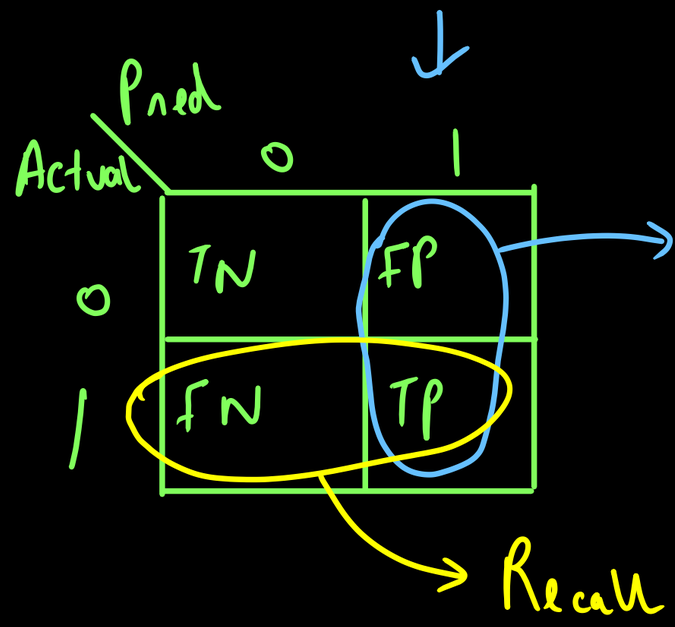
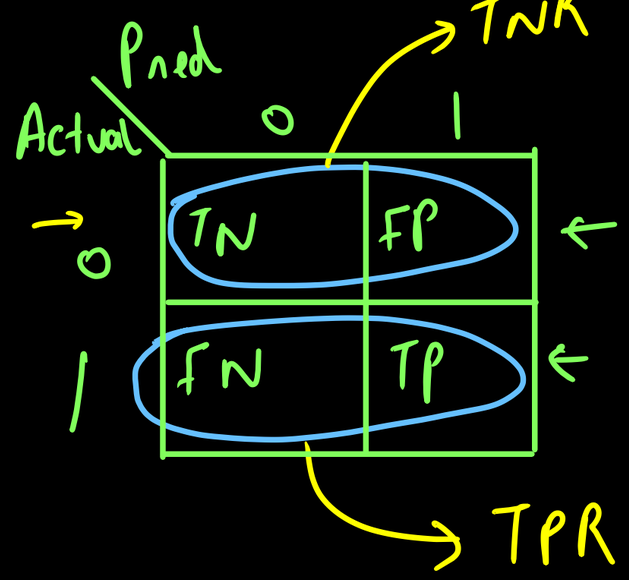
Agenda

- → TPR, TNR, FPR, FNR
- → Roc Auc curve
- → PR curve
- → Imbalance data



$$\uparrow Prec = \frac{TP}{TP + FP} \downarrow \qquad \uparrow Rec = \frac{TP}{TP + FN} \downarrow$$

$$F1 = \frac{2PR}{P + R}$$



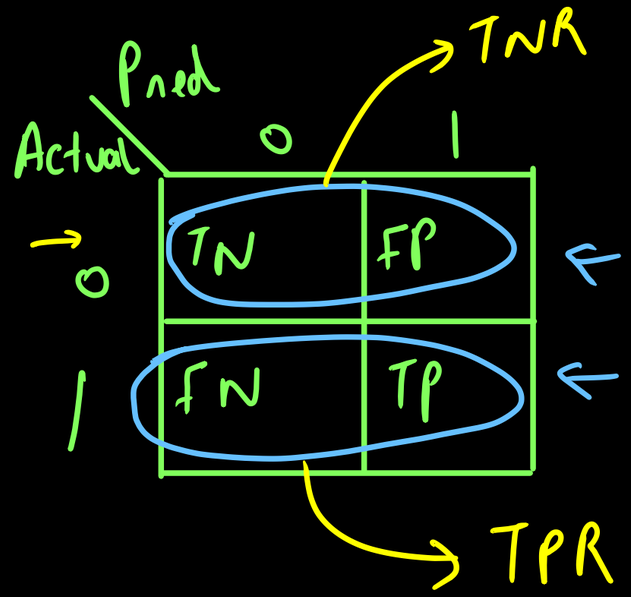
True Positive Rate

$$\uparrow TPR = \frac{TP}{\text{No of Actual} = 1} = \frac{TP}{TP + FN} = \text{Recall}$$

$$\uparrow TNR = \frac{TN}{\text{No of Actual} = 0} = \frac{TN}{TN + FP} = \text{Specivity}$$

→ FPR FNR → False Negative Rate

↳ False Positive Rate



$$\downarrow FPR = \frac{FP\downarrow}{FP + TN} = 1 - TNR$$

$$\downarrow FNR = \frac{FN\downarrow}{FN + TP} = 1 - TPR$$

Min log-loss P, R, F1

$z^{(i)} = w^T x^{(i)} + b$

$\sigma(z^{(i)})$

$\hat{y} = P(y=1/x)$

$\hat{y} \geq \tau$ — τ = threshold = 0.5

$\hat{y} \geq \tau \rightarrow 1$

$< \tau \rightarrow 0$

y	$\hat{y}$
1	0.9
1	0.8
0	0.7
0	0.6
0	0.3

$\hat{y} \geq 0.5$	
1	✓
1	✓
1	x
1	x
0	✓

$\hat{y} \geq 0.8$	
1	✓
1	✓
0	✓
0	✓
0	✓

Actual \ Pred	0	1
0	TN 1	FP 2
1	FN 0	TP 2

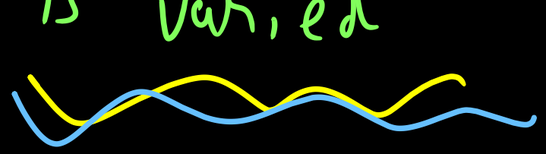
Actual \ Pred	0	1
0	TN 3	FP 0
1	FN 0	TP 2

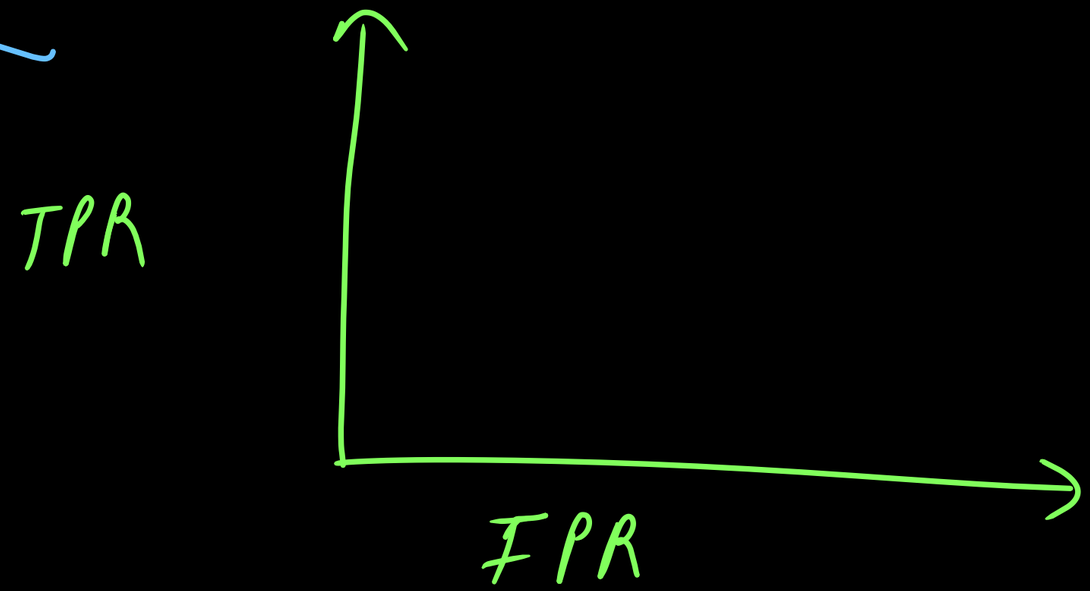
→ Problem of P, R, f1

↳ Dependent on threshold value

ROC (Receiver Operating Characteristics)

↳ Defined as plot b/w TPR (Recall) and FPR as decision threshold is varied





y	$\hat{y}$
1	0.9
0	0.6
1	0.7
0	0.3
1	0.8

① Sort in decreasing probab

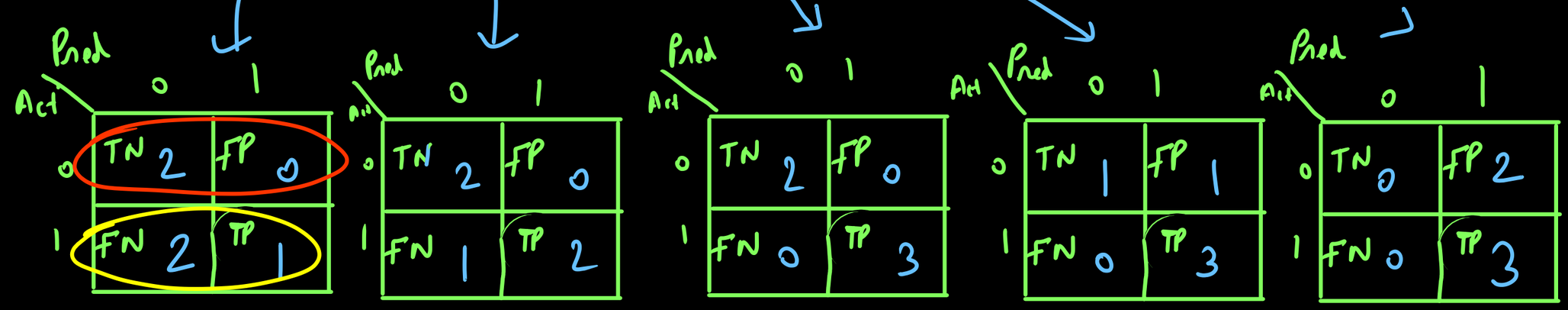
y	$\hat{y}$
1	0.9
1	0.8
1	0.7
0	0.6
0	0.3

2. Choose diff. threshold and calculate prediction

y	$\hat{y}$	$\hat{y} \geq 1$	$\hat{y} \geq 0.9$	$\hat{y} \geq 0.8$	$\hat{y} \geq 0.7$	$\hat{y} \geq 0.6$	$\hat{y} \geq 0.3$	$\hat{y} \geq 0$
1	0.9	0	1	1	1	1	1	1
1	0.8	0	0	1	1	1	1	1
1	0.7	0	0	0	1	1	1	1
0	0.6	0	0	0	0	1	1	1
0	0.3	0	0	0	0	0	1	1

3. Plot confusion matrix & get TPR and FPR

y	$\hat{y}$	$\hat{y} \geq 0.9$	$\hat{y} \geq 0.8$	$\hat{y} \geq 0.7$	$\hat{y} \geq 0.6$	$\hat{y} \geq 0.3$
1	0.9	1	1	1	1	1
1	0.8	0	1	1	1	1
1	0.7	0	0	1	1	1
0	0.6	0	0	0	1	1
0	0.3	0	0	0	0	1



$TRR = 1/1+2 = 1/3 \rightarrow 0.33$ $\quad 2/3 \rightarrow 0.66$ $\quad 3/3+0 = 1$ $\quad 3/3+0 = 1$ $\quad 3/3+0 = 1$

$FPR = 0/2+0 = 0$ $\quad 0/2 = 0$ $\quad 0/2+0 = 0$ $\quad 1/1+1 = 0.5$ $\quad 2/2+0 = 1$

Ideal

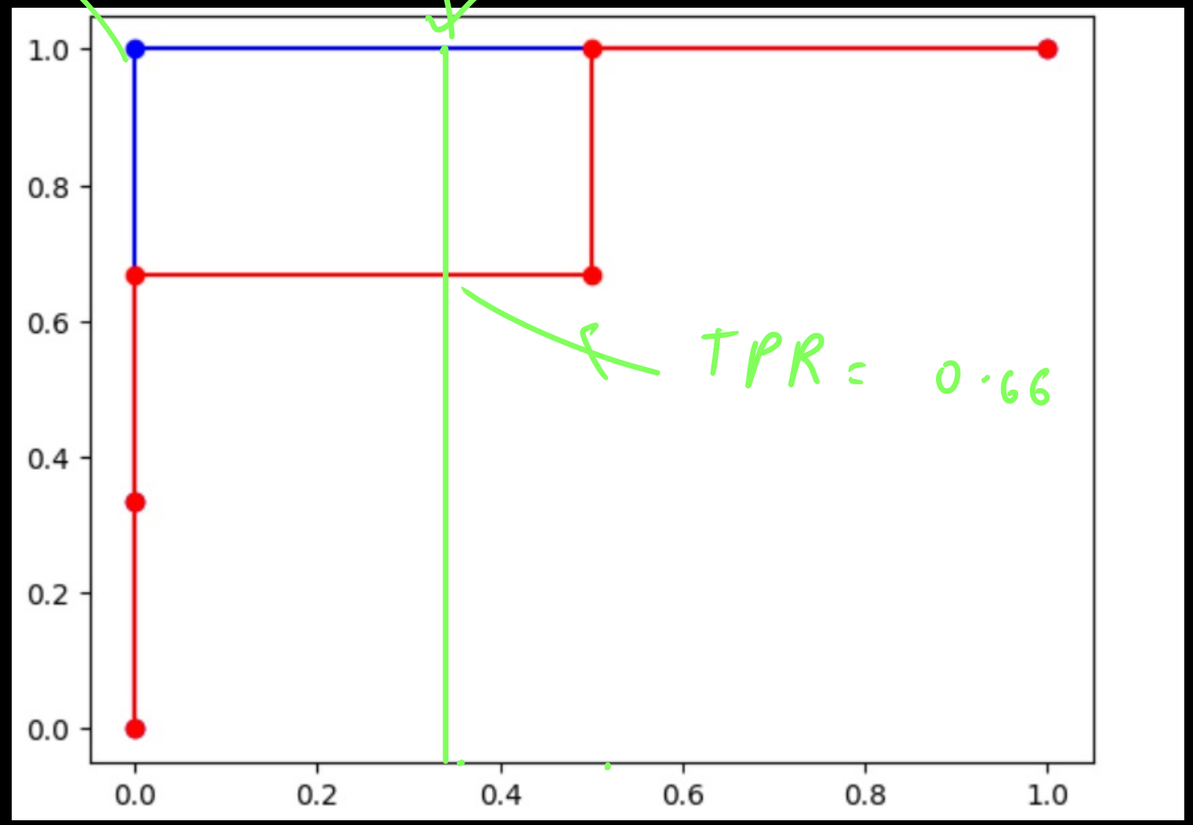
TPR = 1

1) R

2) B

↑ TPR

TPR = 0.66



TPR ↑
FPR ↓

FPR →

TRR = $1/1+2 = 1/3$ → 0.33

FPR = $0/2+0 = 0$

$2/3$ → 0.66

$0/2 = 0$

$3/3+0 = 1$

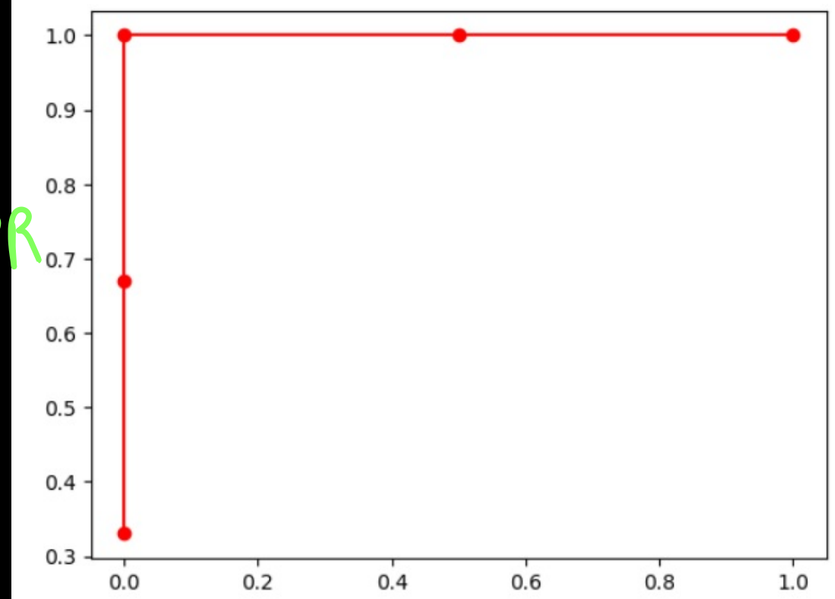
$0/2+0 = 0$

$3/3+0 = 1$ | $3/3+0 = 1$

$1/1+1 = 0.5$ | $2/2+0 = 1$

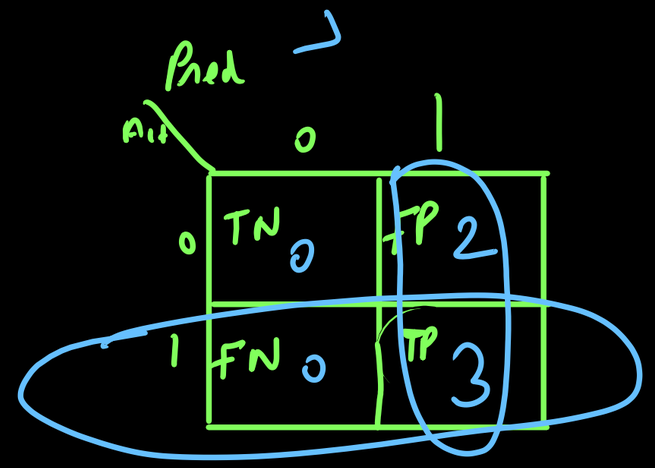
```
import matplotlib.pyplot as plt

FPR = [0,   0,   0,0.5, 1]
TPR = [0.33,0.67,1, 1,  1]
plt.plot(FPR,TPR,color='red' ,marker='o')
```

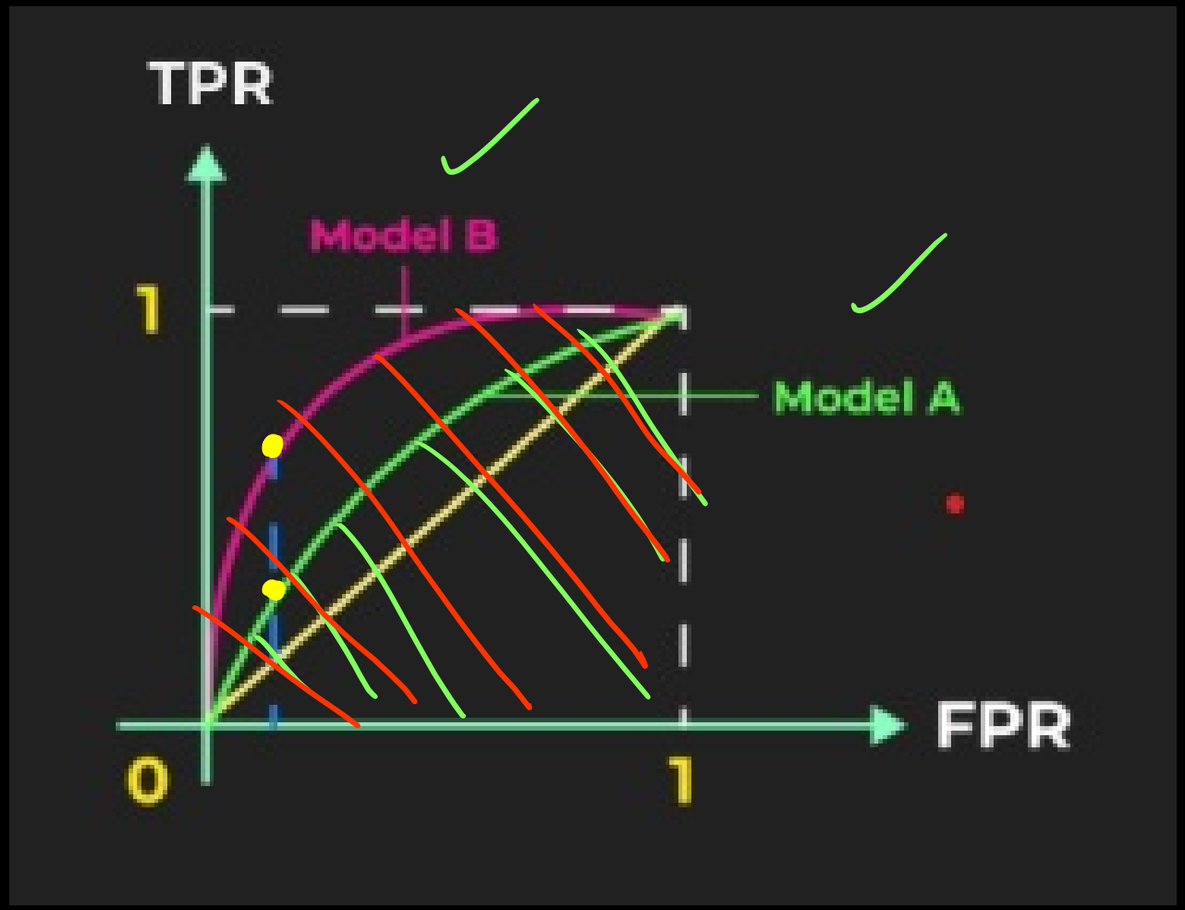
```
[<matplotlib.lines.Line2D at 0x79874c95fc90>]
```



TPR ↑

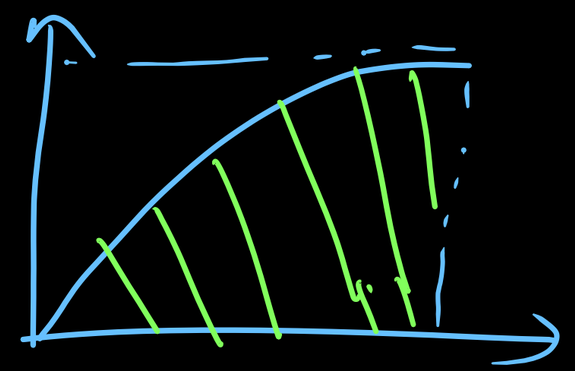
FPR →

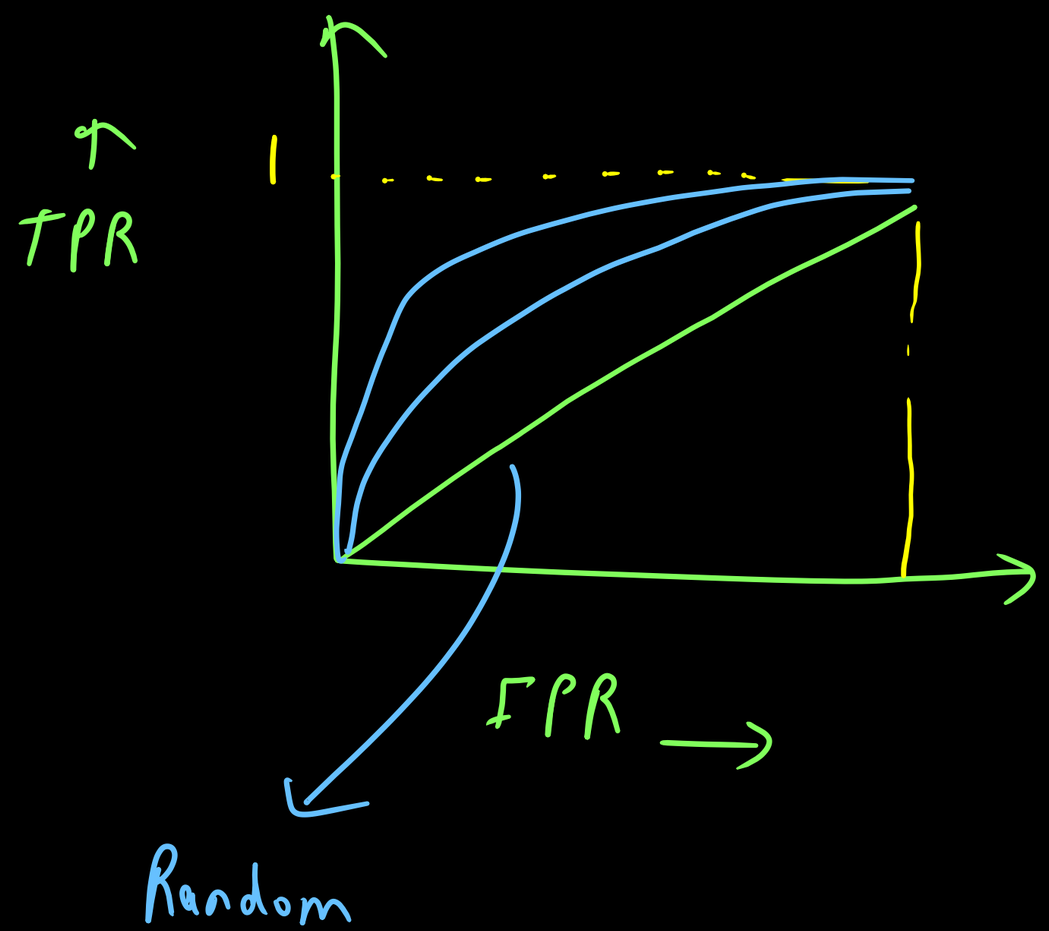




1) A

2) B





↑ AUC
(Area under
curve)

Problem for AUC -ROC

1. Does not work for imbalanced dataset
2. It takes "order of Probab" and not actual probab

Model 1 →

M1

y	P_{m1}	$\hat{y}_\tau \geq 0.95$	$\hat{y}_\tau = 0.92$
1	0.95	1	1
1	0.92	0	1
0	0.80	0	0
1	0.76	0	0
1	0.71	0	0

Same

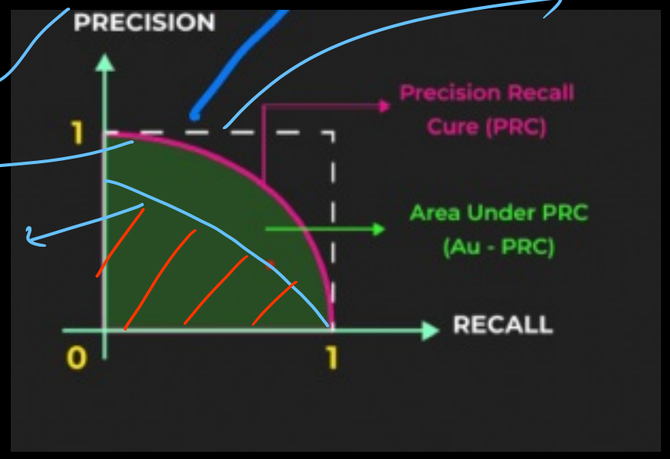
Model 2 →

y	P_{m2}	$\hat{y}_\tau \geq 0.2$	$\hat{y}_\tau = 0.1$
1	0.2	1	1
1	0.1	0	1
0	0.08	0	0
1	0.06	0	0
1	0.02	0	0

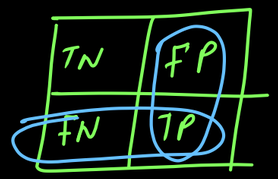
→ Break until 22:33 PM

PR-curve

Ideal case



AUC – PR curve

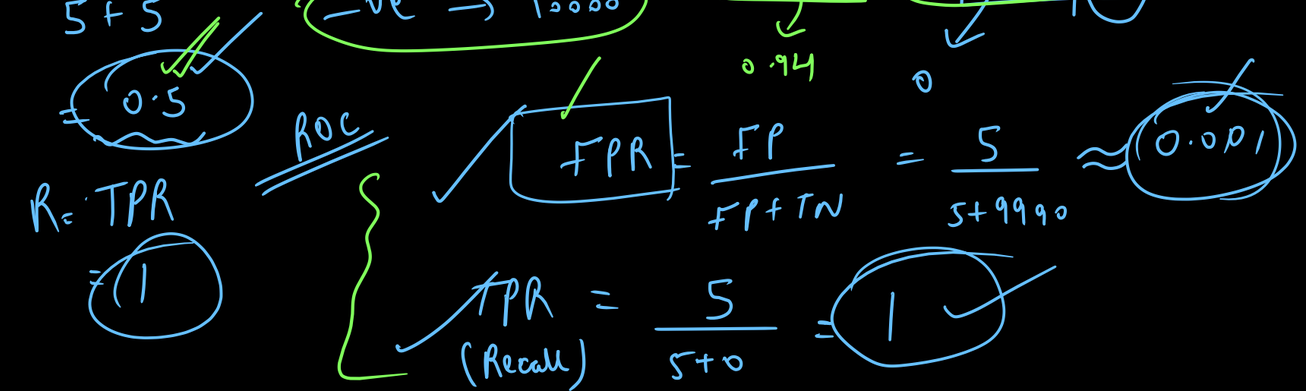


ROC → Balanced

PR → Imbalanced dataset & Balanced Dataset

-ve → 9990

+ve → 10



→ Imbalanced data

No. of samples of **one class** > No. of samples of **other class**

Majority Class — Minority Class

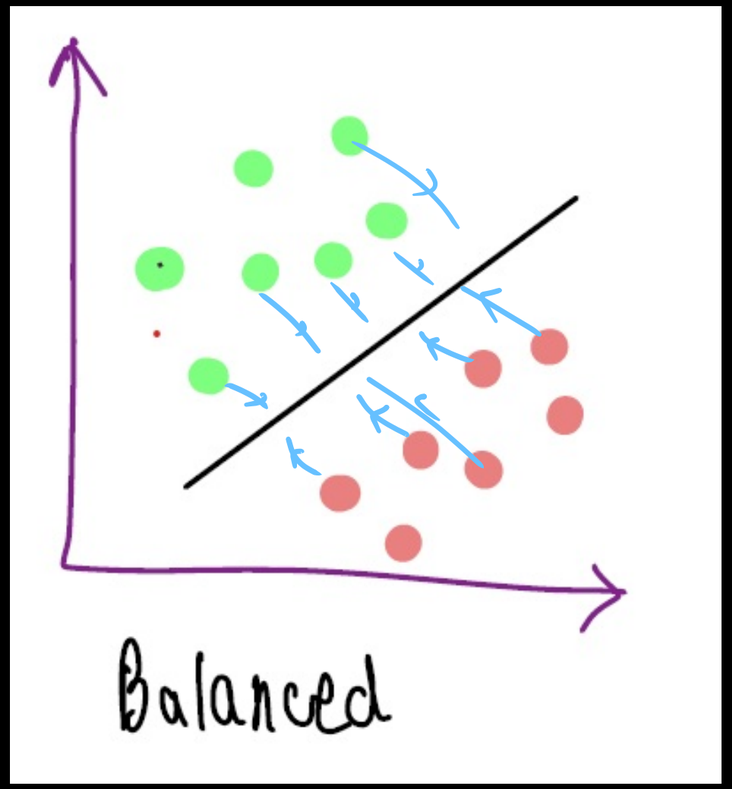
50 % - 50% → **Data Balanced**

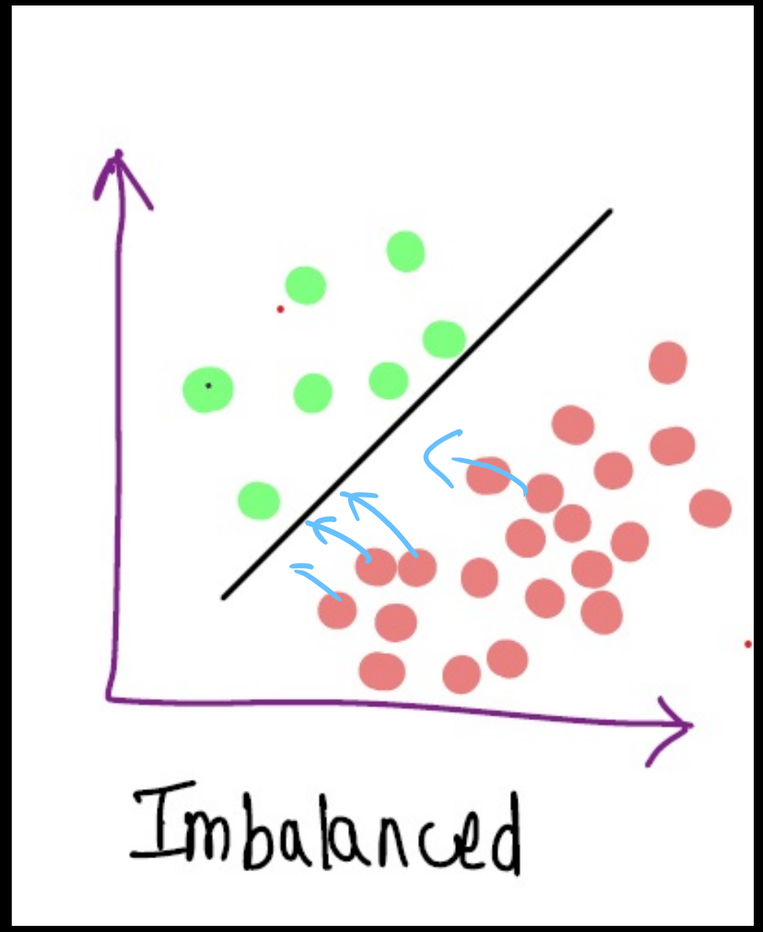
60 % - 40 % → **Slightly Balanced**

70 % - 30% → **Slightly Imbalanced**

80 / 90 % - 20 / 10 % → **Data Imbalanced**

95 % - 5 % → **Extremely Imbalanced**


Balanced


Imbalanced

Spam Classification

1. Spam → 150

2. NS → 850

$$\frac{NS}{S} = \frac{850}{150} = 5.67$$

$$NS = 5.67S$$

→ More weightage to minority class

$$\therefore \; loss = \sum_{i=1}^{n} logloss_i \; W_i + \lambda \sum_{j=1}^{d} w_j^2$$

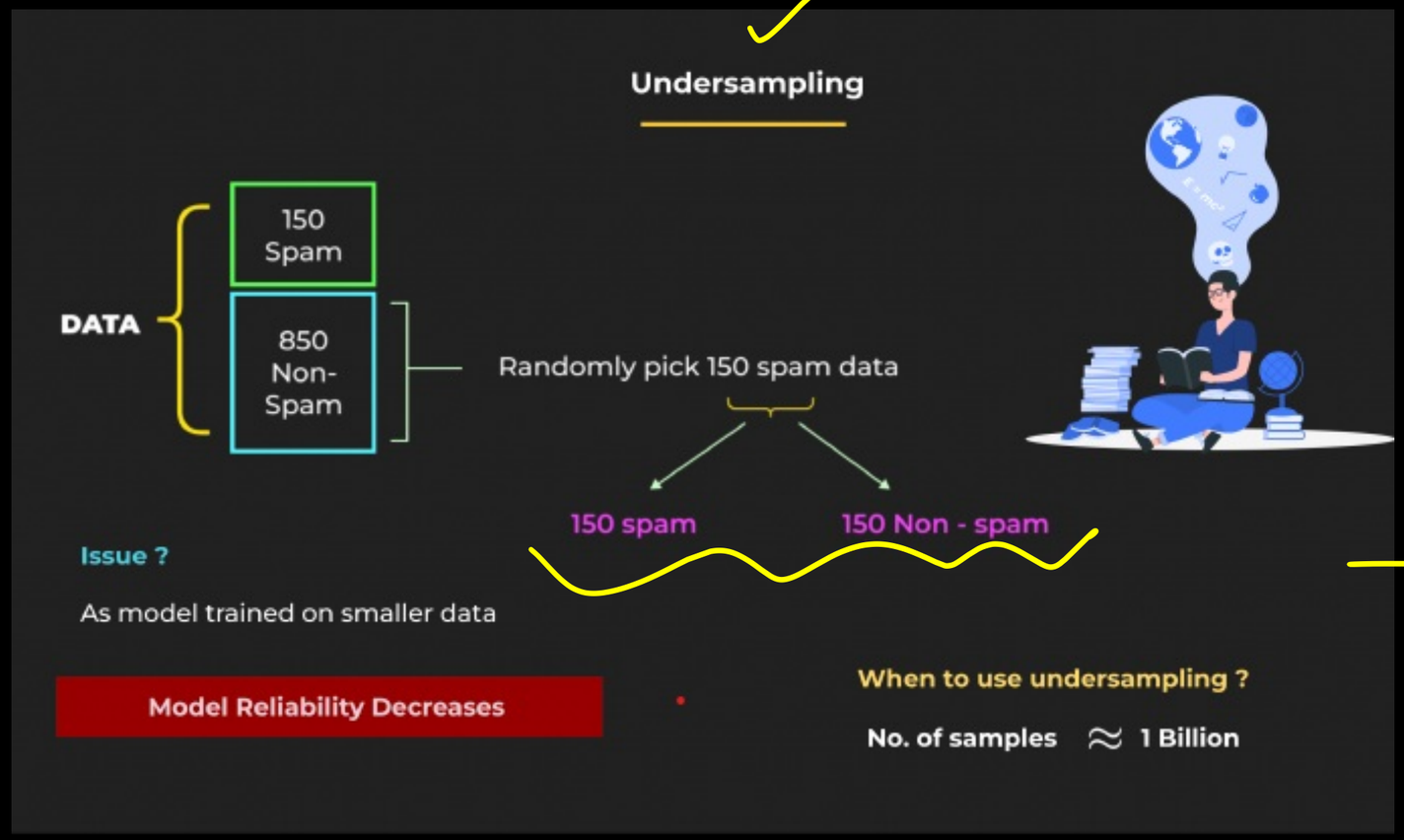
$W_i = 5.67$ when spam

$W_i = 1$ when non - spam

$\frac{NS}{S} = \frac{0}{1} = 5.67$

{0: 1 , 1: 5.67}

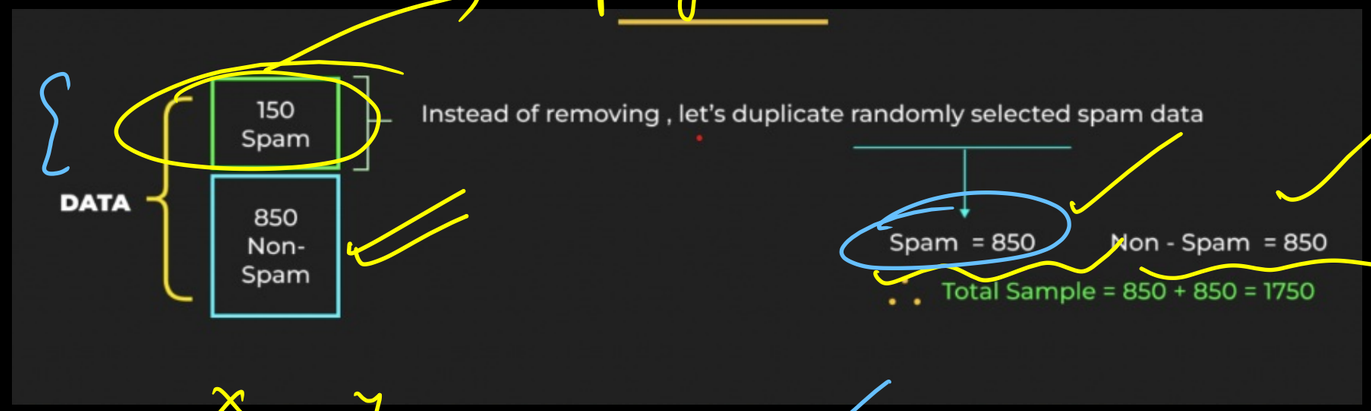
2) Undersampling

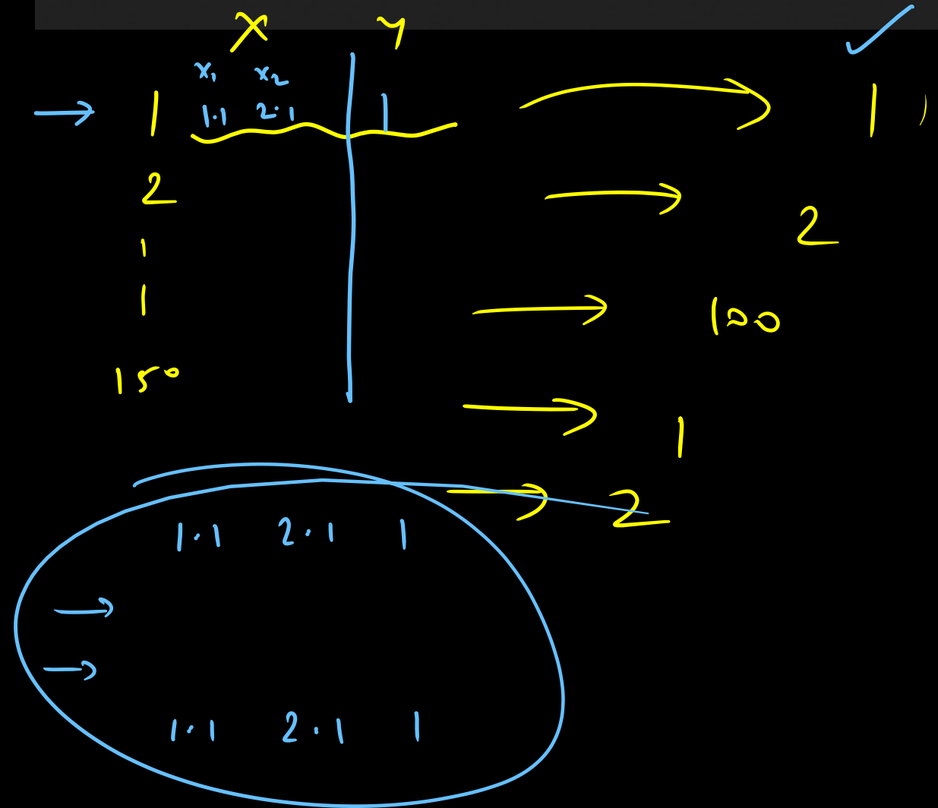


→ Data loss

3) Oversampling

Sampling with replacement





→ Duplicate data

→ Overfitting

SMOTE

↳ Synthetically Minority Oversample technique

↳ Generate Synthetic data

1. Select Random point from minority class
2. Find $K = 3$ nearest point
3. Randomly select one of K - neighbor

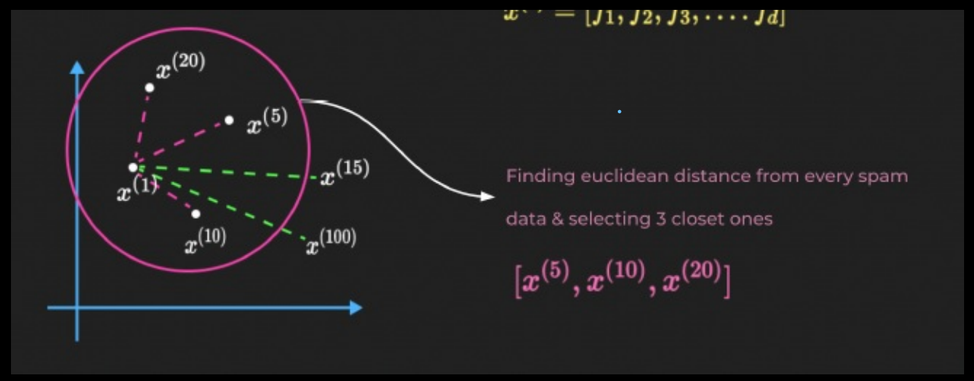
$$X_{syn} = X_i + \lambda (X_j - X_i)$$

1. Select any Random point from minority class
2. Find K = 3 nearest point
3. Randomly select one of K - neighbor

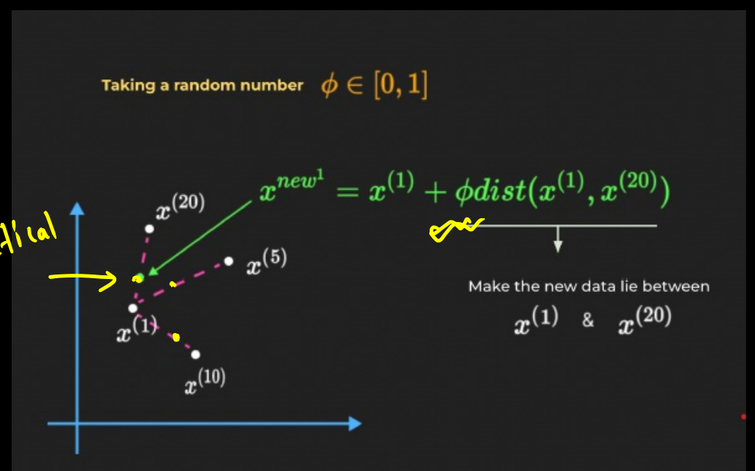
$$X_{syn} = X_i + \lambda (X_j - X_i)$$

0.2

$x^{(1)}$ $x^{(5)}$, $x^{(10)}$, $x^{(20)}$

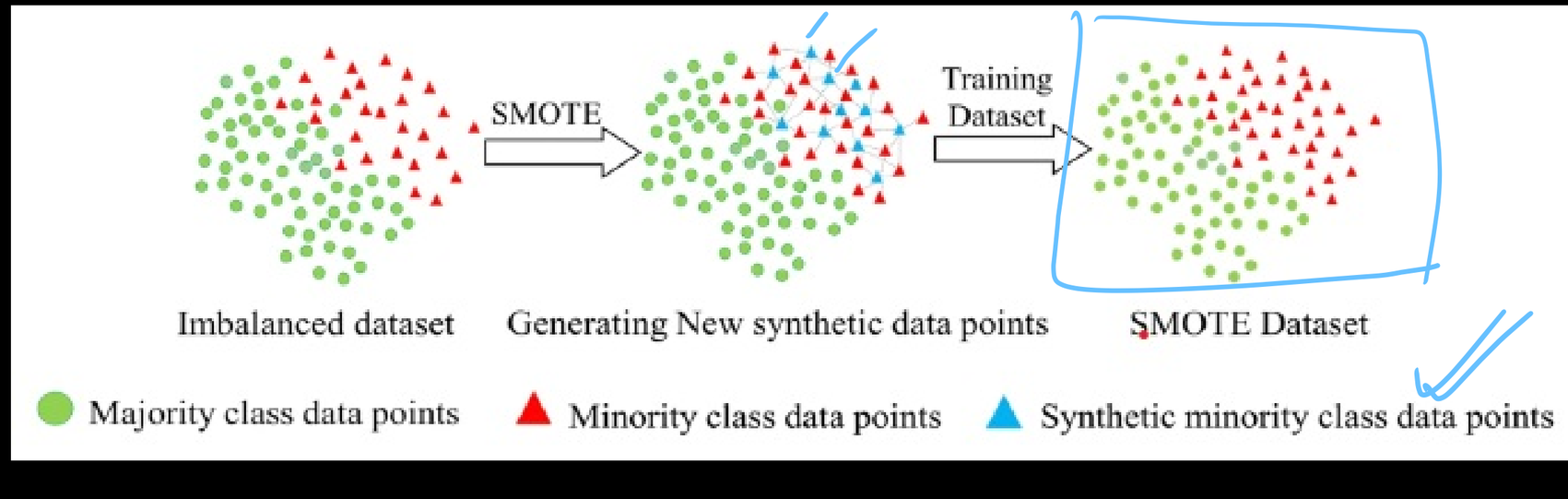


X	Y
	1
	1
	0
	0



Synthetical

0.1
0.2
0.8


SMOTE
Training Dataset
Imbalanced dataset
Generating New synthetic data points
SMOTE Dataset
Majority class data points
Minority class data points
Synthetic minority class data points