Business Case

12 August 2025 17:17

Emails

	text	label_num
0	Subject: enron methanol ; meter # : 988291\r\n...	0
1	Subject: hpl nom for january 9 , 2001\r\n(see...	0
2	Subject: neon retreat\r\nho ho ho , we ' re ar...	0
3	Subject: photoshop , windows , office . cheap ...	1
4	Subject: re : indian springs\r\nthis deal is t...	0

label_num → Spam / Not Spam

1 → Spam

0 → Non-spam

Higher occurrence of non-spam ⇓ 71%

- **Not spam** $\approx \frac{850}{1200} \times 100 = 70.83\ \%$
- **spam** $\approx \frac{350}{1200} \times 100 = 29.17\ \%$

{ imbalanced data } → When the target has one class significantly higher in number

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

4 users have participated

$$\text{data } 1 \Rightarrow \left.\begin{matrix} 1 \Rightarrow 20 \\ 0 \Rightarrow 100 \end{matrix}\right\} \text{ imbalanced}$$

$$\text{data } 2 \Rightarrow \left.\begin{matrix} 1 \Rightarrow 80 \\ 2 \Rightarrow 100 \end{matrix}\right\} \text{ balanced}$$

$$\left\{\begin{matrix} \dfrac{100}{180} & & \dfrac{80}{180} \\ \Downarrow & & \Downarrow \\ 56\% & & 44\% \end{matrix}\right\}$$

$$\left\{\begin{matrix} 75 & & 25 \\ 80 & & 20 \end{matrix}\right\} \Rightarrow \text{imbalanced}$$

Problems with Accuracy

12 August 2025 17:20

Let us say,

In my test data, I have 900 not spam and 100 spam.

Not Spam = 0 => 90 %
{Spam = 1 => 10 %}

Highly imbalanced

Let us say my model has accuracy of 91 %. Seems good ?

Error Analysis

Let me dig further in :-

{91 % accuracy means 910 out of 1000} samples classified correctly.

I want to look into the 90 wrongly classified points.

{All 90 mis-classifications are spam being mis-classified as not spam!}

{This means, out of 100 spams, 90 are mis-classified as non-spam}

$\{ y_{actual} = 1 \quad y_{pred} = 0 \}$

Model performs terribly on spam data, yet gave accuracy of 91%

Mis-leading!

When working with imbalanced sets, never solely depend on accuracy!

1000

Non Spam 900 (0)

spams 100 (1) => In my dataset

91%

910 => spam + non spam

spam ✓ as nonspam

non-spam ✓ as spam

90 (incorrect)

0 non-spam as spam

90 spam predicted as non spam

90 out of 100 spams are not

90 out of 100 spams are not identified by the model!!

{Spam accuracy = 10%}

Non Spam accuracy = 100%

The whole point was to identify spams!!

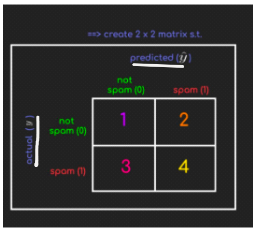
In the evaluation of a classification model, what does the accuracy metric represent?

19 users have participated

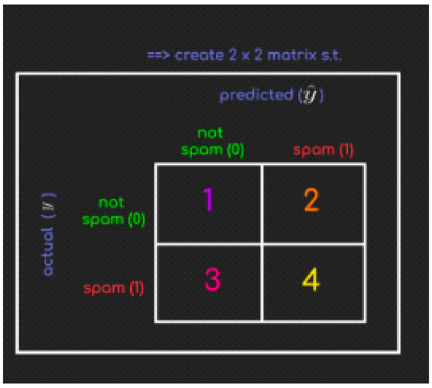
A	The model's ability to handle imbalanced datasets	32%
B	The precision of the model in predicting positive instances	0%
C	The ratio of true positive predictions to the total predictions	26%
D	The overall correctness of the model's predictions across all classes	42%

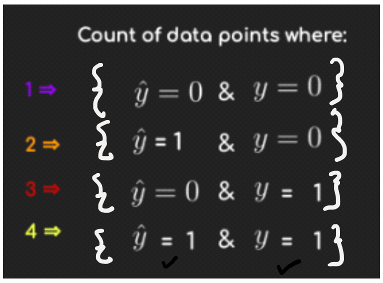
Confusion Matrix

12 August 2025 17:29



To better visualize our results, instead of solely depending on accuracy, we use confusion Matrix.





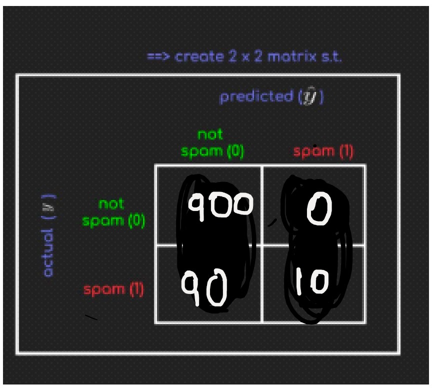
	True / False	Result of pred
$\hat{y} = y = 0 \rightarrow$	True	Negatives
$\hat{y} \neq y \longrightarrow$	False	positives
$\hat{y} \neq y \rightarrow$	False	negatives
$\hat{y} = y = 1 \rightarrow$	True	Positives

If 90 out of 100 Actual spams are misclassified as not spam

900 out of 900 Non spams are classified as Not spam

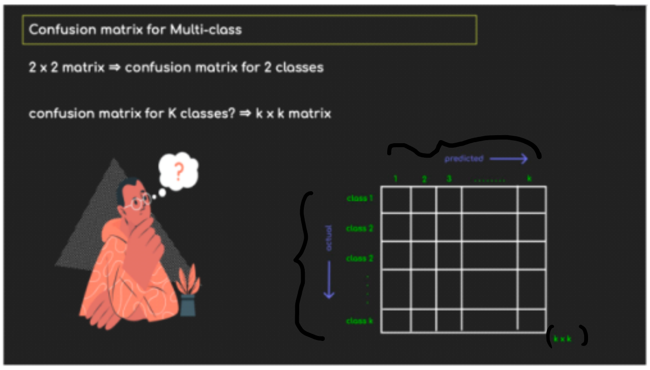
10 out of 100 spams classified as spam

What will Confusion Matrix look like?



$$\left\{ \begin{array}{l} \text{False positives} \Rightarrow 0 \\ \text{False negatives} \Rightarrow 90 \end{array} \right\}$$

FPR => 0



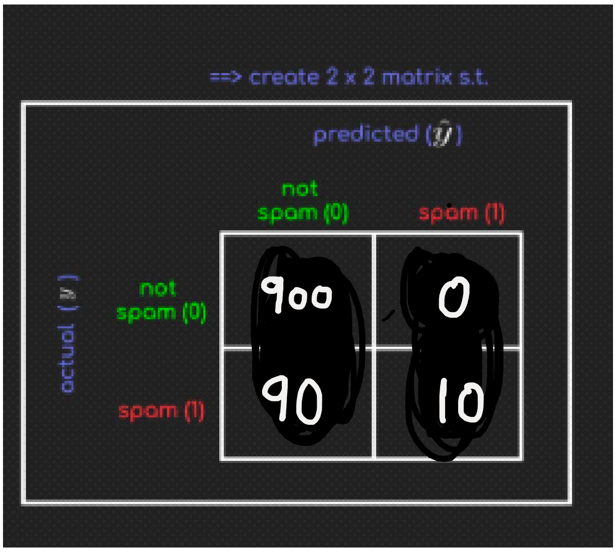
$$\left\{ FNR = \frac{90}{100} = 90\% \right\}$$

$$2 \times 2 = 4$$

$$n \times n$$

Understanding Confusion Matrix

12 August 2025 17:41



True positive rate

$$= \frac{TP}{\text{Total No. of pos in actuals}} = \left\{ \frac{TP}{TP + FN} \right\}$$

True Negative rate

$$= \frac{TN}{\text{Total No. of neg in actuals}} = \frac{TN}{TN + FP}$$

False positive rate } Prediction = + , Actual = −

$$= \frac{FP}{\{\text{Total No. of Actual Negatives}\}} = \frac{FP}{\{FP + TN\}}$$

} {Type 1 Error}

intuition → how many negatives were wrongly classified as positives?

False Negative rate

$$= \frac{FN}{\text{Total No. of Actual positives}} = \frac{FN}{\{FN + TP\}}$$

} {Type 2 Error}

intuition → how many

..., how many
positives were wrongly
classified as negatives?

For Ideal Model, which of the following is true?

26 users have participated

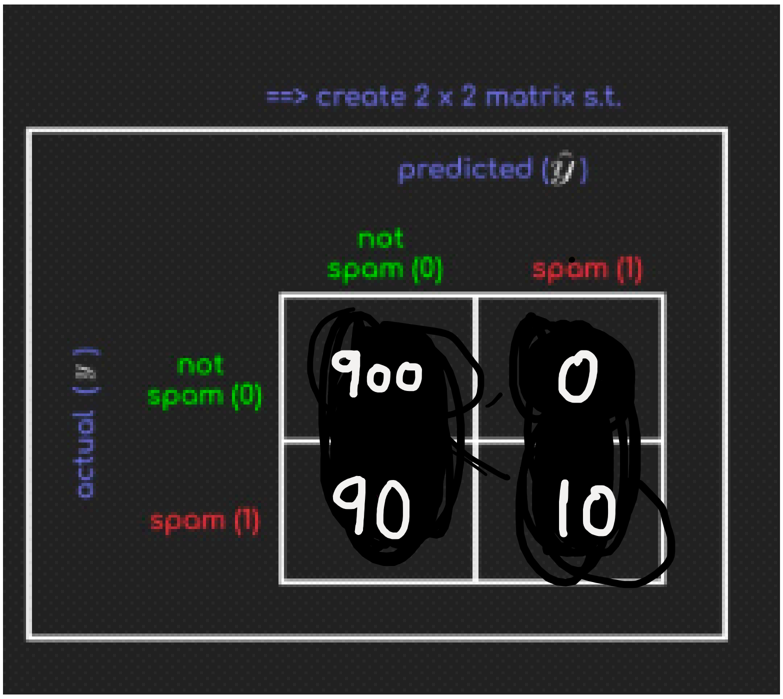
- A — FP and FN ⇓, while TP and TN ⇑ — 77%
- B — TP and TN ⇓, while FP and FN ⇑ — 4%
- C — TP and FN ⇓, while FP and TN ⇑ — 12%
- D — FP and TN ⇓, while TP and FN ⇑ — 8%

High
TP and TN ✓

Low
FP FN ✗

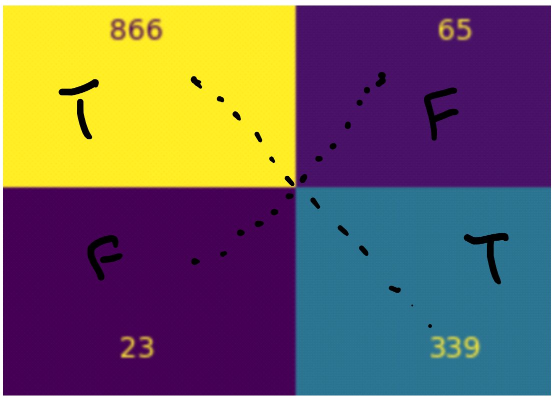
Accuracy through Confusion Matrix

12 August 2025 17:53



$$\frac{900 + 10}{900 + 90 + 10} = \frac{910}{1000} = 91\%$$

$$= \frac{TP + TN}{\text{All cells in Matrix}}$$



$F = 65 + 23$

$= 88$

Precision

12 August 2025 17:54

⇒ Consider 2 scenarios

1. Receiving a spam email in inbox 1
2. Missing out an offer letter email (by categorizing it as spam) 2

Scenario 2

Which amongst the two scenarios is more hazardous?

Which amongst the two scenarios is more hazardous?

⇒ 2nd case (having offer letter in spam)

FP or FN : Having an offer letter email categorised as spam

Actual : not spam (class 0) ⟶ False Positive (FP)

Predicted : spam (class 1)

{ Conclusion : FP is dangerous }

{ Need : Minimize FP }

Metric Needed : FP decreases , TP increases

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

We want a metric that focusses on TP and FP!

{Total pos ✗}

Precision!

$$\left\{ \text{Precision} = \frac{TP \checkmark}{\{TP + FP\}} \right\}$$

→ Total {predicted} positives

→ is a metric based on model prediction

{ intuition → Of all positive predictions from my model, what % are actually truly positive? }

If my precision = 90%

It means that if

} Aims to minimize

It means that if
my model predicts there are
{100 spams} 90 will actually
✓ be spam, and remaining
× 10 are non-spams classified
as spam

Minimize
FP,
Meaning, you
want to
reduce chances
of missing out
on an
important
mail being
misclassified as
spam.

{ 90% of times,
I will not miss my offer letter }

More the precision,
surer I am about what goes
into my spams.

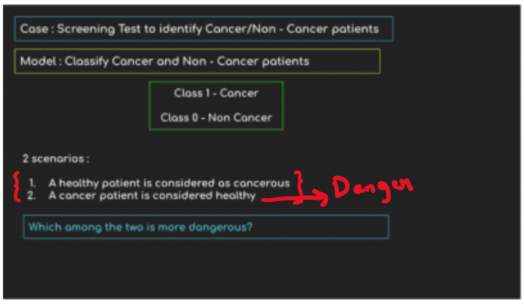
For spam detection,

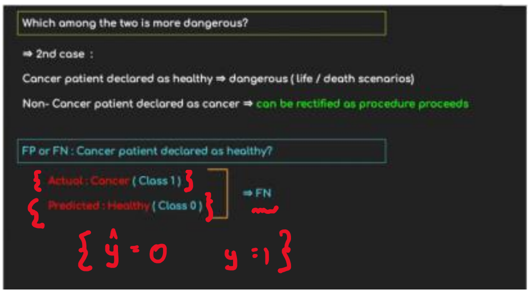
Use precision !!

Not Accuracy !!

Recall

12 August 2025 18:04

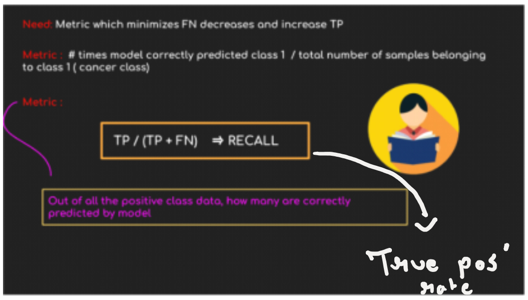




TP → correct detection of cancer

FN → Incorrectly missing detection

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



Recall is 90% → { 90% of those who had cancer were correctly classified }

Precision focusses on predictions } → How many positive pred were correct

Recall focusses on actuals } → How many positive actuals were correctly predicted

Precision we use the total predicted positives in the

Recall: we use the actual positives in the denominator

[Y → Recommend a Movie]

Recall: we use the actual positives in the denominator

For movie recommendation, what would you prioritize more in this case?

17 users have participated

A	High recall	47%
B ✓	High precision	53%

End Quiz Now

{ 1 => Recommend a Movie
0 => do not recommend }

If someone is recommend a wrong movie:

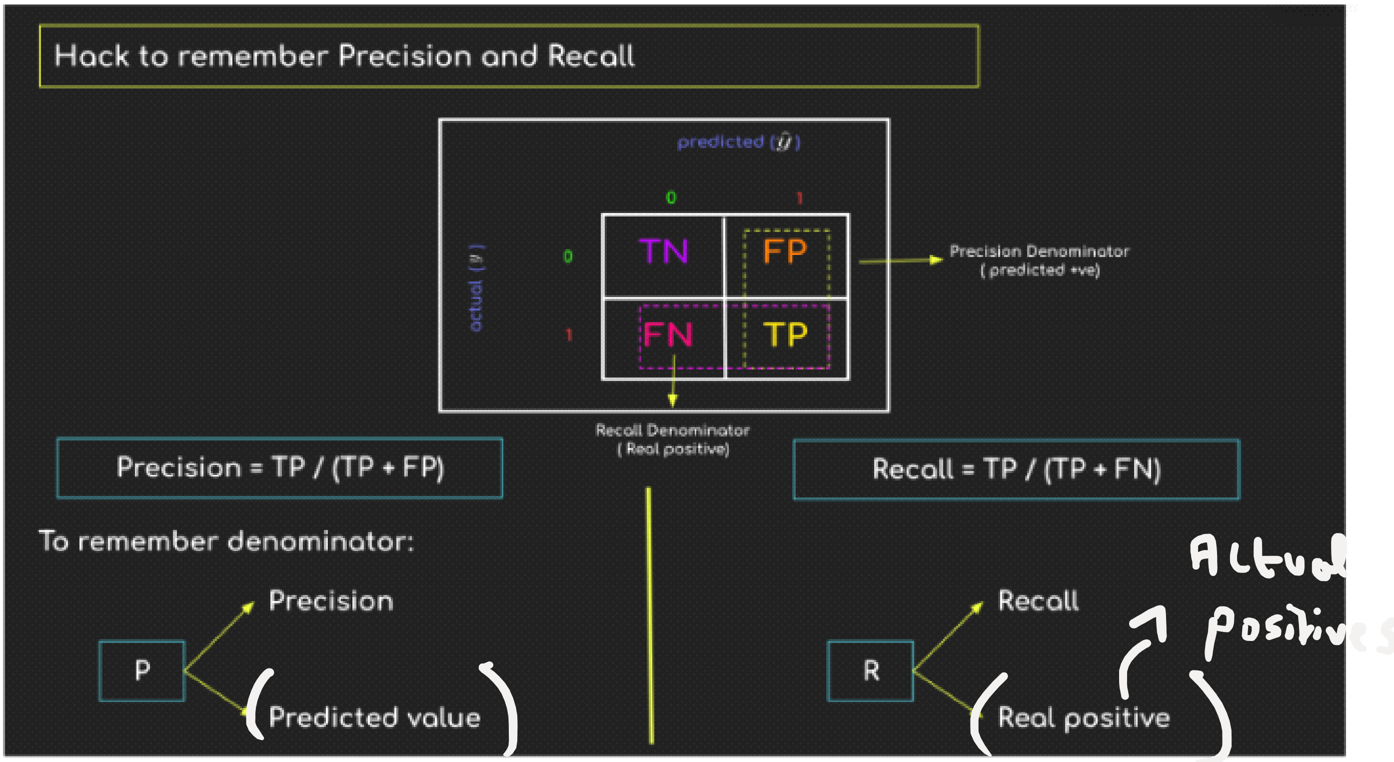
i) Customer is frustrated

ii) The movie will low rating because customer does not prefer such movies

{ Maximize Precision }

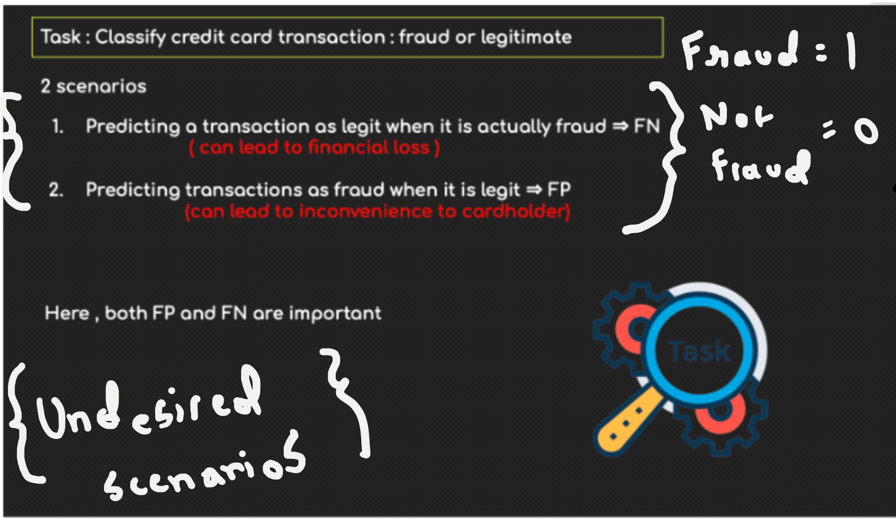
Hack to remember Precision and Recall

12 August 2025 18:06



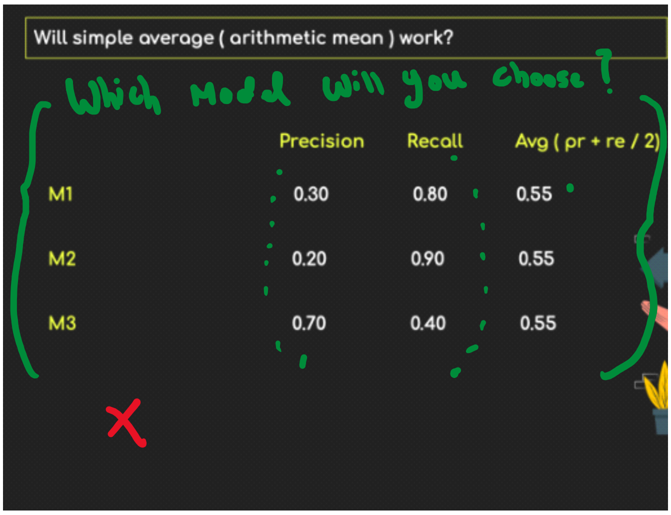
F 1 Score

12 August 2025 18:08



Both precision and recall are important.

Solution ?



Will simple average (arithmetic mean) work?

Which Model will you choose ?

	Precision	Recall	Avg (pr + re / 2)
M1	0.30	0.80	0.55
M2	0.20	0.90	0.55
M3	0.70	0.40	0.55

✗

incorrect way to combine precision and recall

Mean Won't work !

Brings us back to square one.

Will Harmonic mean work?

HM of (Precision , Recall) = $\dfrac{2}{\dfrac{1}{pr.} + \dfrac{1}{re.}} = \dfrac{2\ pr.\ re.}{pr. + re.}$

Note :

—- This HM of precision and recall is called F1 score

$$\left\{ \boxed{\text{F1 score}} = \frac{2\ pr.\ re.}{pr. + re.} \right\}$$

Both should be high

Both should be closer to each other

	Precision	Recall	F1 Score
M1	0.30	0.80	$\dfrac{2 \times 0.8 \times 0.3}{0.3 + 0.3} = 0.44$
	0.5		
M2	0.20	0.90	$\dfrac{2 \times 0.20 \times 0.9}{0.9 + 0.2} = 0.33$
	0.7		
M3	0.70	0.40	$\boxed{\dfrac{2 \times 0.7 \times 0.4}{0.7 + 0.4} = 0.51}$
	0.3		Best model

intuition → Harmonic Mean penalizes the model when both precision and recall are low.

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

KIPIMPS

27 users have participated

A	AM penalizes models the most when even Precision and Recall are low.	19%
B	HM penalizes models the most when even Precision and Recall are low.	48%
C	HM penalizes models the most when even Precision and Recall are high.	22%
D	AM penalizes models the most when even Precision and Recall are high.	11%

End Quiz Now

↳ selects model high precision/recall

Imbalanced dataset

↳ F1 score if you care about negatives / positives

A perfect Model will have F1 score = 1
(When both precision and recall are 1)

The worst model will have F1 score of 0

In the real world, F1 score of 0.7
is considered decent.

50 Spams } ⇒ Oversample

950 non Spam

Create synthetic spam points
⇓
{ Random oversampling
ADASYN
SMOTE }

{ Weight the datapoints }
⇓
Tree based Models