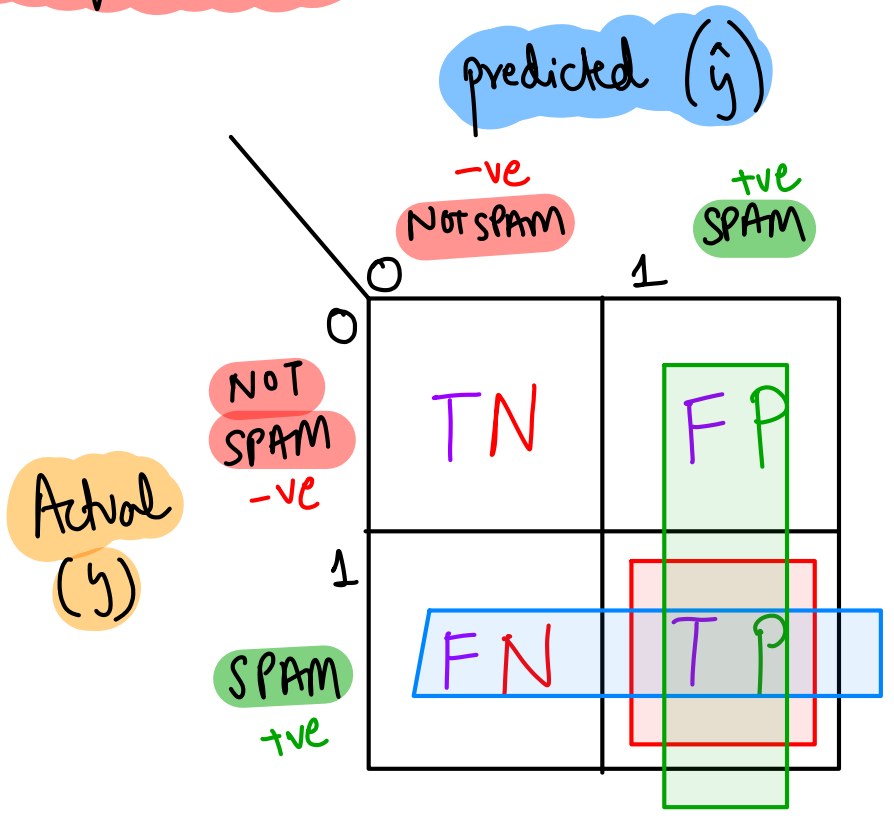
CLASSIFICATION METRICS-2 CLASS IMBALANCE

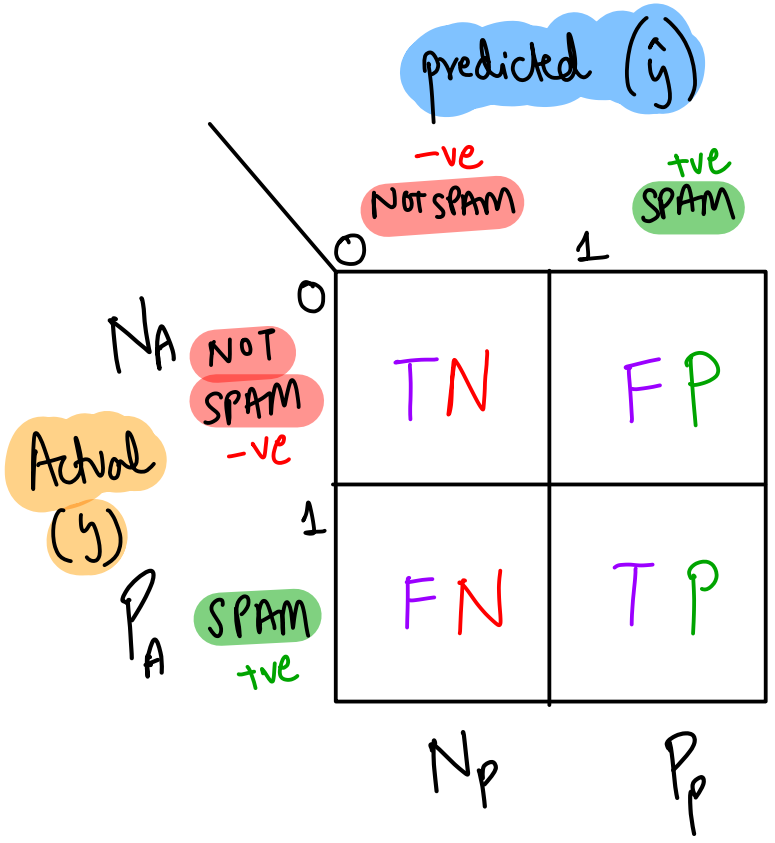
Confusion Matrix



$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP}$$

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

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



Recall

True Positive Rate = Sensitivity. $= \frac{TP}{P_A} = \frac{TP}{TP+FN}$

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

False Positive Rate $= \frac{FP}{N_A} = \frac{FP}{FP+TN}$

False Negative Rate $= \frac{FN}{P_A} = \frac{FN}{FN+TP}$

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

Classification metrics → spam/NS ⟷ $\hat{y}$

log Reg ⇒ x_i → $z_i = (w^T x + w_0)$ → $\sigma(z_i)$ → 1 if ≥ 0.5, 0 if < 0.5

ROC Curve

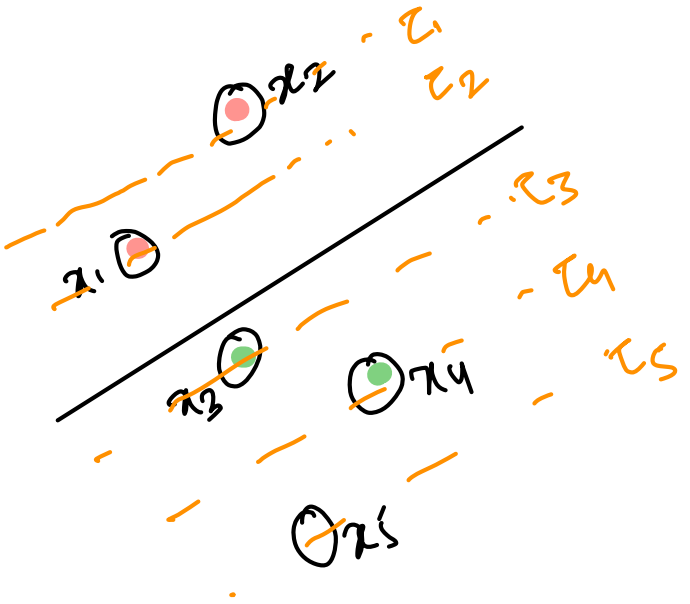
$T = 0.5$ $T = 0.6$

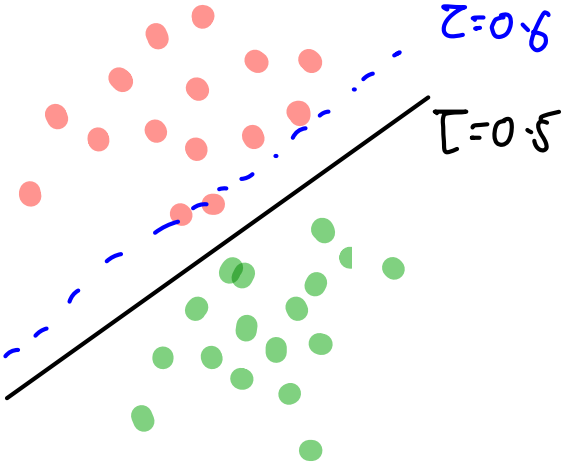
$\sigma(z_i) \geq 0.6 \Rightarrow 1$

$\sigma(z_i) < 0.6 \Rightarrow 0$

Receiver's operating characteristics.

Context : Binary Classification.



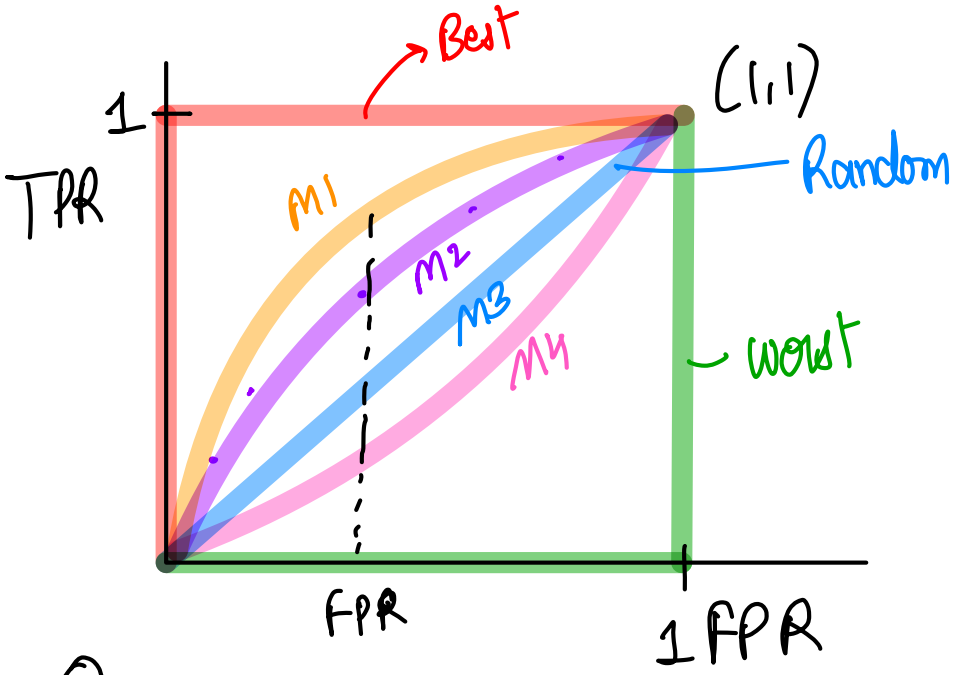


			τ	τ				τ = 0.5
X	Y	σ(xi)	τ=0.7	τ=0.6	τ=0.4	0.3	0.2	
x_2	1	0.7	1	1	1	1	1	
x_1	1	0.6	0	1	1	1	1	
x_3	1	0.4	0	0	1	1	1	
x_4	0	0.3	0	0	0	1	1	
x_5	0	0.2	0	0	0	0	1	

(1) Sort Records based on σ(xi) in descending order

(2) Find TPR and FPR for every threshold.

	P	TPR	FPR	
CM1	0.7	TPR_1	FPR_1	(— , —)
CM2	0.6	⋮		(— , —)
CM3	0.4	⋮		(— , —)
CM4	0.3	⋮		(— , —)
CM5	0.2	TPR_5	FPR_5	(— , —)



1. I
2. Models Comparison
3. [0, 1]

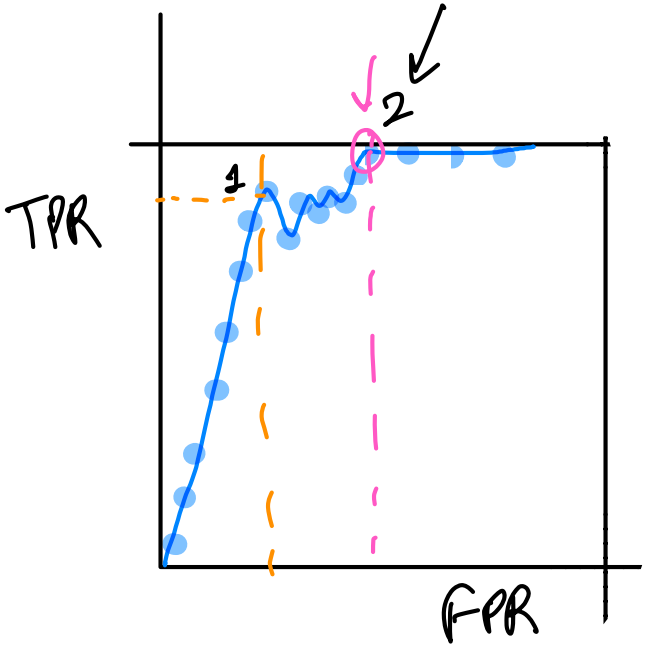
1 — $TPR = \frac{TP}{TP+FN}$

$FPR = \frac{FP}{N} = \frac{FP}{FP+TN}$ — 0

Best ⇒ TPR = 1, FPR = 0

Worst ⇒ TPR = 0, FPR = 1

Random model:



How to choose 2 for model?

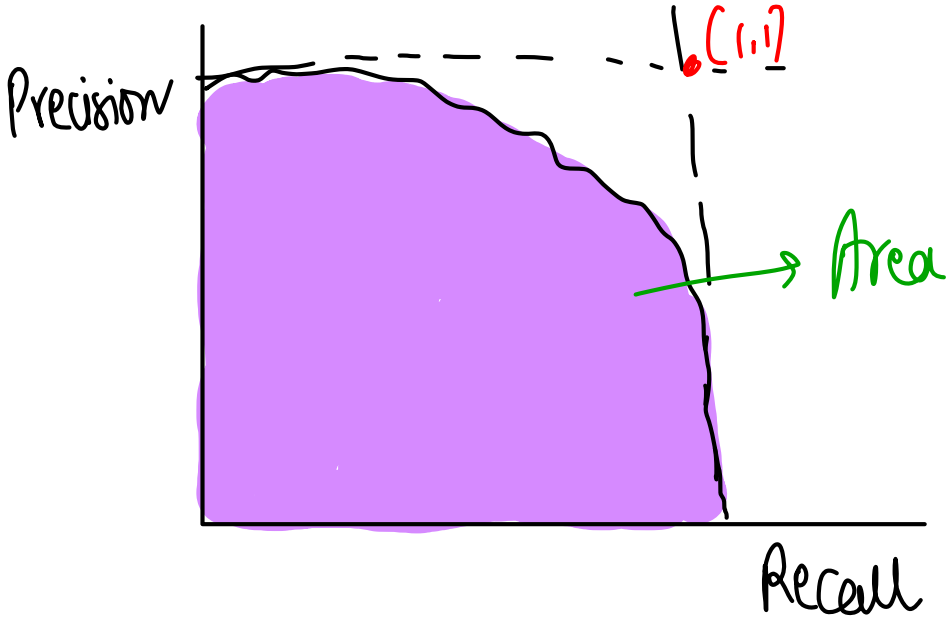
TPR ↑

FPR ↓

(2) ✓

problems with ROC-AUC

(1) doesn't work with imbalanced dataset



Area under PRC → Curve.

precision

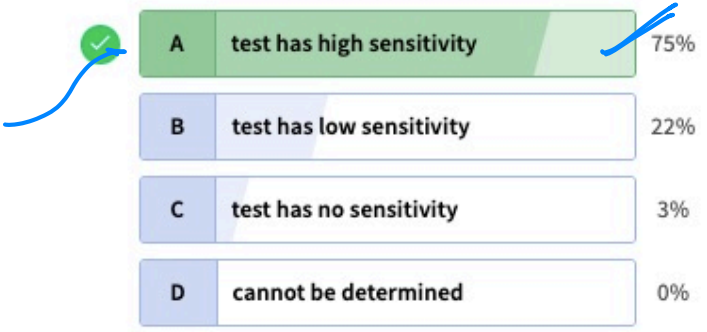
Recall

Quiz time!

Quiz Ended!

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

32 users have participated



Quiz time!

Quiz Ended!

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

28 users have participated

A	2 points (0,0) and (1,1)	18%
B	3 points representing the thresholds 0.25, 0.5, and 0.75	11%
C	10 points equally spaced between 0 and 1	4%
D	Depends on the number of unique threshold values	68%

Quiz time!

Quiz Ended!

Which of the following metrics can be directly derived from the ROC curve?

31 users have participated

A	Accuracy	6%
B	Precision	3%
C	Recall	3%
D	Area Under the Curve (AUC)	87%

Class Imbalanced

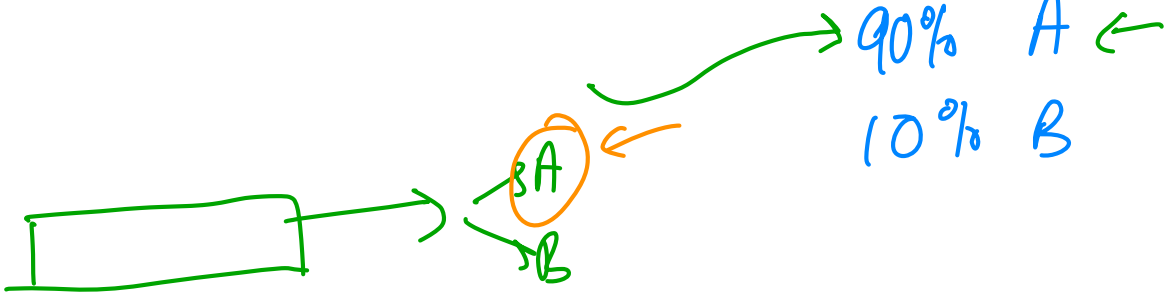
One class → dominated.

A	B	
50%	50%	→ Balanced
60%	40%	→ Slightly Balanced
70%	30%	→ Slightly Imbalanced
80%	20%	→ Imbalanced
90%	10%	→ Highly Imbalanced.
95%	5%	→ Highly Imbalanced.
99%	1%	→ Highly Imbalanced.

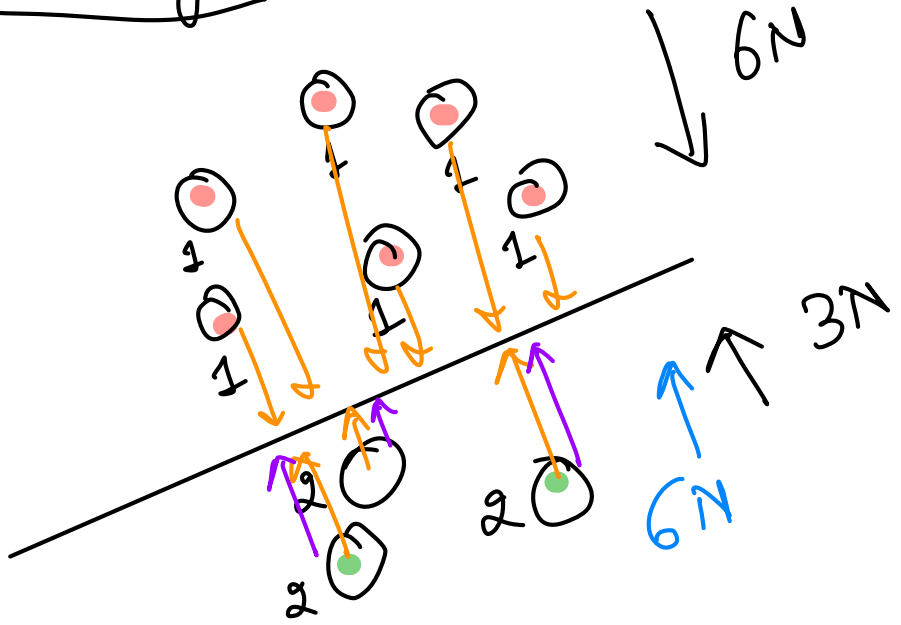
* value counts
* Count plot

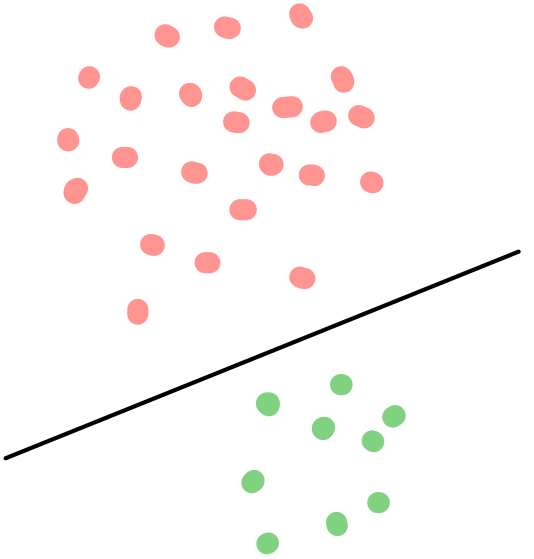
Problems with imbalanced dataset

1. Accuracy / Metrics become unreliable
2. Model becomes biased towards majority class



Class Weights

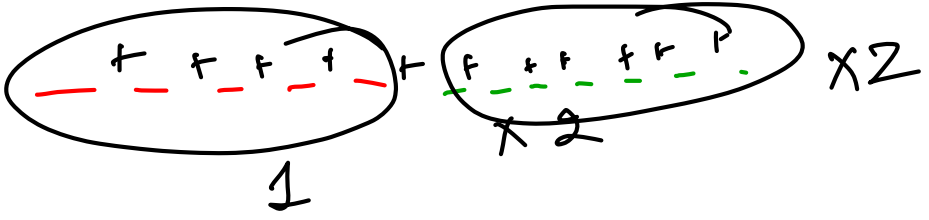




6 Red 3 Green

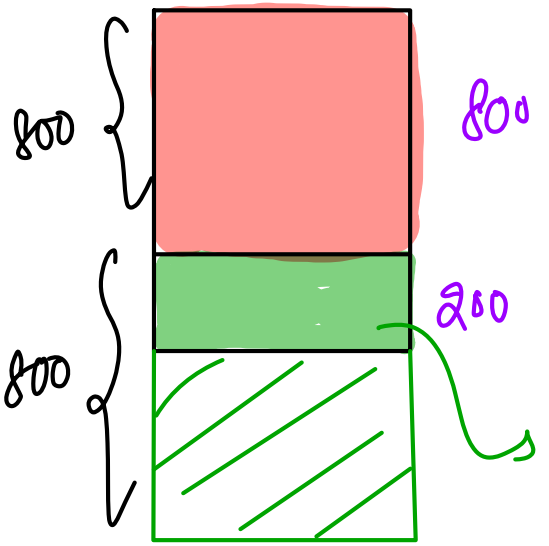
{ "•" : 2 "•" : 1 }

LOSS: $w_j \times \log \text{loss} + \lambda \text{Reg}$



$$-\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)\right]$$

Over Sampling



} majority

} minority

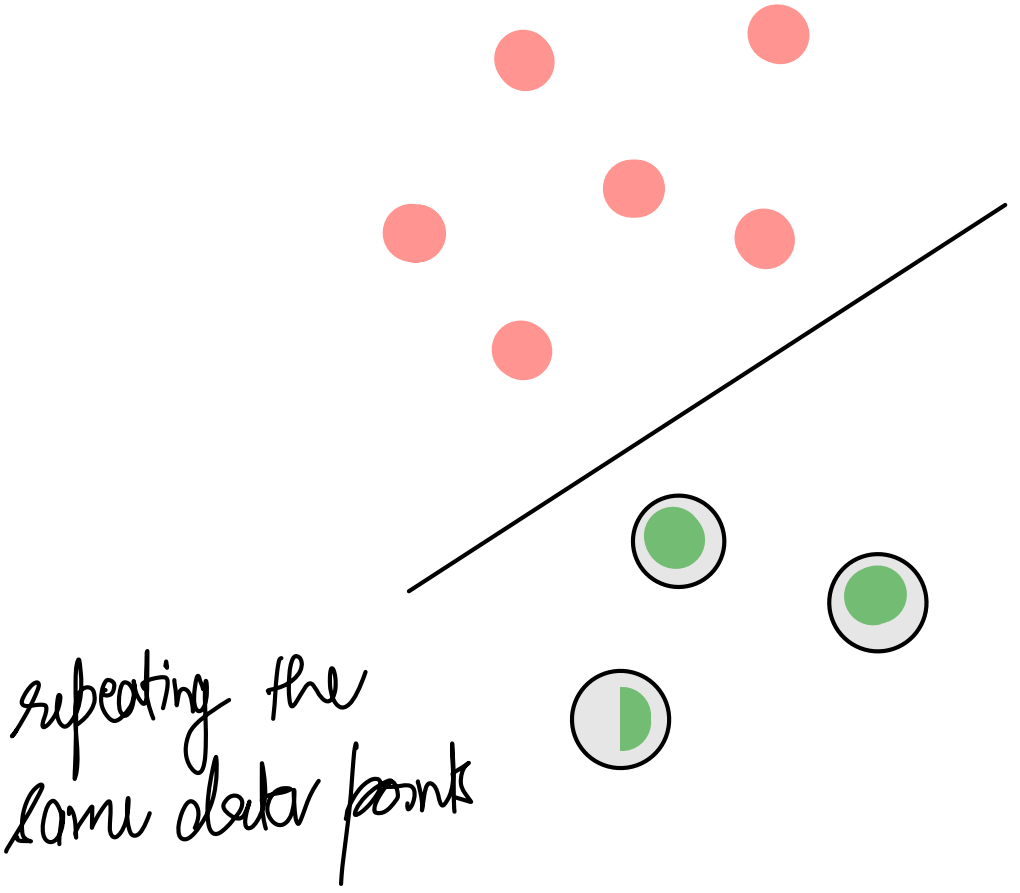
Random Sampling with repetition

disadvantage

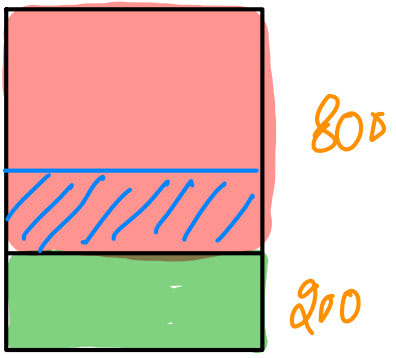
1. possibility of overfitting
2. Duplication of data

Advantage :

1. No data loss
2. Better than Under sampling


repeating the
same data points

Under Sampling



Random Selection Sampling

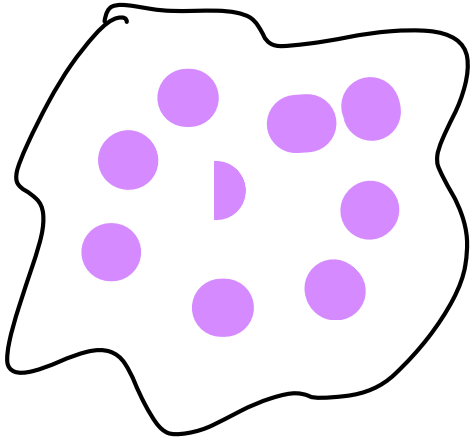
≠ # datapoints in the minority class.

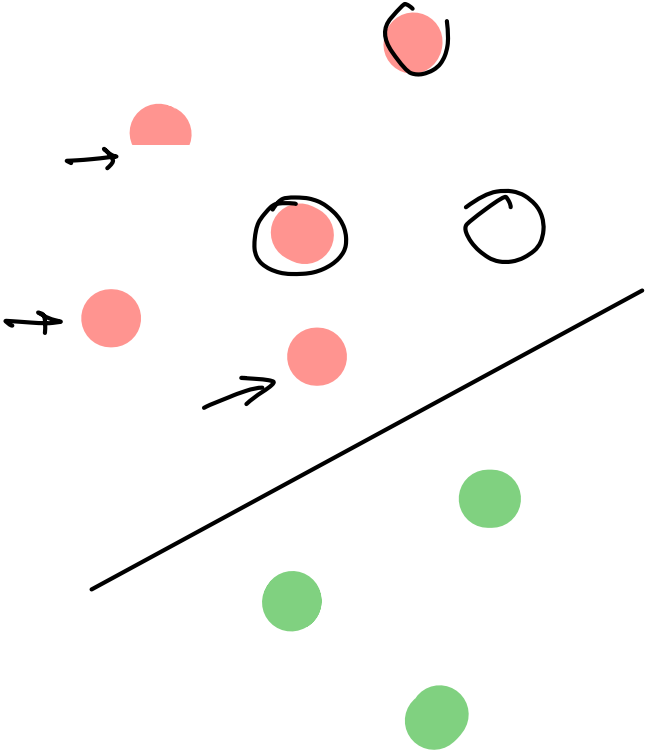
disadvantage

1. Data loss
   Possibility of losing some very imp data point
2. Sample may be biased.

Advantage:

1. time / compute / cost effective





SMOTE → Synthetic Minority Oversampling Technique

Quiz time!

Quiz Ended!

Which among the following is a balanced data ?

31 users have participated

A	50 -ve samples, 50 +ve samples	97%
B	100 -ve samples, 10 +ve samples	0%
C	10 -ve samples, 100 +ve samples	0%
D	2 -ve samples, 98 +ve samples	3%

Quiz time!

Quiz Ended!

Why are class weights important in imbalanced datasets?

31 users have participated

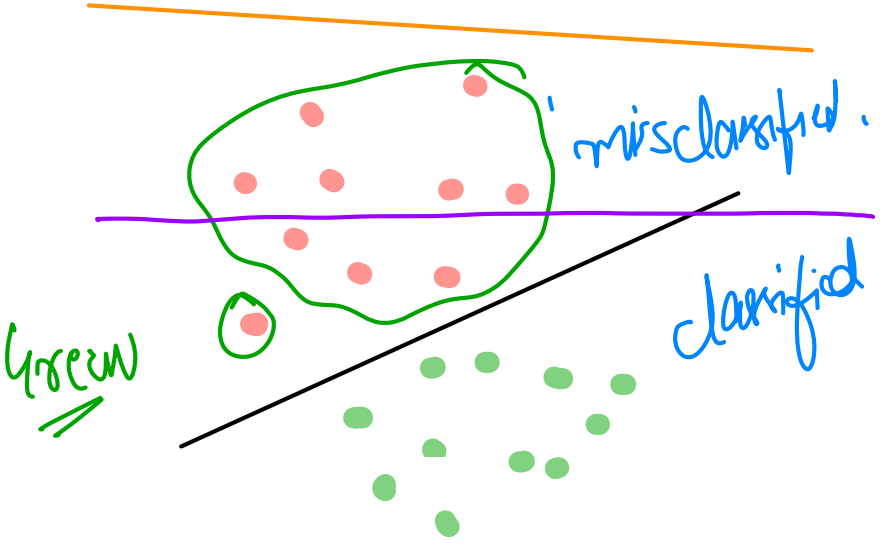
A	They improve the convergence speed of the model	10%
B	They reduce overfitting in the model	13%
C	They help give more importance to the minority class	71%
D	They enhance the model's ability to handle missing values	6%

① Assessment / Problem Solving Class

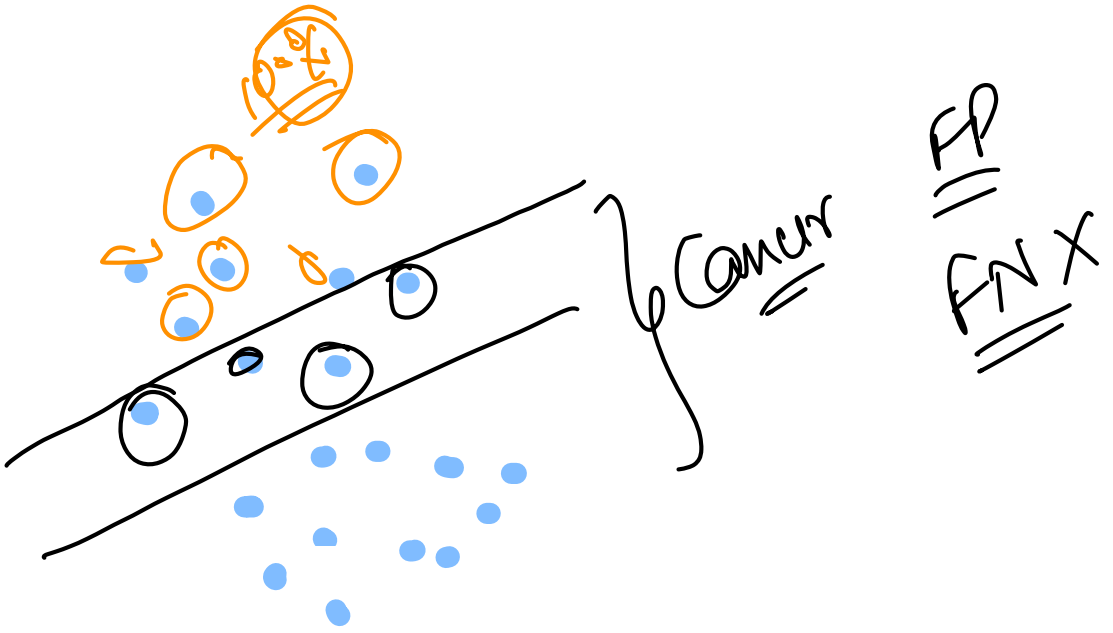
② Code →

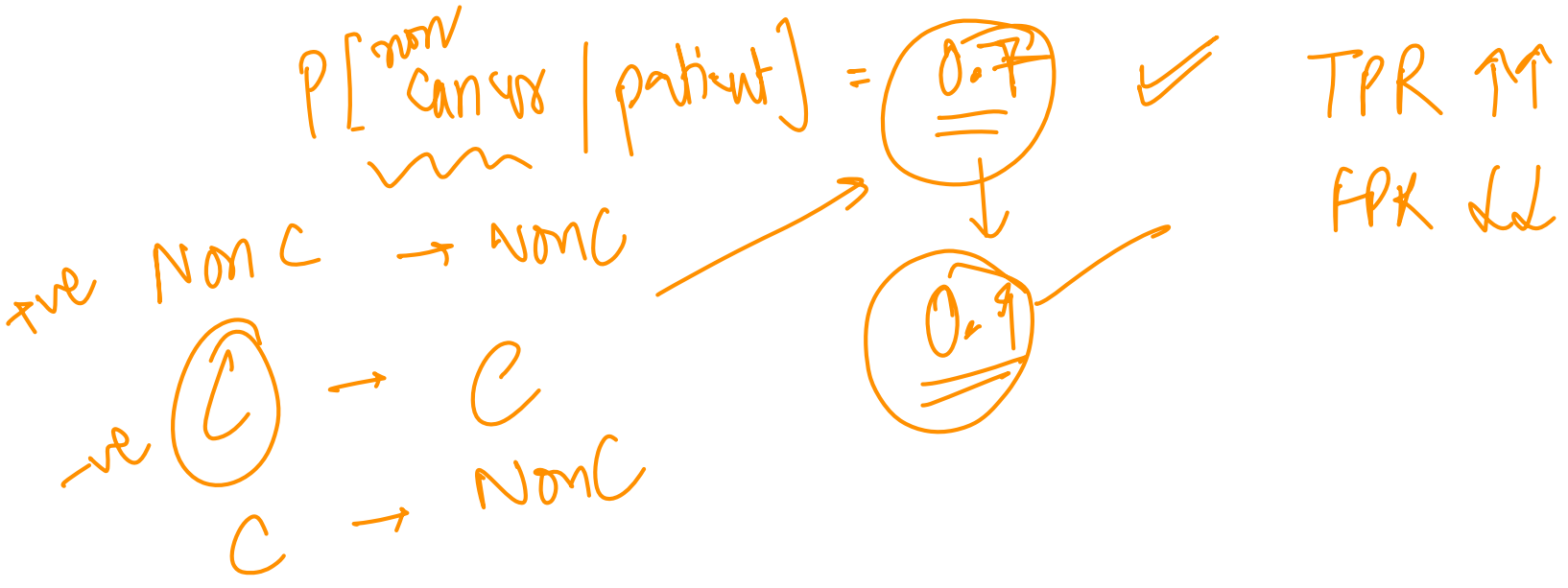
$x_i^0 \rightarrow W^T x_i^0 + w_0 \rightarrow \sigma(z_i) \rightarrow$ if $\sigma(z_i) \geq 0.5$ 1

$(0, 0.5)$ if $\sigma(z_i) < 0.5$ 0



Loss ↑↑↑


Courbe
FP
FN X


P[non cancer | patient] = 0.7
TPR ↑↑
FPR ↓↓
+ve Non C → NonC
-ve C → C
C → NonC
0.9