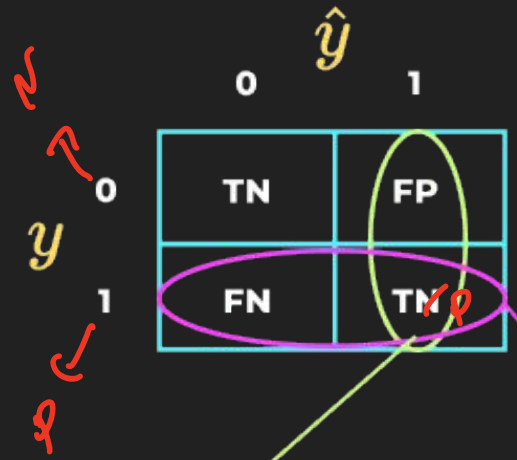
RECAP

Problem with Accuracy - Doesn't work for imbalance data

Confusion Matrix :



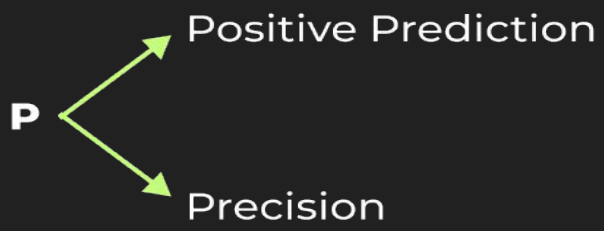
$$\text{Precision} = \frac{\text{Correct +ve Predictions}}{\text{All positive prediction}}$$

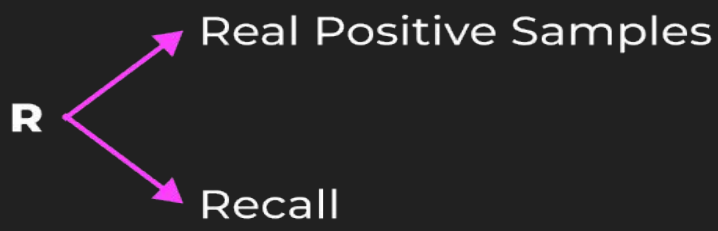
$$\frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{\text{Correct +ve Predictions}}{\text{Real prediction}}$$

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

Quick way for leaving Precision & Recall :





When FP & FN both important.

F1_Score : Harmonic Mean between precision & recall

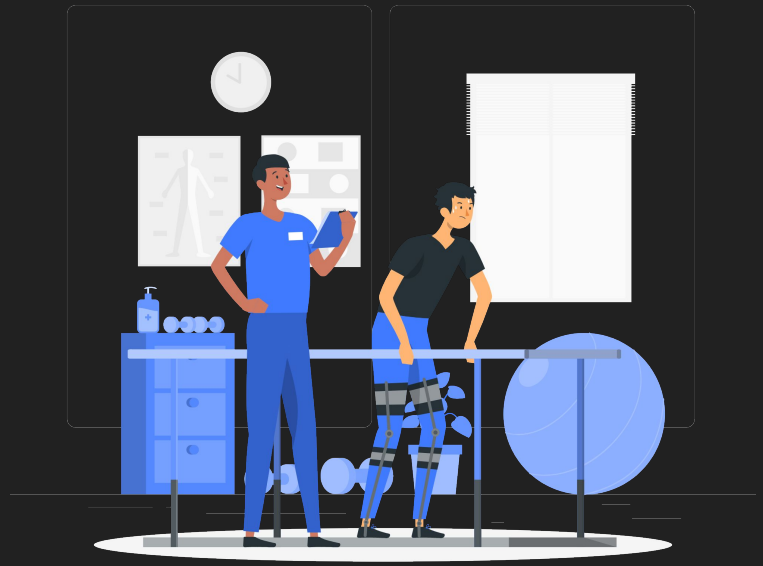
$$F1 = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2\,(\text{Precision})(\text{recall})}{\text{Precision} + \text{Recall}}$$

SENSITIVITY

Suppose you are a Data Scientist at Fortis :

Fortis Hospital data

- 100 Cancer Patient (Class 1)
- 900 non - cancer Patient (Class 0)



AIM ? - Correctly identify all cancer patients

- Correctly predict as many cancer patients (TP ↑)
- Misclassification of cancer patient should be low (FN ↓)

This need of TP ↑ & FN ↓ is called Sensitivity. ≈ Recall

Sensitivity Important ? - Yes, since here failing of model to identify cancer patient can cause death

Formula ? -

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

Same as Recall

Tracks TP, hence also called TPR

TPR = Sensitivity
FNR = 1 - Sensitivity

False Negative Rate (FNR) / Miss rate

What happens if model is insensitive ?

FN ↑ leading to worsening of health condition of cancer patient

This means FN should also be tracked

Formula ? - FNR = 1 - Sensitivity

$$= 1 - \frac{TP}{TP + FN}$$

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

As patients are missed by the MODEL,

- FNR is called Miss Rate.

How ? - Measure change in FN which is called as FNR.

SPECIFICITY

TP

TN

Do TN & FP not play any role in medical firms ?

TN & FP plays important role

- Model should correctly predict non - cancer patient (**TN** ↑)
- Model should keep misclassification of non - cancer patient low (FP ↓)

Need of TN ↑ & FP ↓ is called **Specificity**

Analogy : Specificity is sensitivity for negative class (CLASS 0)

Specificity Important ? - High specificity avoids

1. Fruitless experience treatments
2. Reduces social stigma anxiety for a non - cancer patient

Formula ? - Specificity = $\frac{TN}{TN + FP}$ → Tracks TN , Hence called TNR

→ TNR

False Positive Rate (FPR)

What happens if model is not specific ?

FP ↑ leading to unnecessary treatment , high dosage , anxiety, stress

This means FP should also be tracked

Formula ? - FPR = 1 - Specificity

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

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

How to track ? - Using FPR to measure change in FP

Points to Remember

Sensitivity : TP ↑ & FN ↓ = $\frac{TP}{TP + FN}$ ≃ .92%

FNR : tracks FN = 1 - Sensitivity

Specificity : TN ↑ & FP ↓ = $\frac{TN}{TN + FP}$

(Sensitivity for Negative class)

FPR : tracks FP = 1 - Specificity

70%

Accuracy : 80%

99%

>80%

RECEIVER OPERATING CURVE

How to improve performance of our model ?

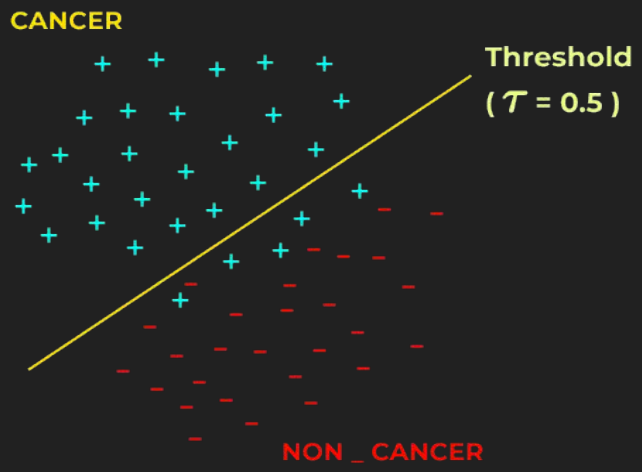
↓

HYPERPARAMETER TUNING

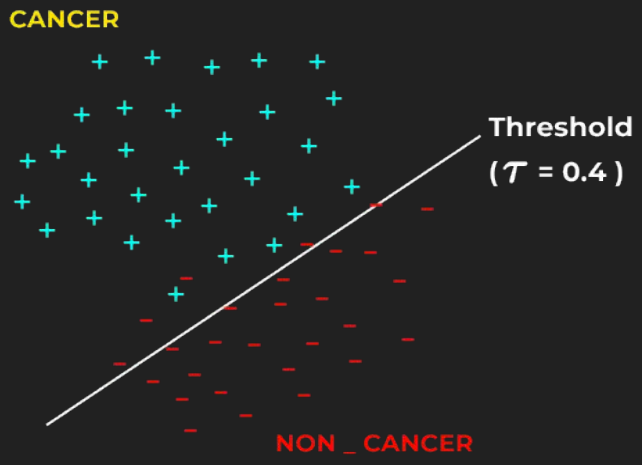
↓

Find the correct threshold (τ)

(the model misclassifies cancer patients
- need of a correct threshold)



Find the correct threshold (τ)



How to find the correct threshold τ ?

- Considering Validation data

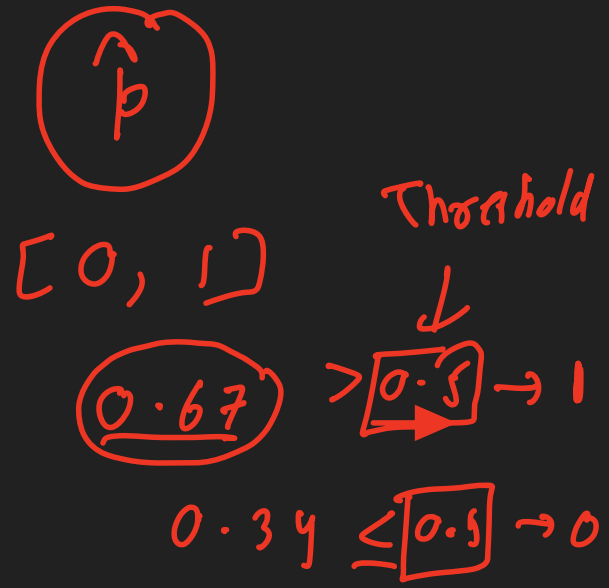
Step : 1

X	Y	$P = P(Y^{(i)} = 1 \mid X^{(i)})$
$x^{(1)}$	$y^{(1)}$	$p^{(1)}$
$x^{(2)}$	$y^{(2)}$	$p^{(2)}$
$x^{(n)}$	$y^{(n)}$	$p^{(n)}$

n

Sort samples in descending order based on value $P^{(i)}$

Complexity : O(nlogn)



Step : 2 *Use every $p^{(i)}$ as threshold $(\tau^{(i)})$ for predicting $\hat{y}^{(i)}$*

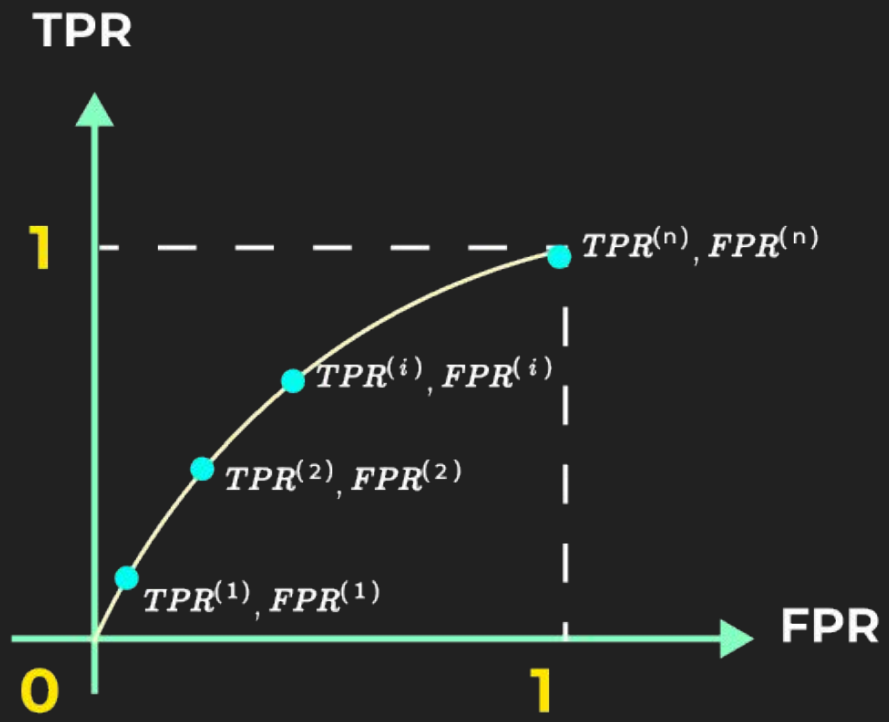
X	Y	P	$\hat{y}_\tau = p^{(1)}$	$\hat{y}_\tau = p^{(2)}$	$\hat{y}_\tau = p^{(n)}$
$x^{(1)}$	$y^{(1)}$	$p^{(1)}$	1	1	1
$x^{(2)}$	$y^{(2)}$	$p^{(2)}$	0	1	1
.	.	.	.	.	.
.	.	.	.	.	.
$x^{(n)}$	$y^{(n)}$	$p^{(n)}$	0	0	1

Based on $\hat{y}_\tau = p^{(i)}$, we find TPR & FPR

P	TPR	FPR
$p^{(1)}$	$TPR^{(1)}$	$FPR^{(1)}$
$p^{(2)}$	$TPR^{(2)}$	$FPR^{(2)}$
.	.	.
.	.	.
$p^{(n)}$	$TPR^{(n)}$	$FPR^{(n)}$

$$\text{TPR} = \frac{\text{TP}}{\text{TP + FN}} \quad , \quad \text{FPR} = \frac{\text{FP}}{\text{FP + TN}}$$

Step : 3 - Plot TPR vs FPR



In spam classifier, we want TP to increase and FP to decrease

Therefore , pick hat threshold where TPR increases & FPR decreases

Understanding ROC with example

Step : 1

X	Y	P
$x^{(1)}$	1	0.65
$x^{(2)}$	1	0.94
$x^{(3)}$	0	0.30
$x^{(4)}$	1	0.92
$x^{(5)}$	0	0.70
$x^{(6)}$	0	0.20

Sort according to P

X	Y	P
$x^{(2)}$	1	0.94
$x^{(4)}$	1	0.92
$x^{(5)}$	0	0.70
$x^{(1)}$	1	0.65
$x^{(3)}$	0	0.30
$x^{(6)}$	0	0.20

Step : 2 $Create\ \hat{y}_\tau = p^{(i)}$

X	Y	P	$\hat{y}_\tau = 0.94$	$\hat{y}_\tau = 0.92$	$\hat{y}_\tau = 0.20$
$x^{(2)}$	1	0.94	1	1	1
$x^{(4)}$	1	0.92	0	1	1
$x^{(5)}$	0	0.70	0	0	1
$x^{(1)}$	1	0.65	0	0	1
$x^{(3)}$	0	0.30	0	0	1
$x^{(6)}$	0	0.20	0	0	1

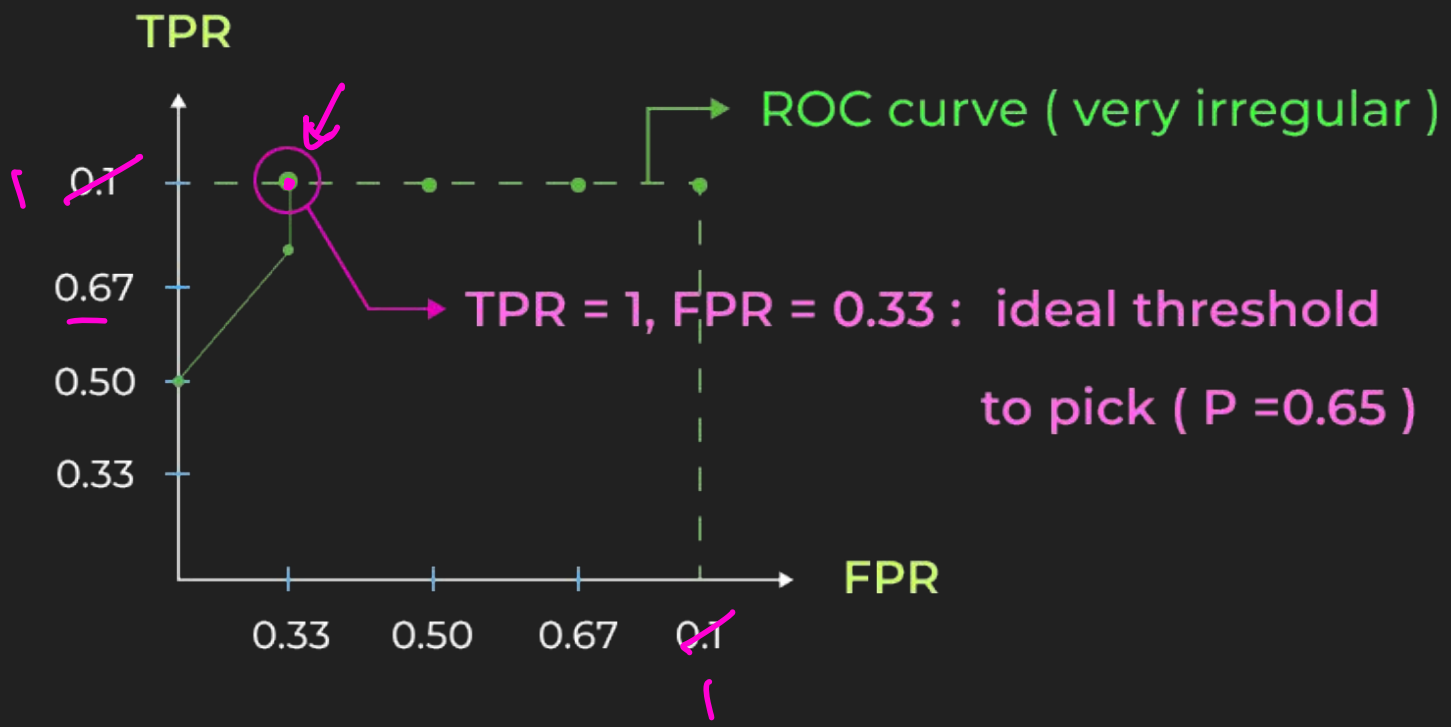
Find TPR , FPR - TP = 1 , FP = 1, FN = 2 , TN = 3

$$\therefore TPR = \frac{1}{1+2} = \frac{1}{3} = 0.33 \qquad FPR = \frac{0}{(0+3)} = 0$$

Repeating step - 2 for each P :

P	TPR	FPR
0.94	0.33	0
0.92	0.50	0
0.70	0.67	0.33
0.65	1	0.33
0.30	1	0.67
0.20	1	1

Pair of TPR & FPR for all 6 probabilities


TPR
0.1
0.67
0.50
0.33
FPR
0.33
0.50
0.67
0.1
ROC curve (very irregular)
TPR = 1, FPR = 0.33 : ideal threshold
to pick (P =0.65)

What will be the ROC curve for Random Model ?

As our example data is **BALANCED**

- **3 +ve data**
- **3 -ve data**

TP + FN = 3

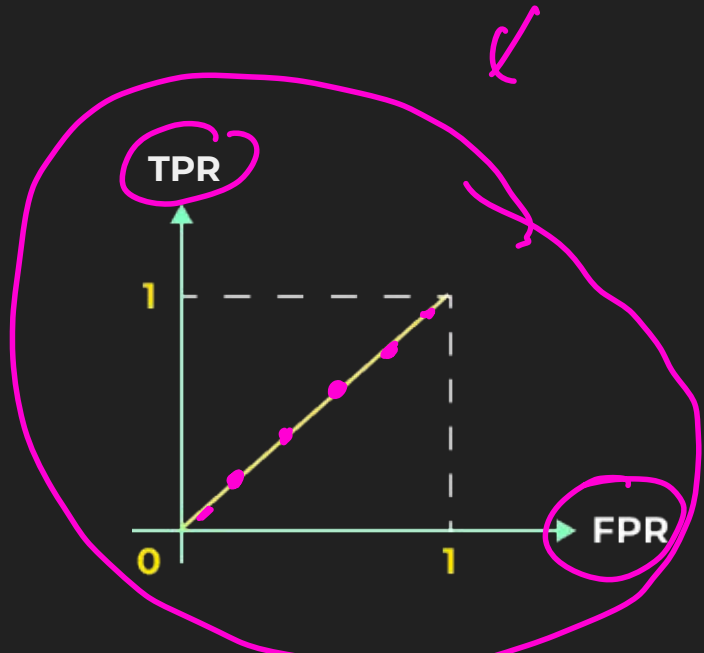
FP + TN = 3

Also as Random Model generates (1, 0) randomly

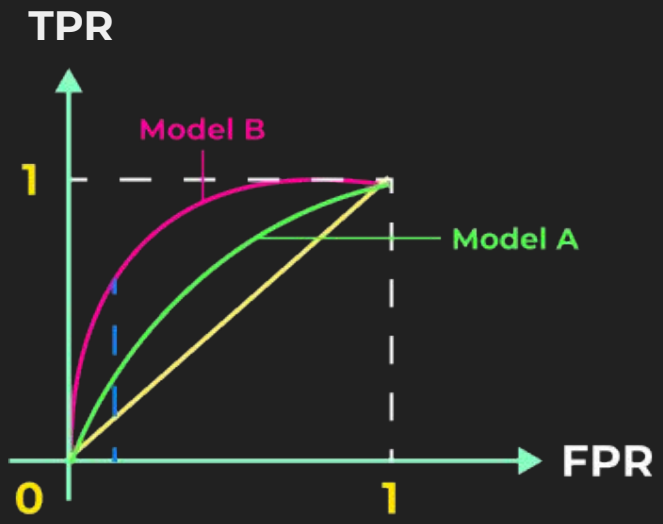
Produces K TP & K FP points

$$TPR = \frac{K}{3}, \quad FPR = \frac{K}{3}$$

Hence, TPR = FPR (y - axis) = (x - axis)



Suppose we have 2 models → **Model A**, **Model B** → Which is better ?

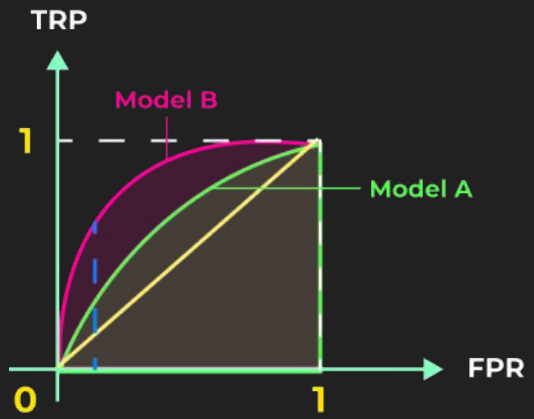


Intuitively,

TPR(B) > TPR(A) for same FPR.

Hence , MODEL B is better

How to mathematically show Model B better ?



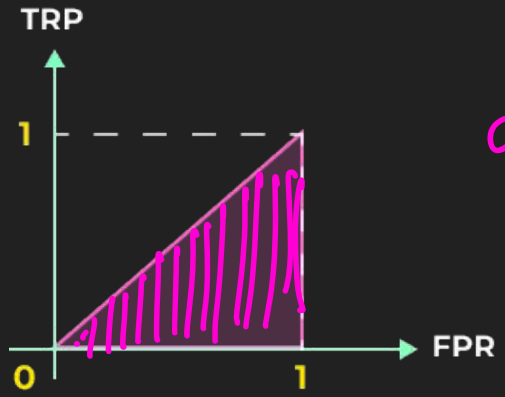
Let's take area under the ROC curve (All - ROC / Auc)

Clearly ,

Auc Model B > Auc Model A

Hence, Model B better

What will be Auc - Random Model ?



0.5 → Random model

$$Auc = \frac{1}{2} * 1 * 1$$

$$= 0.5$$

Area under curve (AUC)

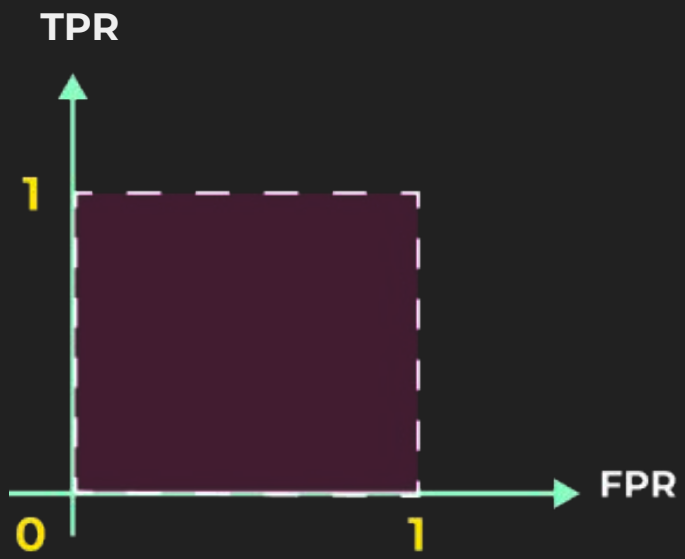
For ideal model

$threshold \approx 1\ (most\ \hat{y} = 0)$

$\therefore TPR \approx 1\ \&\ FPR \approx 0$

$threshold \approx 0\ (most\ \hat{y} = 1)$

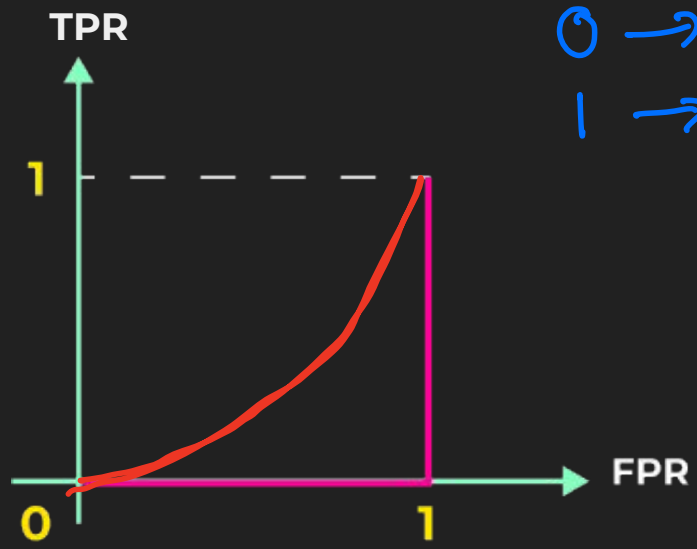
$\therefore TPR \approx 1\ \&\ FPR \approx 1$



Auc = 1 * 1 = 1

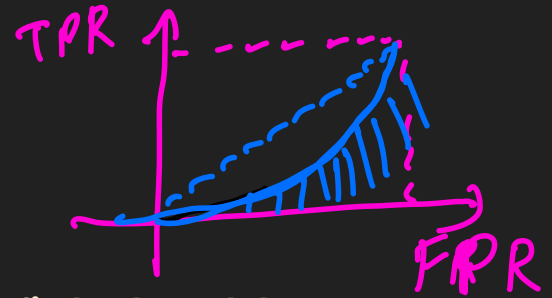
Test

What will be Auc - bad Model ?



0 → 1
1 → 0

< 0.4 → 0
≥ 0.4 → 1



Would (1 - p) fix bad model ?

$$AUC_{new} = 1 - AUC_{old}$$

$$= 1 - 0$$

$$AUC_{new} = 1$$

Means bad model probabilities misclassified every datapoint

Break: 8:11 AM

Hence, for any model which has ROC < 0.5

Can be fixed by revising probability

How Au - ROC different from Precision / Recall / F1 ?

Precision , Recall , F1

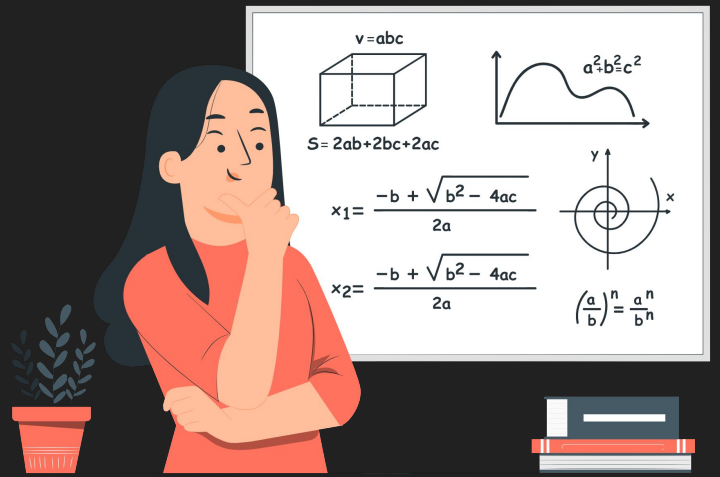
All find metric on a single threshold

(generally, $\tau = 0.5$)

Au - ROC

We find AUC metric on every

possible threshold



Issue with Au - ROC curve :

$Suppose\ y = [1, 1, 0, 1, 1]$ → **Imbalance Data**

$P_{m1} = [\ 0.95.\ \ 0.92.\ \ 0.80.\ \ 0.76.\ \ 0.71\ ]$

$P_{m2} = [\ 0.20.\ \ 0.10.\ \ 0.08.\ \ 0.06.\ \ 0.02\ ]$

then what will be $\hat{y}_{m1}$ & $\hat{y}_{m2}$?

M1

y	P_{m1}	$\hat{y}_\tau = 0.95$	$\hat{y}_\tau = 0.92$
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

M2

y	P_{m2}	$\hat{y}_\tau = 0.2$	$\hat{y}_\tau = 0.1$
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

M1 : 0.95 > 0.92 > 0.80 > 0.76 > 0.71

M2 : 0.2 > 0.1 > 0.08 > 0.06 > 0.02

$$[TPR_{M1}^{(1)}, FPR_{M2}^{(1)}] \ldots [TPR_{M1}^{(n)}, FPR_{M2}^{(n)}]$$

$$=$$

$$[TPR_{M2}^{(1)}, FPR_{M2}^{(1)}] \ldots [TPR_{M2}^{(n)}, FPR_{M2}^{(n)}]$$

Hance, Auc(M1) = Auc(M2)

Observe

- AU-ROC considers how data is ordered rather than the value itself

Hence AUC(M1) and AUC(M2) same

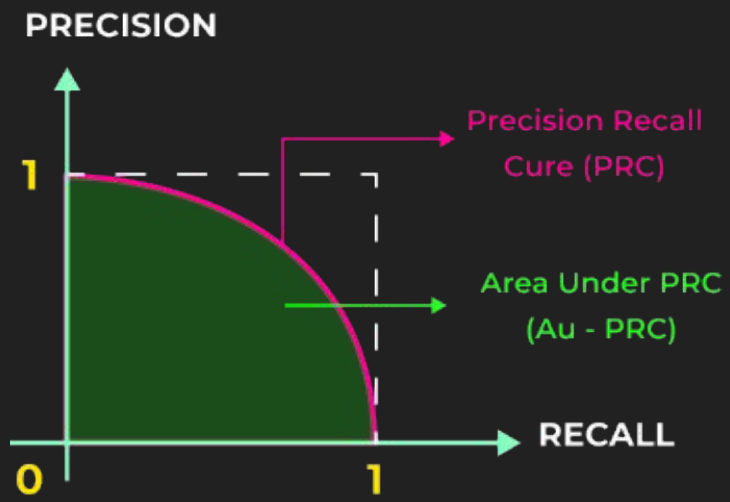
PR - ROC Curve

What can be used instead of Au - ROC when data imbalanced ?

A F1 score works well for imbalance data:

[Precision **Recall]**

Let's take **Precision and Recall** values for every probability instead of **TPR and FPR**



Points to Remember

Sensitivity : TP ↑ & FN ↓ = $\frac{TP}{TP + FN}$

FNR : tracks FN = 1 - Sensitivity

Specificity : TN ↑ & FP ↓ = $\frac{TN}{TN + FP}$

(Sensitivity for Negative class)

FPR : tracks FP = 1 - Specificity

- Au - ROC for random model = **0.5**
- Au - ROC for ideal model = **1**
- Au - ROC does **not work in imbalance setting**
- Au - ROC depends on **order of probabilities.**

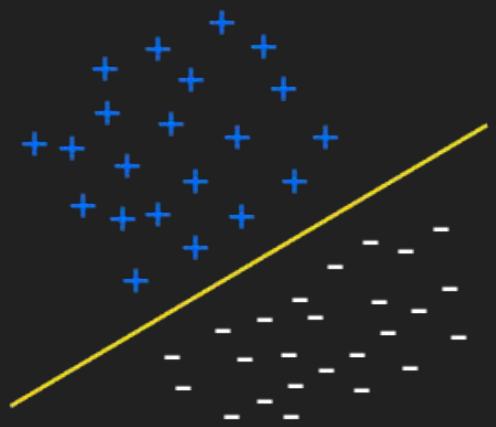
Understanding Imbalance Data

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

How does Logistic - Regression look when data is balanced ?



$$\pi : y = w^T x + w_0$$

Logistic Regression creates

a hyperplane

Which separates the data perfectly

by finding optional weights (w_0, w)

using Gradient Descent

What happens to logistic Regression when data is imbalance ?

$$logloss = -\frac{1}{n}[\sum_{i=1}^{n} y^{(i)} log(p^{(i)}) + (1 - y^{(i)}) log(1 - p^{(i)})]$$

$$\downarrow$$

$$p^{(i)} = P(y^{(i)} = 1 | x^{(i)})$$

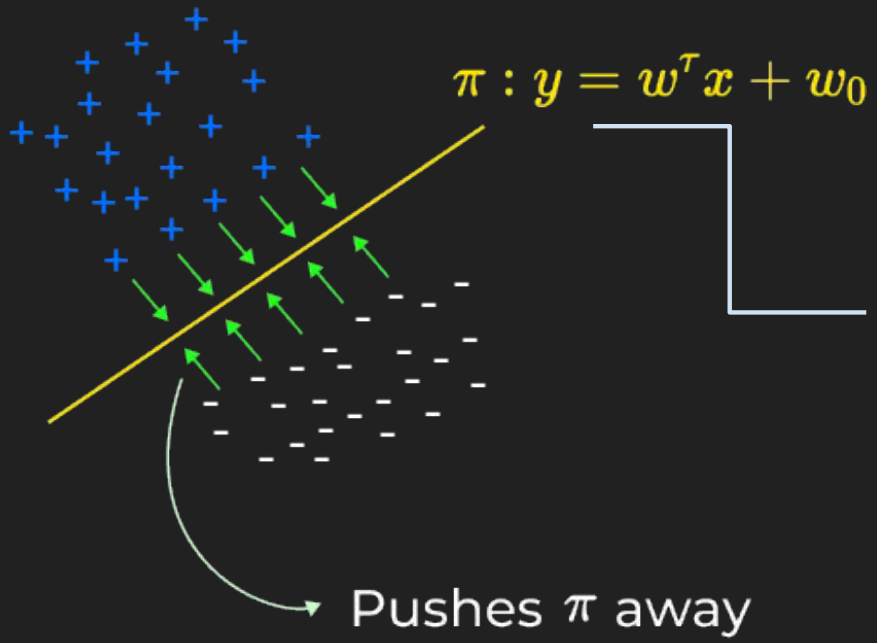
Suppose $p^{(i)} = 0.7$ Means model 70% confident that $x^{(i)}$ has label = 1

$$p^{(1)} = 0.5] \rightarrow \quad logloss = -[(1 - 0) log(1 - 0.5)] = 0.301$$
$$p^{(2)} = 0.1] \rightarrow \quad logloss = -[(1 - 0) log(1 - 0.1)] = 0.045$$

We can conclude, as logloss small for $p^{(2)}$, model is confident $p^{(2)} \neq Class1$

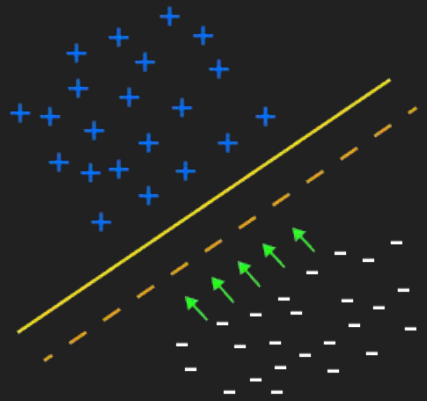
Hence , to conclude

Model is confident → point is further away from it



Lies between (+) &(-) samples
when data is balanced

Logistic Regression



Hyperplane shifts towards minority class

→ Uncertainty when predicting + ve class data

How to make imbalance data balanced ?

Suppose → **1000 data** samples

- **150 SPAM**
- **850 NON - SPAM**

Class Weight

∴ Non - Spam 5.67 times Spam

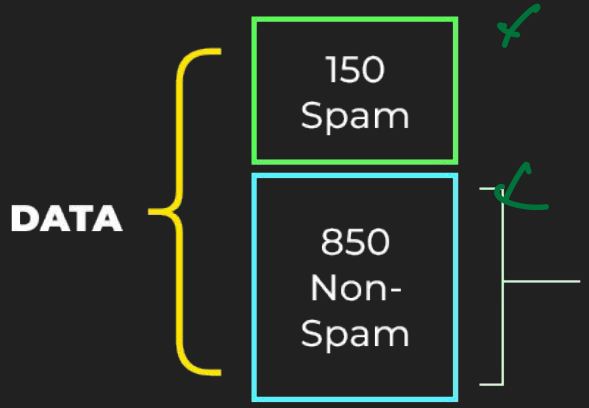
If 1 spam data has weightage of 5.67 non - spam

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

$W_i = 5.67$ when spam

$W_i = 1$ when non - spam

Undersampling



Randomly pick 150 spam data

150 spam

150 Non - spam

Issue ?

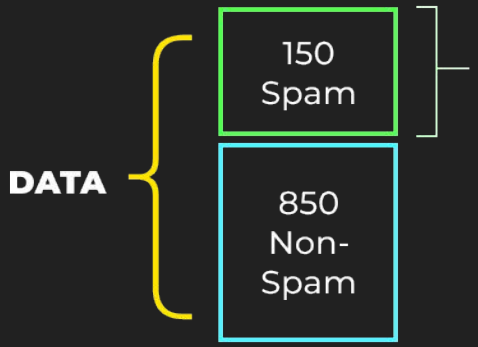
As model trained on smaller data

Model Reliability Decreases

When to use undersampling ?

No. of samples ≈ 1 Billion

Oversampling



Instead of removing , let's duplicate randomly selected spam data

Spam = 850 Non - Spam = 850

∴ Total Sample = 850 + 850 = 1750

Why model overfit when using Oversampling ?

→ Just repeating minority samples

↓

[Leads to model learning patterns]

↓

Causes overfitting

To avoid overfitting

↓

Creates synthetic new data samples of minority class

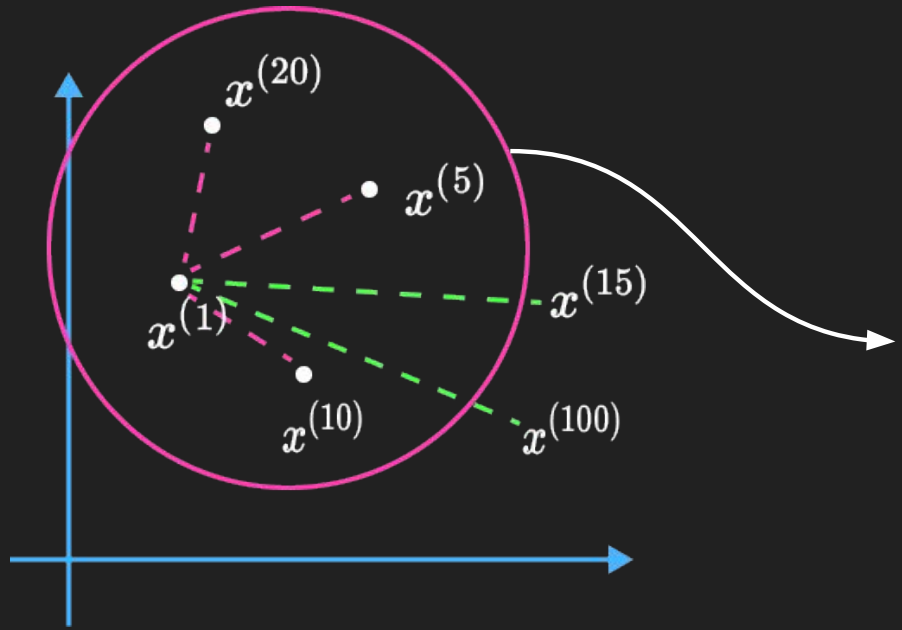
SMOTE

KNN 2

How to create synthetic data ?

$x^{(1)} \rightarrow random\ spam\ data$

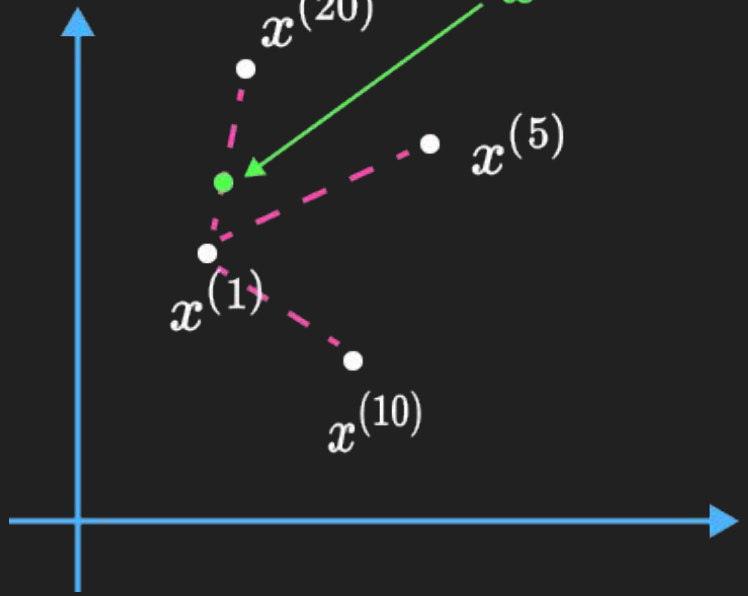
$x^{(1)} = [f_1, f_2, f_3, \ldots . f_d]$



Finding euclidean distance from every spam data & selecting 3 closet ones

$[x^{(5)}, x^{(10)}, x^{(20)}]$

Taking a random number $\phi \in [0, 1]$



$$x^{new^1} = x^{(1)} + \phi dist(x^{(1)}, x^{(20)})$$

Make the new data lie between

$x^{(1)}$ & $x^{(20)}$

Points to Remember

Sensitivity : TP ↑ & FN ↓ = $\frac{TP}{TP + FN}$

FNR : tracks FN = 1 - Sensitivity

Specificity : TN ↑ & FP ↓ = $\frac{TN}{TN + FP}$

(Sensitivity for Negative class)

FPR : tracks FP = 1 - Specificity

- Au - ROC for random model = **0.5**
- Au - ROC for ideal model = **1**
- Au - ROC does **not work in imbalance setting**
- Au - ROC depends on **order of probabilities.**

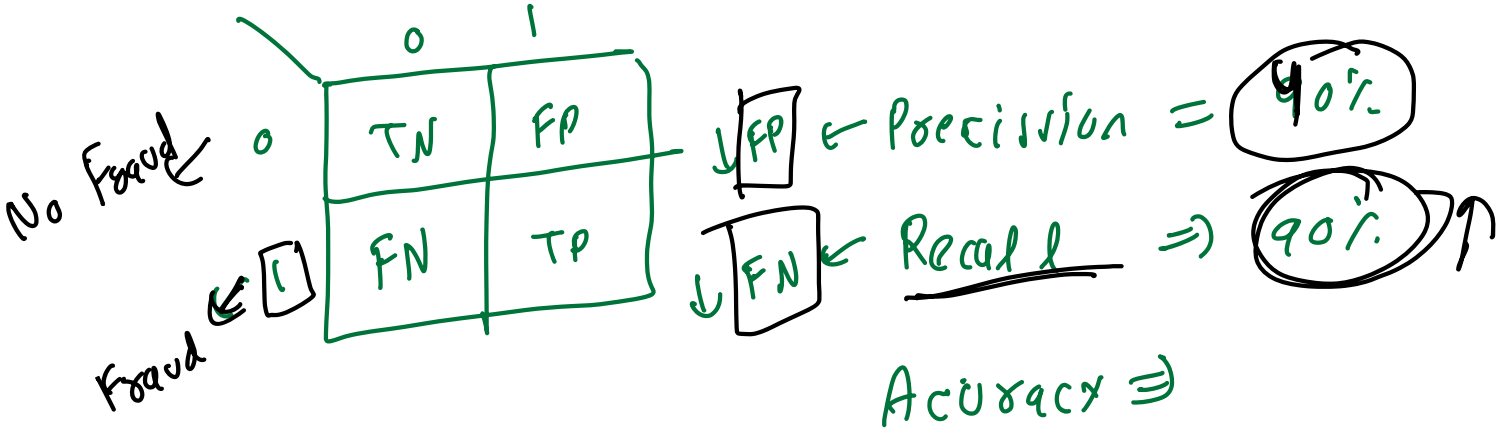
Points to Remember

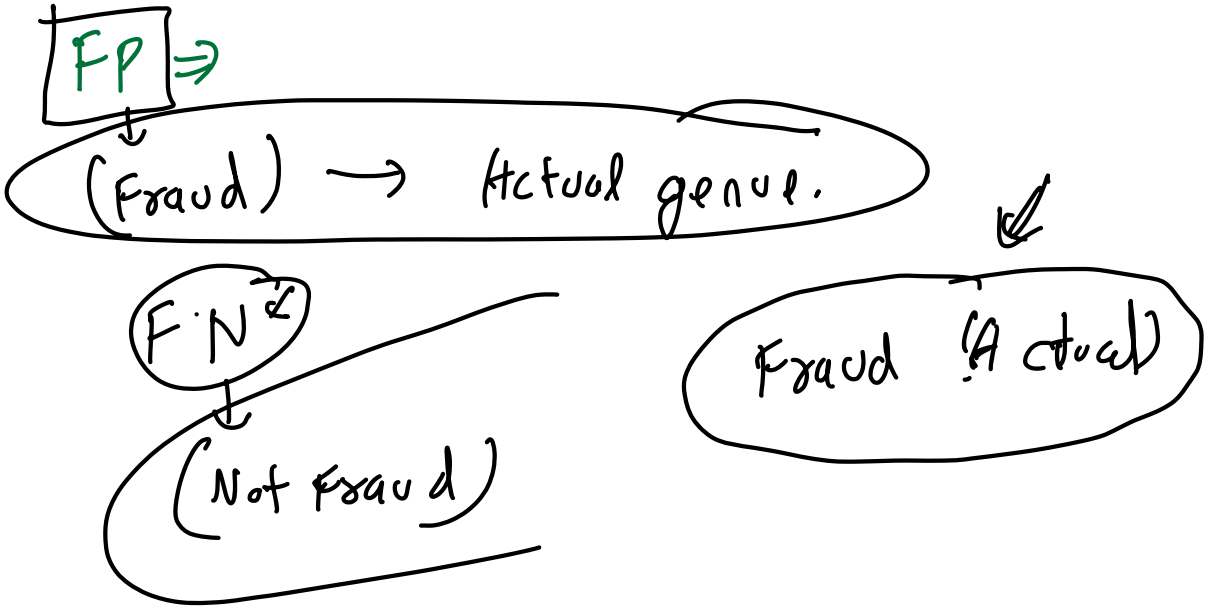
Hyperplane of Logistic Regression is pushed away by majority class

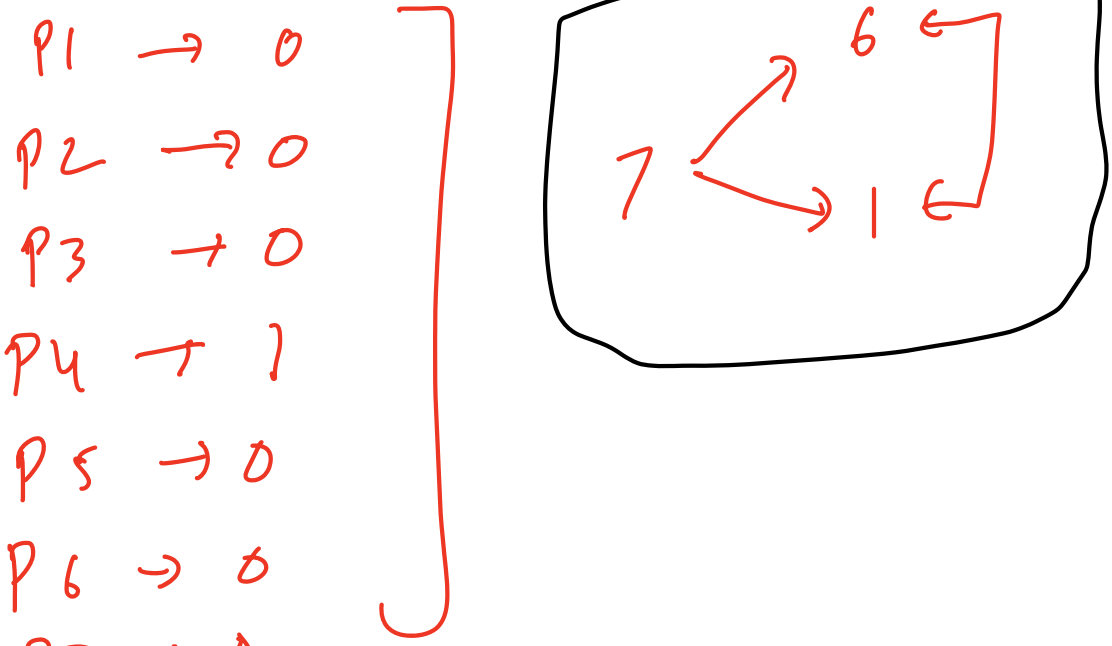
Ways to balance , imbalance data :

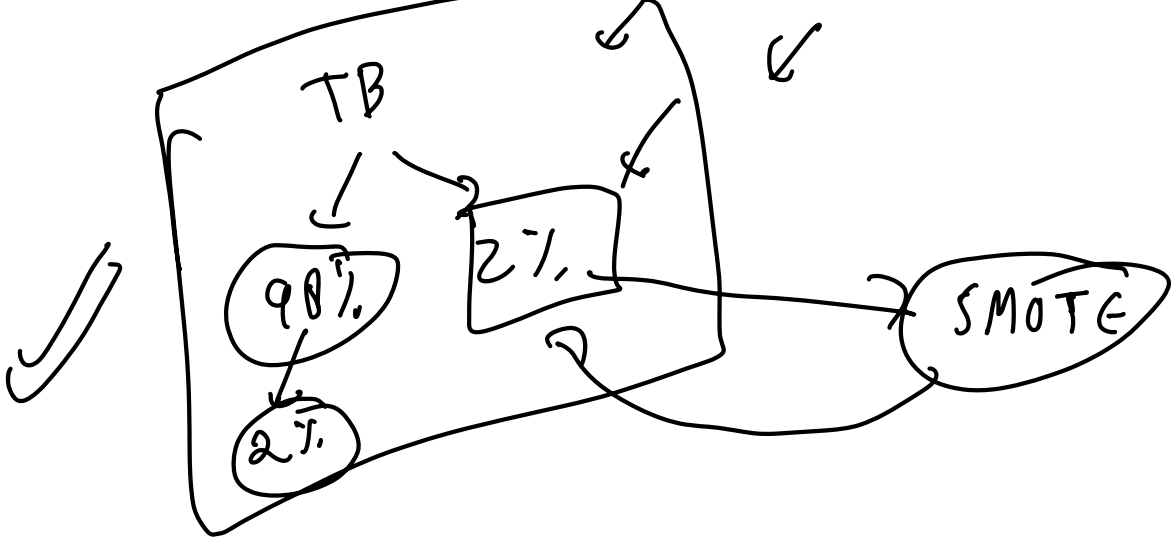
- Class Weight : Add weight to minority class samples
- Undersampling : Repeat minority class samples
- SMOTE : Generate synthetic minority class samples

CME


0
1
No Fraud
Fraud
TN
FP
FN
TP
FP
Precision = 90%
FN
Recall ⇒ 90%
Accuracy ⇒


FP ⇒
(Fraud) → Actual genuine
FN
(Not Fraud)
Fraud (Actual)


P1 → 0
P2 → 0
P3 → 0
P4 → 1
P5 → 0
P6 → 0
P7 → 0
6
7
1



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