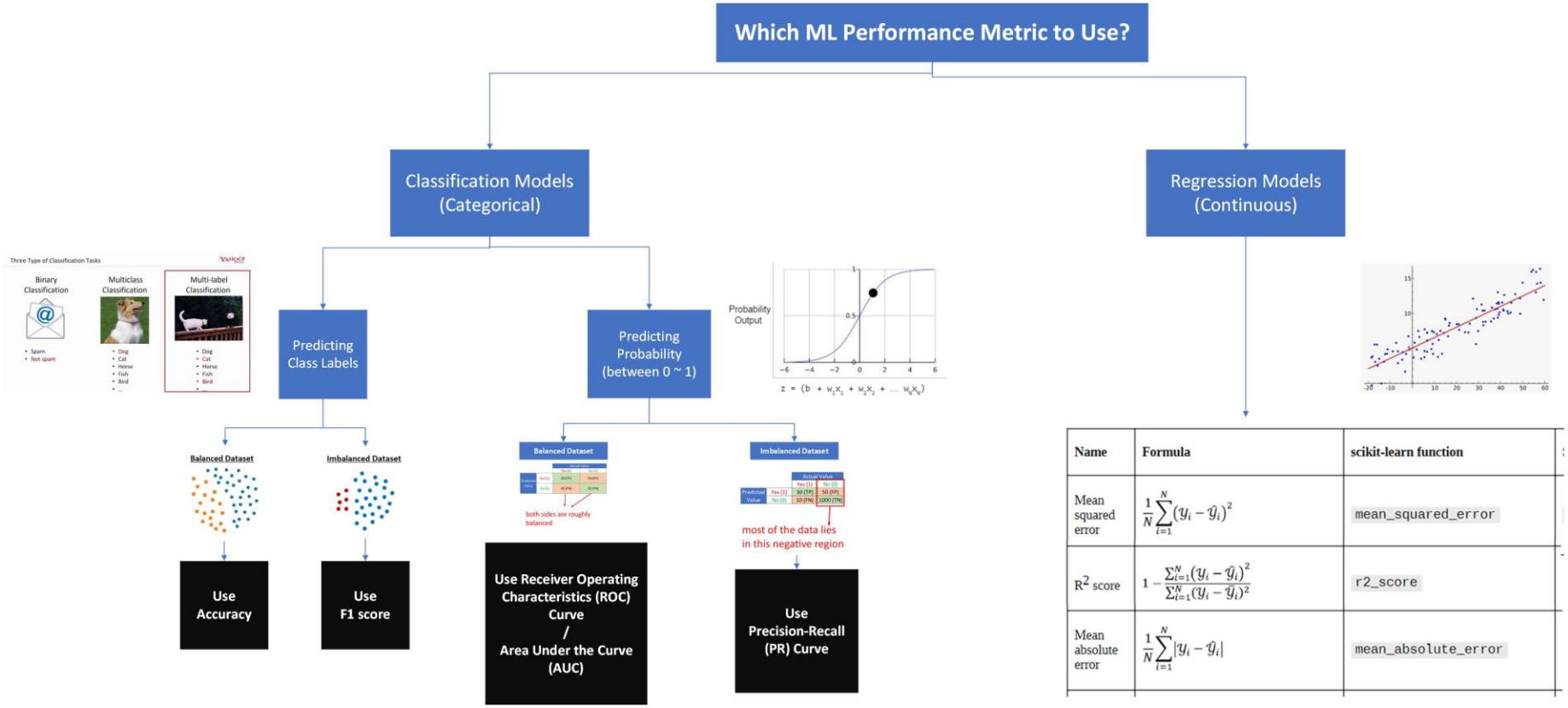
PERFORMANCE METRICS FOR MACHINE LEARNING MODELS

DR. ALVIN ANG

TABLE OF CONTENTS



Name	Formula	scikit-learn function
Mean squared error	$\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$	`mean_squared_error`
R^2 score	$1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2}$	`r2_score`
Mean absolute error	$\frac{1}{N}\sum_{i=1}^{N}\lvert y_i - \hat{y}_i \rvert$	`mean_absolute_error`

Most Popular one widely used is ROC / AUC for Classification.

MOST Popular because it gives more weight to larger errors.

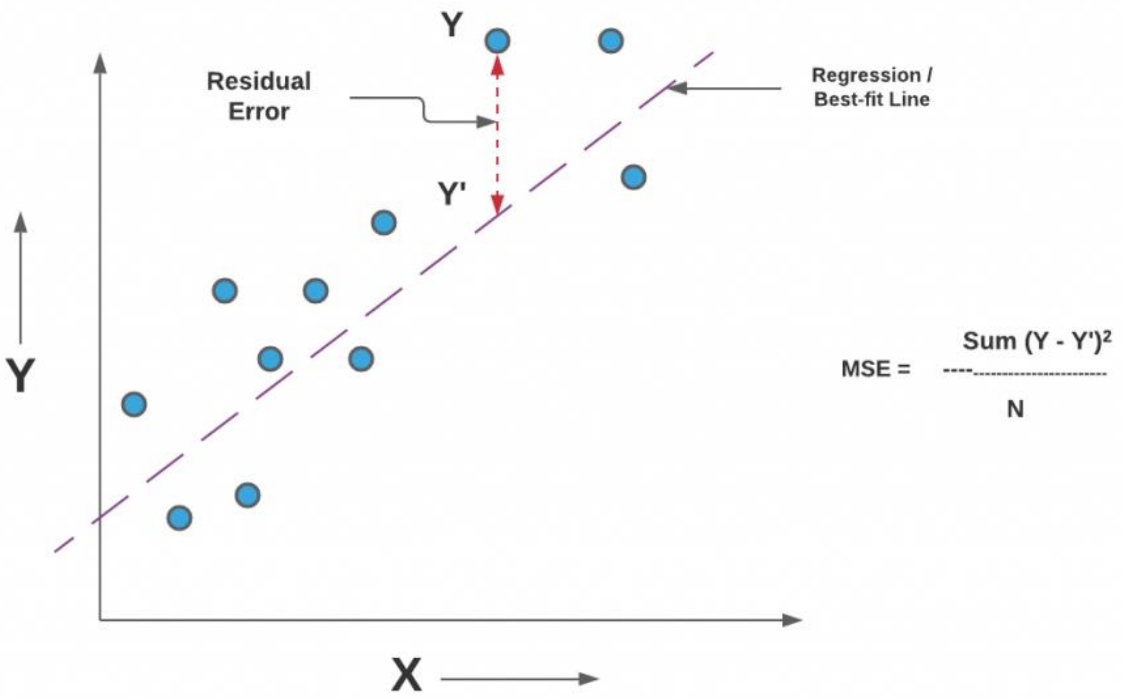
Also popular because it measures the "goodness of fit" of the model.

Least Popular because it does not give more weight to larger errors.

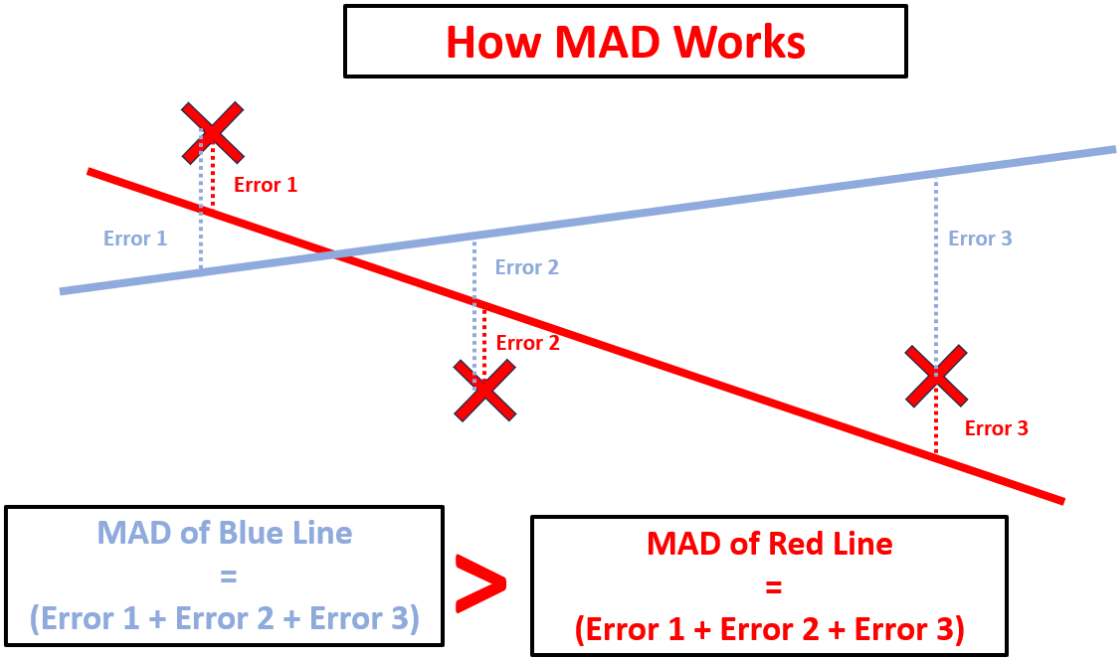
Name	Formula	scikit-learn function
Mean squared error	$\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$	`mean_squared_error`
R^2 score	$1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2}$	`r2_score`
Mean absolute error	$\frac{1}{N}\sum_{i=1}^{N}\lvert y_i - \hat{y}_i\rvert$	`mean_absolute_error`

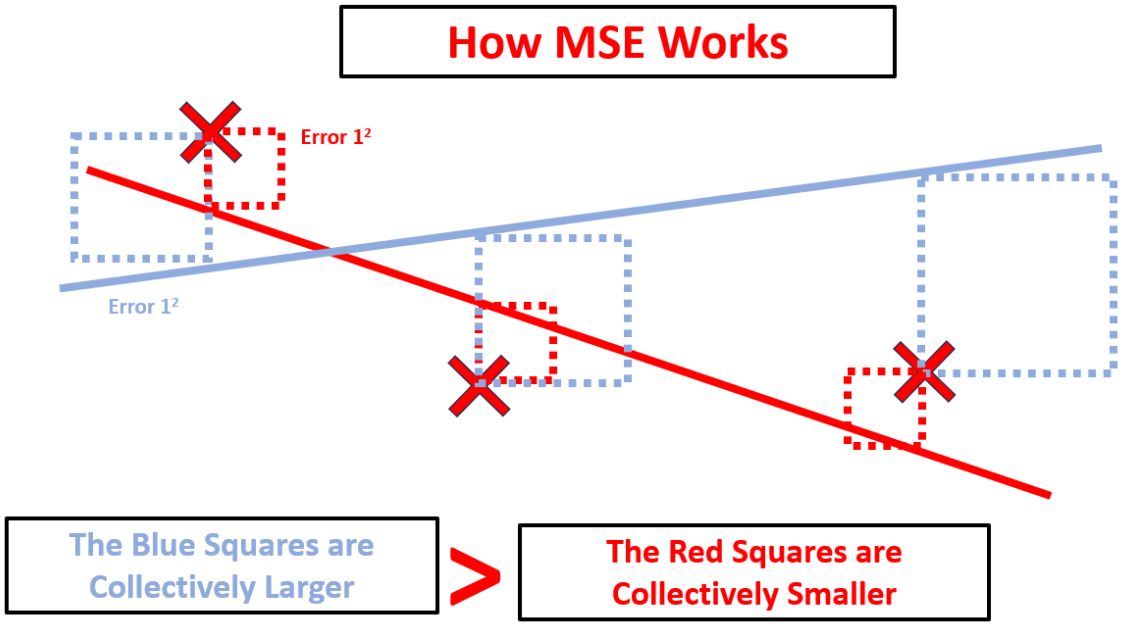
A. OBTAINING THE MSE

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$



B. COMPARING MSE VS MAD





Even though in both cases, you can see that the Red line fits better (because the Error is lower),

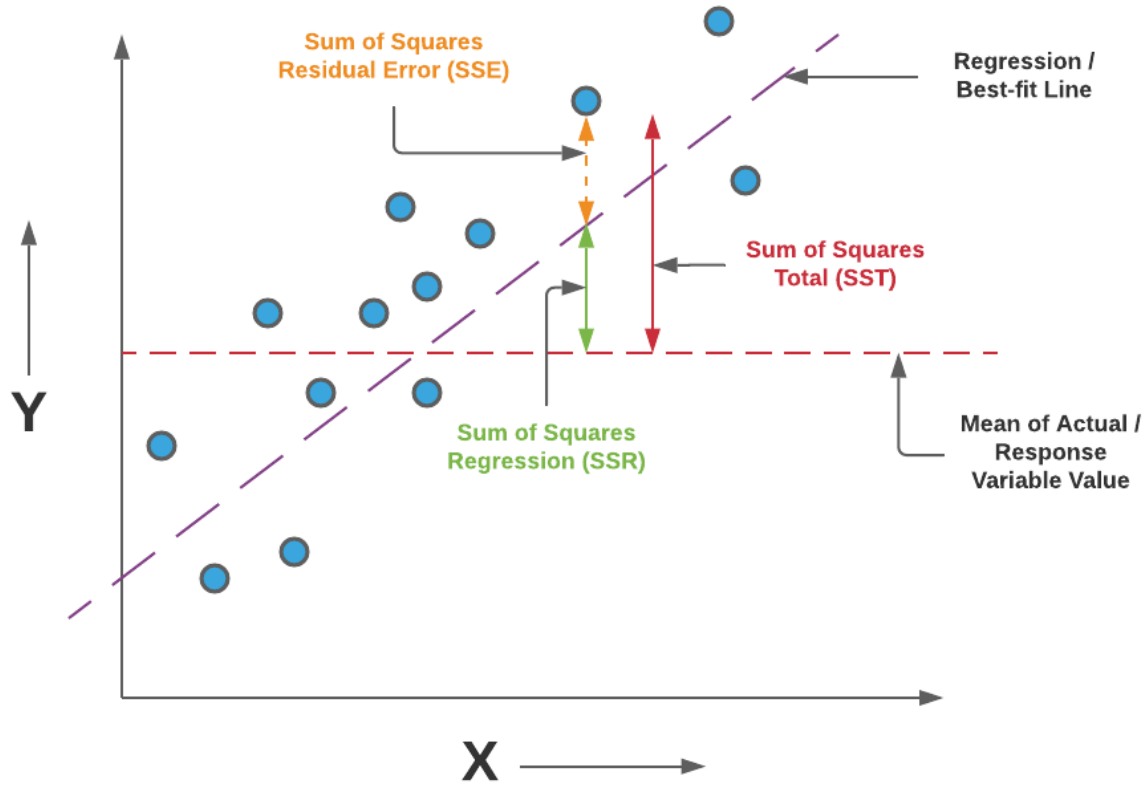
Its easier to compare the Squares in the MSE.

MSE Amplifies the Errors to make the difference more significant.

Else, through the MAD lines, you can't really tell which Line / Model is better.

C. OBTAINING THE R2

$$1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2}$$



Regression Statistics	
Multiple R	0.9205436
R Square	0.84740053
Adjusted R Square	0.82298461
Standard Error	1.39577331
Observations	30

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	270.4615184	67.6153796	34.706891	7.22088E-10
Residual	25	48.70457828	1.94818313		
Total	29	319.1660967			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-1.6685399	1.408315529	-1.18477702	0.24724684	-4.56902	1.23194026	-4.56902	1.23194026
Population (000,000)	0.55190731	0.050629062	10.9009981	5.468E-11	0.447634803	0.65617981	0.4476348	0.65617981
Percent Unemployed	0.20316264	0.117086169	1.73515493	0.09502688	-0.03798084	0.44430612	-0.03798084	0.44430612
Advertising Expense (000)	0.03135496	0.016062075	1.95211152	0.06221113	-0.0017255	0.06443543	-0.0017255	0.06443543
Mall Location	0.21979032	0.540028513	0.40699763	0.68747343	-0.89241922	1.33199987	-0.89241922	1.33199987

Coefficient of Multiple Determination $R^2 = \frac{SSR}{SS\ total}$

https://towardsdatascience.com/top-10-model-evaluation-metrics-for-classification-ml-models-a0a0f1d51b9

this is known as the Confusion Matrix

		Actual Value	
		Yes (1)	No (0)
Predicted Value	Yes (1)	True Positive (TP)	False Positive (FP)
	No (0)	False Negative (FN)	True Negative (TN)

A. TYPE I VS TYPE II ERROR

		Actual Value	
		Yes (1)	No (0)
Predicted Value	Yes (1)	500 (TP)	100 (FP)
	No (0)	200 (FN)	200 (TN)

Type I Error occurs here

Type II Error occurs here

B. ACCURACY

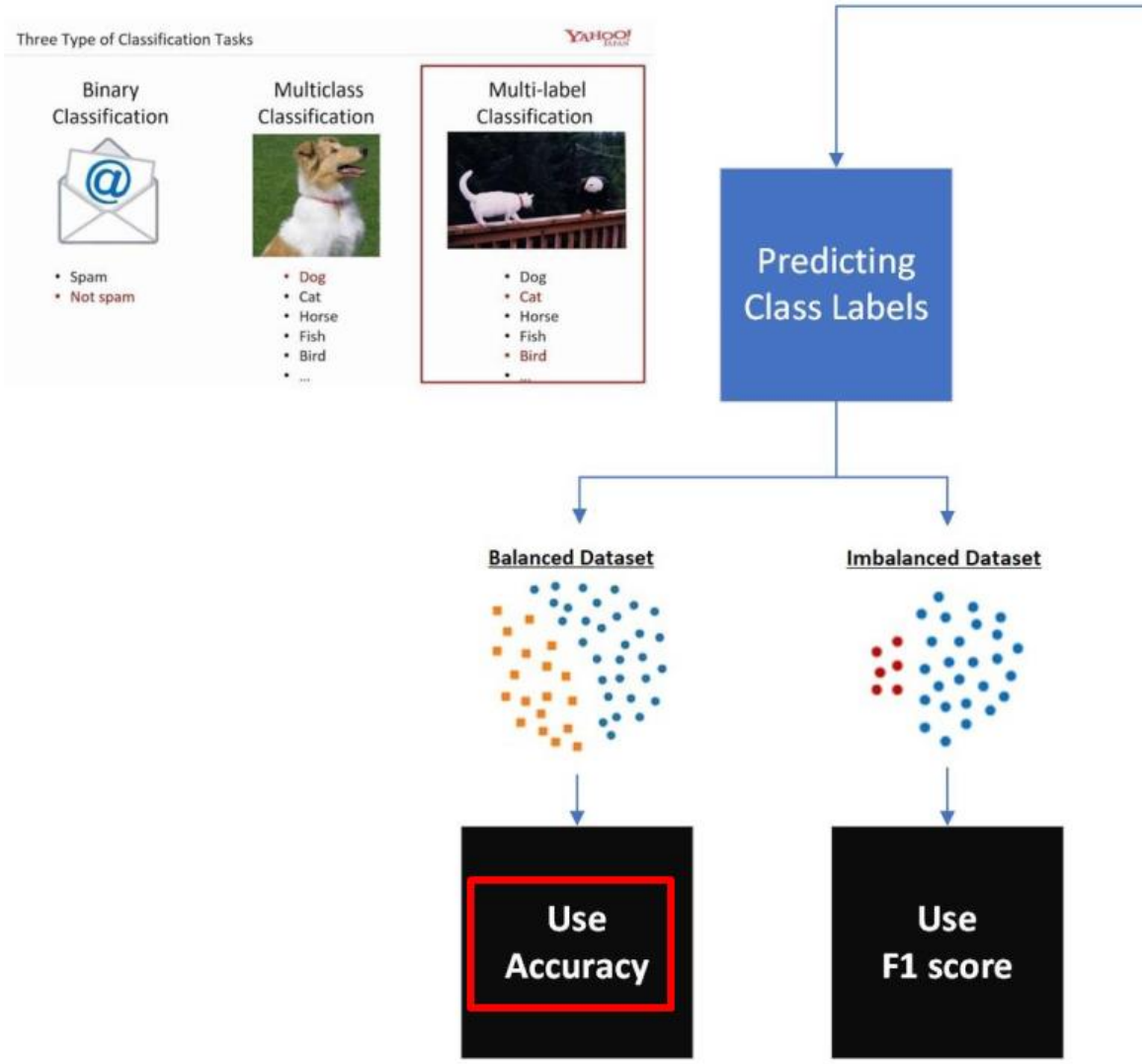
		Actual Value		
		Yes (1)	No (0)	Total
Predicted Value	Yes (1)	500 (TP)	100 (FP)	J
	No (0)	200 (FN)	200 (TN)	K
	Total	X	Y	TOTAL

$$\text{ACCURACY} = \frac{TP + TN}{TOTAL}$$

Accuracy is between 0 to 1

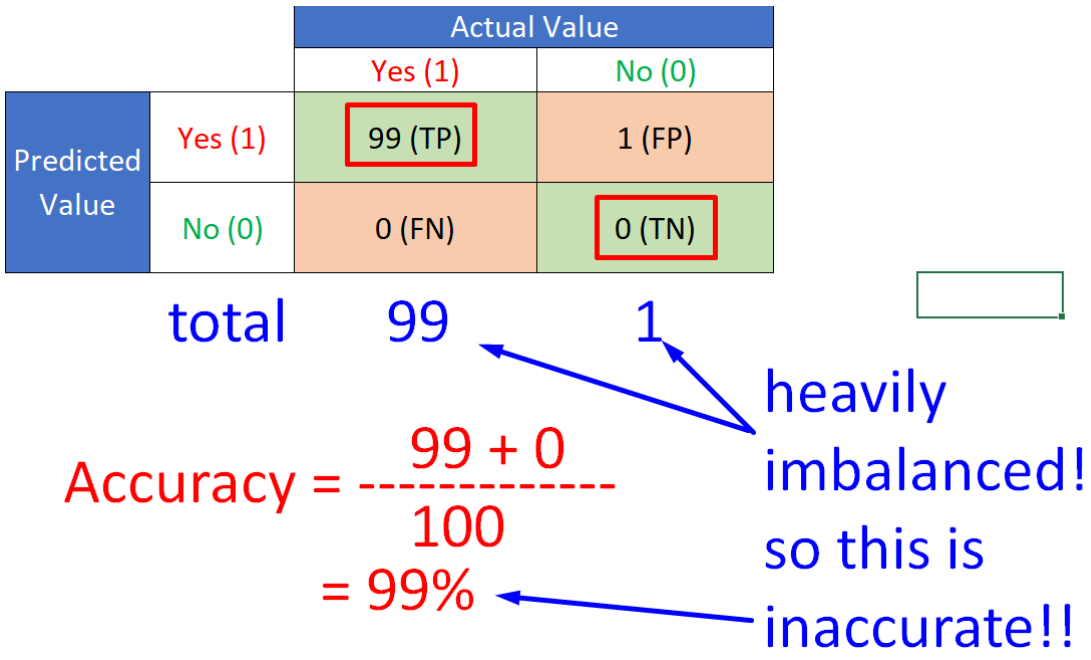
0 = BAD

1 = GOOD



Accuracy Metric can only be used on Balanced Dataset.

If its used on Imbalanced Dataset, the result could still be very Accurate even though its NOT.



To couunter the Imbalanceness, we use the F1 score (which comprises of both Precision + Recall).

C. PRECISION

		Actual Value		
		Yes (1)	No (0)	Total
Predicted Value	Yes (1)	500 (TP)	100 (FP)	J
	No (0)	200 (FN)	200 (TN)	K
	Total	X	Y	TOTAL

$$\text{PRECISION} = \frac{TP}{J}$$

- Precision is between 0 to 1
- 0 = BAD
- 1 = GOOD
- Both Precision and Recall are needed to obtain the F1 score (as we will see later).

D. RECALL / SENSITIVITY

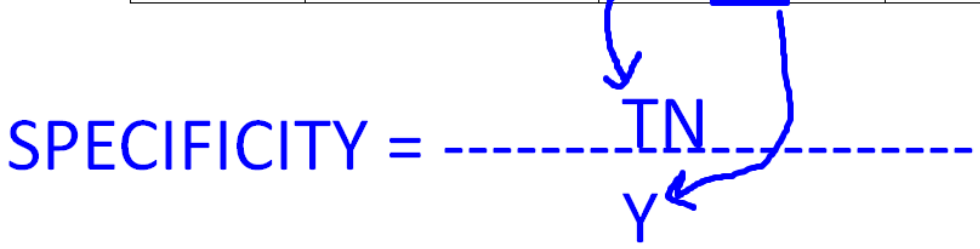
		Actual Value		
		Yes (1)	No (0)	Total
Predicted Value	Yes (1)	500 (TP)	100 (FP)	J
	No (0)	200 (FN)	200 (TN)	K
	Total	X	Y	TOTAL

$$\text{RECALL} = \frac{TP}{X}$$

- RECALL = SENSITIVITY = TRUE POSITIVE RATE (TPR)
- They are all the same name!
- Sensitivity is between 0 and 1
- 0 = BAD
- 1 = GOOD
- Both Precision and Recall are needed to obtain the F1 score (as we will see later).

E. SPECIFICITY

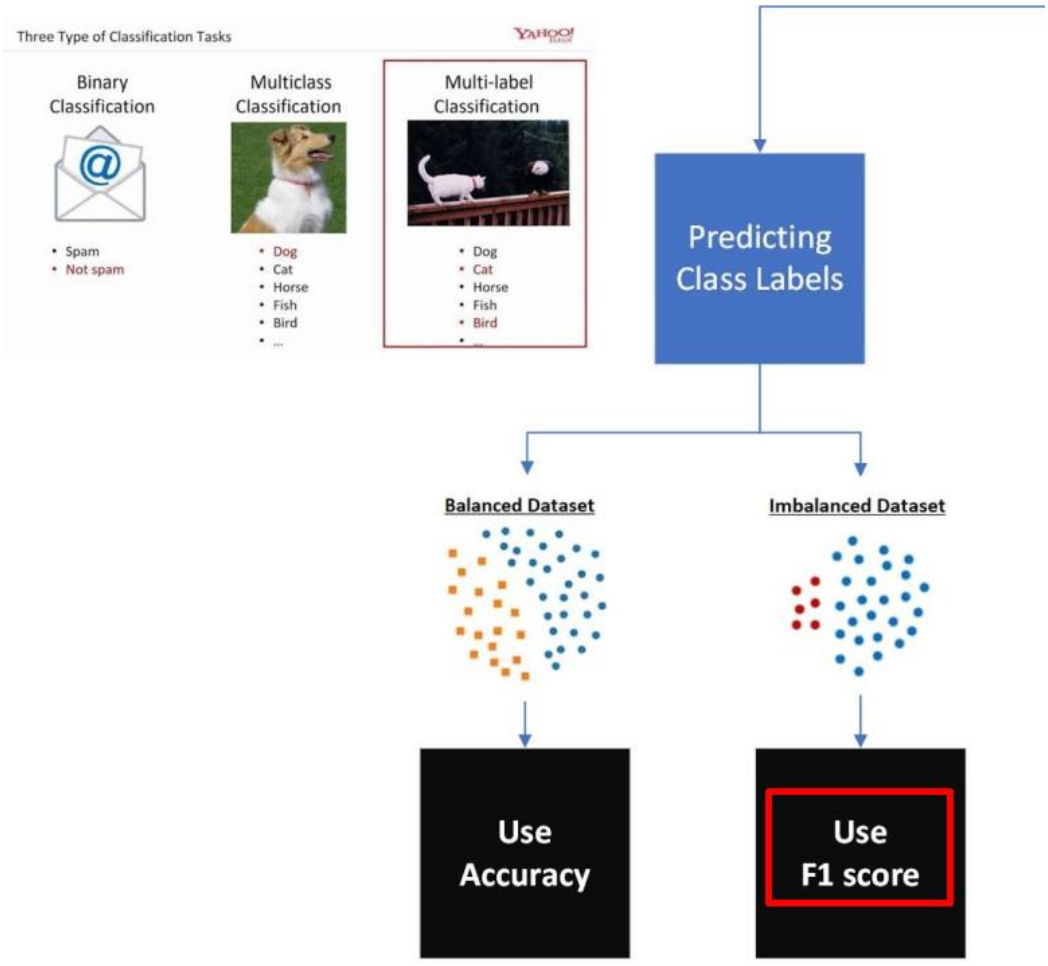
		Actual Value		
		Yes (1)	No (0)	Total
Predicted Value	Yes (1)	500 (TP)	100 (FP)	J
	No (0)	200 (FN)	200 (TN)	K
	Total	X	Y	TOTAL

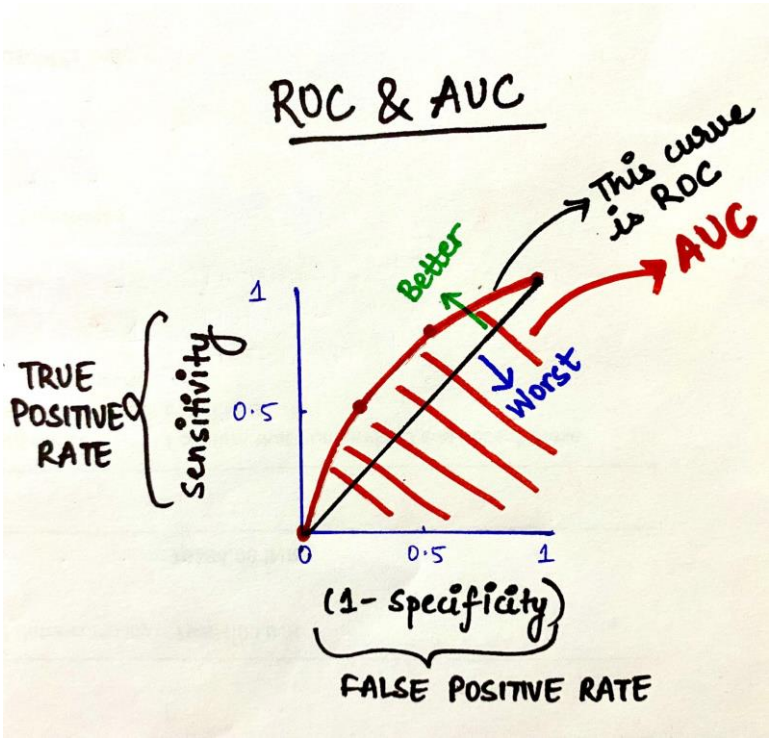


- Specificity is between 0 to 1
- 0 = BAD
- 1 = GOOD
- Specificity is required for the ROC / AUC curve as we will see later.

$$F1\,score = 2 * \left(\frac{Precision * Recall}{Precision + Recall} \right)$$

- Don't have to focus on the formula above…
- What's more important to note is that: F1 score is dependent on BOTH PRECISION and RECALL





- ROC = Receiver Operating Characteristics Curve
- AUC = Area Under the Curve = Area under the ROC curve
- Axis are SENSITIVITY vs (1 – SPECIFICITY)
- The further away the ROC curve from the middle linear line, the better your model is.
- The greater the area taken up by the AUC the better your model is.
- AUC score is between 0.5 to 1.
    - 0.5 = BAD model → cannot be used to classify
    - 1 = GOOD model → Good at classifying
    - 0.5 = not good not bad.

A. HOW TO PLOT THE ROC CURVE?

Example:

- We have 10 persons.
- A probability score of near 0 means low chance that the person is NOT OBESE.
- A probability score of near 1 means high chance the person IS OBESE.
- If THRESHOLD = 0.5 →
    - Anything <=0.5 is classified as NOT OBESE
    - Anything > 0.5 is classified as OBESE

1. SAY FOR EXAMPLE, WE SET THE THRESHOLD = 0

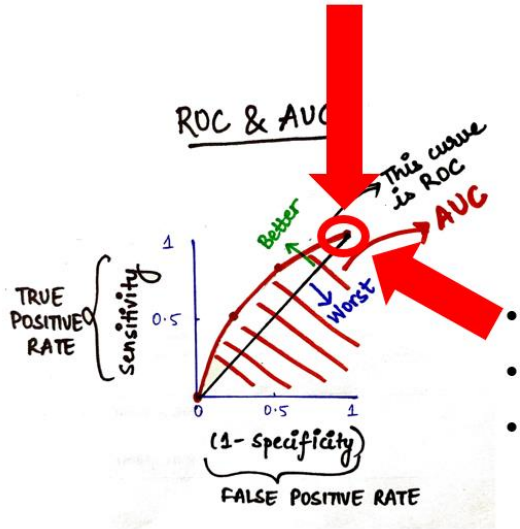
		Actual Value		
		Yes (1)	No (0)	Total
Predicted Value	Yes (1)	6 (TP)	4 (FP)	J
	No (0)	0 (FN)	0 (TN)	K
	Total	X	Y	TOTAL

Sensitivity
= 6 / 6
= 100%

Specificity
= 0 / 4
= 0%

1 - Specificity = 100%

You have plotted this point!



- Sensitivity = 1
- 1 – Specificity = 1
- Threshold = 0

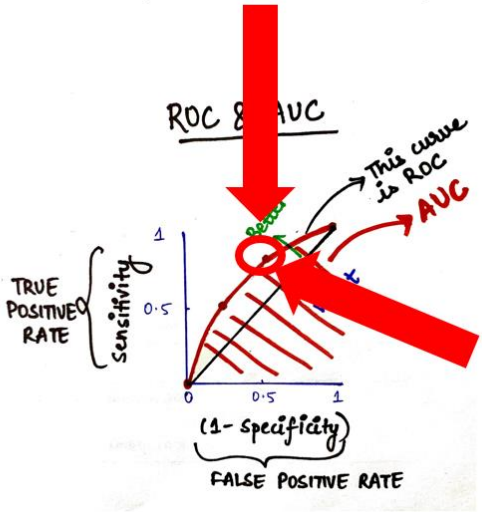
2. SAY FOR EXAMPLE, WE SET THE THRESHOLD = 0.3

		Actual Value		
		Yes (1)	No (0)	Total
Predicted Value	Yes (1)	5 (TP)	2 (FP)	J
	No (0)	1 (FN)	2 (TN)	K
	Total	X	Y	TOTAL

sensitivity
= 5 / 6
= 83%

Specificity
= 2 / 4
= 50%
1 - Specificity = 50%

You have plotted this point!



- Sensitivity = 0.83
- 1 – Specificity = 0.5
- Threshold = 0.3

3. SAY FOR EXAMPLE, WE SET THE THRESHOLD = 0.6

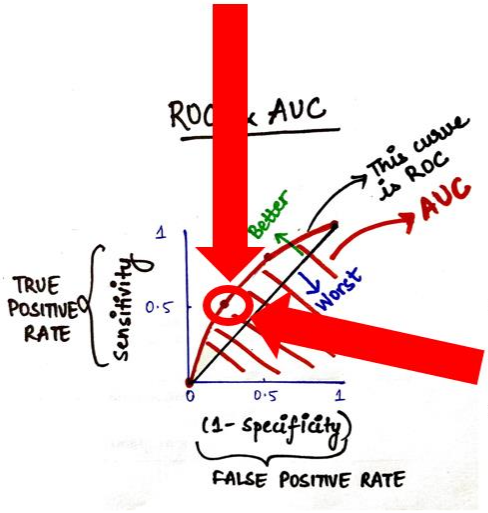
		Actual Value		
		Yes (1)	No (0)	Total
Predicted Value	Yes (1)	3 (TP)	1 (FP)	J
	No (0)	3 (FN)	3 (TN)	K
	Total	X	Y	TOTAL

Sensitivity
=3 / 6
= 50%

Specificity
= 3 / 4
= 75%

1-Specificity = 25%

You have plotted this point!



- Sensitivity = 0.5
- 1 – Specificity = 0.25
- Threshold = 0.6

4. SAY FOR EXAMPLE, WE SET THE THRESHOLD = 0.9

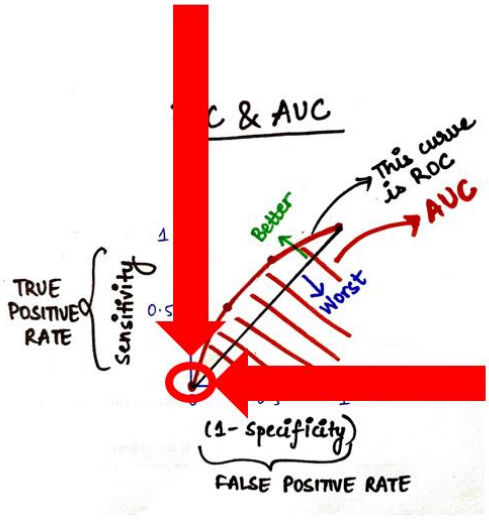
		Actual Value		
		Yes (1)	No (0)	Total
Predicted Value	Yes (1)	0 (TP)	0 (FP)	J
	No (0)	6 (FN)	4 (TN)	K
	Total	X	Y	TOTAL

Sensitivity
=0/6
=0%

Specificity
= 4 / 4
= 100%

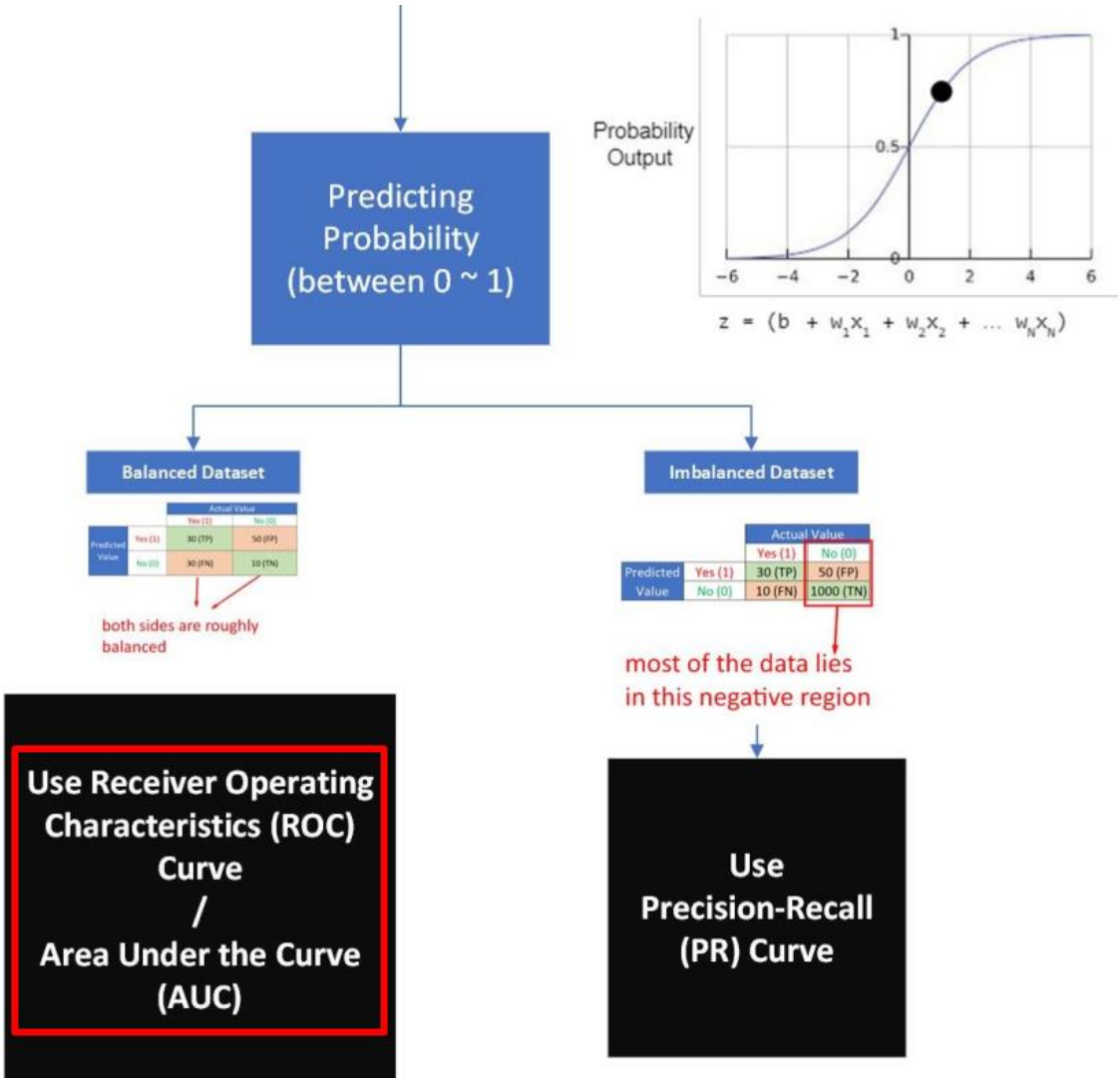
1 - Specificity = 0%

You have plotted this point!

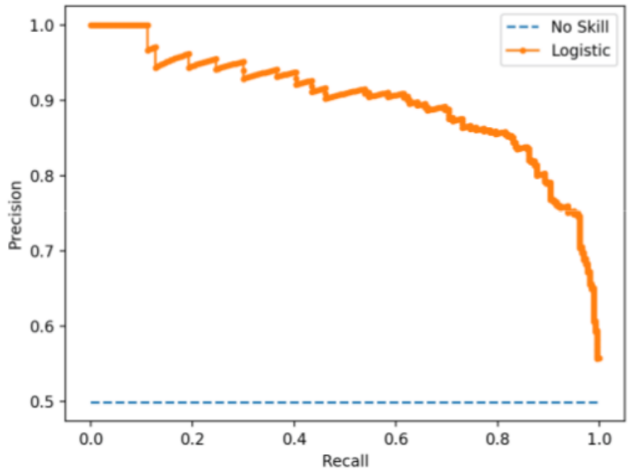


- Sensitivity = 0
- 1 – Specificity = 0
- Threshold = 0.9

B. WHEN TO USE THE ROC / AUC?

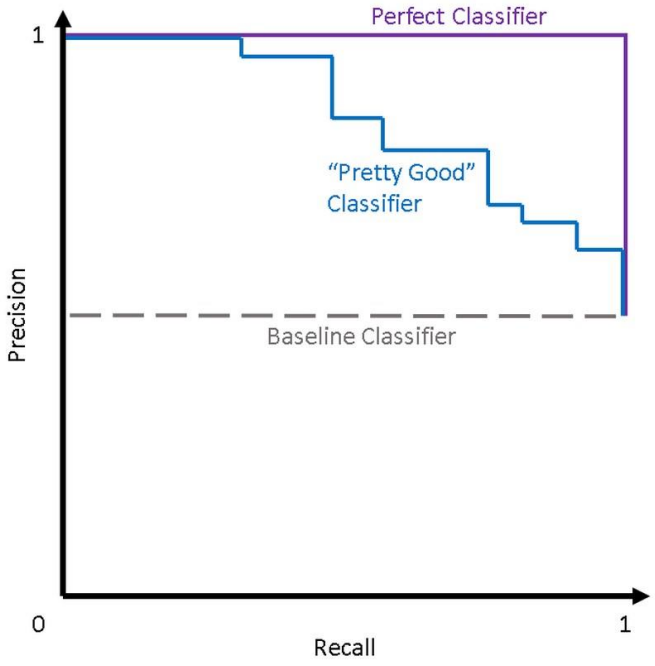


- ROC / AUC is the MOST POPULAR metric.

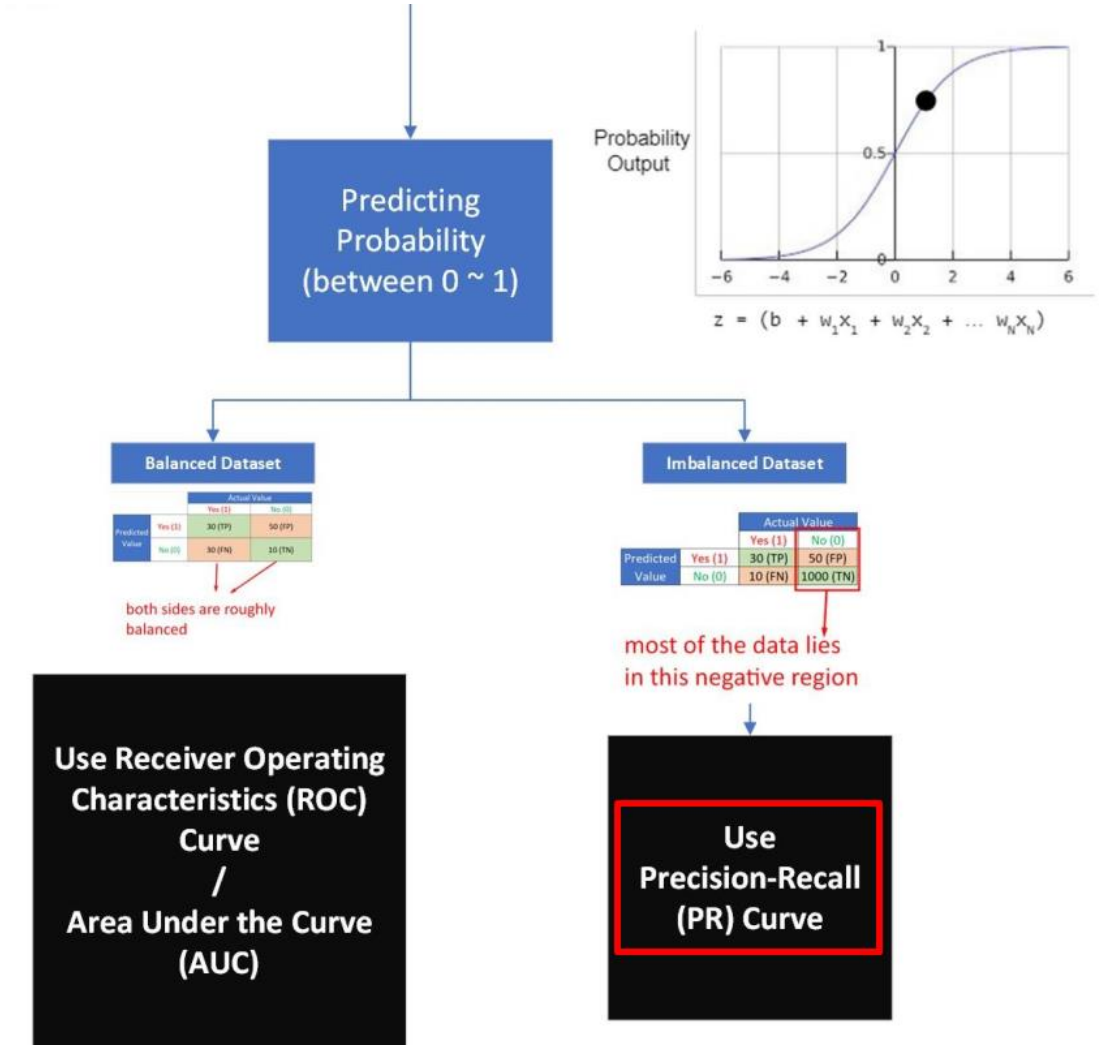


Source: www.machinelearningmastery.com

- A typical PR curve looks like this.
- Its simply the opposite of the ROC / AUC curve.



- The further away the AUC-PR curve from the Baseline Classifier, the better your model is.
- If AUC-PR = 1 → it's a Perfect Classifier → EXCELLENT.
- If AUC-PR = 0.5 → its Baseline Classifier → BAD!



Dr. Alvin Ang earned his Ph.D., Masters and Bachelor degrees from NTU, Singapore. He was an Assistant Professor, Data Scientist and Financial Consultant. Currently, he owns multiple self-started businesses and is a Trainer cum Business Advisor.

More about him at www.AlvinAng.sg