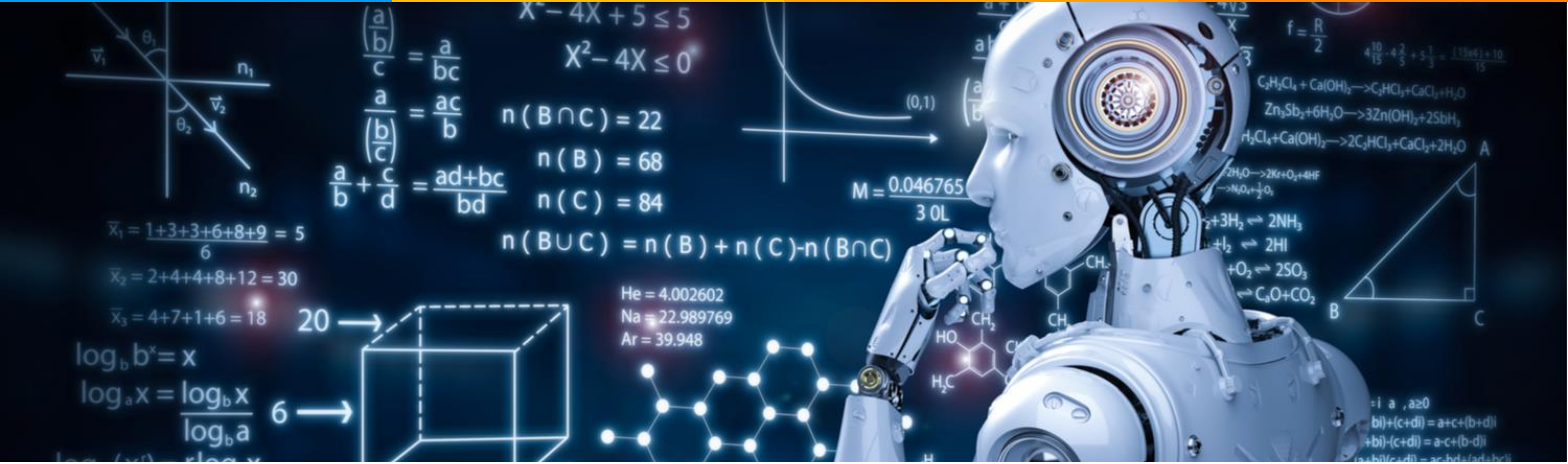
Duration: 60 min | Srikanth Shenoy | AML 5102 | M.E. AI & ML



Lecture 19: Evaluation Metrics

Recap

- Ensemble Learning
- Random Forest

Why evaluation metrics

Three type of binary classification models

- Categorization based on model output
- Models that output categorical class
  - K Nearest Neighbors, Decision Tree
- Models that output a real valued score
  - SVM
- Models that output a probability
  - Logistic Regression, Neural Networks
- Raw output (scores) across models cannot be compared

Reasons for having metrics

- Machine Learning task has a real world objective
- The ML algorithm + cost function is only a proxy for the real world objective
  - Different algorithms give different loss values
  - Comparing loss values across algorithms is meaningless
- Different distributions in data favor different algorithms
- Quantify gap between
  - Baseline model & a better model across algorithms
  - Desired performance and current performance

Eval metrics categories

Confusion Matrix

- Not a metric by itself
- Captures raw prediction type
- TP, TN, FP, FN
- Comes in many flavors
- Stick to one

	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

Confusion matrix for binary classification

Actual value	A	TP	FN
	B	FP	TN
		A	B
		Predicted value	

Predicted classes

Actual classes

	Negative 0	Positive 1
Negative 0	TN	FP
Positive 1	FN	TP

ACTUAL VALUES

PREDICTED VALUES

	POSITIVE	NEGATIVE
POSITIVE	TP	FP
NEGATIVE	FN	TN

Format used in this lecture

Predicted

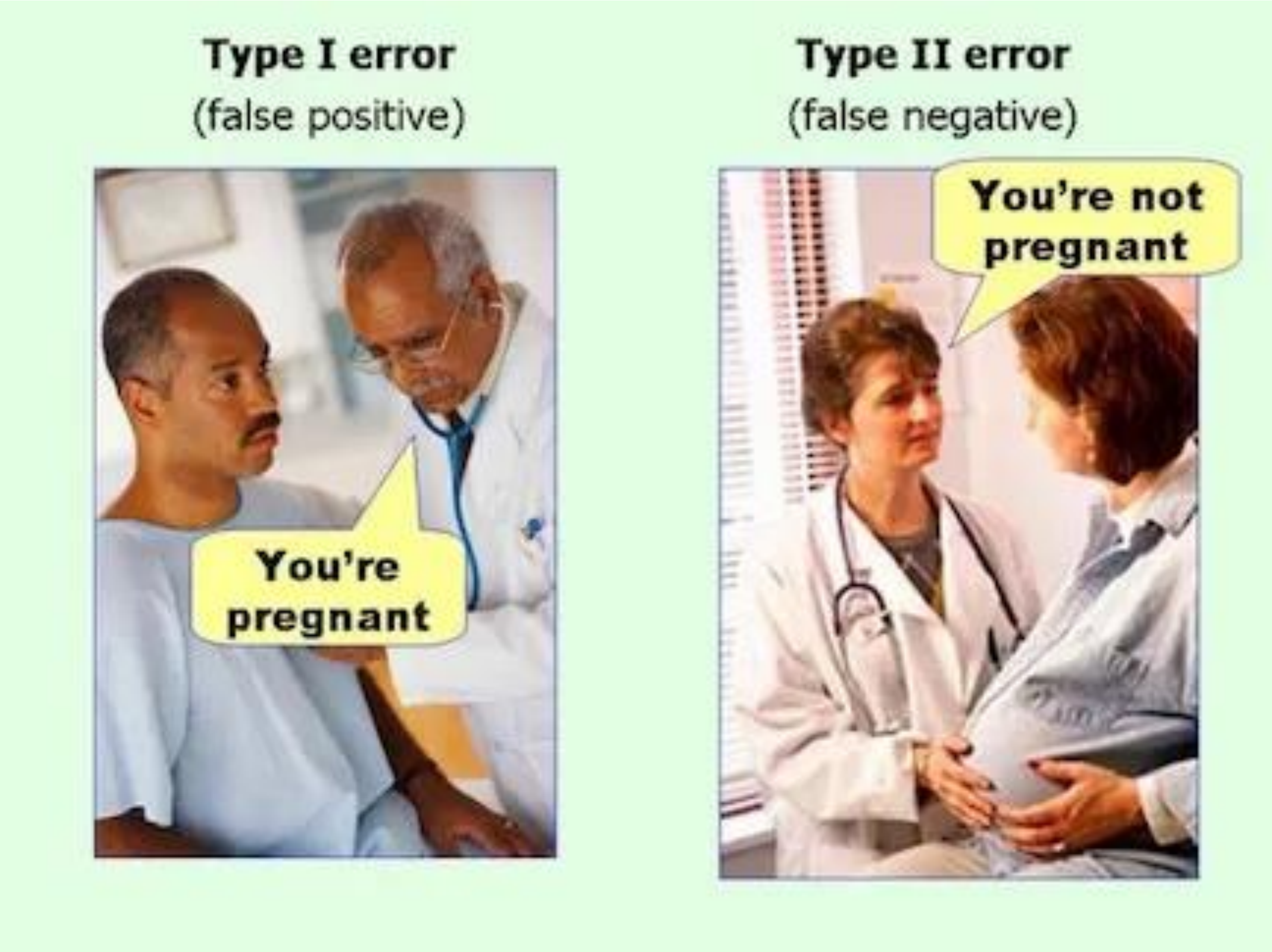
Actual

Positive/ Negative
1/0
1/-1

		Assigned class	
		Positive	Negative
Real class	Positive	TP	FN
	Negative	FP	TN

Confusion Matrix elements as joint probabilities

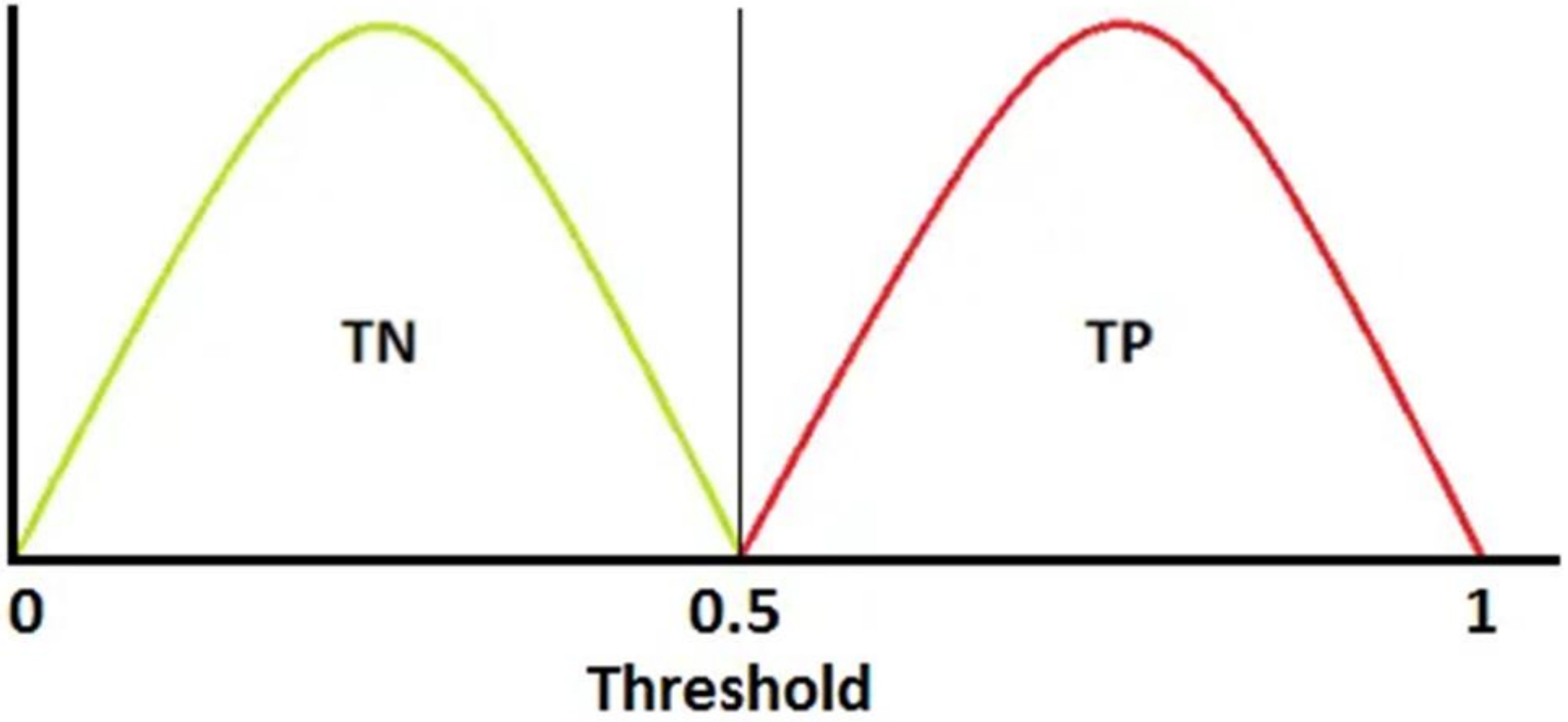
		Assigned class		
		Positive	Negative	
Real class	Positive	TP $P(\hat{y} = 1 \cap y = 1)$	FN $P(\hat{y} = 0 \cap y = 1)$	$P(y = 1)$
	Negative	FP $P(\hat{y} = 1 \cap y = 0)$	TN $P(\hat{y} = 0 \cap y = 0)$	$P(y = 0)$
		$P(\hat{y} = 1)$	$P(\hat{y} = 0)$	

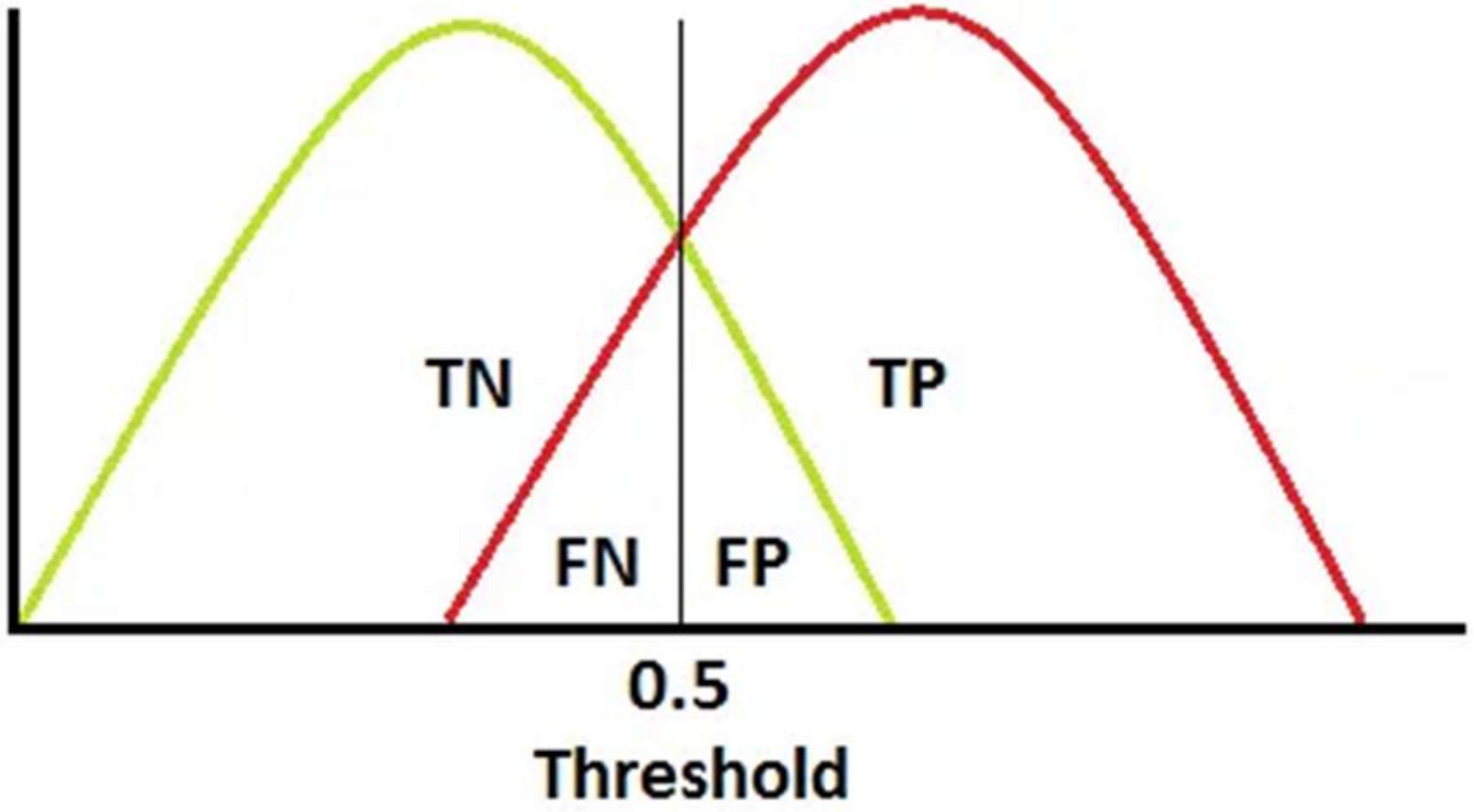

Type I error
(false positive)
You're pregnant
Type II error
(false negative)
You're not pregnant

Understanding Type I and II errors

		Assigned class	
		Positive	Negative
Real class	Positive	TP	FN
	Negative	FP	TN

For a given algorithm & hyperparams, FN + FP remains constant





Types of Metrics for Classification

- Point Metrics
  - Accuracy, Precision/Recall
- Composite Metrics
  - F-Score (F-1, F-Beta), Balanced Accuracy
- Summary Metrics
  - AU-ROC, AU-PRC

Point metrics

Accuracy, Precision, Recall

		Assigned class	
		Positive	Negative
Real class	Positive	TP	FN
	Negative	FP	TN

Not a good measure when +ve class is minority& its prediction is impt

Recall

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

Not a good measure for imbalanced datasets

Accuracy

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

Not a good measure when -ve class is minority& its prediction is impt

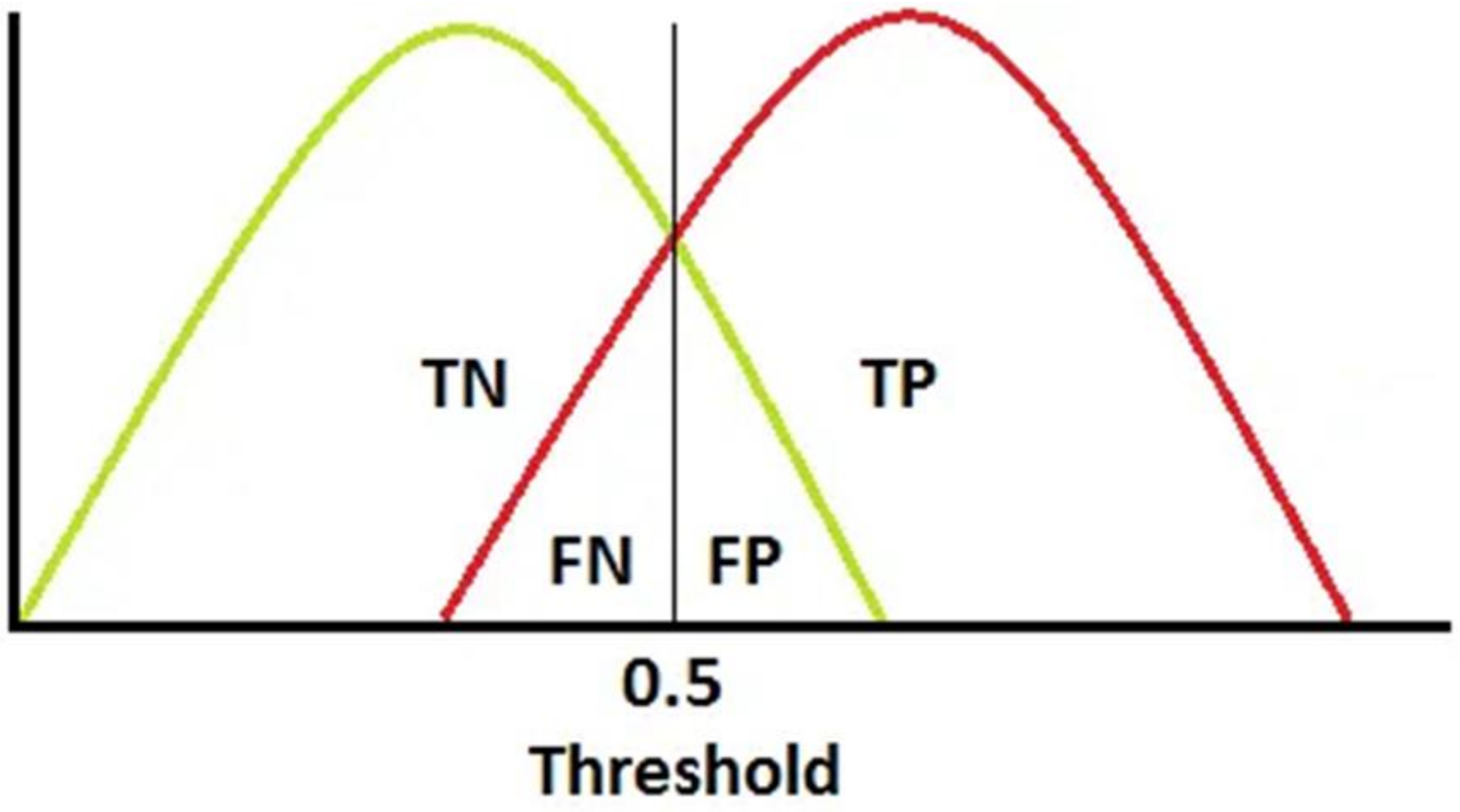
Precision

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

Precision

- TP/(TP+FP)
- P(actual=1 | predicted=1)
- Higher the Precision lesser the false positives
- Reducing FP leads to increase in FN
  - FP can be reduced by increasing threshold

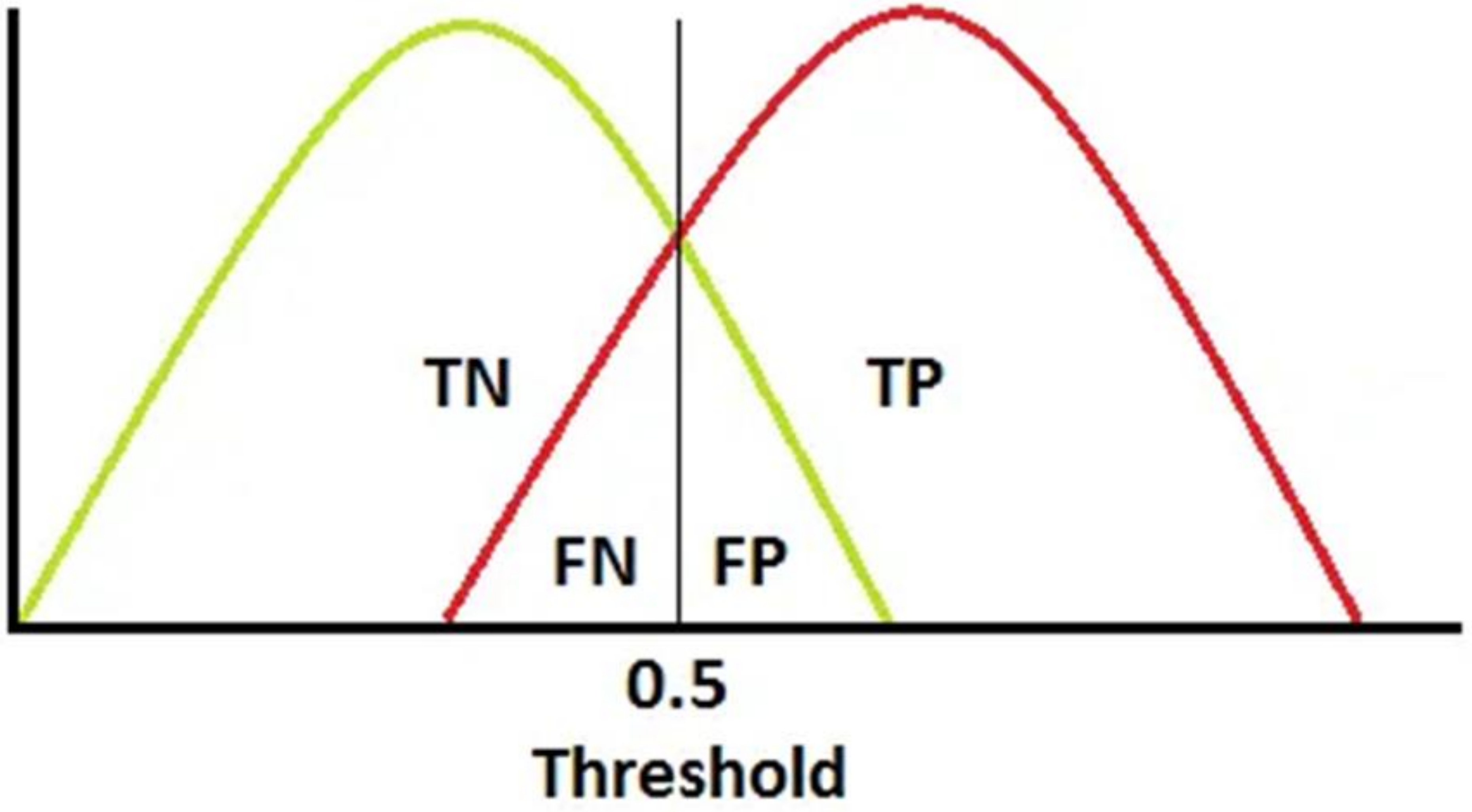
		Assigned class	
		Positive	Negative
Real class	Positive	TP	FN
	Negative	FP	TN



Recall / Sensitivity / True Positive Rate

- TP/(TP+FN)
- P(predicted=1 | actual=1)
- Higher the Recall lesser the false negatives
- Reducing FP leads to increase in FN
  - FN can be reduced by decreasing threshold
- Always Tug of war between Precision & Recall

		Assigned class	
		Positive	Negative
Real class	Positive	TP	FN
	Negative	FP	TN



Specificity

		Assigned class	
		Positive	Negative
Real class	Positive	TP	FN
	Negative	FP	TN

Recall

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

False positive rate

$$\frac{FP}{TN+FP}$$

Precision

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

Specificity

$$\frac{TN}{TN+FN}$$

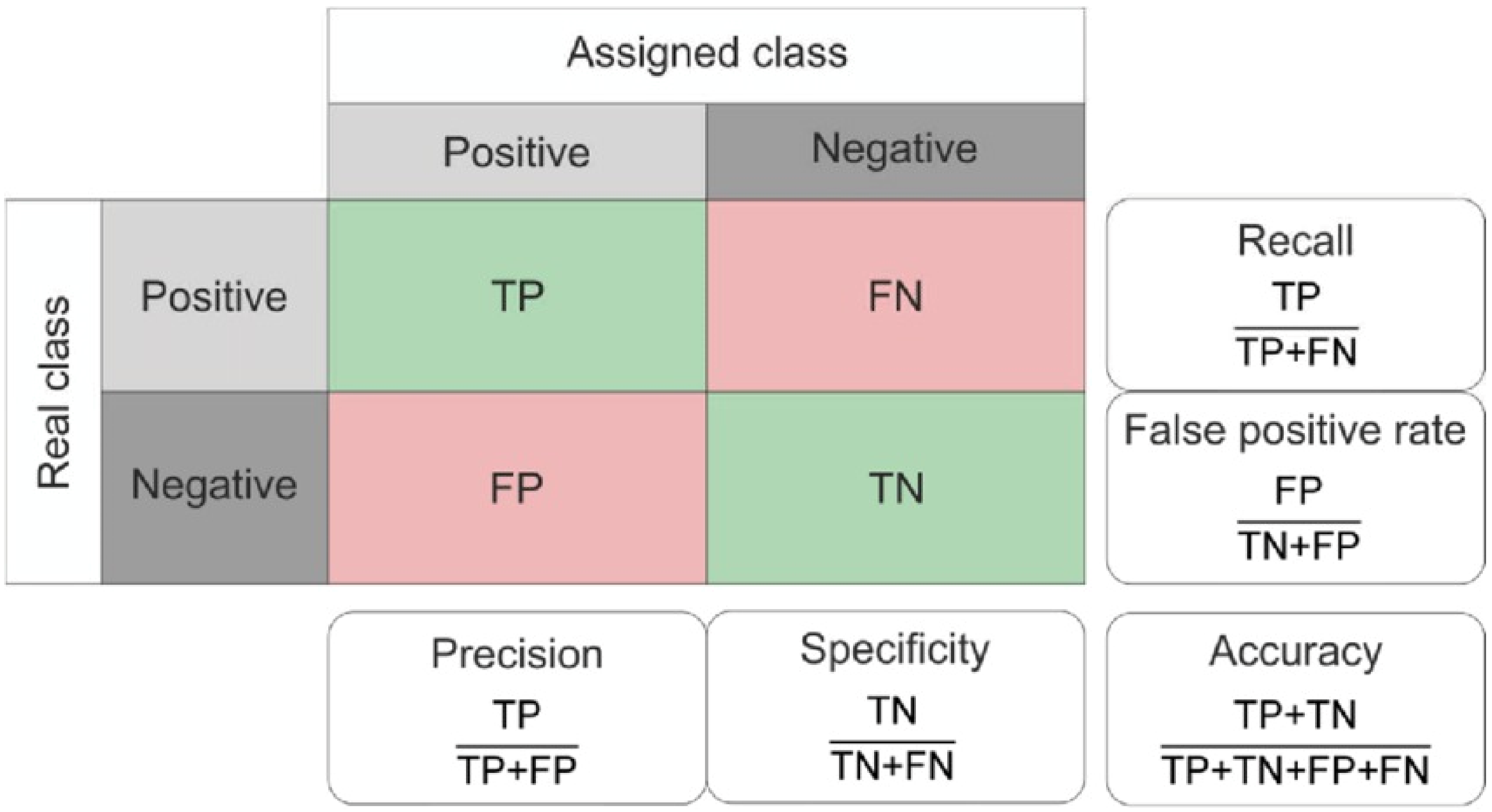
Accuracy

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

Composite metrics

Balanced Accuracy

- For binary classification: (Sensitivity + Specificity)/2
- For multiclass: Average of all recall
- For balanced datasets accuracy ~ balanced accuracy
- Imbalanced dataset
  - Accounts for imbalance



F-1 and F-Beta

- Harmonic mean

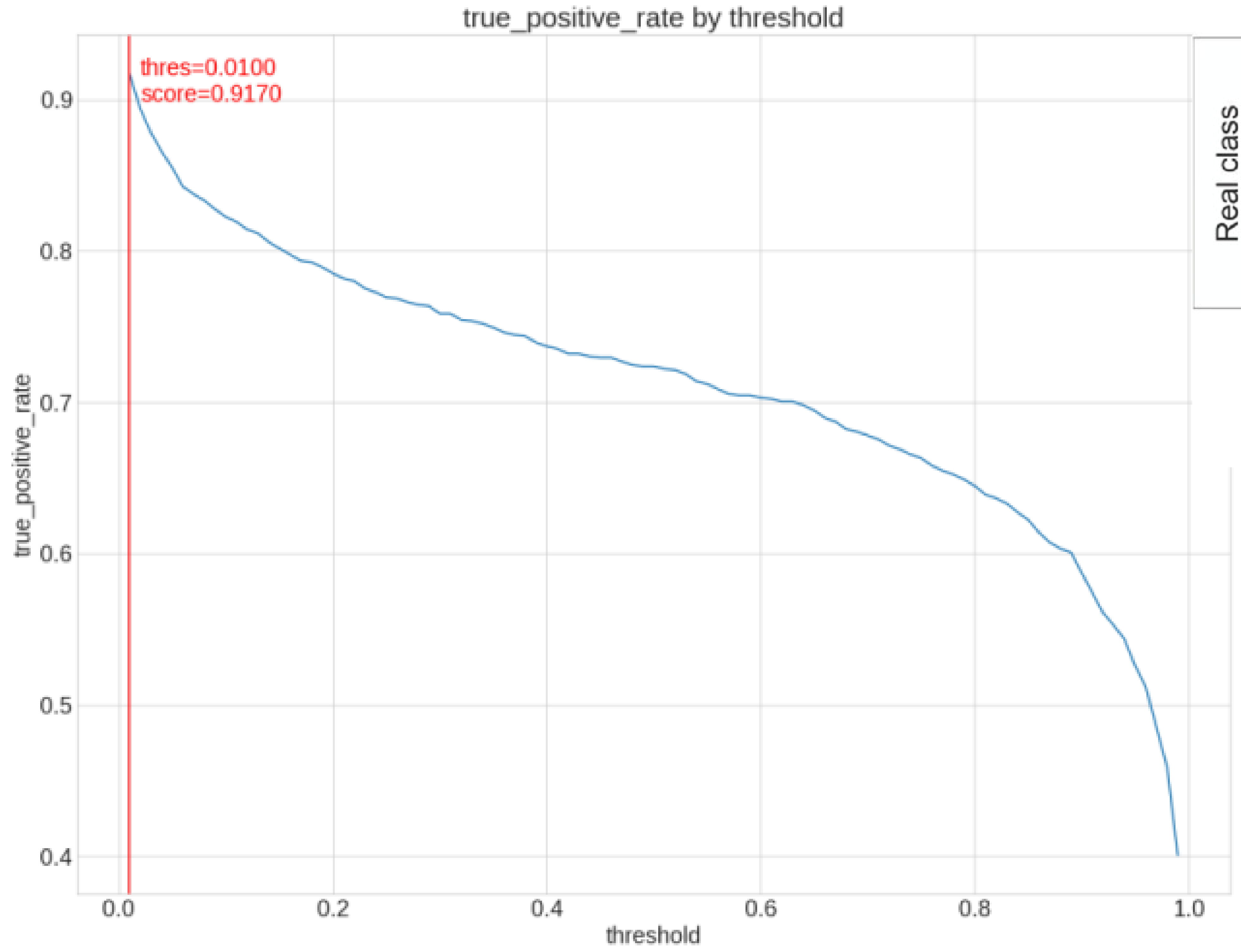
$$\frac{1}{\frac{\frac{1}{Precision}+\frac{1}{Recall}}{2}} \qquad \frac{2}{\frac{1}{Precision}+\frac{1}{Recall}}$$

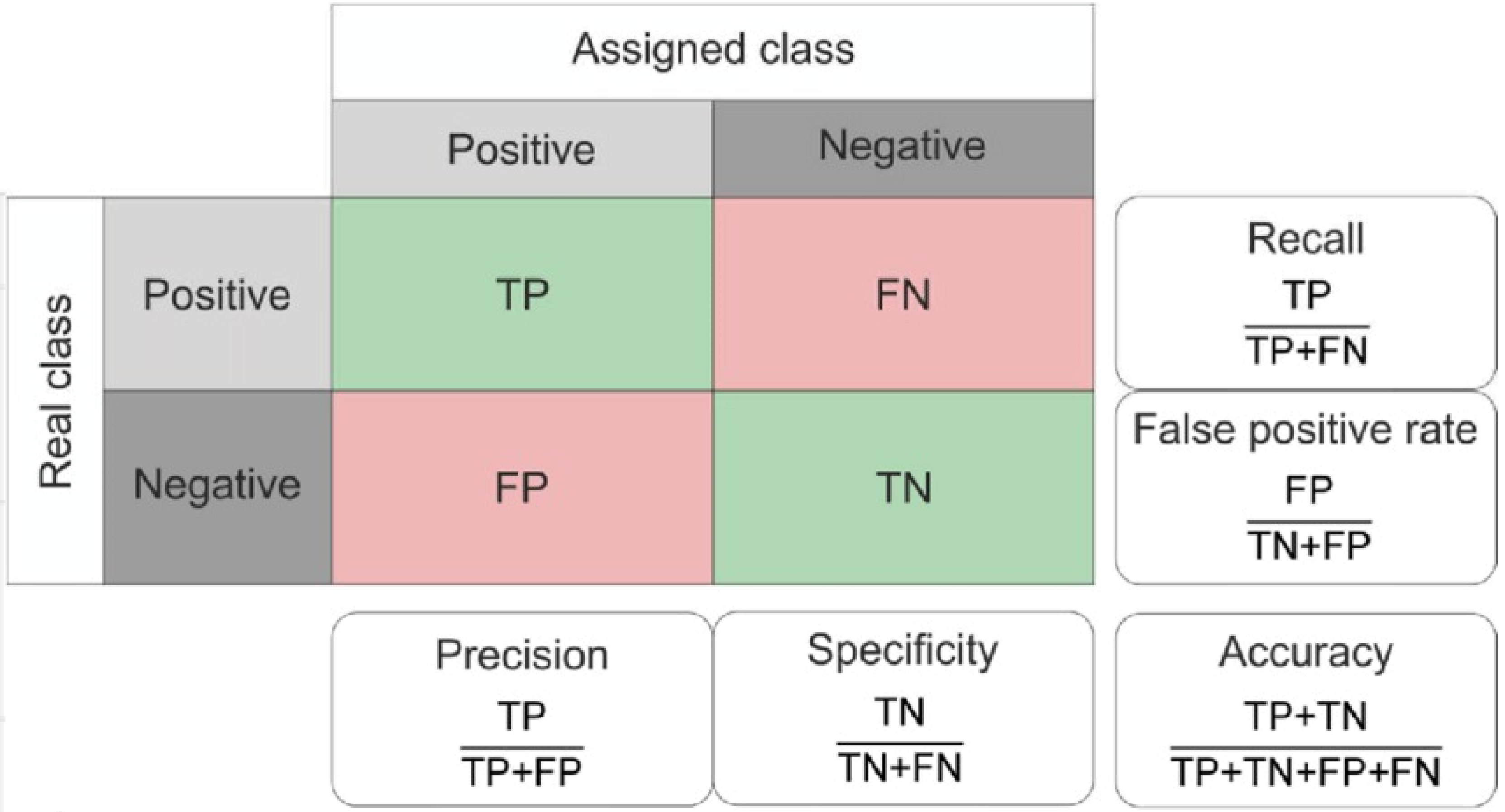
$$2 * \frac{precision \times recall}{precision + recall} \qquad = (1+\beta^2) * \frac{precision \times recall}{\beta^2 * precision + recall}$$

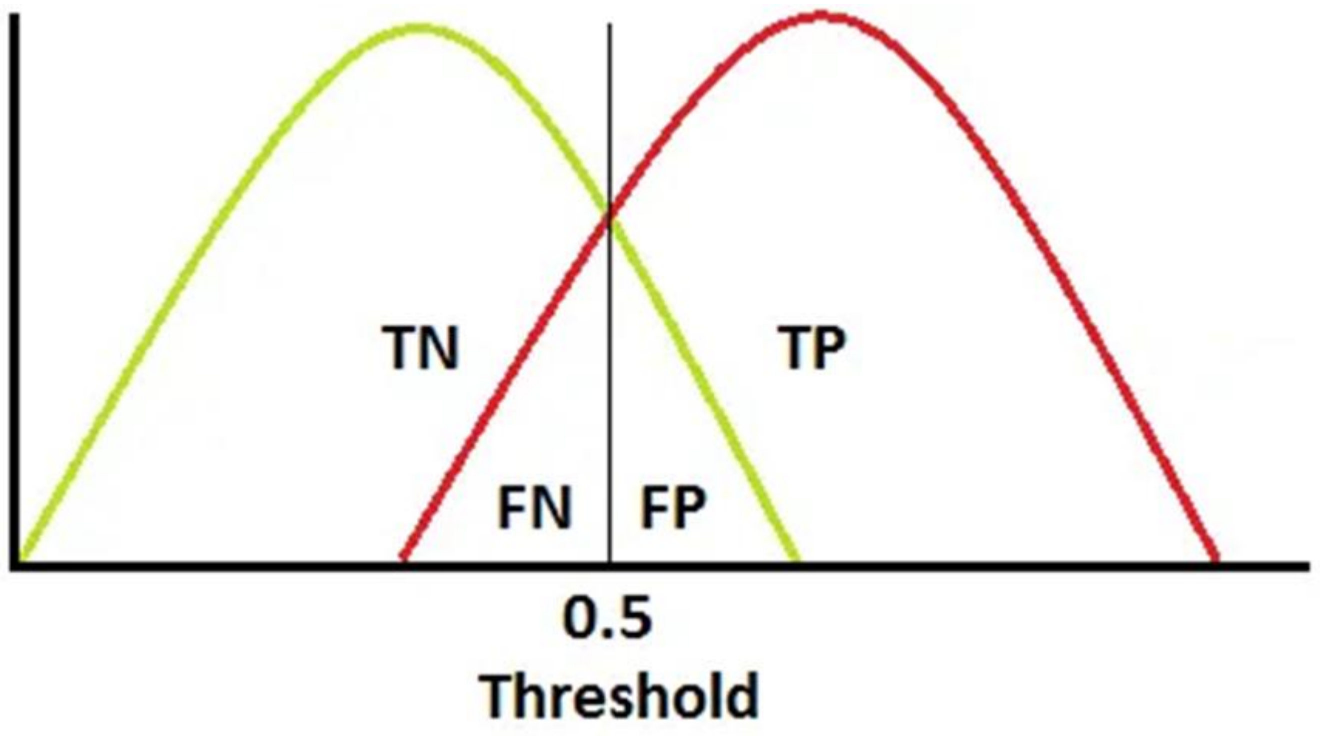
- Why harmonic mean?
  - Penalizes extreme values of either Precision or Recall
- Beta = 1 F-1
- Beta < 1 favors Precision (i.e. ok to have False Negative)
- Beta > 1 favors Recall (i.e. ok to have False Positive)

Summary metrics

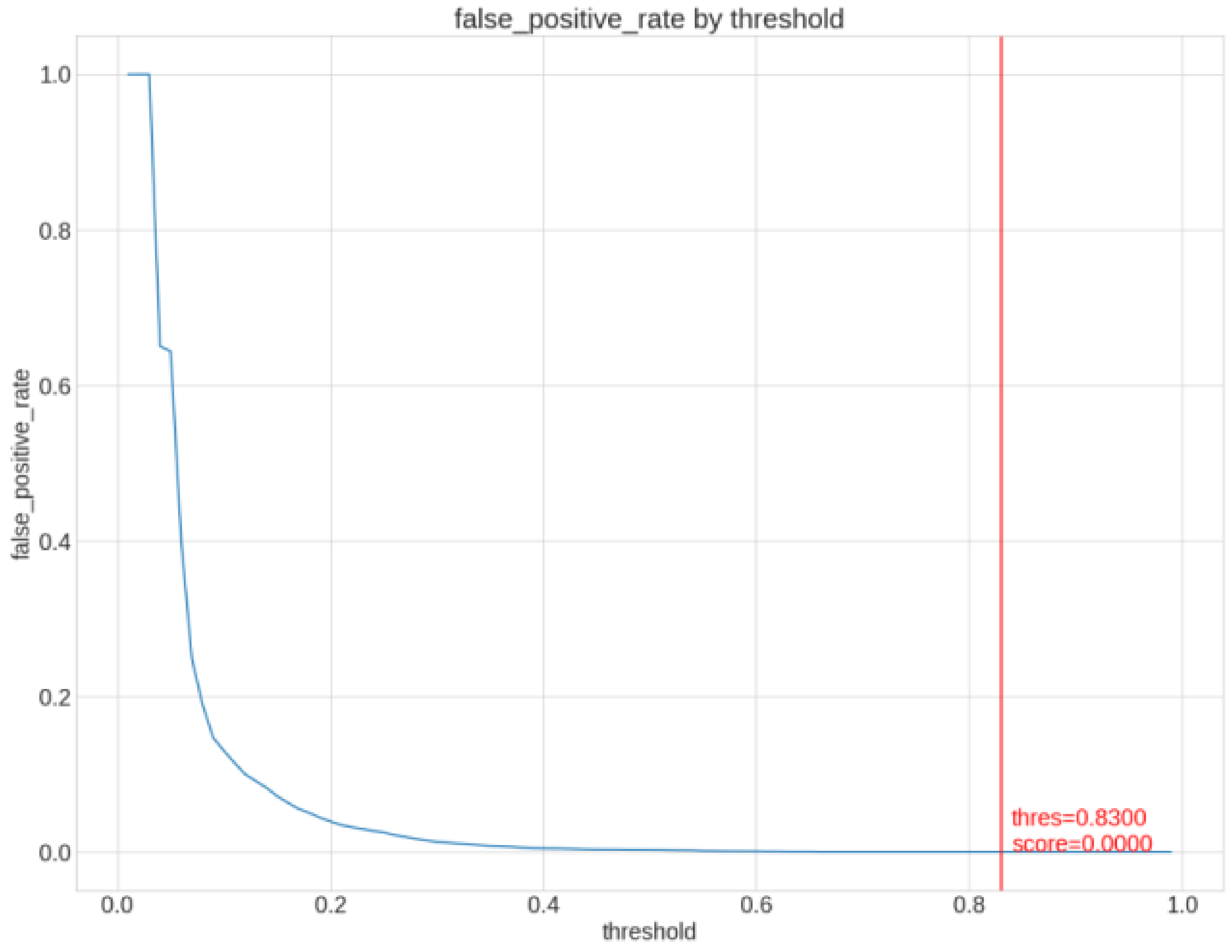
True positive rate versus threshold

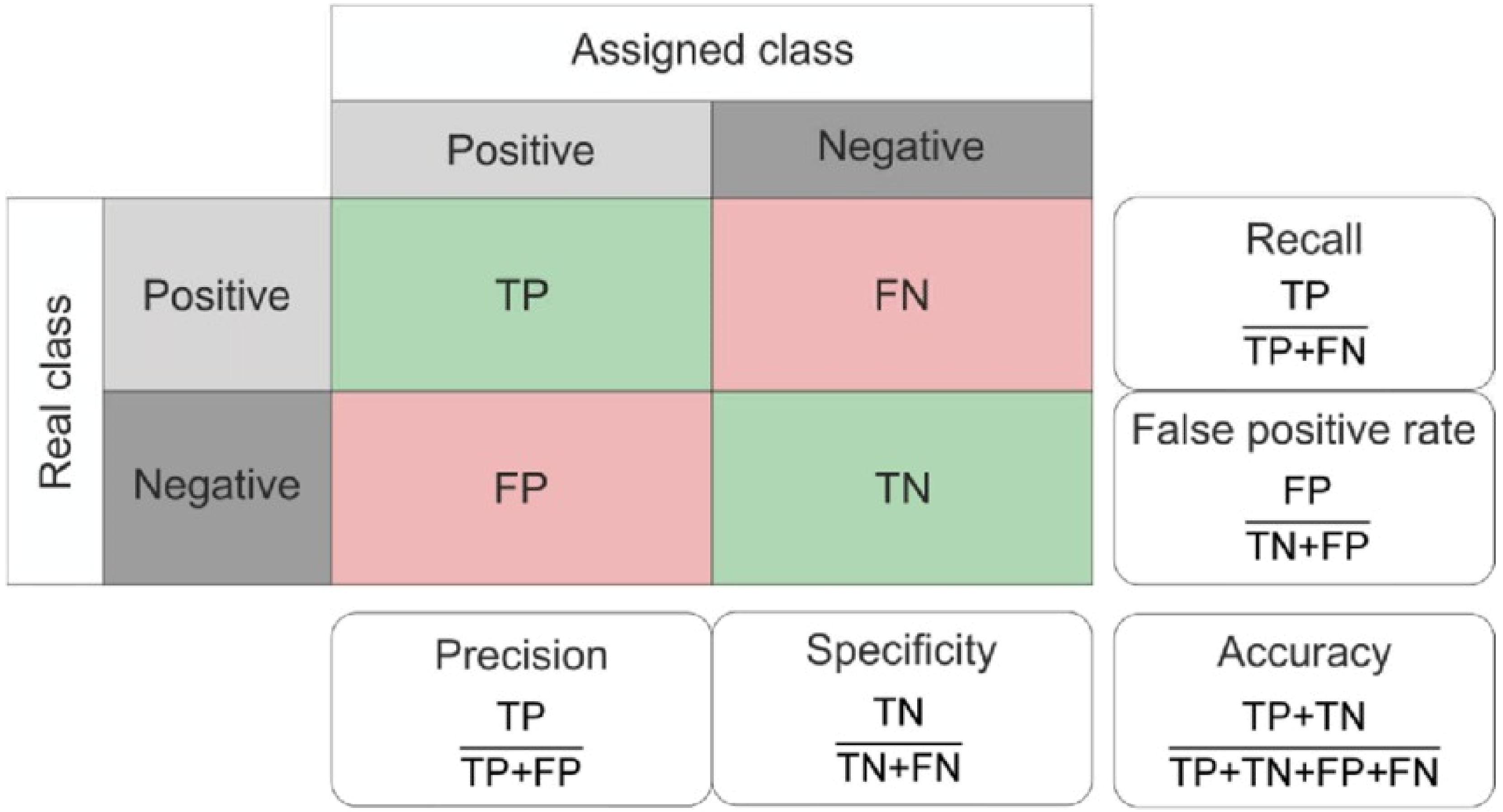


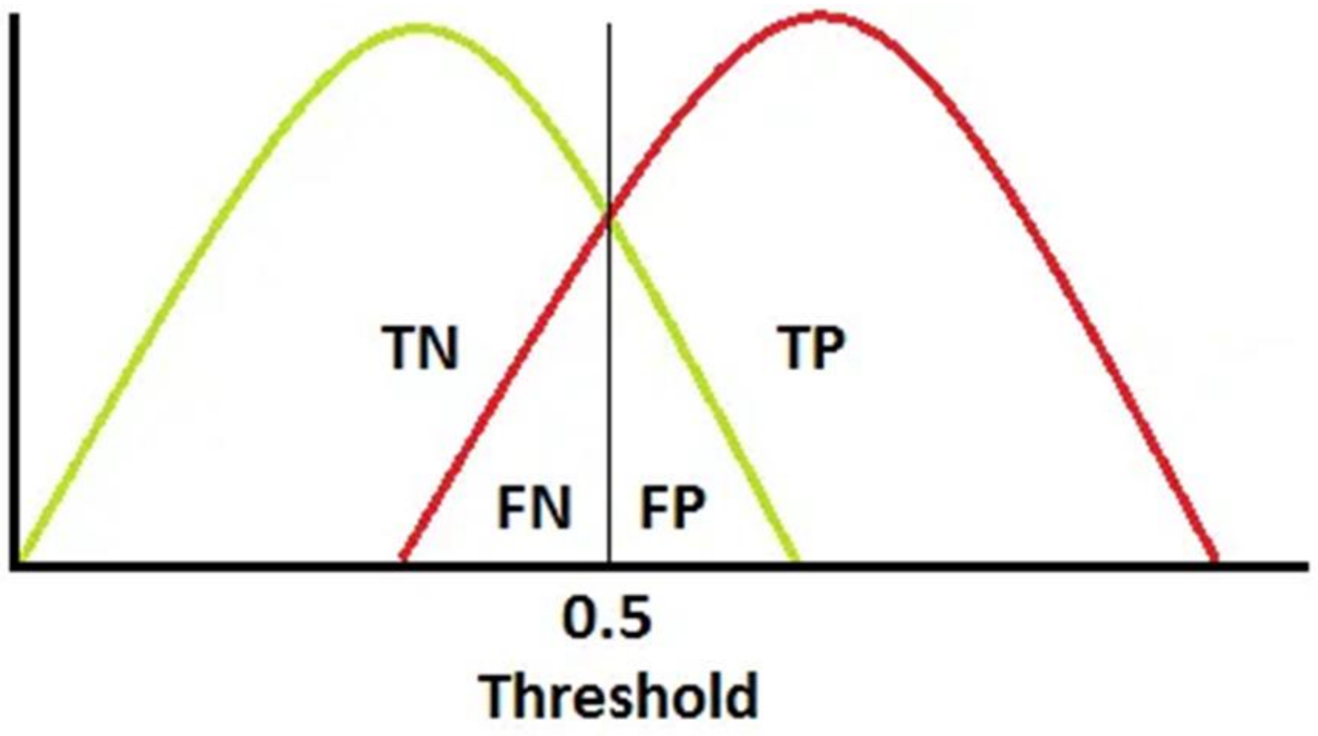




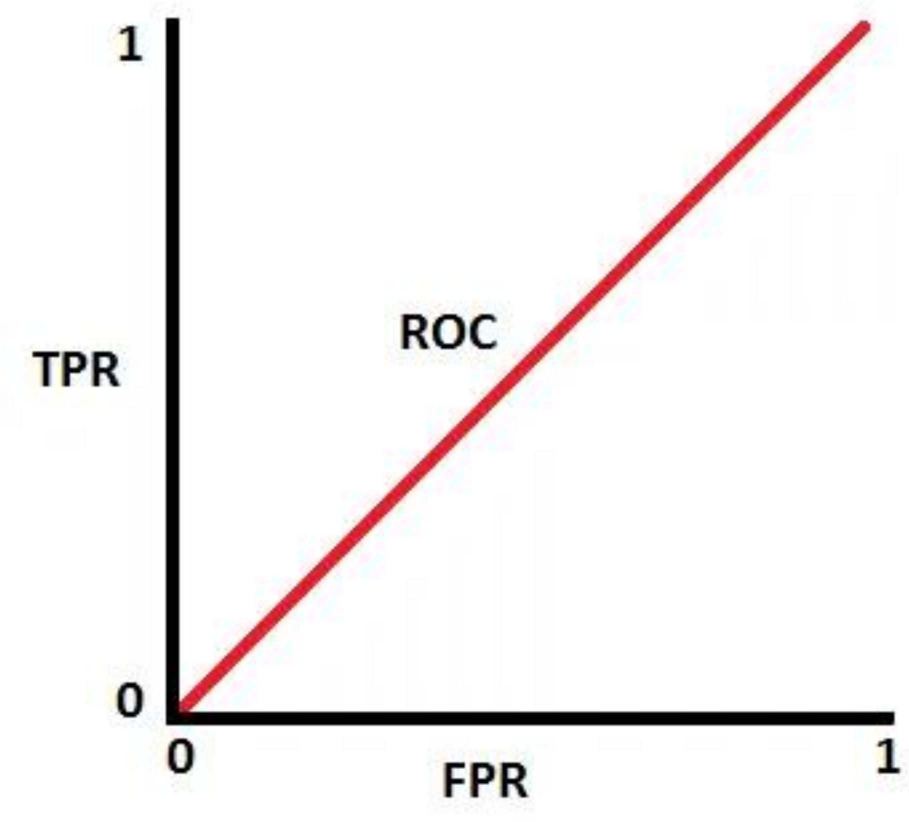
False positive rate versus threshold

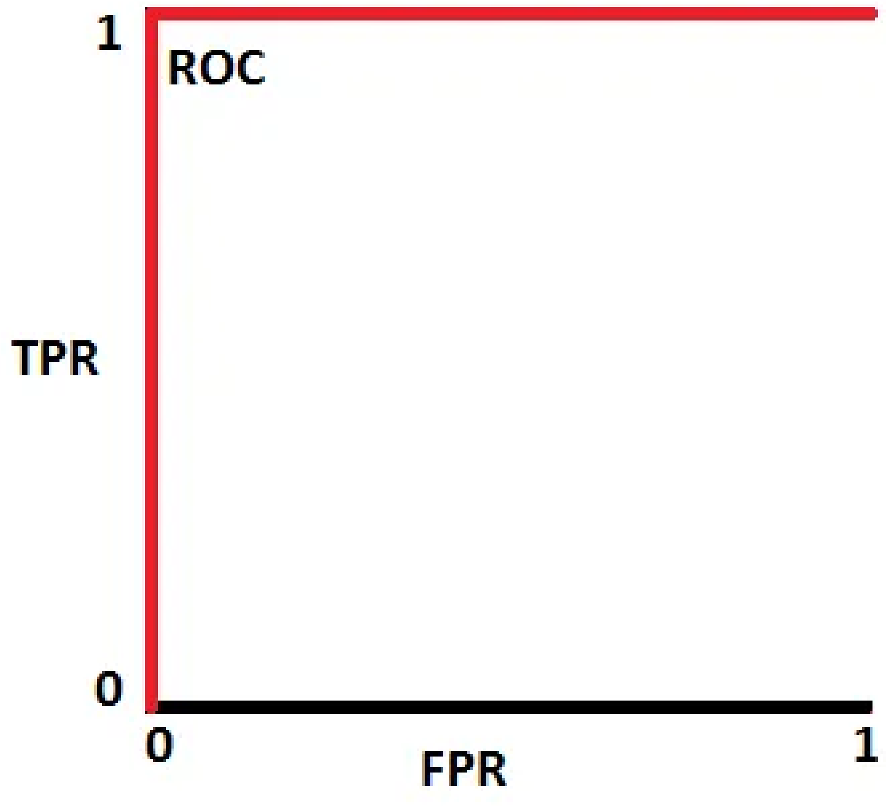


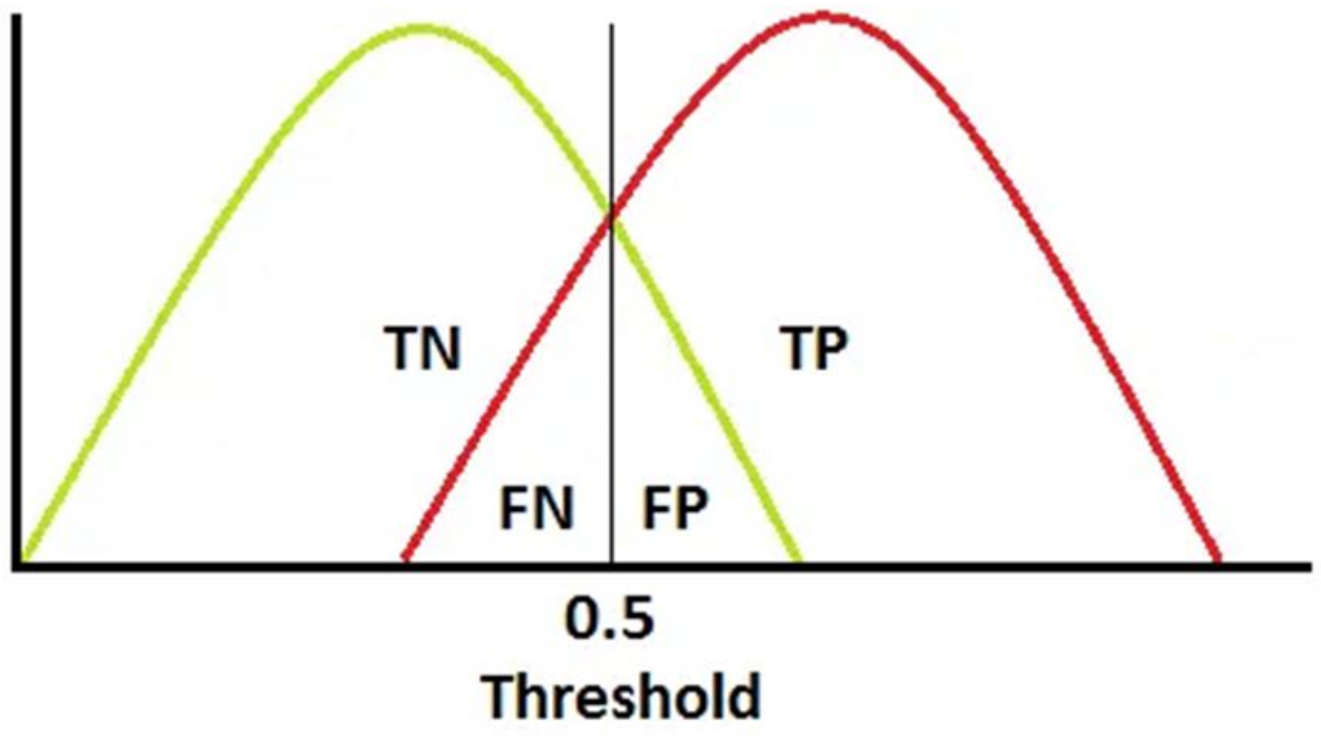


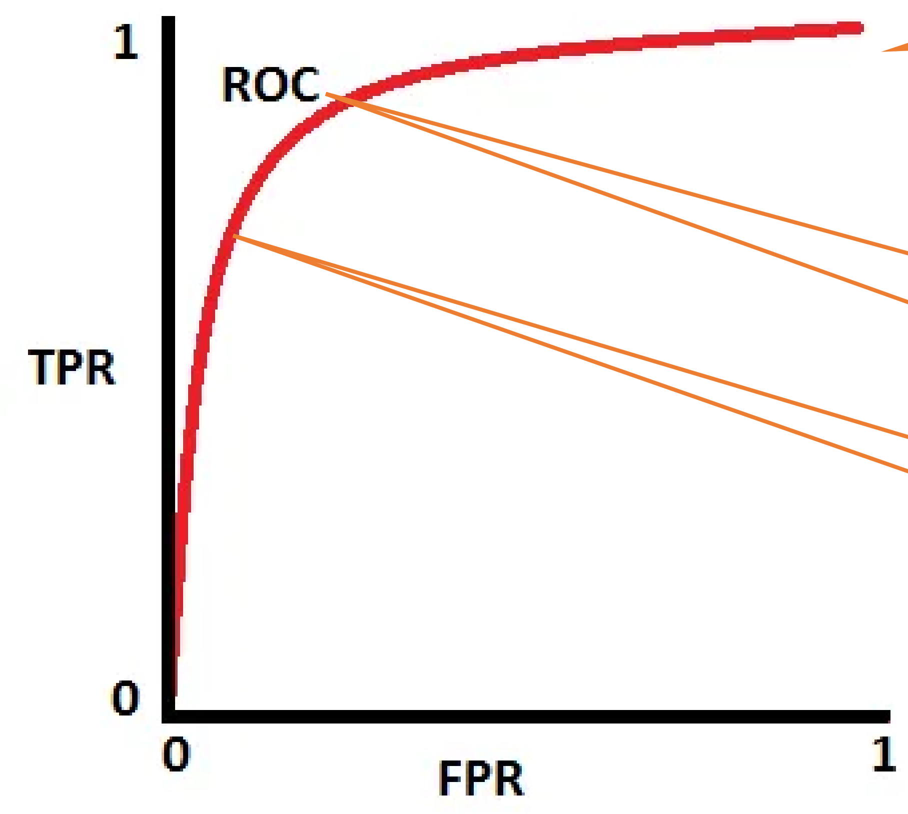


AU ROC





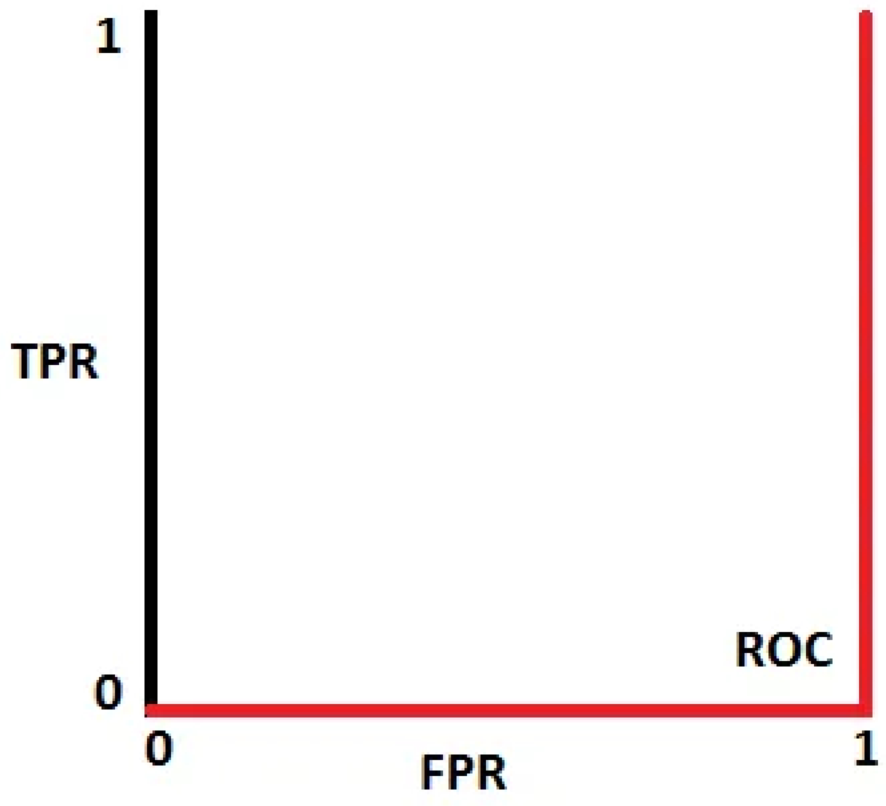




When data is balanced

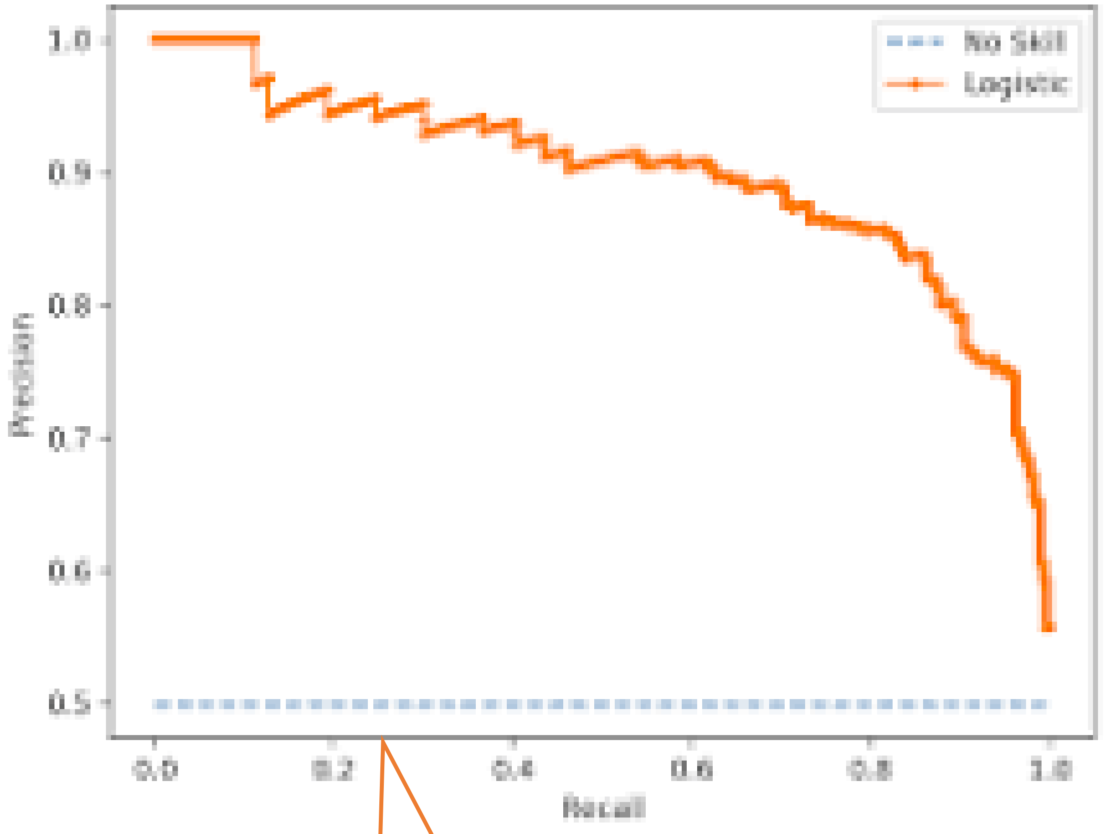
Measure of robustness

Left is better, Up is better

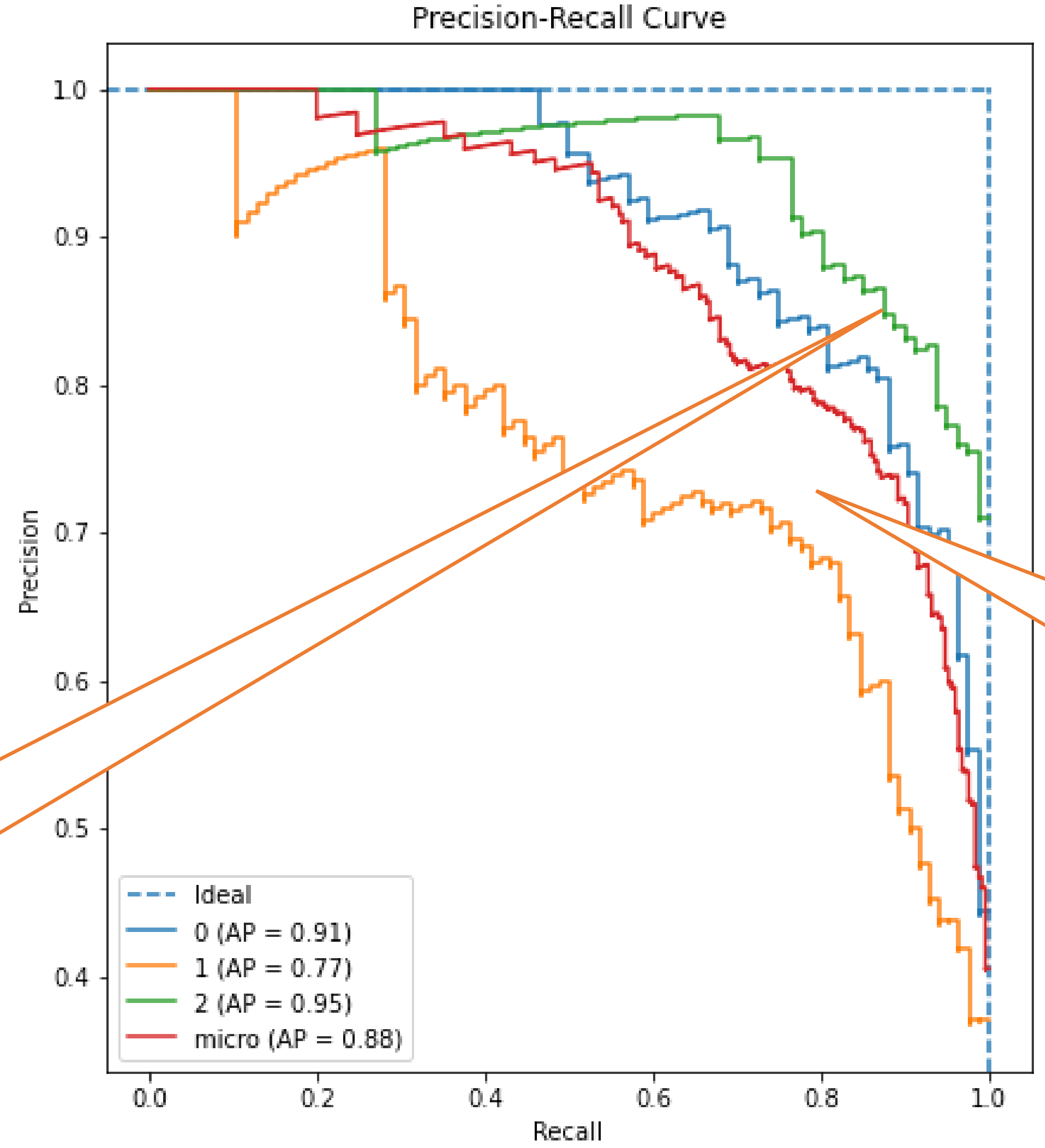


AU PRC

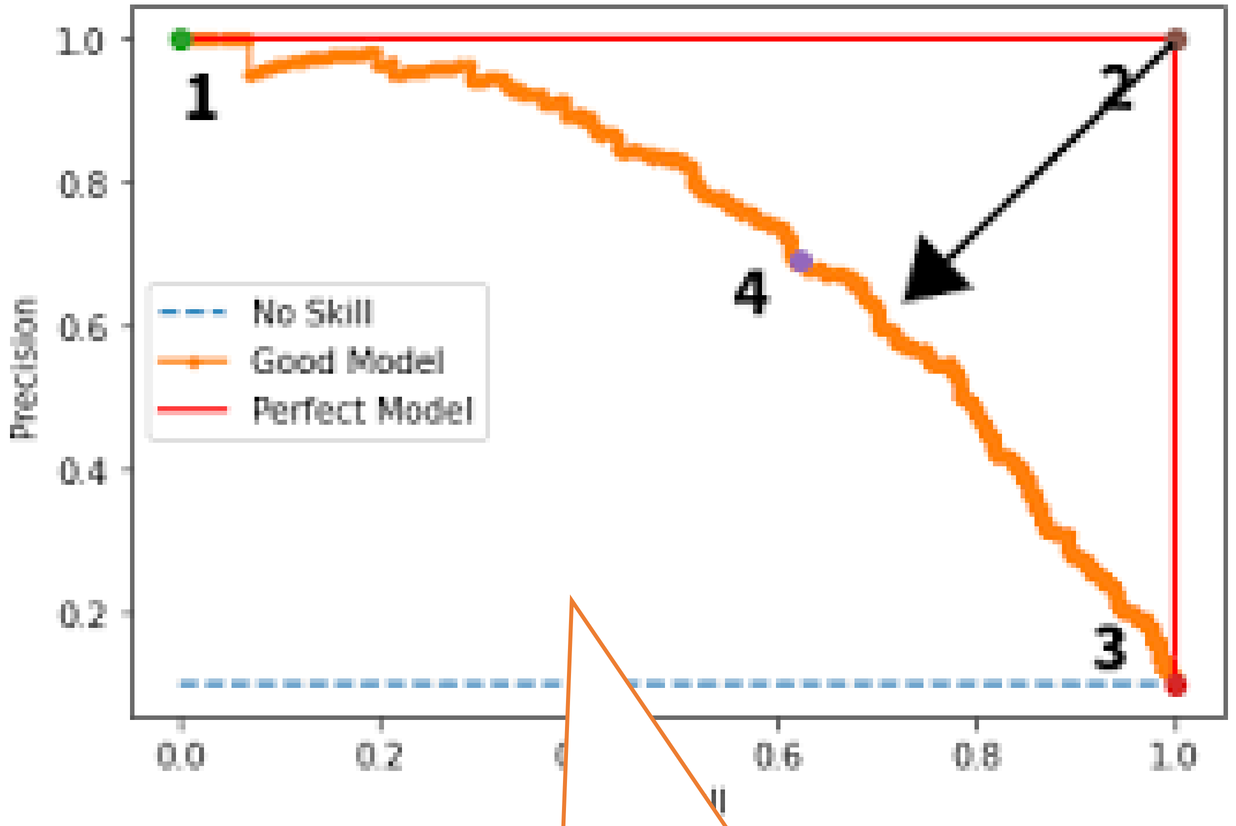
Use when data is imbalanced



Not very good



Right is better,
Up is better



But better than this

Use for comparing multiple models robustness over a range of thresholds

Multi class evaluation metrics

Multi class confusion matrix

		Predicted		
		Airplane	Boat	Car
Actual	Airplane	2	1	0
	Boat	0	1	0
	Car	1	2	3

Classification report

	precision	recall	f1-score	support
Aeroplane	0.67	0.67	0.67	3
Boat	0.25	1.00	0.40	1
Car	1.00	0.50	0.67	6
accuracy			0.60	10
macro avg	0.64	0.72	0.58	10
weighted avg	0.82	0.60	0.64	10

Macro average

Label	Per-Class F1 Score	Macro-Averaged F1 Score
Airplane	0.67	$\frac{0.67 + 0.40 + 0.67}{3}$ **= 0.58**
Boat	0.40	
Car	0.67	

Weighted average

<table>
<tr><th>Label</th><th>Per-Class F1 Score</th><th>Support</th><th>Support Proportion</th><th>Weighted Average F1 Score</th></tr>
<tr><td>Airplane</td><td>0.67</td><td>3</td><td>0.3</td><td rowspan="4">$(0.67 * 0.3) +$
$(0.40 * 0.1) +$
$(0.67 * 0.6)$
= 0.64</td></tr>
<tr><td>Boat</td><td>0.40</td><td>1</td><td>0.1</td></tr>
<tr><td>Car</td><td>0.67</td><td>6</td><td>0.6</td></tr>
<tr><td>Total</td><td>-</td><td>10</td><td>1.0</td></tr>
</table>


QUESTIONS