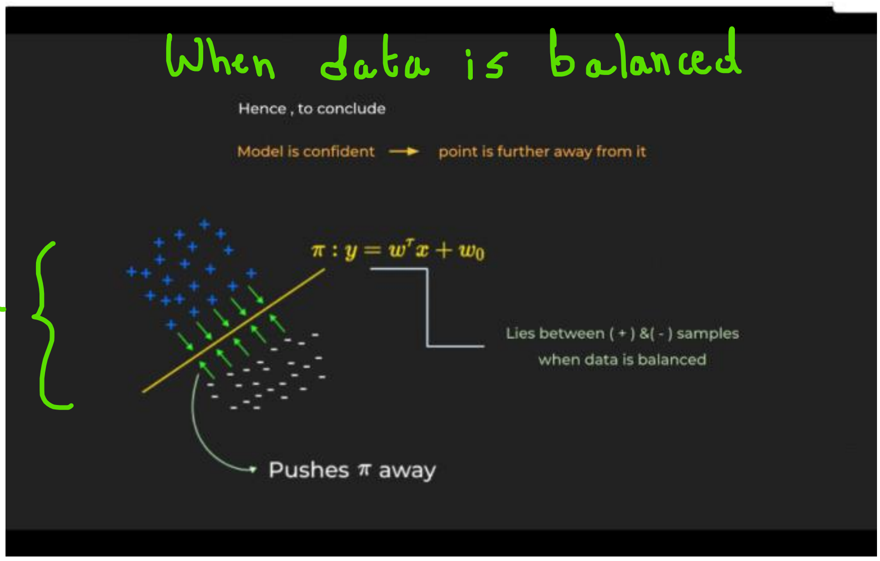
Effect of imbalance on model

15 August 2025 22:00

Goal of optimization algorithms like Gradient Descent is to minimize log-loss.

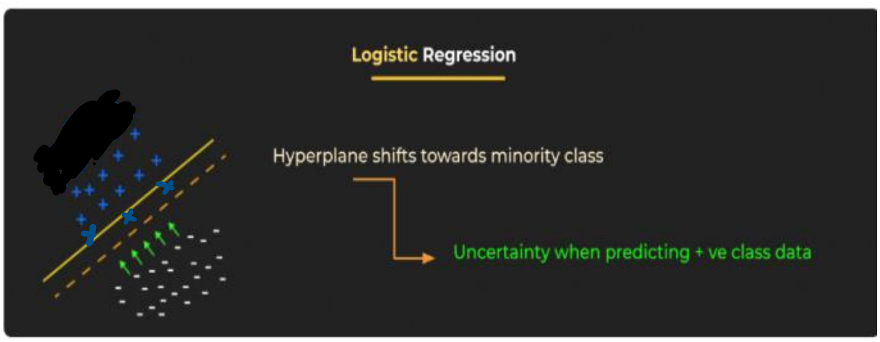
{ Less log loss ⇒ Higher likelihood of points }

{ Higher likelihood ⇒ farther points are from the boundary! }



Equal forces from each end balance the boundary

Imbalanced data



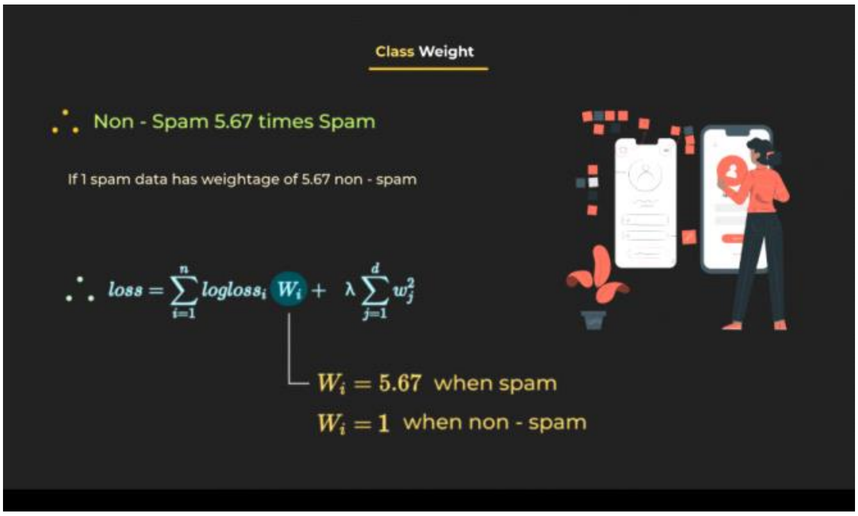
More force from one side. So, boundary pushed closer to

Minority class → This causes uncertainity

Many points in and around

$p = 0.5$

One solution :- Add weights to the loss function




```
model = LogisticRegression(class_weight={0:1,1:2.37})
```

Sampling

15 August 2025 22:12

Ways to handle imb data

Undersampling

Spam → 10000 points

{Not Spam → 90000 points}

Randomly pick 10000 Not spam points.

{Use All 10000 spam and 10000 not spam for Modelling. Undersample majority class}

{Use this technique when you have enough data}

Oversampling :-

Too few points

Spam => 100 points

Not Spam => 900 points

100 vs 100 => too few points!!

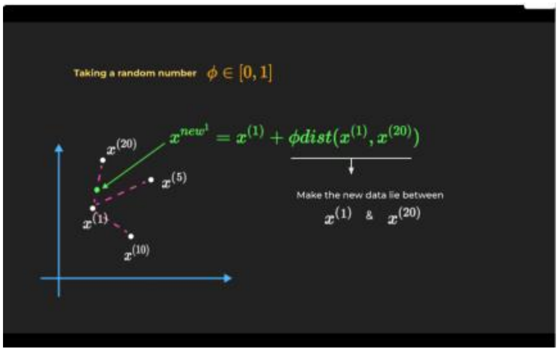
Create {800 synthetic spam points}!!

{Making spam = 900 not spam = 900}

SMOTE

15 August 2025 22:21

(Synthetically Minority Oversampling Technique)



i) pick a point x_1 from Minority class

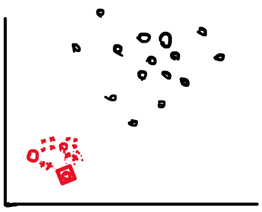
ii) pick n closest points to x_1 → x_{10}, x_5, x_{20}

iii) Put n synthetic points b/w x_1 and other n points.

x_{new1} b/w x_1 and x_{20}

x_{new^2} b/w x_1 and x_5

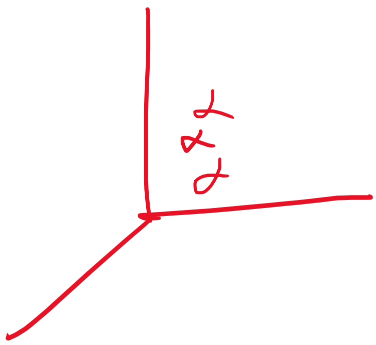
x_{new^3} b/w x_1 and x_{10}



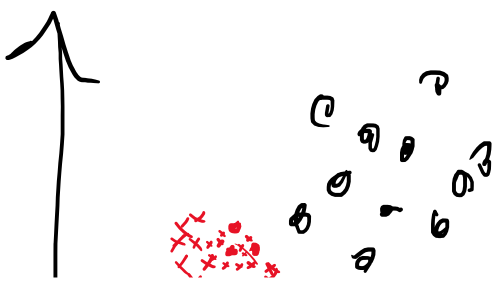
Drawback : SMOTE introduces more noise and irregularities into the dataset.

Avoid when you have many features

High dimensionality



A	SMOTE can increase the risk of underfitting 9%
B	SMOTE introduces noise and may not work for high-dimensional features 63%
C	Cannot use F1-score, Accuracy or any metric after using SMOTE 9%
D	SMOTE has no limitations 20%



→ Non spam