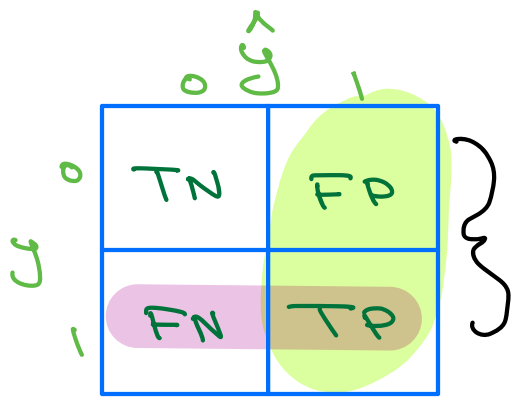
Agenda

- Recap
- Sensitivity and Specificity
- FNR and FPR
- ROC and AU-ROC
- PR Curve
- Data Imbalance
- Techniques to Handle
  - Adjusting Weights y Loss function
  - UnderSampling and OverSampling
  - SMOTE

Recap



$$\text{precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

- Precision ⇒ How precise is model in pred y=1
- Recall ⇒ Out of all points belonging to y==1, How many model was able to predict.

F1-Score

F-Beta Score

→ When Both Precision and Recall are equally important, we can take mean of P and R

① Simple mean $= \frac{P+R}{2}$ ✗ (0, 1)

② Harmonic Mean $= \frac{2}{\frac{1}{P}+\frac{1}{R}} \Rightarrow \frac{2 \times P \times R}{P+R}$ ✓

* F1-Score for bad Model ⇒ ?

$P \Rightarrow 0$, $R \Rightarrow 0$

$F1\text{-Score} = \frac{0 \times 0 \times 2}{0+0+0.000\ldots1} \Rightarrow 0$

* F1-Score for ideal Model ⇒ ?

$P = 1$, $R = 1$

$F1\text{-Score} = \frac{1 \times 1 \times 2}{1+1+0.000\ldots1} \Rightarrow \frac{2}{2+\text{less}} \approx 1$

Why Harmonic Mean?

	Precision	Recall	Mean	H.M
m1	0.30	0.80	0.55	0.436
m2	0.20	0.90	0.55	0.327
m3	0.70	0.40	0.55	0.51

$\text{mean} = \frac{P+R}{2}$

$\text{H.M} \Rightarrow \frac{2 \times (P \times R)}{P+R}$

$\frac{2 \times 0.3 \times 0.9}{0.3+0.9}$

Conclusion :

→ F1-Score gives equal importance to both precision and Recall

→ It adds a lot of penalty when either of P or R is very LOW

Sensitivity

Fortis Hospital → 100 Cancer Patient
Fortis Hospital → 900 Non-Cancer Patient

Aim: Correctly identify all Cancer patients

- TP ↑ : Catch as many Cancer patients as possible
- FN ↓ : Reduce mis-classifications of Cancer patients

TP ↑ and FN ↓ : Sensitivity

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

→ TPR
→ Recall
→ Sensitivity

Why sensitivity is important?
High Sensitivity is crucial as failing to detect disease at Early stage can be Fatal

Specificity

Role of TN and FP in Above scenario?

- TN ↑ : Model should be able to identify Non-Cancer patient
- FP ↓ : Model should keep mis-classification of NON-Cancer LOW

$$\text{Specificity} \Rightarrow \frac{TN}{TN + FP}$$

→ TNR (True Negative Rate)

Specificity is Sensitivity for Class 0

Why specificity is important?

- Costly and fruitless Tests / procedures
- Stress and anxiousness to patients

FNR and FPR

FNR ⇒ False Negative Rate

- ⇒ Miss Rate
- ⇒ 1 − Sensitity
- ⇒ $1 - \frac{TP}{TP+FN}$ ⇒ $\frac{TP + FN - TP}{TP + FN}$
- ⇒ $\frac{FN}{TP+FN}$

FN ↑

* Insensitive Model ⇒ Model with high FNR is considered in-sensitive Model

FPR = 1 − Specificity

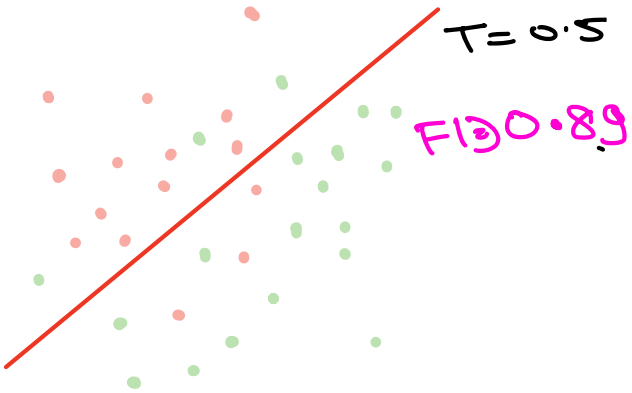
⇒ $1 - \frac{TN}{TN+FP}$ ⇒ $\frac{TN+FP-TN}{TN+FP}$

False positive Rate ⇒ $\frac{FP}{TN+FP}$

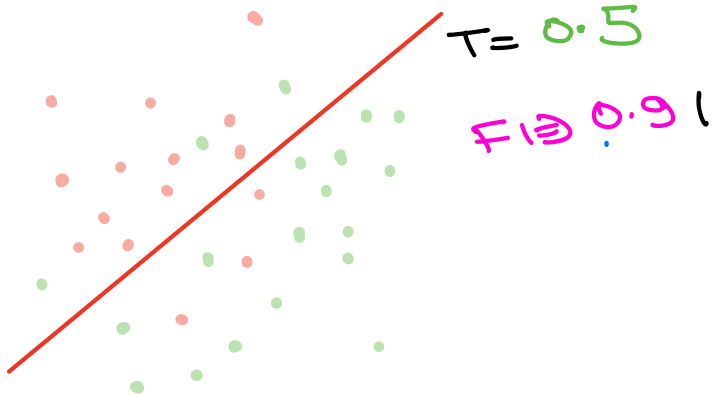
* Not Specific Model : When model has high FPR

Scenario

Q: Suppose you have a model



m_1 ⇒ Threshold (0.5)



m_2 ⇒ Threshold 0.5

To improve Model performance

1) Hyper paramater tuning (L1, L2R)

* 2) Find the best value of Threshold ✓

Find the best value of Threshold

Compare different Model independent of Threshold.

Q-1 How do we find the Best Threshold

Q-2 How do we compare two different type of Models?

⇒ Receiver operating Curve

TPR vs FPR

⇒ PR Curve

Precision vs Recall

ROC

Step 1 : Get all probabilities and Sort

x	$y \in (0,1)$	$p = P(y_i = 1 / x_i)$
x_1	y_1	p_1
x_2	y_2	p_2
x_3	y_3	p_3
⋮	⋮	⋮
x_n	y_n	p_n

nlogn

Sort the table based on $P_{y=1/x}$ (descending)

Step 2 : Calculate $\hat{y}$ for all prob-scores

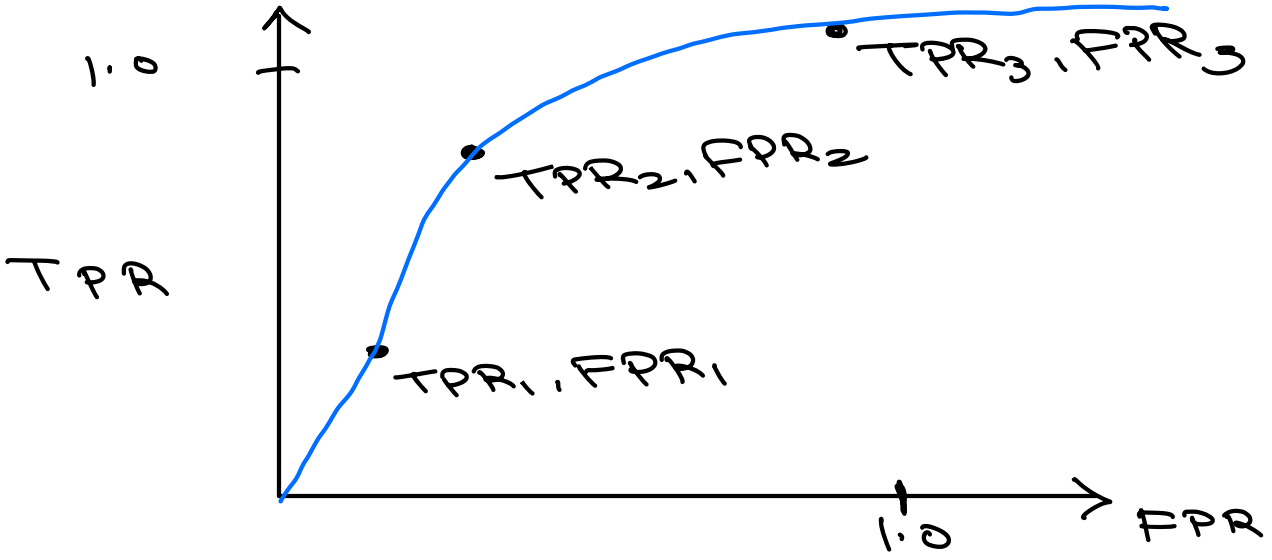
x	y	p	$\hat{y} \to p_1$	$\hat{y} \to p_2$	$\hat{y} \to p_3$	$\hat{y} \to p_n$
x_1	y_1	p_1	1	1	----	1
x_2	y_2	p_2	0	1		1
•	•	•	0	0		1
•	•	•	0	0		1
x_n	y_n	p_n	0	0		1

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

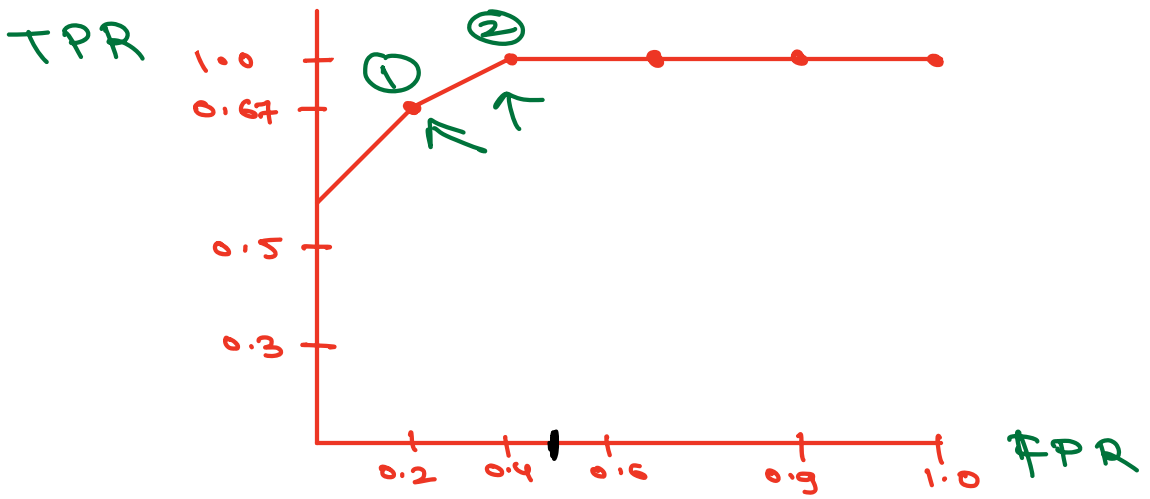
$$FPR = \frac{FP}{FP + TN}$$

p	TPR	FPR
p_1	TPR_1	FPR_1
p_2	TPR_2	FPR_2
p_3	TPR_3	FPR_3
⋮	⋮	⋮
p_n	TPR_n	FPR_n

Step 3: Plot TPR vs FPR



(1) Ideal Threshold:



HomeWork: Plot ROC and Find opt T

X	y	P
x_1	1	0.65
x_2	1	0.94
x_3	0	0.3
x_4	1	0.92
x_5	0	0.7
x_6	0	0.2

y	p	T = 0.94	T = 0.92
1	0.94	1	1
1	0.92	0	1
0	0.7	0	0
1	0.65	0	0
0	0.3	0	0
0	0.2	0	0

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

$$FPR = \frac{FP}{FP + TN}$$

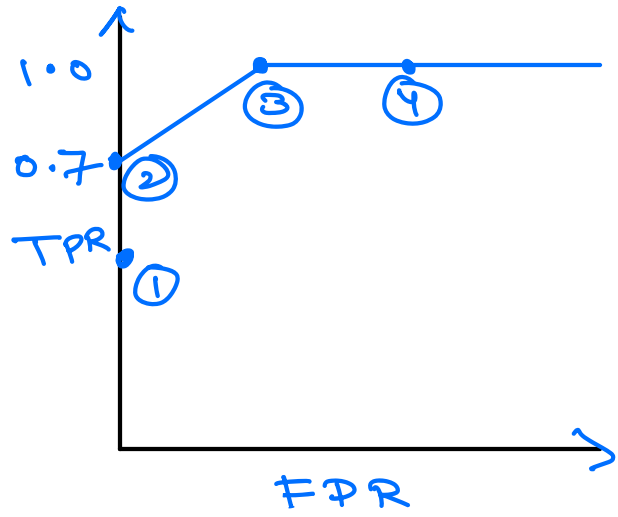
3 ⇒ 0.5
1.0

$TPR_{0.94}$ ⇒ $\frac{1}{1+2}$ ⇒ 0.33

$FPR_{0.94}$ ⇒ ⇒ 0

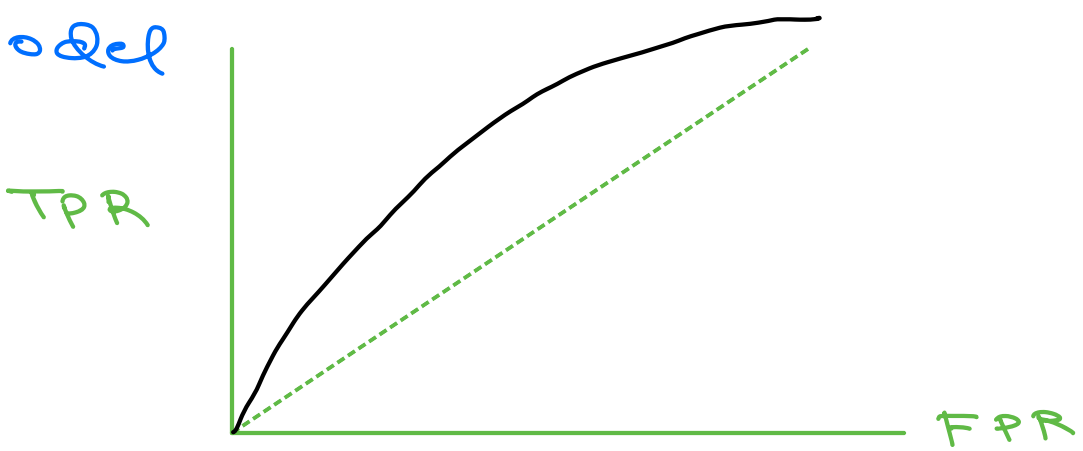
2 ⇒ 0 FP R
0.9 TPR

Thresh	TPR	FPR
0.94	0.33	0



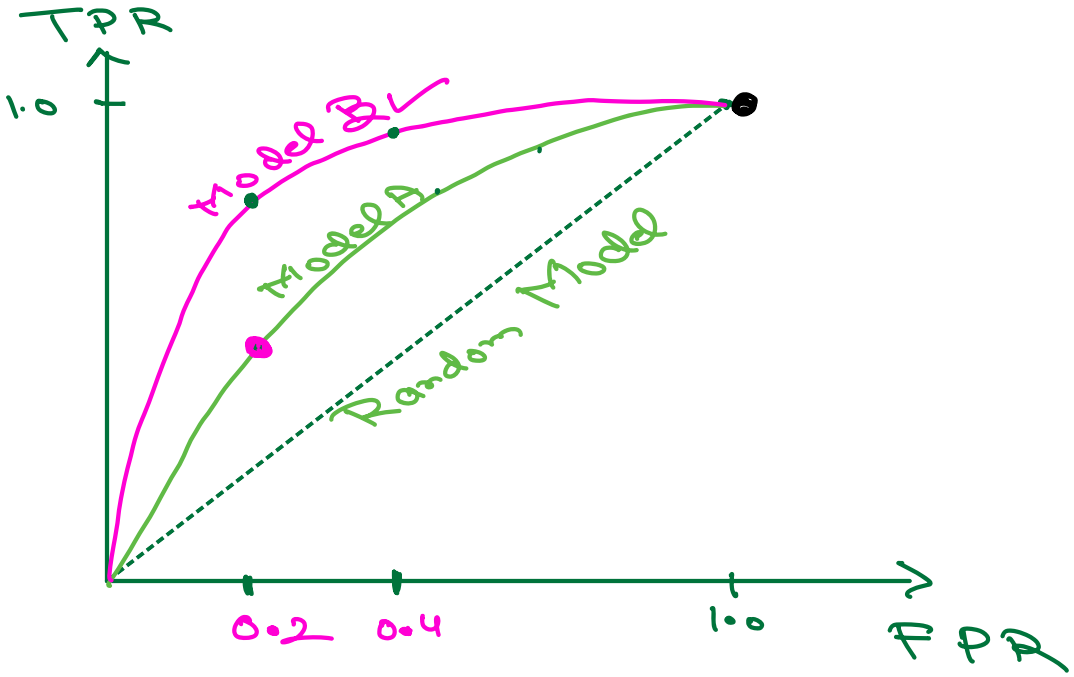
② Comparing different Models

ROC curve for a Random Model



AU-ROC

⇒ Area Under ROC curve

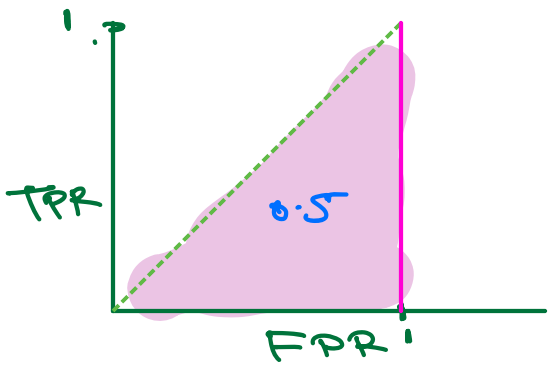


AU of Model A < AU of Model B

* For Random Model

$\frac{1}{2} \times b \times h$

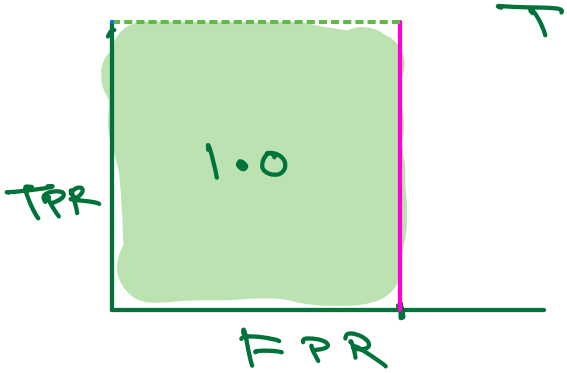
$\frac{1}{2} \times 1 \times 1 \Rightarrow 0.5$



* For ideal Model

$b \times h$

$1 \times 1 \Rightarrow 1.0$



TPR always 1.0

Issue with AU-ROC

① Dependent on Order of probability

$y \Rightarrow [1, 1, 0]$

$M1 \Rightarrow [0.95, 0.92, 0.80]$

$M2 \Rightarrow [0.20, 0.10, 0.08]$

$y_{pred} = 0.5$

y	P	$\hat{y} \geq 0.95$	$\hat{y} \geq 0.92$	$\hat{y} \geq 0.80$
1	0.95	1	1	1
1	0.92	0	1	1
0	0.80	0	0	1

M1

M1

y	P	$\hat{y} \geq 0.2$	$\hat{y} \geq 0.1$	$\hat{y} = 0.08$
1	0.2	1	1	1
1	0.1	0	1	1
0	0.08	0	0	1

M2

M2

- Will not work very well for Data Imbalance

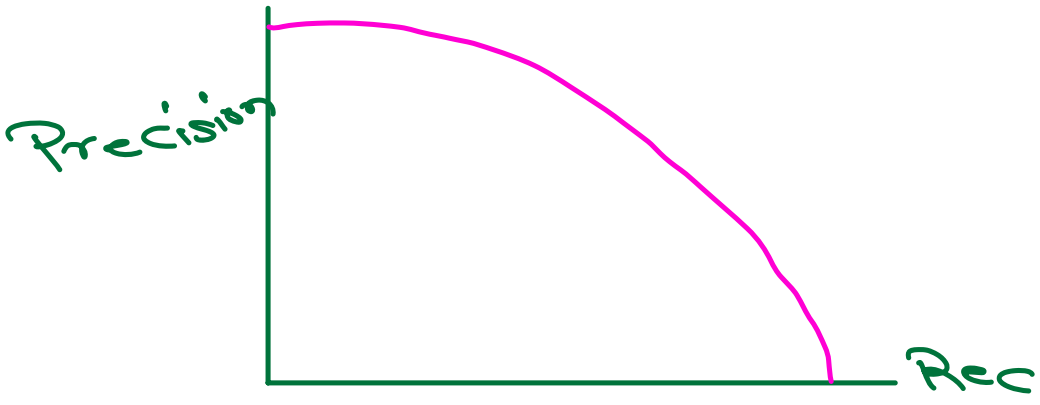
$$TPR = \frac{TP}{TP+FN}$$

$$FPR = \frac{FP}{TN+FP}$$

- PR-Curve and AU PR are better choice

all Precision Recall Curve

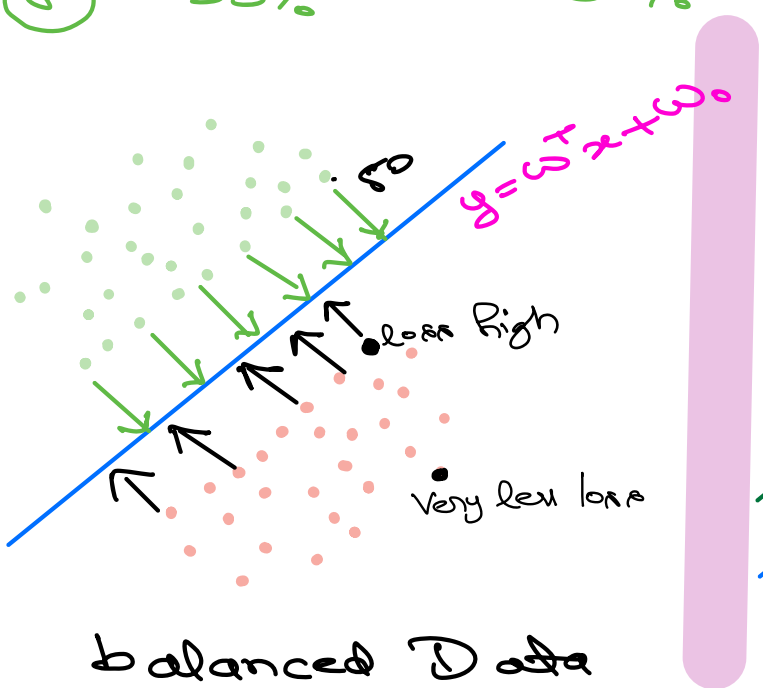
- Same steps as AU-ROC



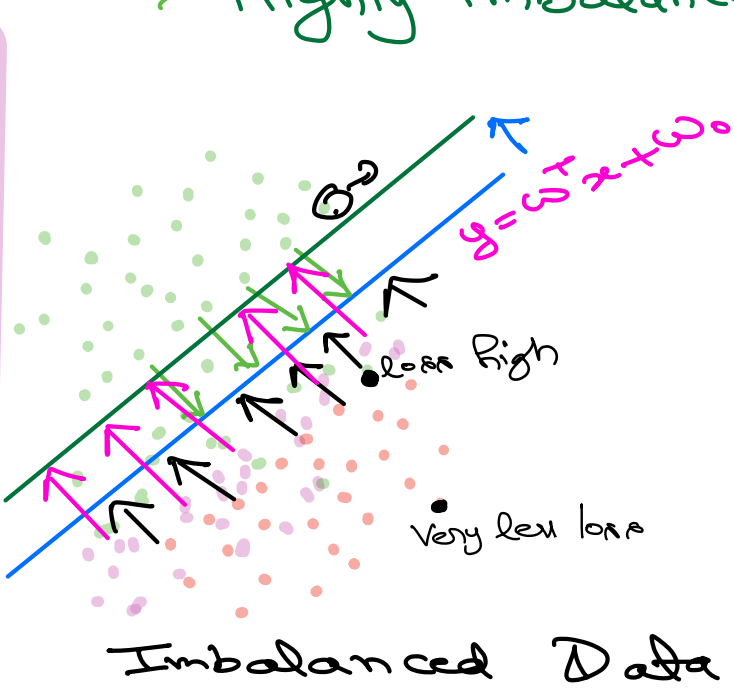
Data Imbalance

- When frequency of one category is higher than other categories

	Majority %	Minority %	Category
1	50 %	50 %	Balanced
2	60 %	40 %	Slightly Balanced
3	70 %	30 %	Slightly imbalance
4	80 %	20 %	Imbalance
5	95 %	5 %	Highly imbalanced



balanced Data



Imbalanced Data

Handling Data Imbalance

Suppose → 1000 Emails → 150 Spam, 850 Non Spam

Balanced Data

$$-\frac{1}{n}\sum \underbrace{y_i \times \log(y_p)}_{50\%} + \underbrace{(1-y) \times \log(1-y_p)}_{50\%} + \lambda w_i^2$$

balanced ⇒

Imbalanced ⇒ 25% (1) 75% (0)

+

$$-\frac{1}{n}\sum 2\left(y_i \times \log(y_p)\right) + (1-y) \times \log(1-y_p) + \lambda$$

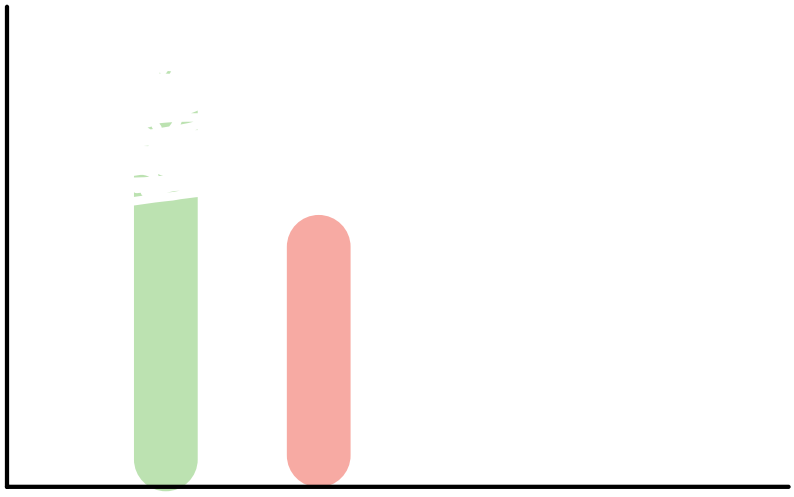
Strategy 1 : Set Class Weights

$$NLL \times [2.1] + \lambda w^2$$

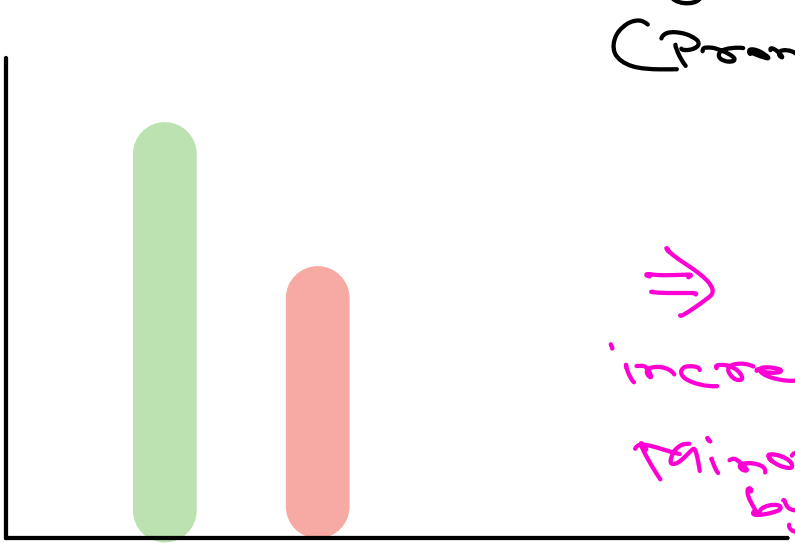
$$NLL \times \text{Class-Weights}$$

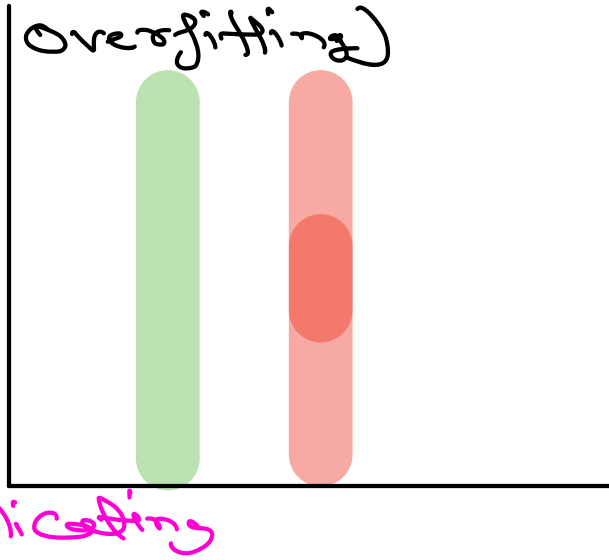
Strategy 2: Modify the Data
UnderSampling and OverSampling

① UnderSampling (Prone to Underfitting)

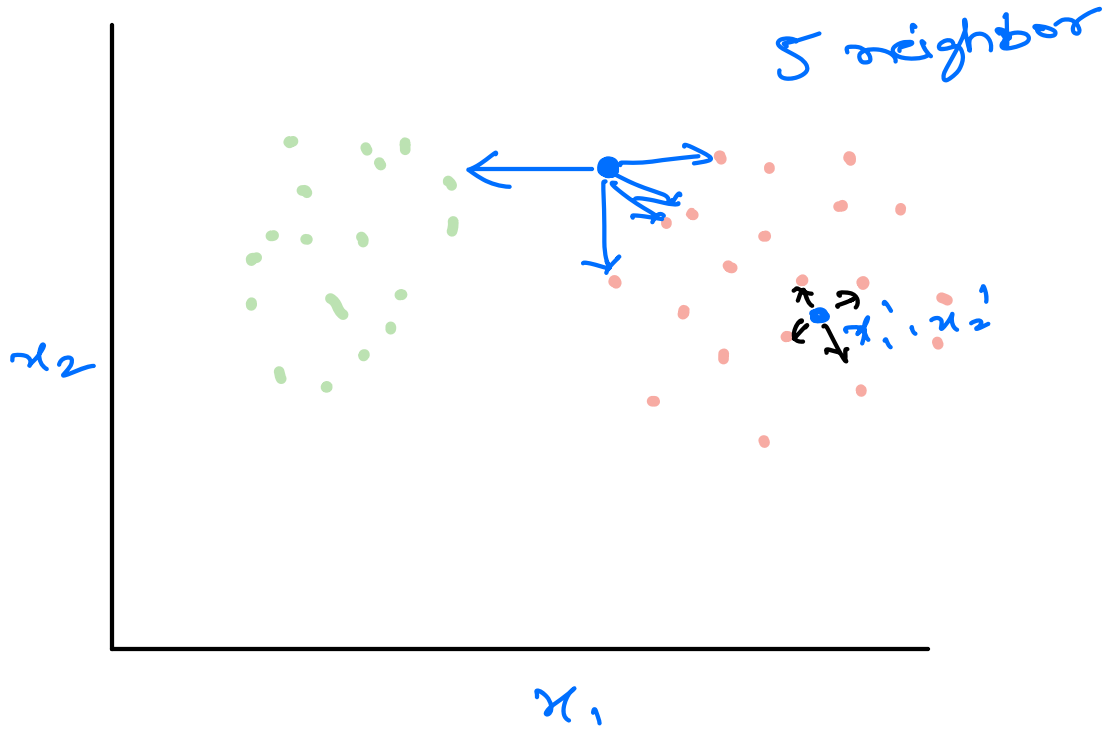


② Oversampling by duplication
(Prone to Overfitting)





Strategy 3: Synthetic Data Generation



$$x_{new} = x_1 + \lambda \, d(x_1, x_2)$$

λ → Random Number between (0, 1)

SMOTE → Synthetic Minority Oversampling Technique

0.5