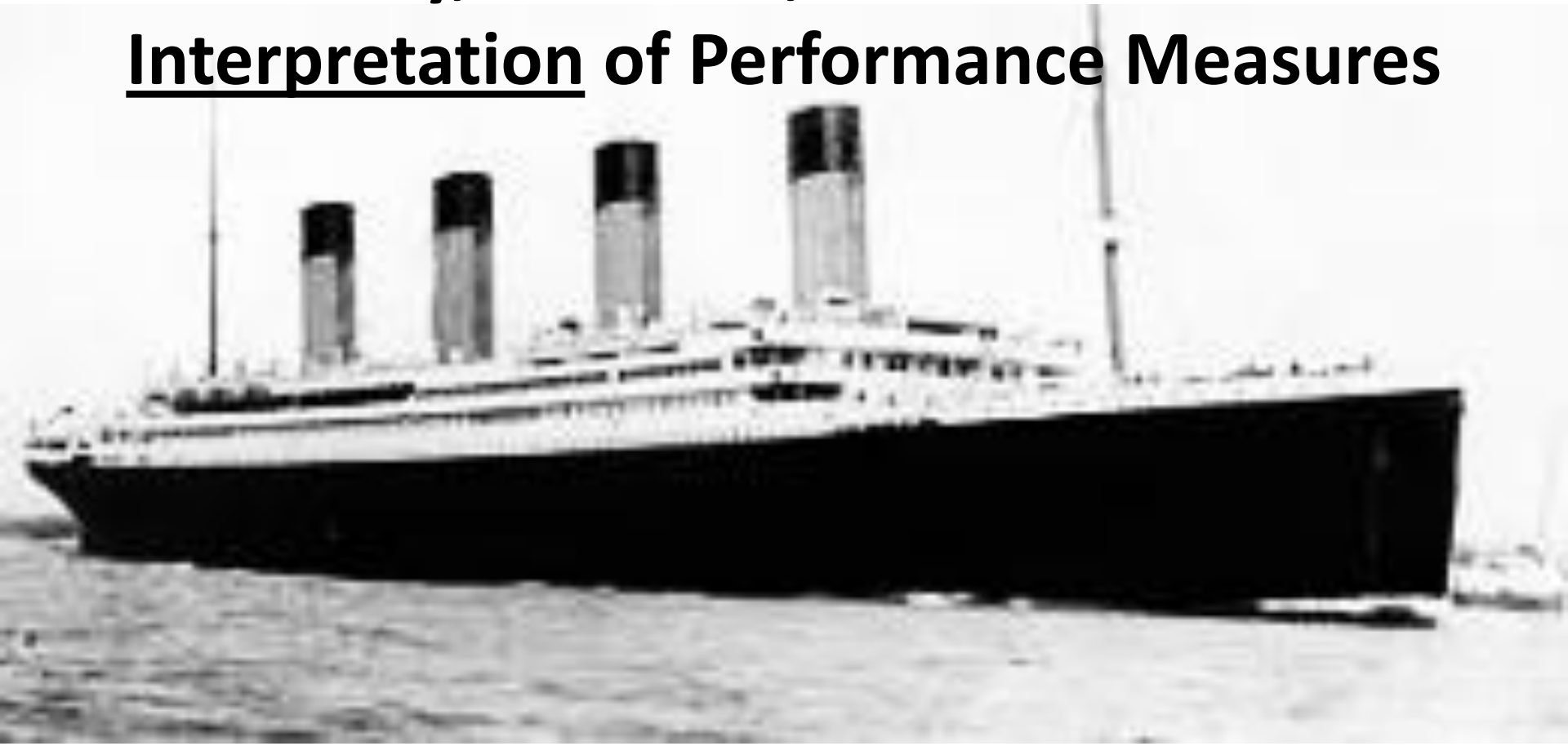
Accuracy, Precision, Recall & F1 Score: Interpretation of Performance Measures



Sample data : titanic data set

Algorithm Used : Decision Tree

Goal : to predict the survival of the passengers on the Titanic.

Once you have built your model, the most important question that arises is how good is your model?

So, evaluating your model is the most important task in the data science project than "how good your predictions are".

Sample O/p :

True Positive	False Negative	Accuracy	Precision	Threshold
41	24	0.803	0.788	0.5

False Positive	True Negative	Recall	F1 Score
11	102	0.631	0.701

Confusion Matrix

Classification report

Lets understand the above data.

Confusion Matrix

	Predicted class		
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

True positive and true negatives are the observations that are correctly predicted and therefore shown in green. We want to minimize false positives and false negatives so they are shown in red color.

True Positives (TP) - E.g. if actual class value indicates that this passenger survived and predicted class tells you the same thing.

True Negatives (TN) - E.g. if actual class says this passenger did not survive and predicted class tells you the same thing.

False Positives (FP) – E.g. if actual class says this passenger did not survive but predicted class tells you that this passenger will survive.

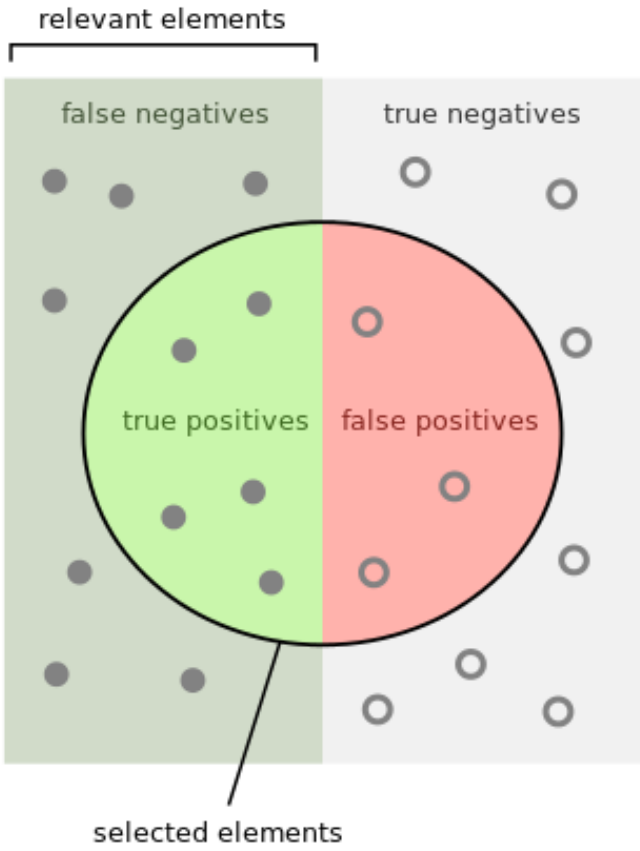
False Negatives (FN) – E.g. if actual class value indicates that this passenger survived and predicted class tells you that passenger will die.

Accuracy - Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same. Therefore, you have to look at other parameters to evaluate the performance of your model. For our model, we have got 0.803 which means our model is approx. 80% accurate.

$$\text{Accuracy} = (TP+TN) / (TP+FP+FN+TN)$$

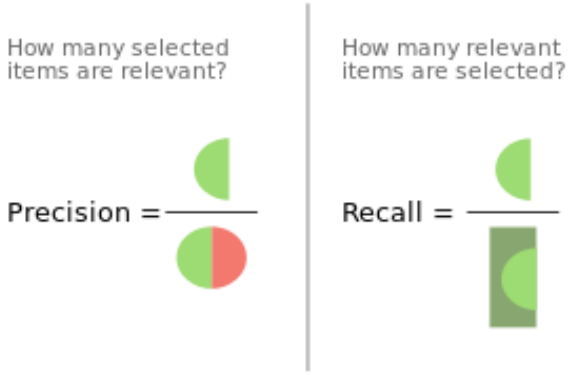
Precision - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is ***of all passengers that labeled as survived, how many actually survived?*** High precision relates to the low false positive rate. We have got 0.788 precision which is pretty good.

Precision = TP / (TP+FP)



Recall (also called Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes. The question recall answers is: ***Of all the passengers that truly survived, how many did we label?*** We have got recall of 0.631 which is good for this model as it's above 0.5.

Recall = TP / TP+FN



F1 score - F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account. Intuitively it is not as easy to understand as accuracy, but F1 is usually more useful than accuracy, especially if you have an uneven class distribution. ***Accuracy works best if false positives and false negatives have similar cost***. If the cost of false positives and false negatives are very different, it's better to look at both Precision and Recall. In our case, F1 score is 0.701.

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

Evaluate classification models using F1 score

https://www.ritchieng.com/machinelearning-f1-score/

Remember

CONFUSION MATRIX

TP = True Positives
TN = True Negatives
FP = False Positives
FN = False Negatives

	p' (Predicted)	n' (Predicted)
p (Actual)	True Positive	False Negative
n (Actual)	False Positive	True Negative

$$\textbf{Precision} = \frac{TP}{TP + FP}$$

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

$$\textbf{F1-score} = \frac{2 * precision * recall}{precision + recall}$$