Prediction

Actual results	0 (non-Span)	1 (Span)
0 (non-Span)	True -ve	False +ve ← Type 1 error
1 (Span)	False -ve	True +ve

Type 2 error

∴ Confusion matrix summarizes the results of model.

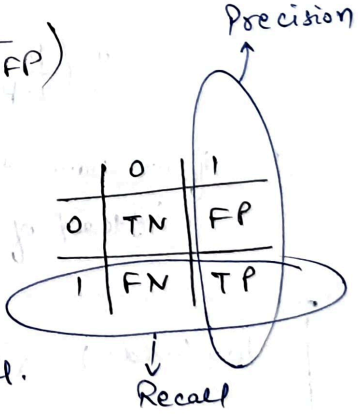
dumb model → FP, TP = 0

Excellent model → ~~0000~~ FP & FN ↓, while TP & TN ↑ Ideal

Accuracy = $\frac{\text{correct prediction}}{\text{total sample}}$ $\left(\frac{TP + TN}{TP + TN + FN + FP}\right)$

Recall = $\frac{\text{Predicted True}}{\text{Actual true}}$ $\left(\frac{TP}{FN + TP}\right)$

Precision = $\frac{\text{Actual true}}{\text{Predicted to be true}}$ $\left(\frac{TP}{TP + FP}\right)$



[x = np.linspace (-10, 10, 100)
Return evenly spaced no' over a specified interval.

F1-score ⇒ Harmonic mean of Precision and Recall

$\frac{2 * P * R}{P + R}$

$$z = wx + b$$

$$p = \sigma(z) = \frac{1}{1+e^{-z}} = \frac{1}{1+\frac{1}{e^z}} = \frac{e^z}{e^z+1}$$

$$1-p = 1-\frac{e^z}{1+e^z} = \frac{1+e^z-e^z}{1+e^z} = \frac{1}{1+e^z}$$

$$\text{odds}\left(\frac{p}{1-p}\right) = \frac{\frac{e^z}{e^z+1}}{\frac{1}{1+e^z}} = e^z$$

$$\log \text{odds} = \log e^z = z$$

$$z = 2, \quad \sigma(z) = \frac{1}{1+e^{-2}} = \frac{e^2}{1+e^2} = 0.88$$

$$\frac{p}{1-p} = \frac{0.88}{1-0.88} = 7.33$$

$$\log\left(\frac{p}{1-p}\right) = \log(7.33) = 1.99 \simeq 2$$

∴ if we have taken the actual value of e instead of 2.71 then it would be 2.

$\log(\text{odds}) \propto$ sigmoid

if $\log(\text{odds})$ is -ve,

then $\frac{p}{1-p} < 1$

$\therefore p < 1-p$

if data1: 20 Cancer Patients and 100 non-Cancer Patients and data2: 80 Cancer Patients 100 non-Cancer Patients, then:

A) Data1 = Imbalance, Data2 = balance
B) Data1 = balance, Data2 = Imbalance
C) Data1 = balance, Data2 = balance
Data1 = Imbalance, Data2 = Imbalance

Ans: option(D)

While Data 2 is relatively balanced, the term "imbalance" is often used when there's a noticeable difference in class distribution, even if it's not as severe as in Data 1.

Why is accuracy a bad metric when dealing with imbalanced data?

a) It focuses only on the minority class.

b) It is sensitive to the distribution of positive instances only.

c) It overestimates the model's performance when the majority class dominates.

d) Accuracy is insensitive to imbalance data

> Ans: option(c)

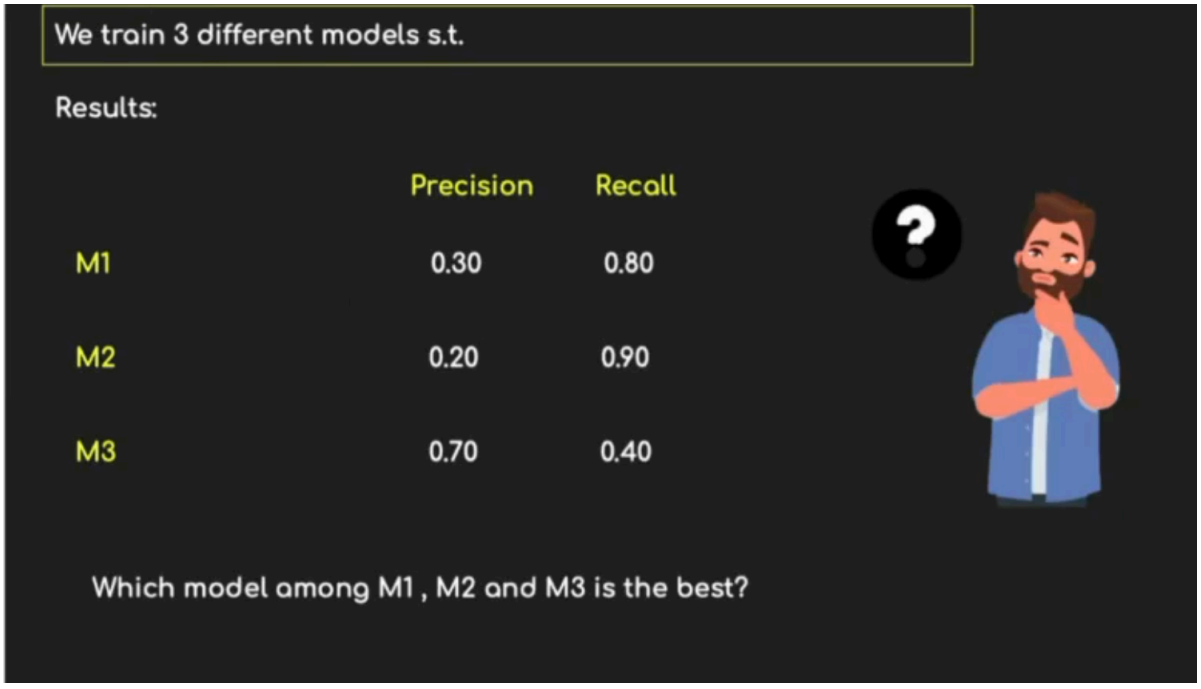
Understanding the Metrics

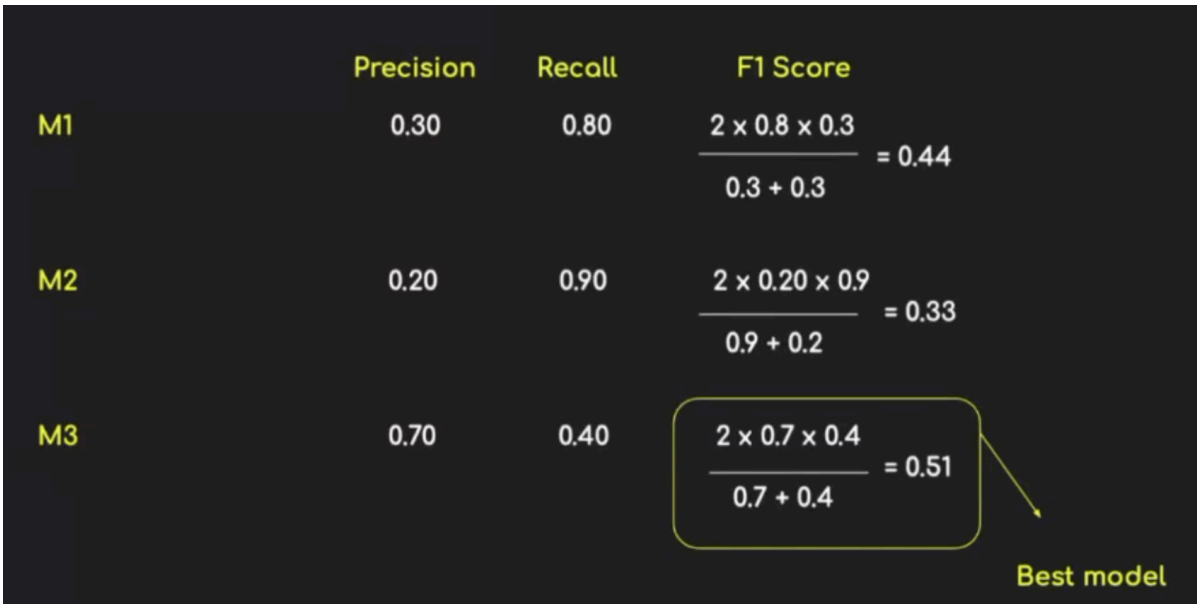
Accuracy: This measures the proportion of correct predictions (both fraudulent and non-fraudulent) out of all transactions. While high accuracy is important, in fraud detection, it can be misleading if the dataset is imbalanced (i.e., there are many more legitimate transactions than fraudulent ones).

Precision: Precision measures how many of the transactions predicted as fraudulent are actually fraudulent. High precision means fewer false positives, which is critical for reducing the cost of unnecessarily blocking or investigating legitimate transactions.

Recall: Recall measures how many actual fraudulent transactions the model successfully identifies. In this case, our model has a recall of 30%, meaning it catches 30% of all fraudulent transactions.

F1 Score: The F1 score is the harmonic mean of precision and recall, balancing the trade-off between them. In this case, the F1 score is low, reflecting the relatively low recall despite high precision





Why does the F-1 score use Harmonic Mean (HM) instead of Arithmetic Mean (AM) ?

a. AM penalizes models the most when even Precision and Recall are low.
b. HM penalizes models the most when even Precision and Recall are low.
c. HM penalizes models the most when even Precision and Recall are high.
d. AM penalizes models the most when even Precision and Recall are high.

> Ans: option(b)

FlashCards

> Pandas

A library in Python used for data manipulation and analysis

> Logistic Regression

A model for binary classification problems

> Confusion matrix

A confusion matrix is used to evaluate the performance of a classification algorithm. It is a table with four different combinations of predicted and actual values:

- **True Positive (TP):** Correctly predicted positive class.
- **True Negative (TN):** Correctly predicted negative class.
- **False Positive (FP):** Incorrectly predicted positive class (Type I error).
- **False Negative (FN):** Incorrectly predicted negative class (Type II error).

> **A table used to evaluate the performance of a classification model**

> odds :
>
> **Ratio of the probability of success to the probability of failure**

> sigmoid :
>
> **A mathematical function that maps any value to a value between 0 and 1【4:5†transcript.txt】**

> Imbalance data :
>
> **When the classes in a dataset are not equally represented**

> Model accuracy
>
> **The ratio of correct predictions to the total predictions made by the model【4:8†transcript.txt】**

The importance of selecting the right evaluation metric cannot be understated, especially with imbalanced datasets where the majority class can dominate accuracy results. Practitioners should focus on precision, recall, and the F1 score to better capture the model's effectiveness in identifying minority classes