Classification Metrics

Business Case : Spam vs Not Spam

You are working in Google & your Task is : to create an email spam detection model

Here,

Not spam ⇒ Class 0 (Negative class)
Spam ⇒ Class 1 (Positive class)

Is model accuracy of 93% a good one?

Assume, we have

Dumb model ⇒ predicts every mail as not spam

Test data = 1200

Not spam = 850

Spam = 350

Imbalanced Data

Accuracy = 850/ 1200 x 100 = 70.83%

Seems good

Let’s increase the not spam data to 1100

i.e , Spam ⇒ 100
Not spam ⇒ 1100

Dumb Model Accuracy = 1100/1200 x 100 = 91.67%

Observe :

⇒ As number of not spam samples increase, bad model accuracy also increases.

⇒ But, model is not able to classify spam emails

Issue with Accuracy as metric

1. When data is imbalanced ⇒ Accuracy is bad metric
2. Fails to capture class wise (granular) performance.

Failing to classify spam data

Points to remember

Accuracy is bad metric for imbalance data

How to overcome the issues of accuracy?

Need : Metric which measures number of data points being

1. Correctly predicted in each class
2. Incorrectly predicted in each class

==> create 2 x 2 matrix s.t.

predicted ($\hat{y}$)

actual (y)		not spam (0)	spam (1)
	not spam (0)	1	2
	spam (1)	3	4

Count of data points where:

1 $\Rightarrow$ $y = 0 \;\; \& \;\; \hat{y} = 0$

2 $\Rightarrow$ $y = 0 \;\; \& \;\; \hat{y} = 1$

3 $\Rightarrow$ $y = 1 \;\; \& \;\; \hat{y} = 0$

4 $\Rightarrow$ $y = 1 \;\; \& \;\; \hat{y} = 1$

Terminologies :

==> create 2 x 2 matrix s.t.

		predicted ($\hat{y}$)	
		not spam (0)	spam (1)
actual (y)	not spam (0)	TN	FP
	spam (1)	FN	TP

Count of data points where:

True Neg (TN) $\Rightarrow$ $y = 0 \;\; \& \;\; \hat{y} = 0$

False Positive (FP) $\Rightarrow$ $y = 0 \;\; \& \;\; \hat{y} = 1$

False Negative (FN) $\Rightarrow$ $y = 1 \;\; \& \;\; \hat{y} = 0$

True Positive (TP) $\Rightarrow$ $y = 1 \;\; \& \;\; \hat{y} = 1$

Hacks to remember TP , TN , FP, FN

1st Term
(True / False)

- T if $\hat{y} = y$
- F if $\hat{y} \neq y$

2nd Term
(Positive/ negative)

- P if $\hat{y} = 1$
- N if $\hat{y} = 0$

Confusion matrix for Multi-class

2 x 2 matrix ⇒ confusion matrix for 2 classes

confusion matrix for K classes? ⇒ k x k matrix

Confusion matrix for Dumb model

Given : Dumb Model , Test Data → 400 data points

360 not spam

40 spam

Predicts all emails as non spam (class 0)

		predicted ($\hat{y}$)	
		0	1
actual (y)	0	360 TN	0 FP
	1	40 FN	0 TP

Observe :

⇒ Both TP and FP = 0 for dumb model

⇒ correctly classified 360 samples as not spam (TN = 360)

⇒ incorrectly classified 40 spam as not-spam (FN = 40)

Confusion matrix for ideal model

Given : Test Data

400 data points

360 not spam

40 spam

actual (y) \ predicted ($\hat{y}$)	0	1
0	360 TN	0 FP
1	0 FN	40 TP

Observe :

⇒ ideal model will correctly classify each datapoint (TN = 360 , TP = 40)

⇒ FP is also called as Type 1 error

⇒ FN is also called as Type 2 error

⇒ FP = FN = 0 i.e there are no errors / misclassification

predicted ($\hat{y}$) →

actual (y) ↓

actual \ predicted	0	1
0	TN	Type 1 FP
1	Type 2 FN	TP

How to find accuracy using confusion matrix ??

Given : confusion matrix i.e TN , FN , FP , TP

To find : Accuracy

Accuracy = Correct predictions / total samples

= TP + TN / (TP + TN + FP + FN)

		predicted ($\hat{y}$)	
		not spam (0)	spam (1)
actual (y)	not spam (0)	TN	FP
	spam (1)	FN	TP

Points to remember

⇒ Accuracy is bad metric for imbalance data

⇒ Confusion matrix :

		predicted ($\hat{y}$)	
		not spam (0)	spam (1)
actual (y)	not spam (0)	TN	FP
	spam (1)	FN	TP

⇒ FP – Type 1 error , FN – Type 2 error

Confusion matrix still doesn’t solve the issues of accuracy

⇒ Consider 2 scenarios

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

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

Need : Metric which measures FP & TP

Metric : # times model correctly predicted class 1 / # times model predicted class 1

Metric :

$$\boxed{TP/(TP + FP) \Rightarrow \text{Precision}}$$

Intuitively,

– It tells how precise model is to detect spam mail

Precision for dumb model

Given : Test Data

360 not spam

40 spam

Confusion matrix =

predicted ($\hat{y}$)

actual (y)	0	1
0	360 TN	0 FP
1	40 FN	0 TP

$$\text{Precision} = TP \; / \; (TP + FP)$$

$$= 0 \; / \; (0 + 0)$$

Math error

(undefined)

$$= 0 \; / \; (0 + 0 + 10^{-6}) \quad = 0$$

Add small value

Precision for ideal model

Given : Test Data

360 not spam

40 spam

Confusion matrix =

actual (y) \ predicted ($\hat{y}$)	0	1
0	360 TN	0 FP
1	0 FN	40 TP

Precision = TP / (TP + FP)

= 40 / (40 + 0) = 1

Note : Range of precision [0, 1]

Points to remember

⇒ Accuracy is bad metric for imbalance data

⇒ Confusion matrix :

		predicted ($\hat{y}$)	
		0	1
actual (y)	0	TN	FP
	1	FN	TP

⇒ FP – Type 1 error , FN – Type 2 error

⇒ Precision = TP / (TP + FP)

⇒ Precision minimizes FP

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

Model : Classify Cancer and Non - Cancer patients

Class 1 - Cancer

Class 0 - Non Cancer

2 scenarios :

1. A healthy patient is considered as cancerous
2. A cancer patient is considered healthy

Which among the two is more dangerous?

Which among the two is more dangerous?

⇒ 2nd case :

Cancer patient declared as healthy ⇒ dangerous (life / death scenarios)

Non- Cancer patient declared as cancer ⇒ can be rectified as procedure proceeds

FP or FN : Cancer patient declared as healthy?

Actual : Cancer (Class 1)

Predicted : Healthy (Class 0)

⇒ FN

Need: Metric which minimizes FN decreases and increase TP

Metric : # times model correctly predicted class 1 / total number of samples belonging to class 1 (cancer class)

Metric :

$$\boxed{TP / (TP + FN) \Rightarrow \text{RECALL}}$$

Out of all the positive class data, how many are correctly predicted by model

Recall for dumb model

Given : Test Data (100 mails)

55 not spam   45 spam

Recall $= TP / (TP + FN)$

$= 0 / (0 + 45)$

$= 0$

actual (y) \ predicted ($\hat{y}$)	0	1
0	55 TN	0 FP
1	45 FN	0 TP

Recall for ideal model

Given : Test Data (100 mails)

55 not spam   45 spam

		predicted ($\hat{y}$) 0	predicted ($\hat{y}$) 1
actual (y)	0	55 TN	0 FP
actual (y)	1	0 FN	45 TP

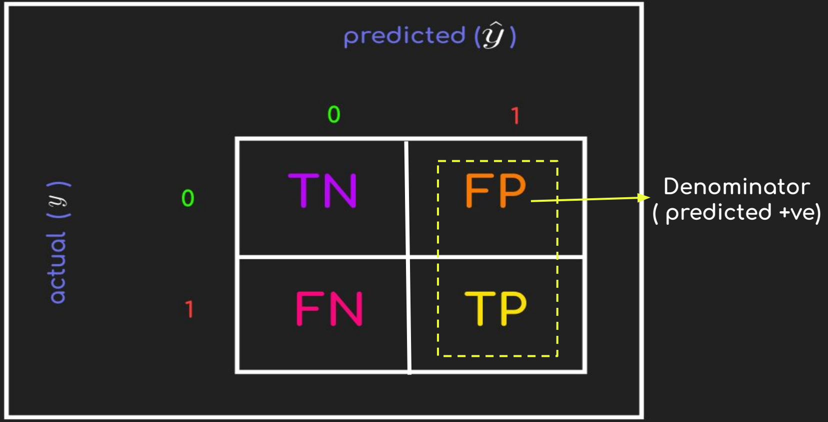
Recall = TP / (TP + FN)

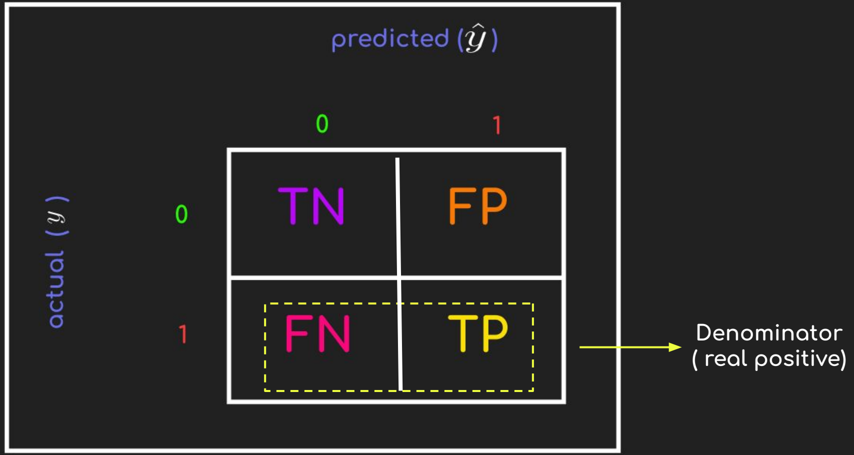
= 45 / (45 + 0) = 1

Recall = 1

Note : Range of recall ⇒ [0 , 1]

Hack to remember Precision and Recall





Precision =

Correctly predicted class 1 / total samples predicted as class 1

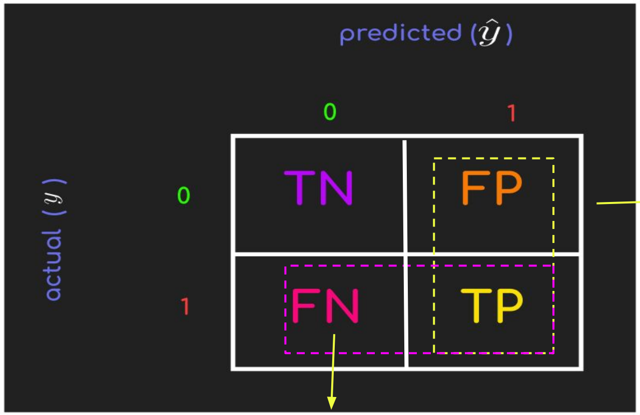
TP / (TP + FP)

Recall =

Correctly predicted class 1 / total samples actual class 1

TP / (TP + FN)

Hack to remember Precision and Recall



Precision Denominator
(predicted +ve)

Recall Denominator
(Real positive)

Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

To remember denominator:

P → Precision
P → Predicted value

R → Recall
R → Real positive

Points to remember

⇒ Accuracy is bad metric for imbalance data

⇒ Confusion matrix :

predicted ($\hat{y}$)

actual (y)	not spam (0)	spam (1)
not spam (0)	TN	FP
spam (1)	FN	TP

⇒ FP – Type 1 error , FN – Type 2 error

⇒ Precision = TP / (TP + FP)

⇒ Precision minimizes FP

⇒ Recall = TP / TP + FN

⇒ Recall minimizes FN

Task : Classify credit card transaction : fraud or legitimate

2 scenarios

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

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

Here , both FP and FN are important

We train 3 different models s.t.

Results:

	Precision	Recall
M1	0.30	0.80
M2	0.20	0.90
M3	0.70	0.40

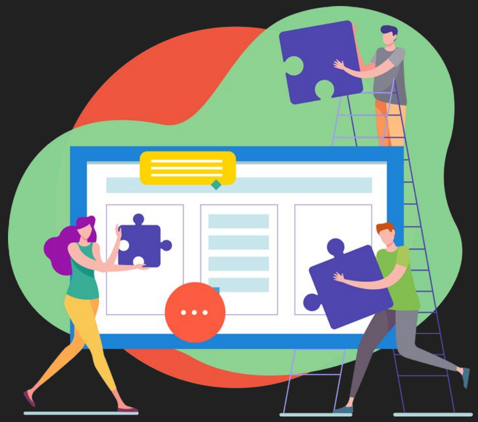
Which model among M1 , M2 and M3 is the best?

⇒ Based on precision

⇒ M3 is the best model

⇒ Based on recall

⇒ M2 is the best model



Which one to choose??

NEED : a way to combine precision and recall

Will simple average (arithmetic mean) work?

	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

Will Harmonic mean work?

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

Note :

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

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

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

Best model

F1 Score of bad model

For bad model ,

Precision = recall = 0

$$\text{F1 score} = \frac{2.\ \text{Precision}\ .\ \text{recall}}{\text{Precision} + \text{recall}}$$

$$= \frac{2.\ 0\ .\ 0}{0 + 0}$$

Math error

$$= \frac{2.\ 0\ .\ 0}{0 + 0 + 10^{-6}} = 0$$

Add small value

F1 score for ideal model

For ideal model ,

Precision = recall = 1

$$= \frac{2.1.1}{1+1} = \frac{2}{2} = 1$$

Conclusion : Range of F1 ⇒ [0 ,1]

Points to remember

⇒ Accuracy is bad metric for imbalance data

⇒ Confusion matrix :

		predicted ($\hat{y}$)	
		not spam (0)	spam (1)
actual (y)	not spam (0)	TN	FP
	spam (1)	FN	TP

⇒ FP – Type 1 error , FN – Type 2 error

⇒ Precision = TP / (TP + FP)

⇒ Precision minimizes FP

⇒ Recall = TP / TP + FN

⇒ Recall minimizes FN

Points to remember

⇒ F1 Score combines precision and recall

$$\text{F1 score} = \frac{2.\ \text{Precision}\ .\ \text{recall}}{\text{Precision} + \text{recall}}$$