→ Class start at 9:05 PM

Agenda

- → Accuracy
- → Confusion Matrix
- → Precision, Recall, F1-score

Linear Reg $\longrightarrow$ Adj R^2, R^2

Logistic Reg $\longrightarrow$ Accuracy

$$\text{Accuracy} = \frac{\text{Correct Pred}}{\text{Total Prediction}}$$

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

+ve class = 3

-ve class = 4

Balanced : $\frac{3}{4}$

$= 4/7$

y

+ve : 5

-ve : 95

← Imbalance (1:9)

$\hat{y}$

-ve : 100

→ Acc = $\frac{\text{No. of Correct Pred}}{\text{Total Pred}}$ = $\frac{95}{100}$

= 0.95

= 95%

-ve (0) : +ve (1)

80 : 20

90 : 10

95 : 5

Accuracy → Cannot be used for Imbalanced data

Precision, Recall, f1 score → For Imbalanced data

Confusion Matrix

↳ It tells us how well model prediction match actual class

(y) Actual \ Prediction ($\hat{y}$)	0	1
0	TN	FP
1	FN	TP

Connect → $y = \hat{y} = 0$ (TN)

→ $y = \hat{y} = 1$ (TP)

Inconnect $y \neq \hat{y}$ →

$y = 0,\ \hat{y} = 1$ (FP)

$y = 1,\ \hat{y} = 0$ (FN)

$= \quad \neq$

↓ T or F

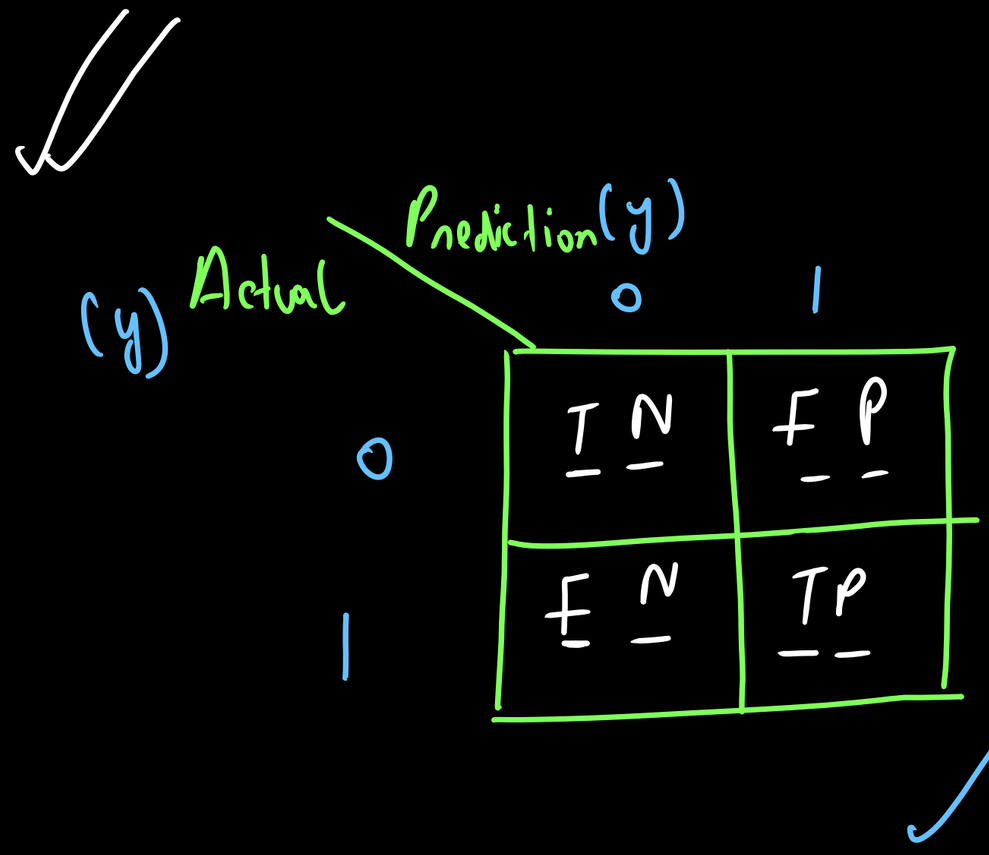
$y = \hat{y}$

$y \neq \hat{y}$

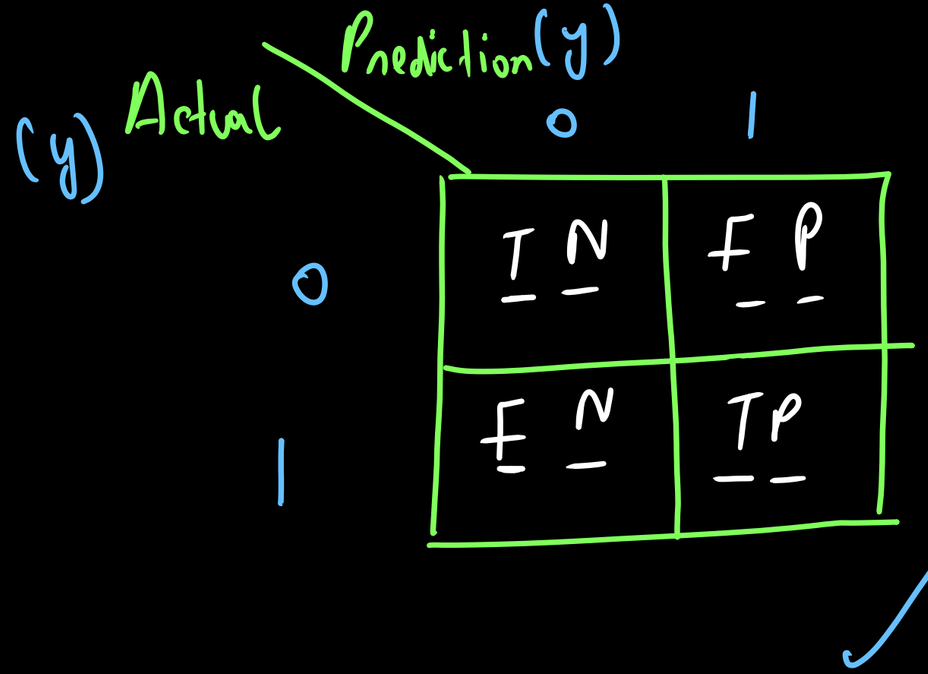
↓ P or N

$\hat{y} = 1$

$\hat{y} = 0$


(y) Actual
Prediction(y)
0
1
0
1
T N
F P
F N
T P

Predicted
Actual
1
0
1
0
T P
F N
F P
T N



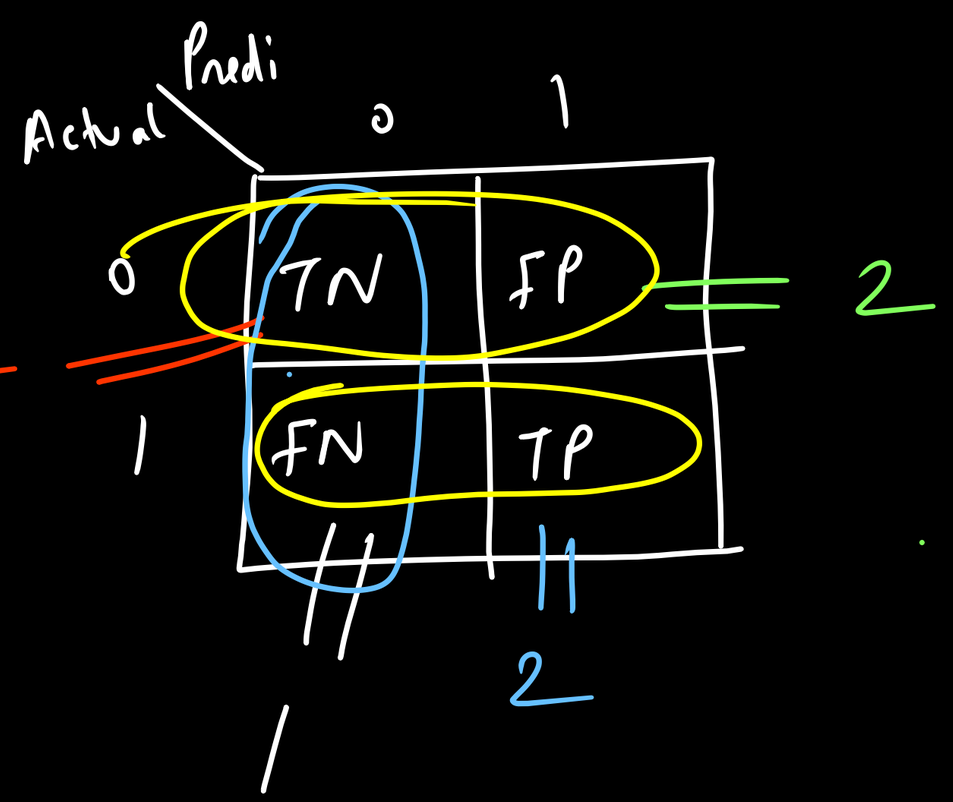
FP = Type 1

FN = Type 2

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

y	$\hat{y}$	$y = \hat{y}$
1	1	✓
0	0	✓
0	1	✓
1	0	✓
1	1	✓
0	1	✓
0	0	✓

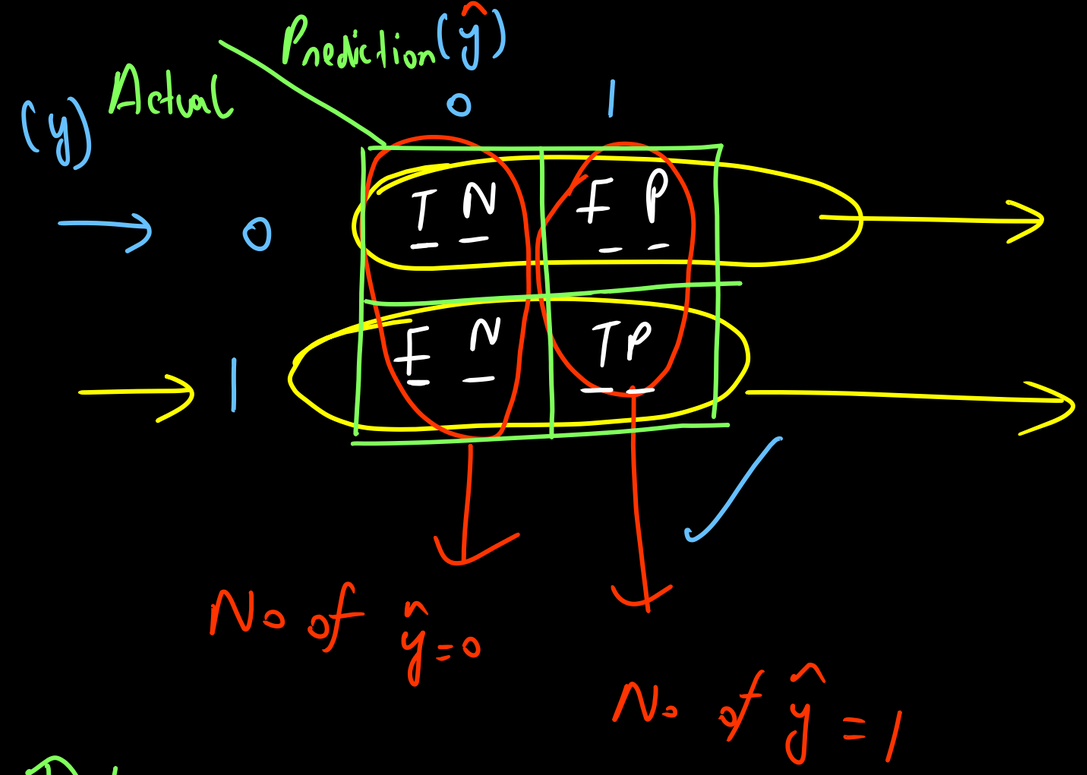
$$Acc = \frac{4}{7}$$



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

$$= \frac{2+2}{2+2+2+1}$$

$$= \frac{4}{7}$$



FP ↑↓

FN ↑↓

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

$$= \frac{100\%}{} \quad \frac{0}{} \quad \frac{0}{}$$

Business Case

→ Create a Email Spam detection system

↳ Spam = 1

Not Spam = 0

→ Precision → Spam Classification → Minority

Spam $y = 1$

Non Spam $y = 0$

⇒ Consider 2 scenarios

1. Receiving a spam email in inbox ($y = 1$, $\hat{y} = 0$) → FN
2. Missing out an offer letter email (by categorizing it as spam) ($y = 0$, $\hat{y} = 1$) → FP

Which amongst the two scenarios is more hazardous?

→ FP

Spam 1
Ns 0

No of $\hat{y}=1$ → TP + FP

TP
↑
$\hat{y}=y=1$

$\hat{y}=1$

↳ Out of all predicted spam, how many are Connectly predicted spam ✓

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{TP(y=\hat{y}=1)}{\text{Total No of Prediction}=1}$$

Act \ Pred	0 (Ns)	1 (S)
0 (Ns)	TN	FP
1 (S)	FN	TP

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

FP = 0

↳ Precision = 1 ✓

→ Act Data
↓ ↘
360 Ns 40 S

Dumb Model
↓ ↘
400 Ns 0 S

	0	1
0	360	0
1	40	0

↳ $\frac{0}{0+0+10^{-6}} = 0$ ✓

→ Precision → [0, 1]

Break until 22:30

→ Recall

Case : Screening Test to identify Cancer/Non - Cancer patients

Model : Classify Cancer and Non - Cancer patients

- Class 1 - Cancer → $y = 1$
- Class 0 - Non Cancer → $y = 0$

2 scenarios :

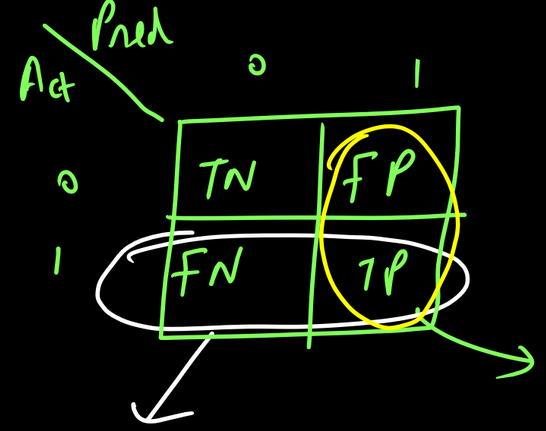
1. A healthy patient is considered as cancerous ($y=0$, $\hat{y}=1$) (FP)
2. A cancer patient is considered healthy ($y=1$, $\hat{y}=0$) FN → Recall

Which among the two is more dangerous?

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

① $FN \to 0$

$\text{Recall} = 1$



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

② $\hat{y}$

$C = 1$
$NC = 0$

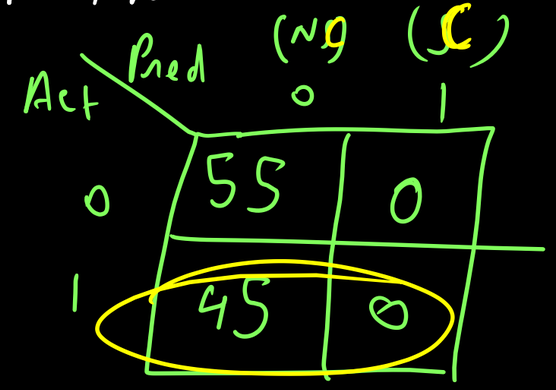
55 NC 45 C

$\hat{y}$

100 NC 0 C

$100 \cdot \infty \to \infty$

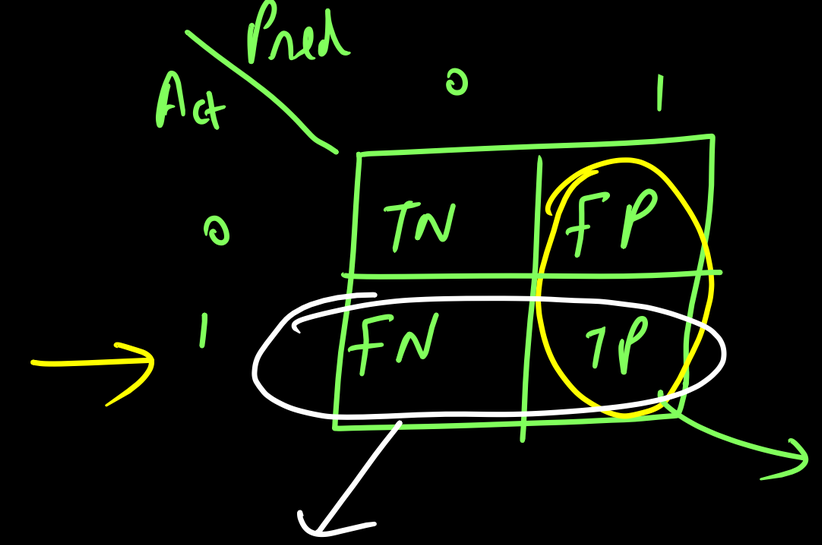
$\text{Recall} = \frac{TP}{TP + FN}$



$\frac{0}{0 + 45} = 0$

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

$$= \frac{TP}{\text{No of Act} = 1}$$



= Out of all +ve class data, (y=1 cancerous) how many are correctly predicted (predicted as cancerous) → $y = \hat{y} = 1$

Recall $\rightarrow [0, 1]$

Precision ↑ $\quad FP \downarrow \quad \rightarrow \quad S/NS$

Recall ↑ $\quad FN \downarrow \quad \rightarrow \quad C/NC$

→ Fraud Detection

Fraud = 1
NF = 0

2 scenarios

1. Predicting a transaction as legit when it is actually fraud ⇒ FN
   (can lead to financial loss)

$\hat{y} = 0$

$y = 1$

2. Predicting transactions as fraud when it is legit ⇒ FP
   (can lead to inconvenience to cardholder)

$\hat{y} = 1$

$y = 0$

Customer

FP ↓ FN ↓

P ↑ R ↑

	Prec	Recall	Avg = $\frac{P+R}{2}$	F1
M1	0.5	0.4	0.45	0.444
→ M2	0	1	0.5	0
→ M3	1	0.01	0.505	0.009
M4	0.7	0.1	0.4	0.175

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

→ Penalize more when either P or R is low

→ Ensure that both P & R are reasonably high.

→ Range of F1-score

1) P = 0, R = 1, F1_score = 0

2) P = 1, R = 0, F1_score = 0

3) P = 1, R = 1, F1_score = 1

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

✓ [0, 1]

-, -, -

P, R, F1 → Depends on business problem

SMS → ↓ FP

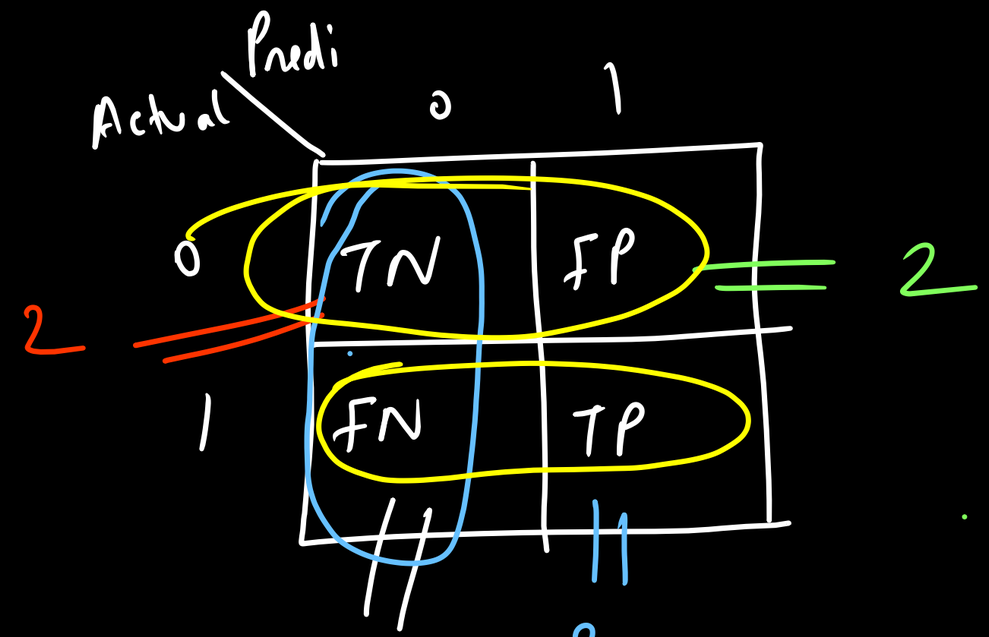
C/NC → ↓ FN

Fraud → ↓ FP & FN

Detection

P & R

y	$\hat{y}$
1	1
0	0
0	1
1	0
1	1
0	1
0	0



$$\frac{2}{4} = \text{Prec} = \frac{TP}{TP + FP} = \frac{2}{2+2} = 0.5$$

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{2}{2+1} = \frac{2}{3} = 0.66$$

$$\frac{2}{3}$$

Actual

990 Ns 10s

	0	1
0	990	0
1	10	0

$= \frac{TN}{TN + FP} = \frac{990}{990 + 10}$

1000 Ns 0s

$\frac{0}{0 + 0 + 10^{-6}} \approx 0$

$TN = 990$