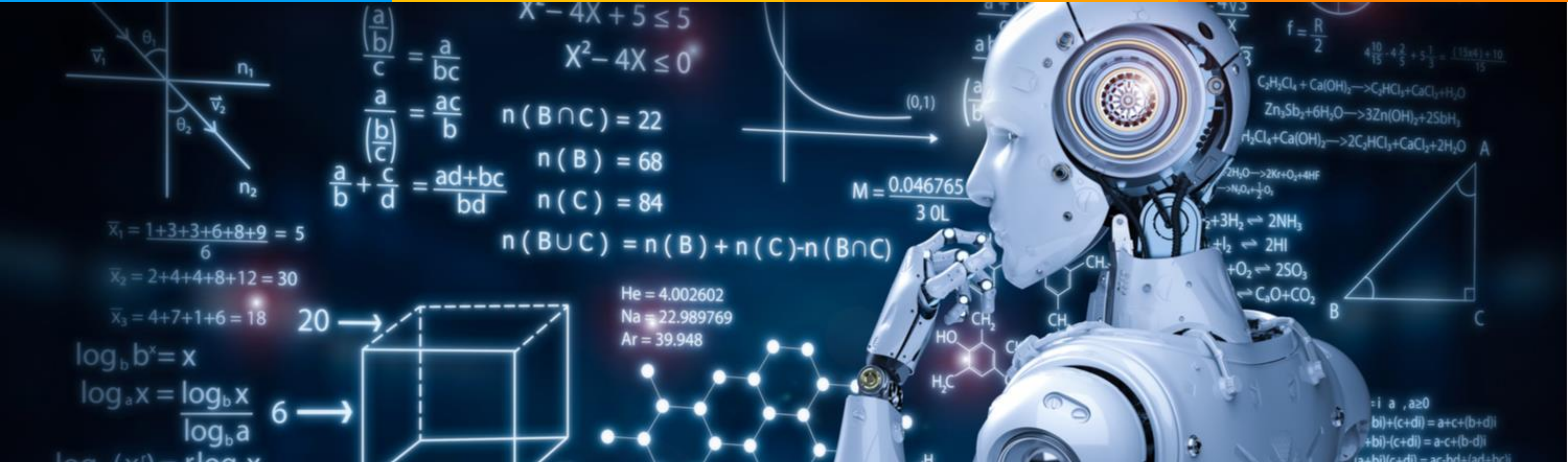
Duration: 60 min | Srikanth Shenoy | AML 5102 | M.E. AI & ML



Lecture 20: Evaluation Metrics

Recap

- Ensemble Learning
- Bagging
- Random Forest
- Boosting

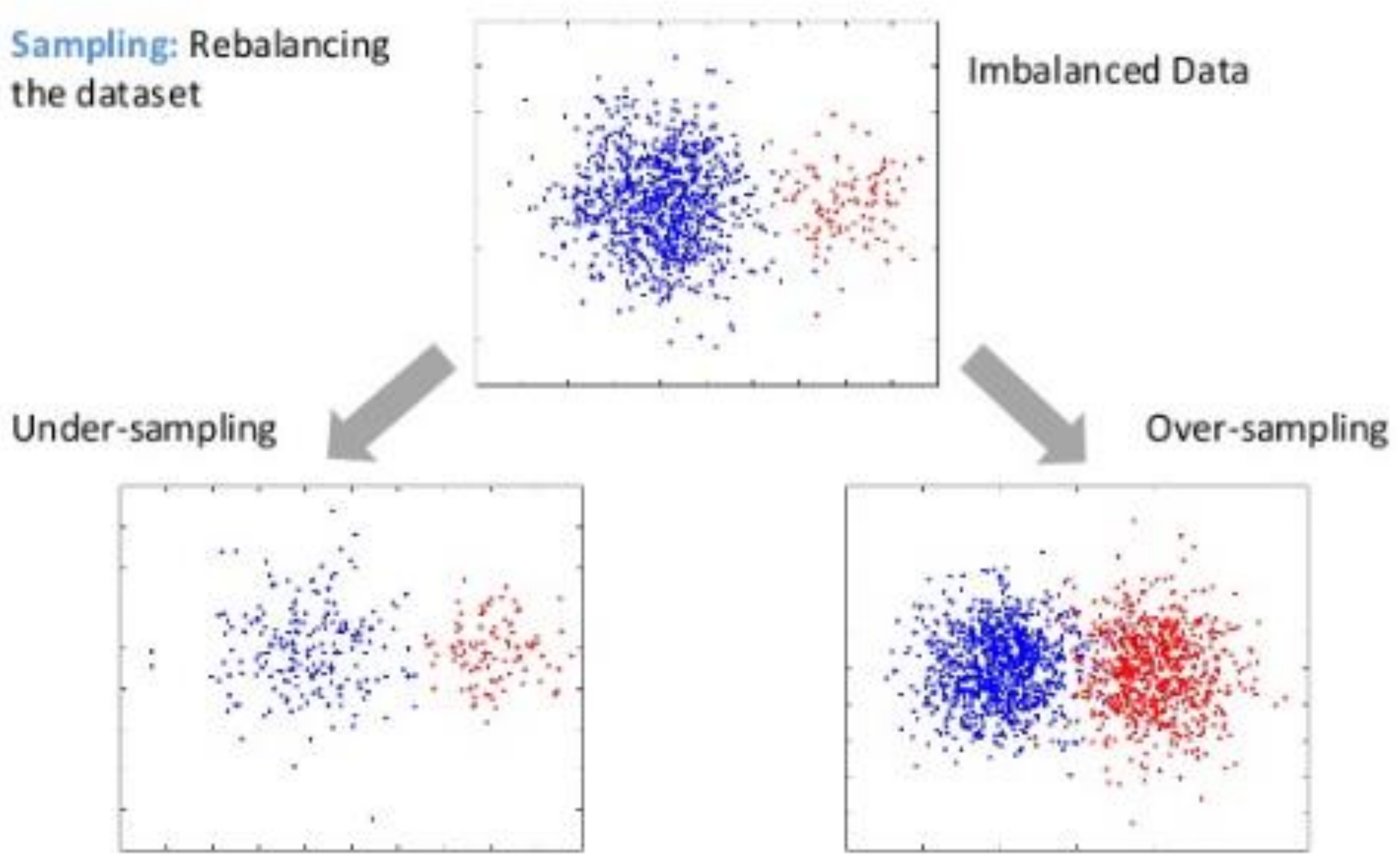
Imbalanced data

Accuracy metric and Imbalanced dataset

- A dataset has 98% majority class and 2% minority class
- A model was developed and it gave accuracy of 0.9499
- Is it a good model?

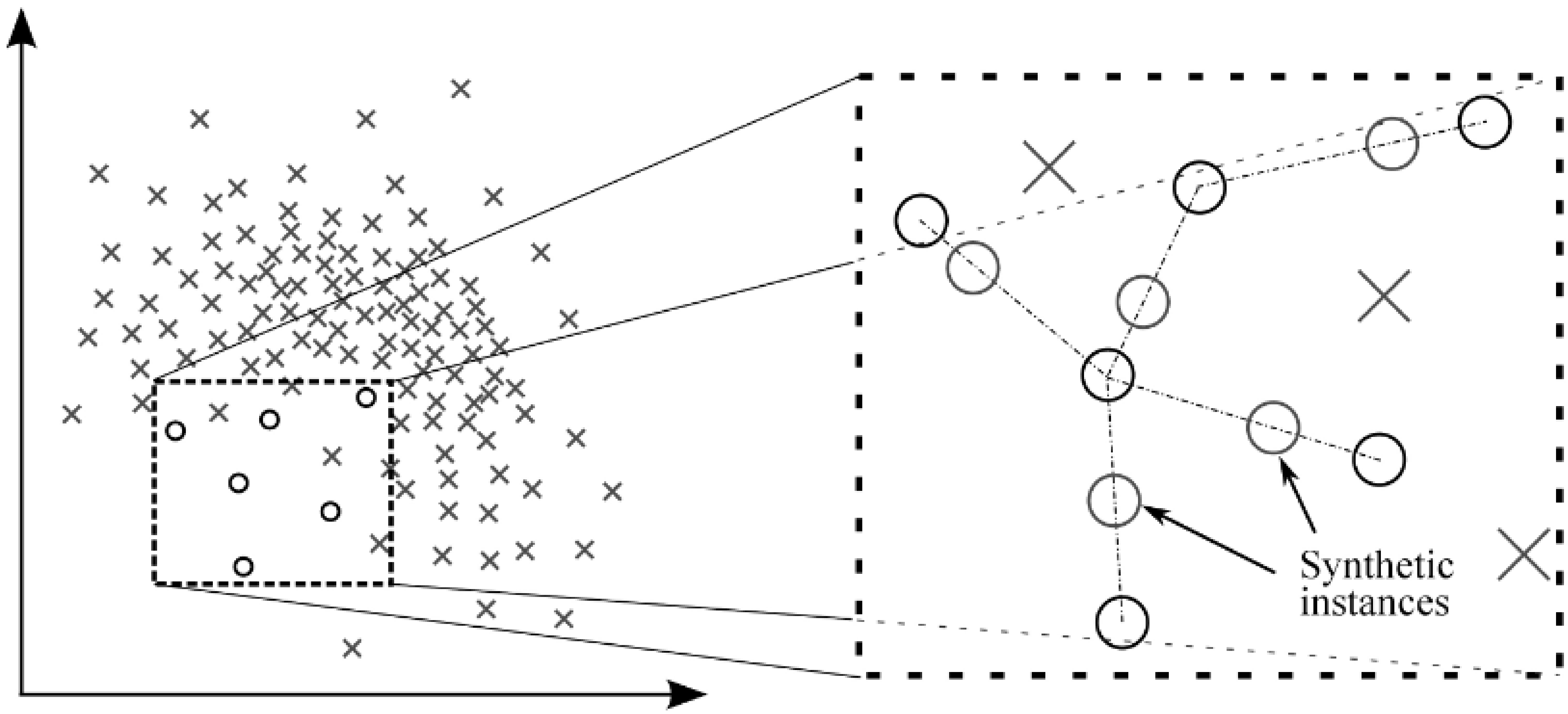
Oversampling minority class

- Synthetic Minority Oversampling Technique: SMOTE



SMOTE (Continued)

- Uses KNN on minority class

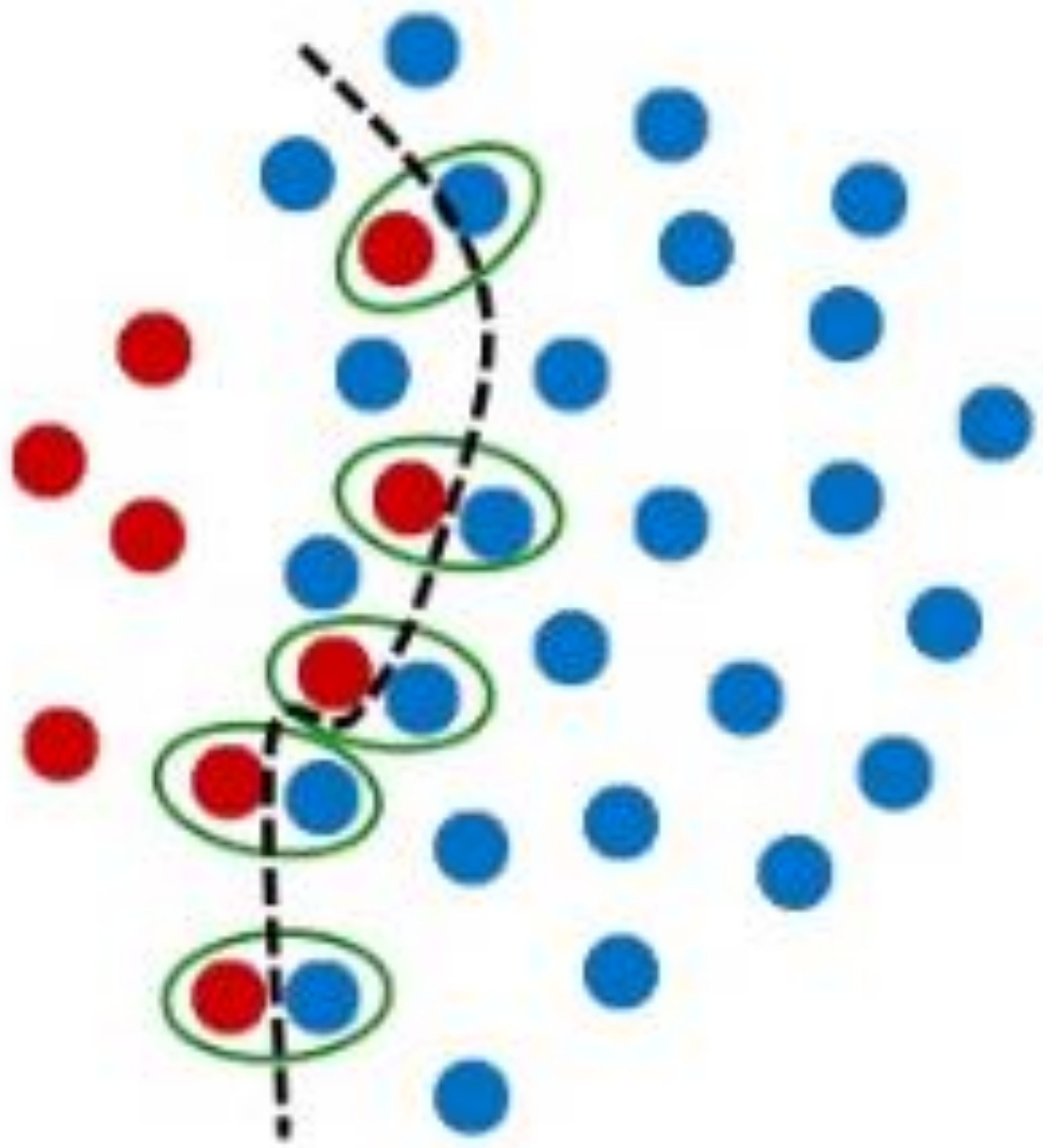


SMOTE (Continued)

- kNN SMOTE
- DBSMOTE: Uses DBSCAN clustering algorithm for SMOTE

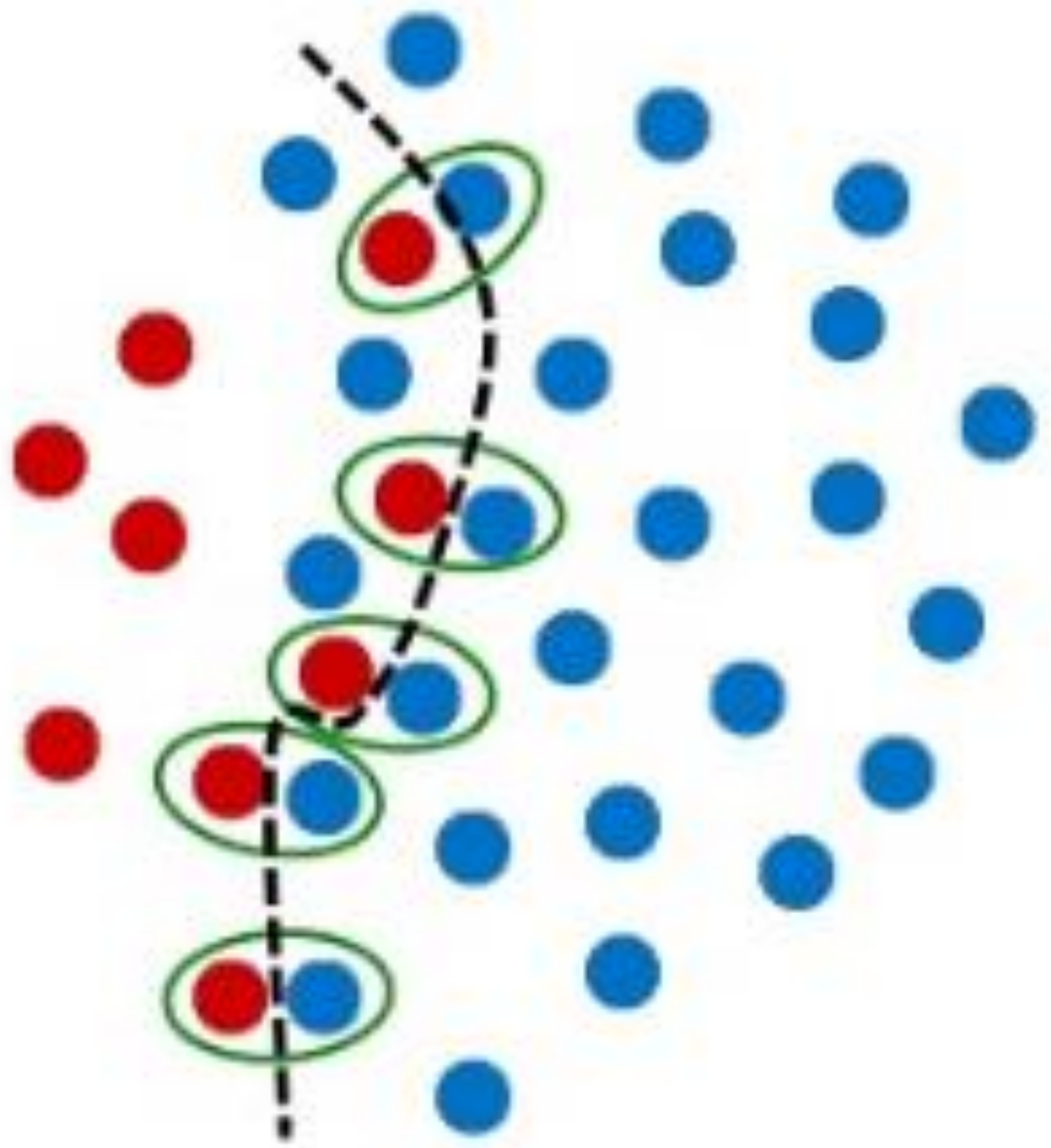
Undersampling majority class

- Random Undersampling
- Tomek Links



Combined oversampling and undersampling

- SMOTE Tomek



Handling Imbalanced dataset in Random Forest

- Class weight
  - Reciprocal of proportion of records per class
  - Balanced – default value
  - Balanced sub sample – each tree gets sub samples based on class weight

- Refer to lab Jupyter notebook

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

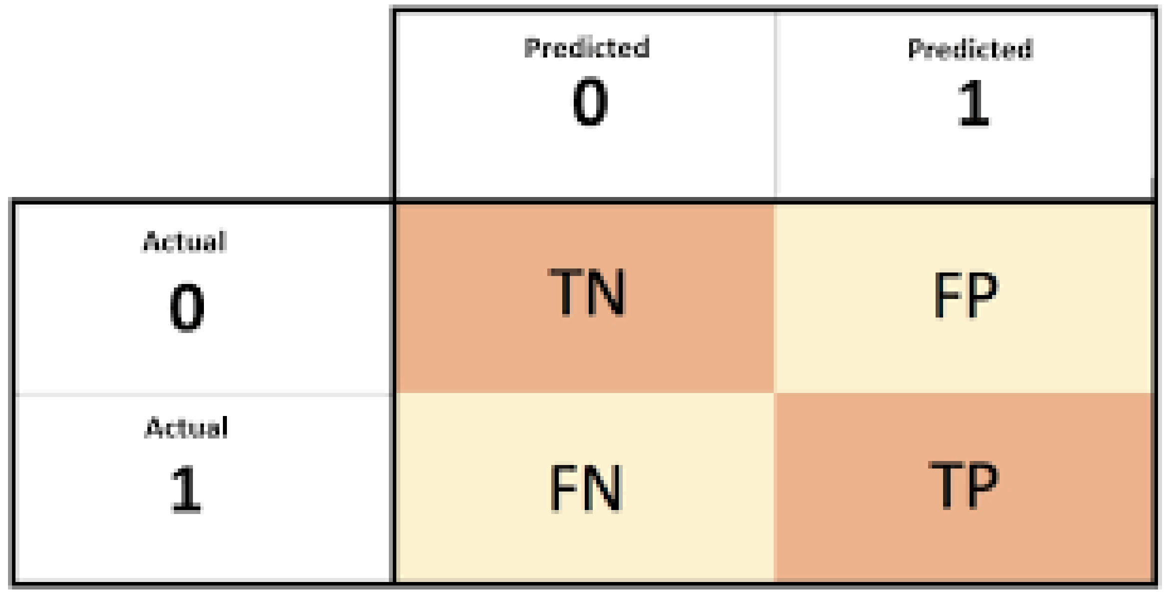
Reasons for having metrics

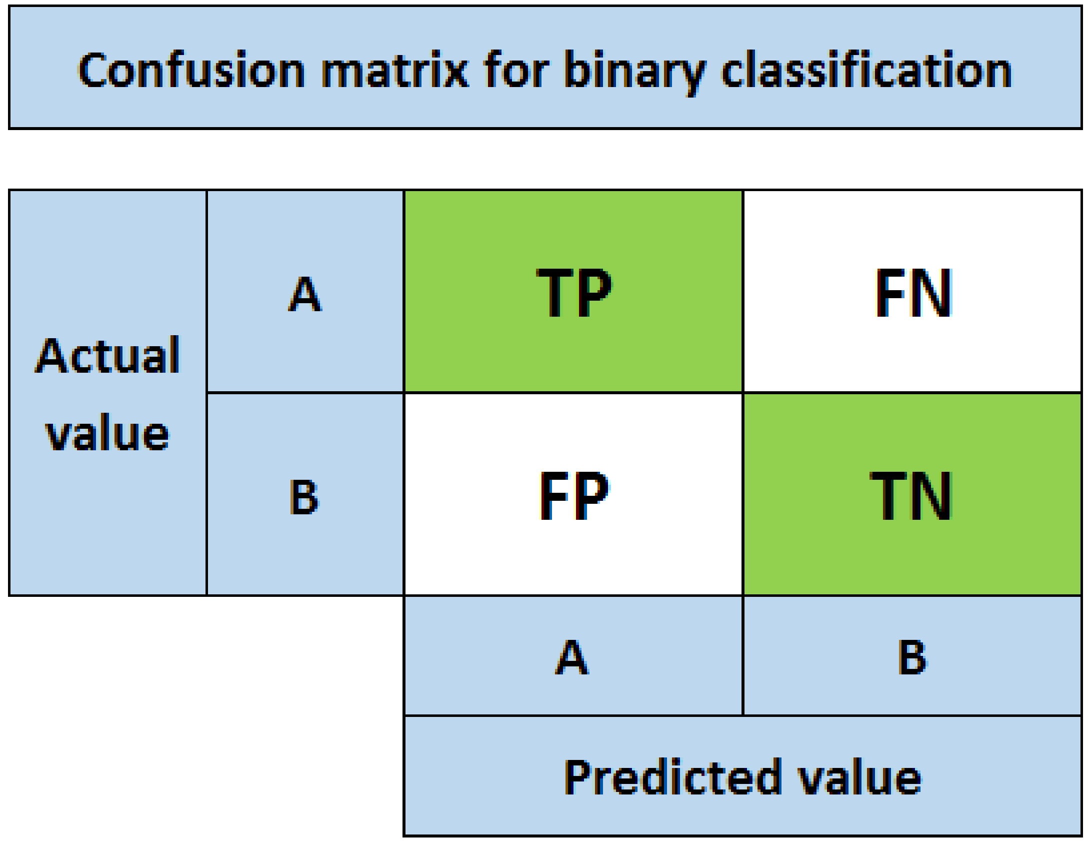
- Machine Learning task has a real world objective
- The ML algorithm + cost function is only a proxy for the real world objective
  - Different distributions in data favor different algorithms
  - Different algorithms give different loss values
  - Comparing loss values across algorithms is meaningless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

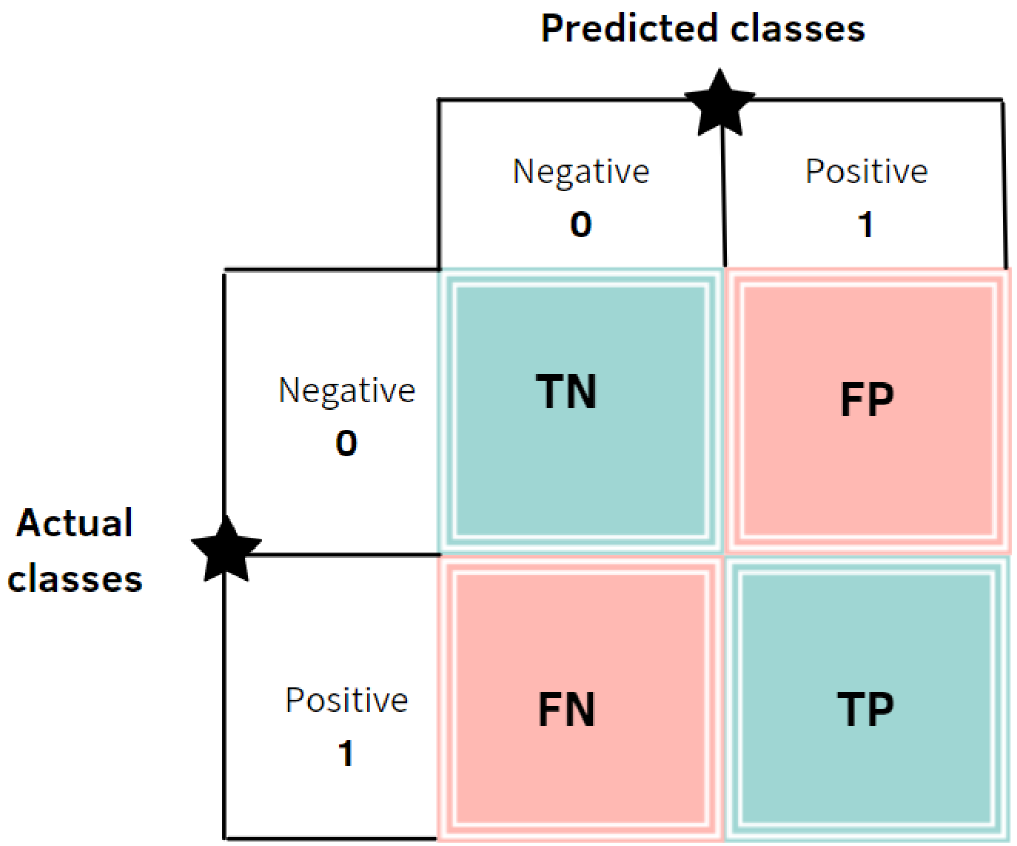
Eval metrics categories

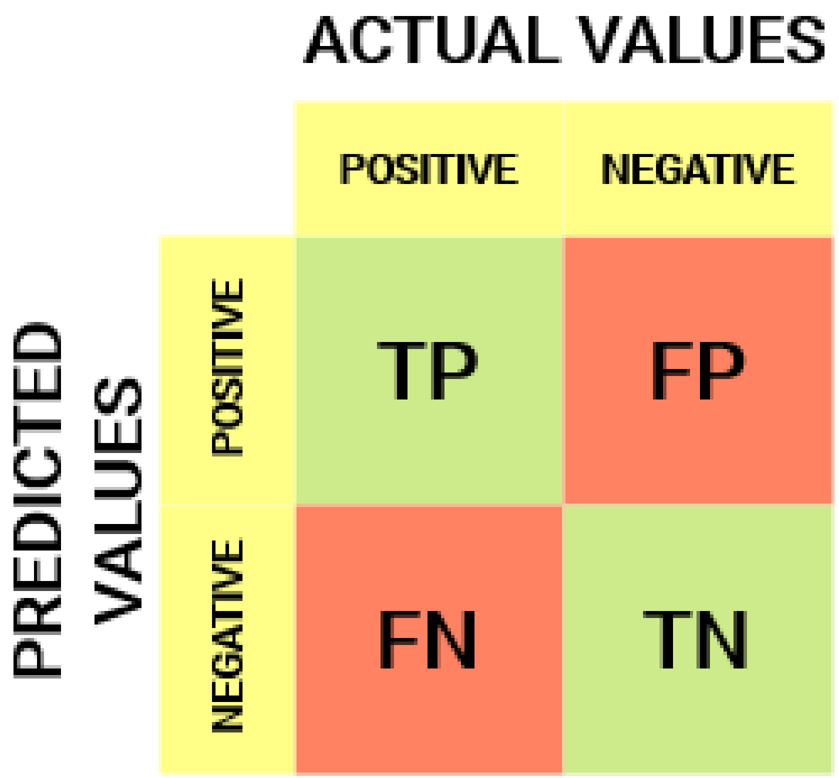
Confusion Matrix

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

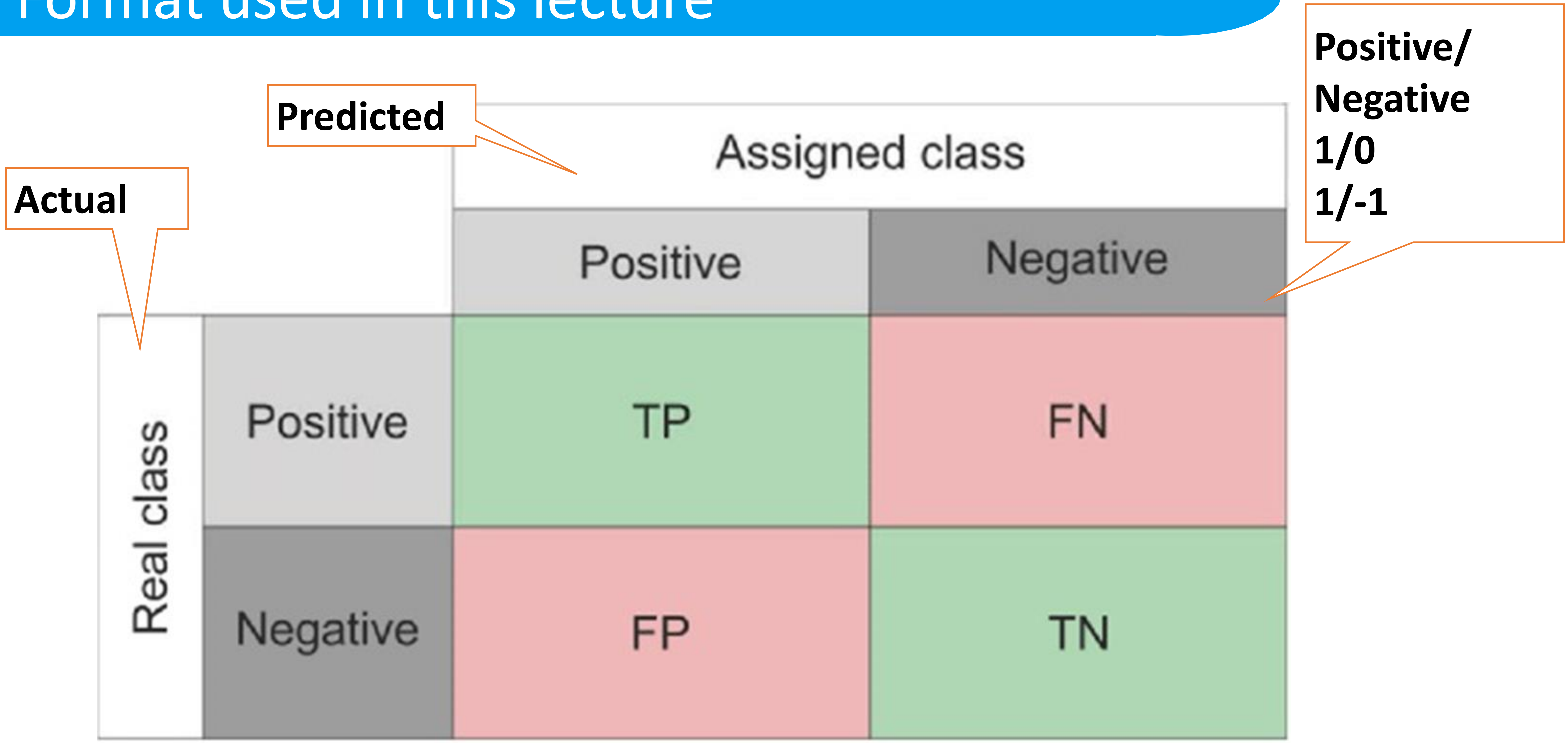






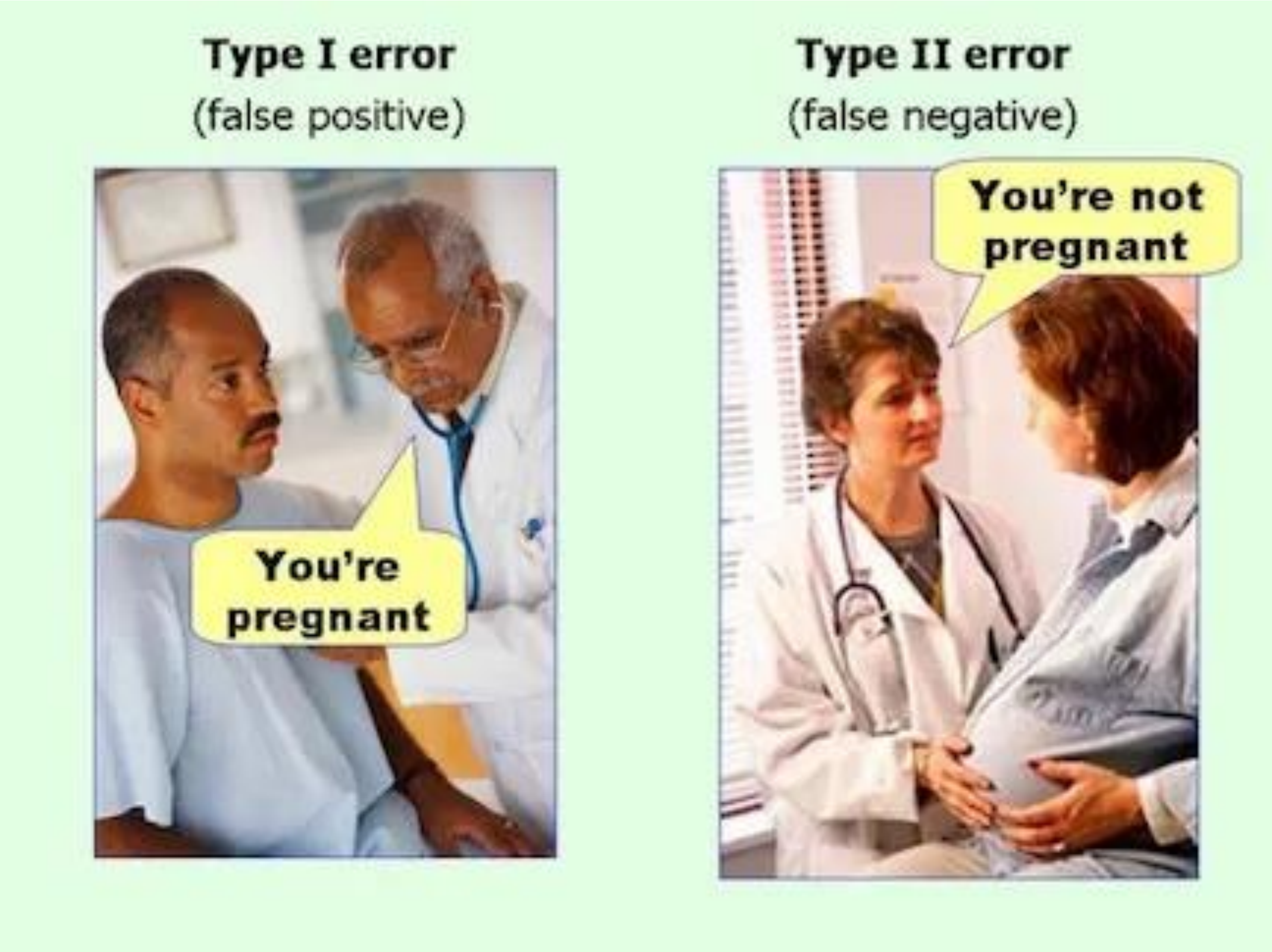


Format used in this lecture



Confusion Matrix elements as joint probabilities

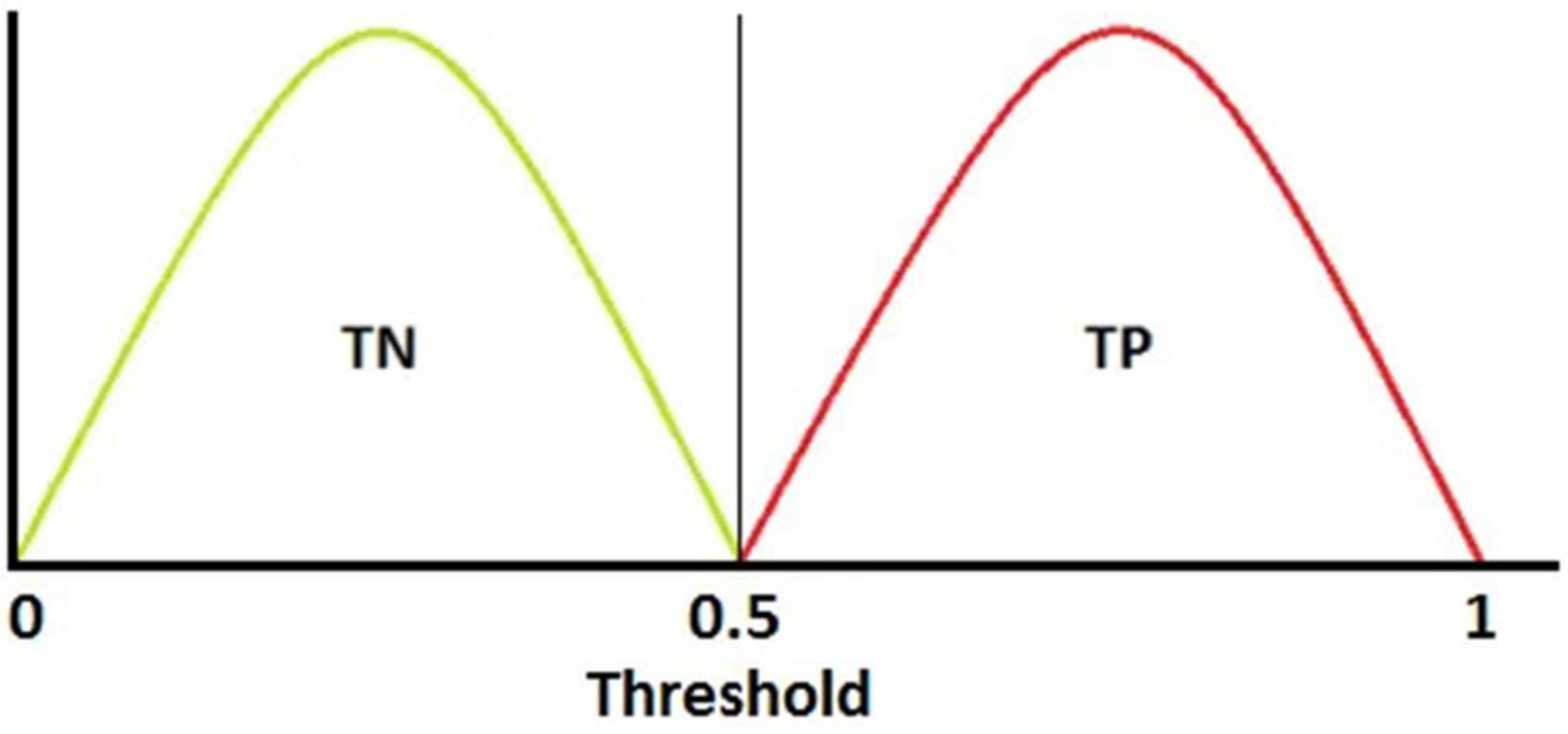
		Assigned class		
		Positive	Negative	
Real class	Positive	TP $P(\hat{y}=1 \cap y=1)$	FN $P(\hat{y}=0 \cap y=1)$	$P(y=1)$
	Negative	FP $P(\hat{y}=1 \cap y=0)$	TN $P(\hat{y}=0 \cap y=0)$	$P(y=0)$
		$P(\hat{y}=1)$	$P(\hat{y}=0)$	

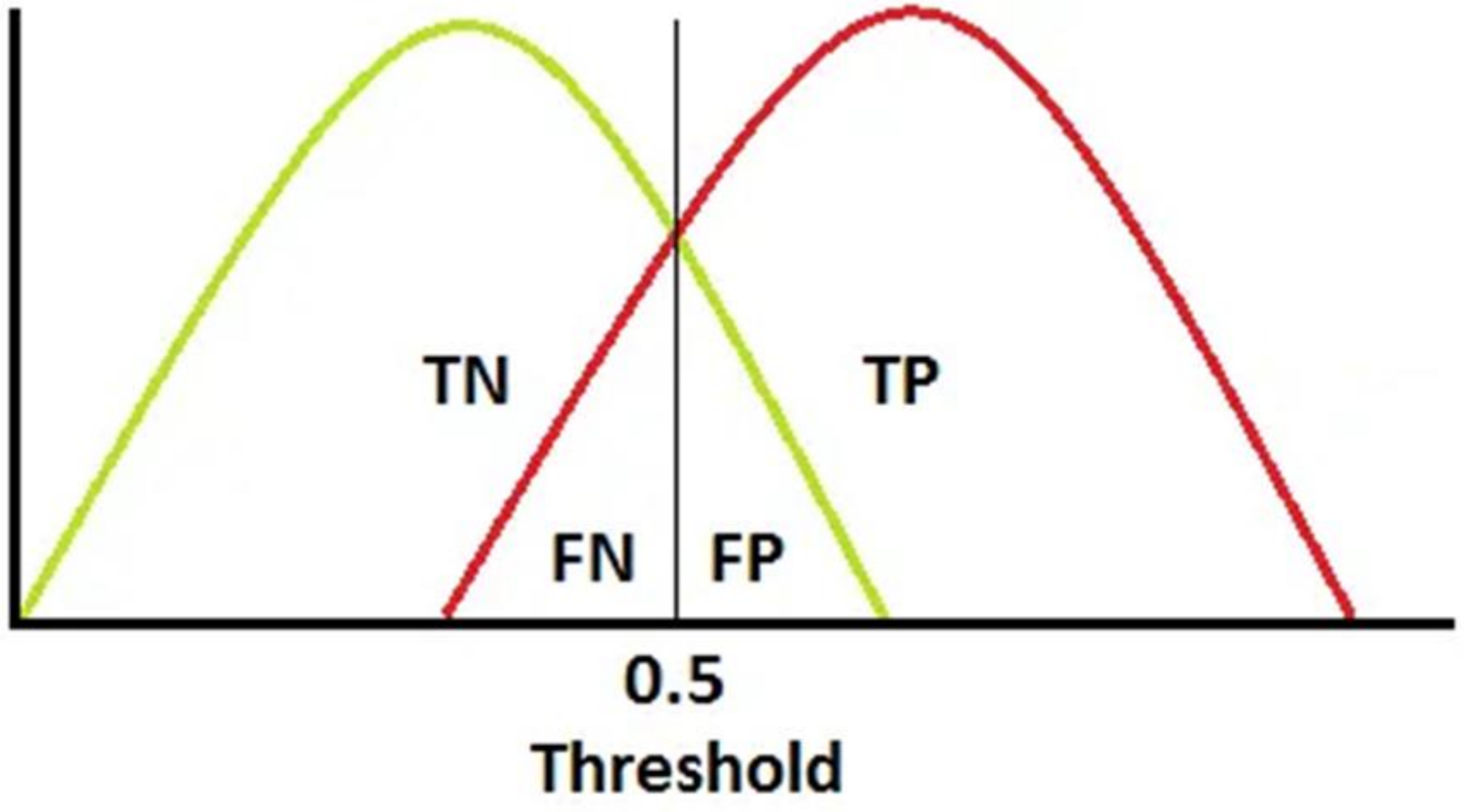

Type I error
(false positive)
You're pregnant
Type II error
(false negative)
You're not pregnant

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

Good measure when +ve class is minority or its prediction is impt

Recall

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

Not a good measure for imbalanced datasets

Accuracy

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

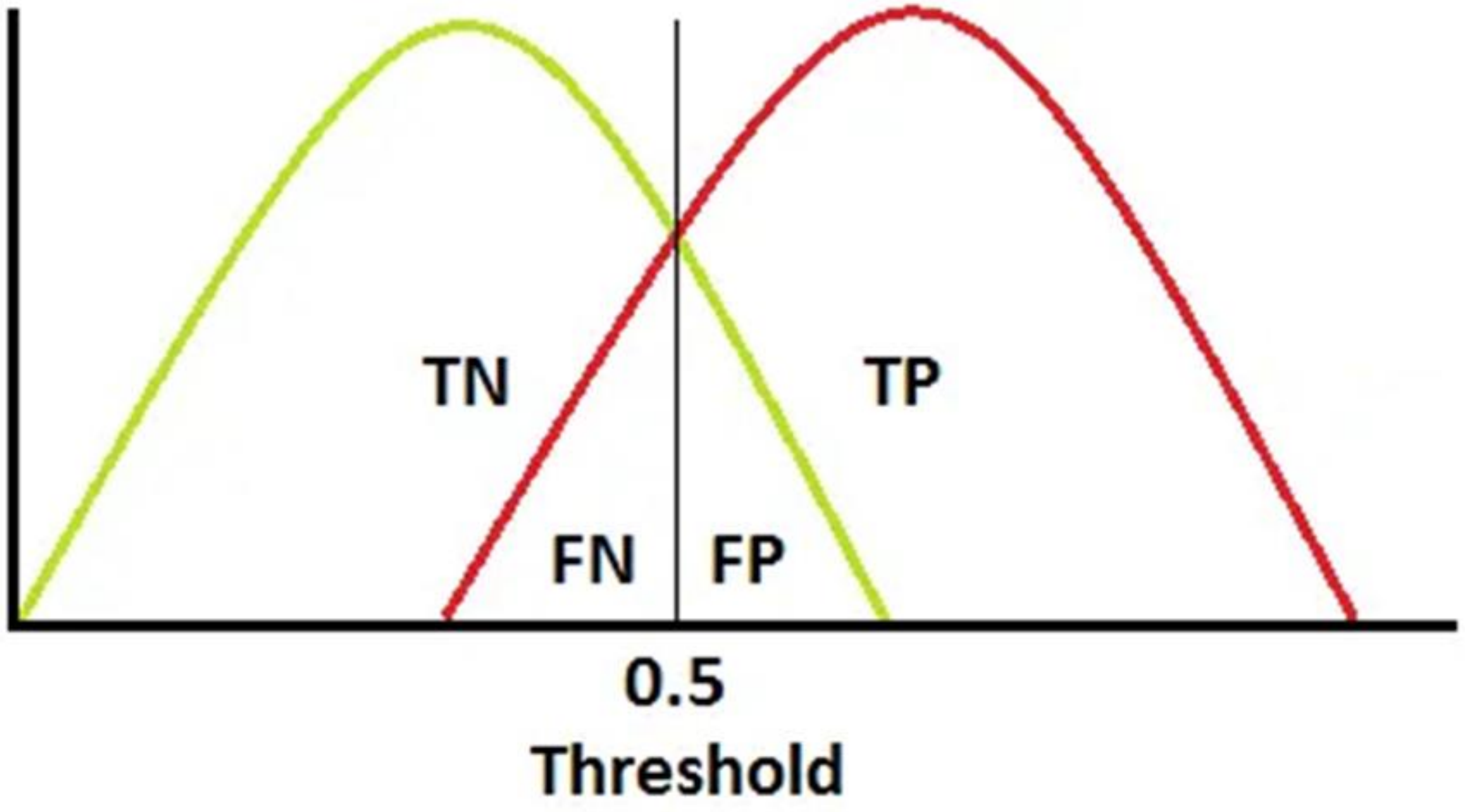
Precision

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

Precision

- TP/(TP+FP)
- P(actual=1 | predicted=1)
- Higher the Precision lesser the false positives
- Reducing FP leads to increase in FN
- How to reduce FP?
- Ans: By increasing threshold

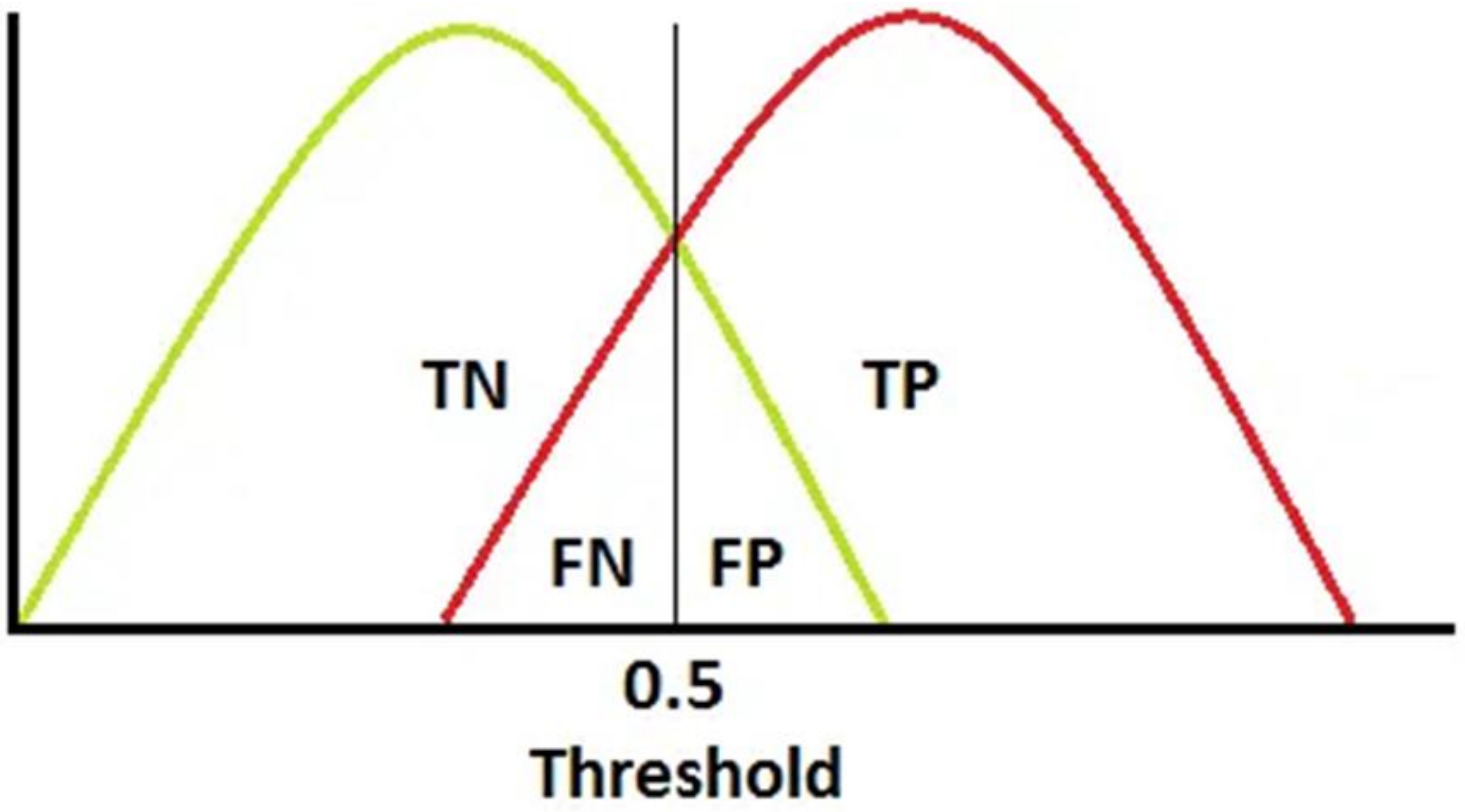
		Assigned class	
		Positive	Negative
Real class	Positive	TP	FN
	Negative	FP	TN



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
- How to increase Recall?
- By reducing FN, i.e. 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

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

Precision

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

Specificity

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

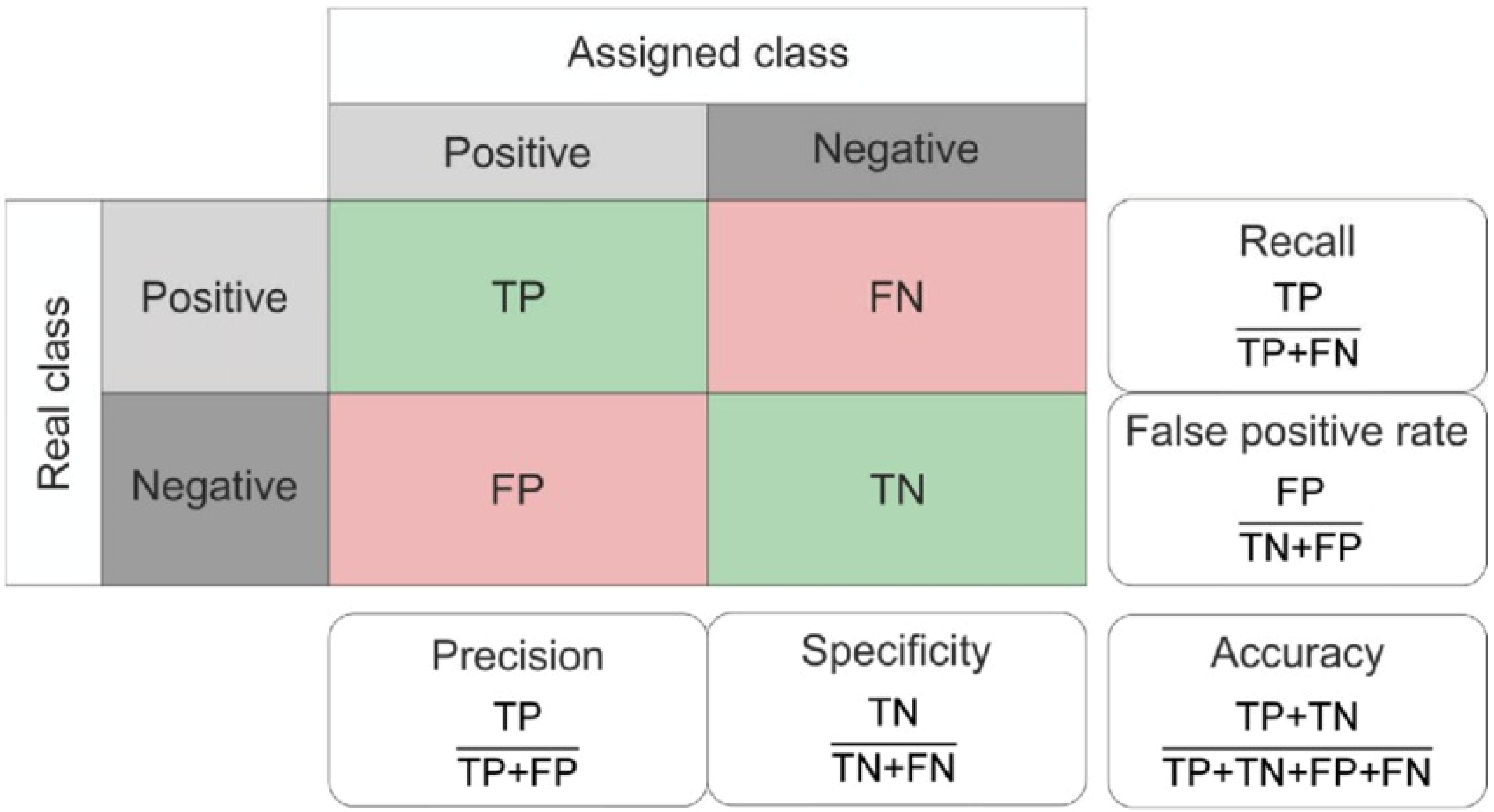
Accuracy

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

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

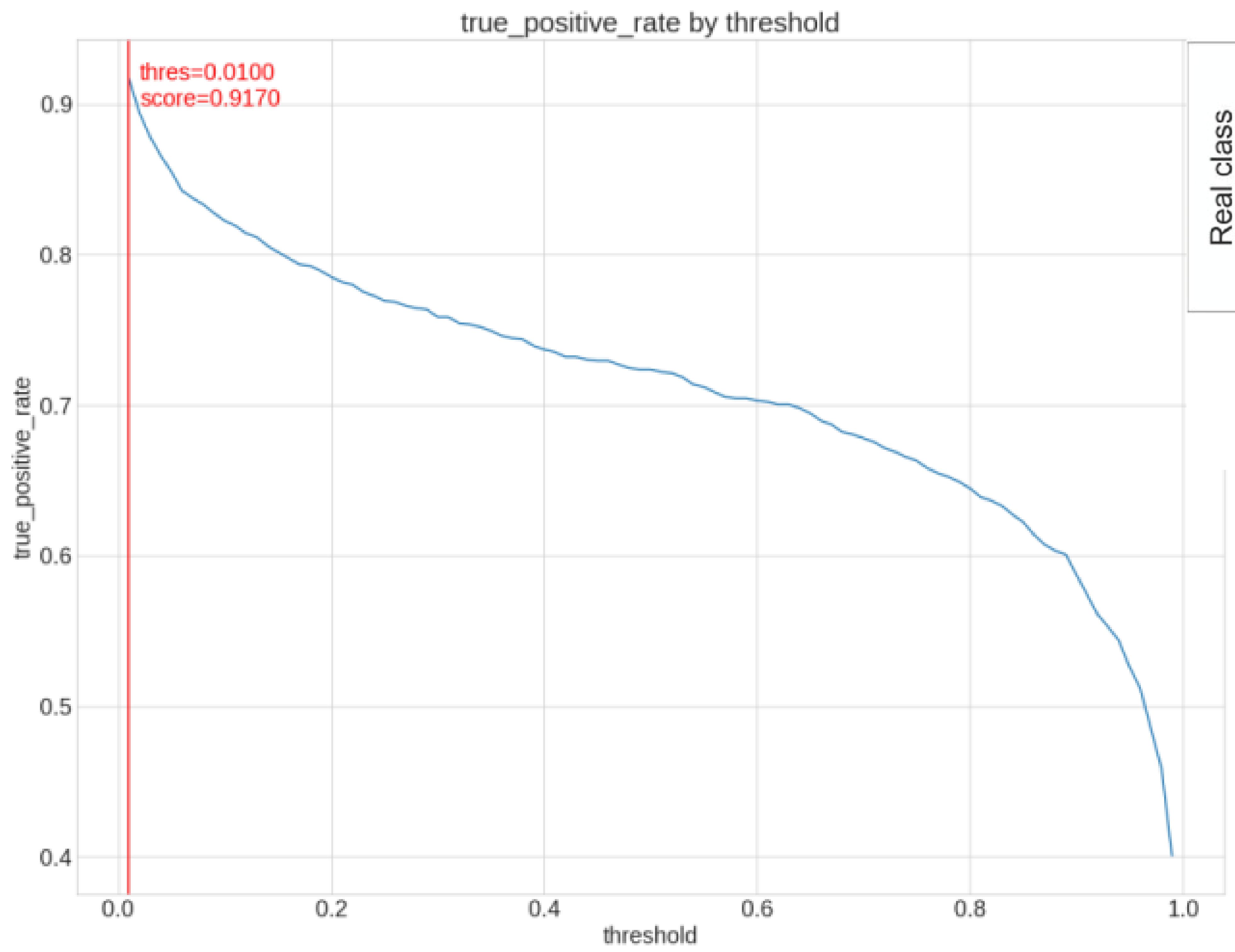
- Harmonic mean $$\frac{1}{\frac{\frac{1}{Precision}+\frac{1}{Recall}}{2}} \qquad \frac{2}{\frac{1}{Precision}+\frac{1}{Recall}}$$

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

- 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



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

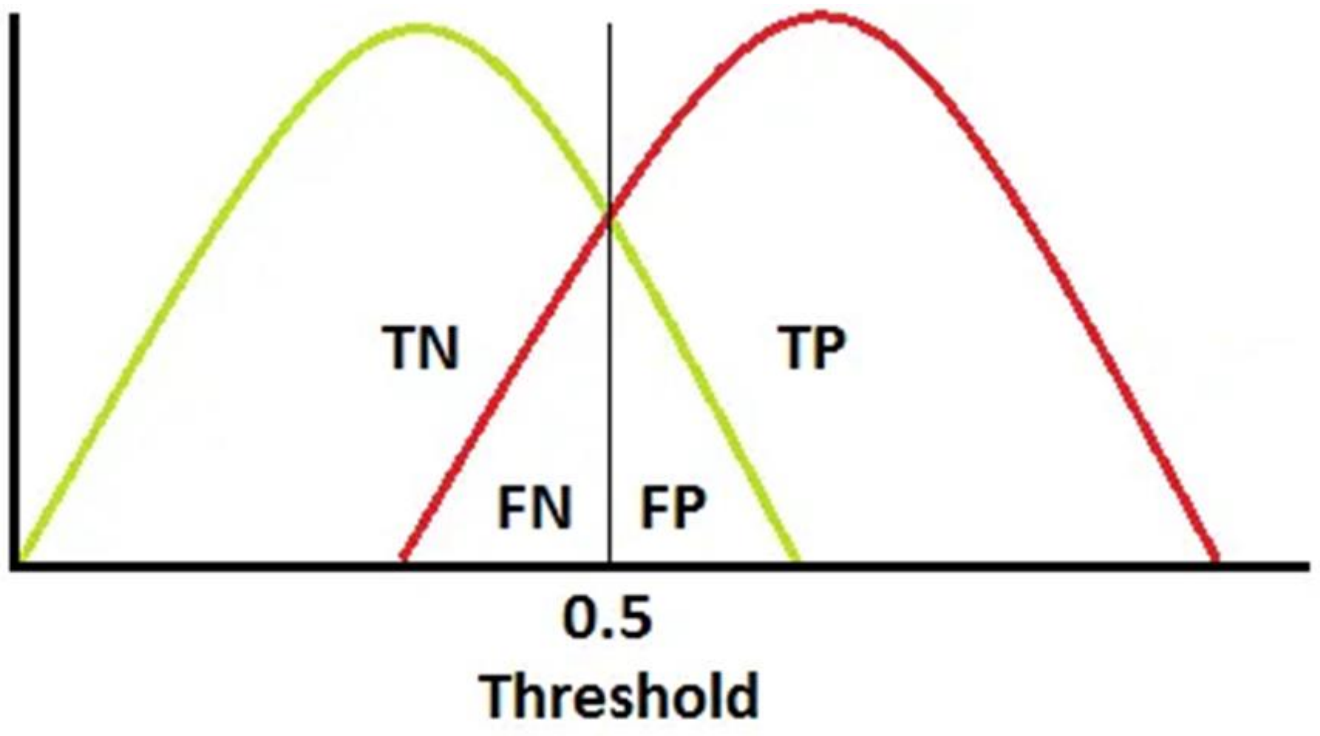
Recall $\frac{TP}{TP+FN}$

False positive rate $\frac{FP}{TN+FP}$

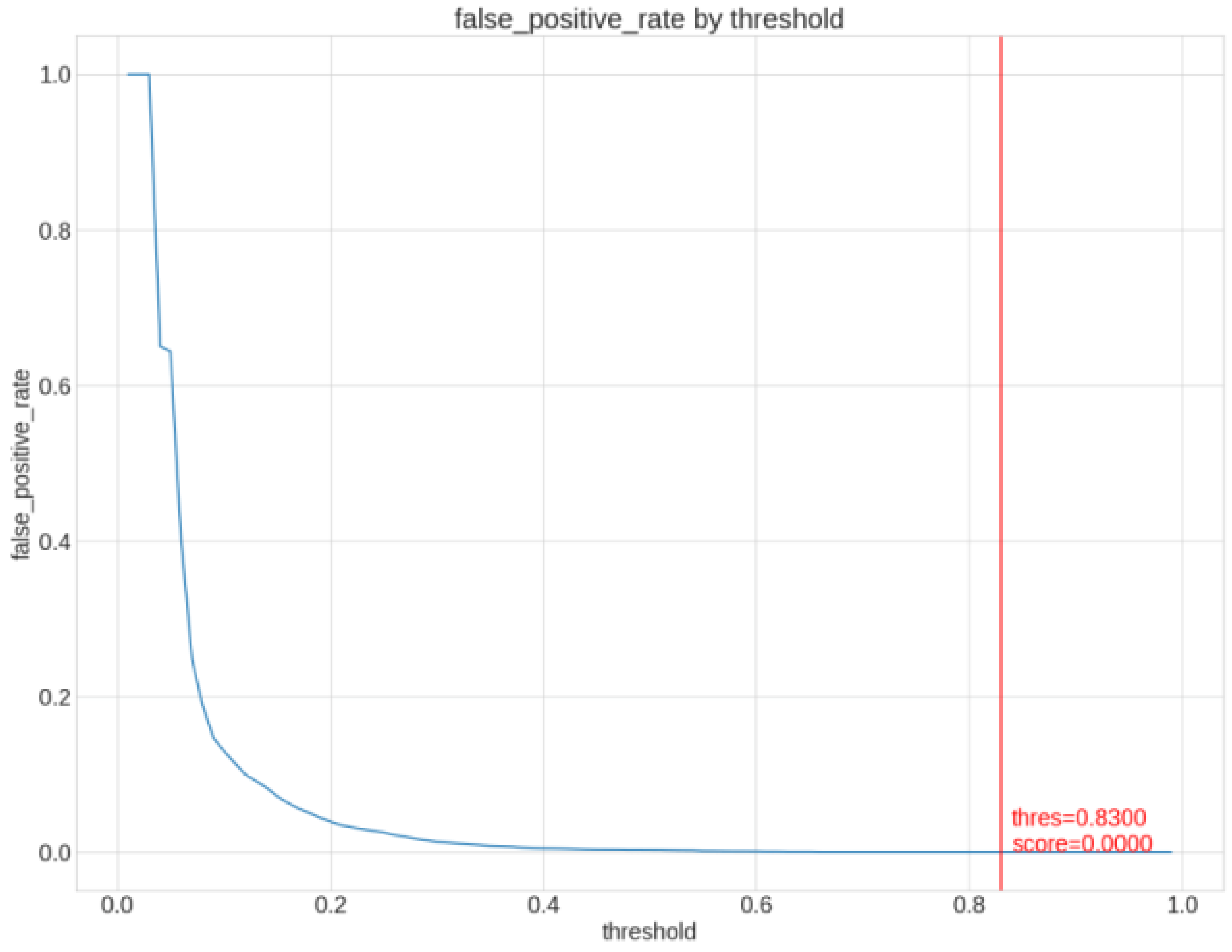
Precision $\frac{TP}{TP+FP}$

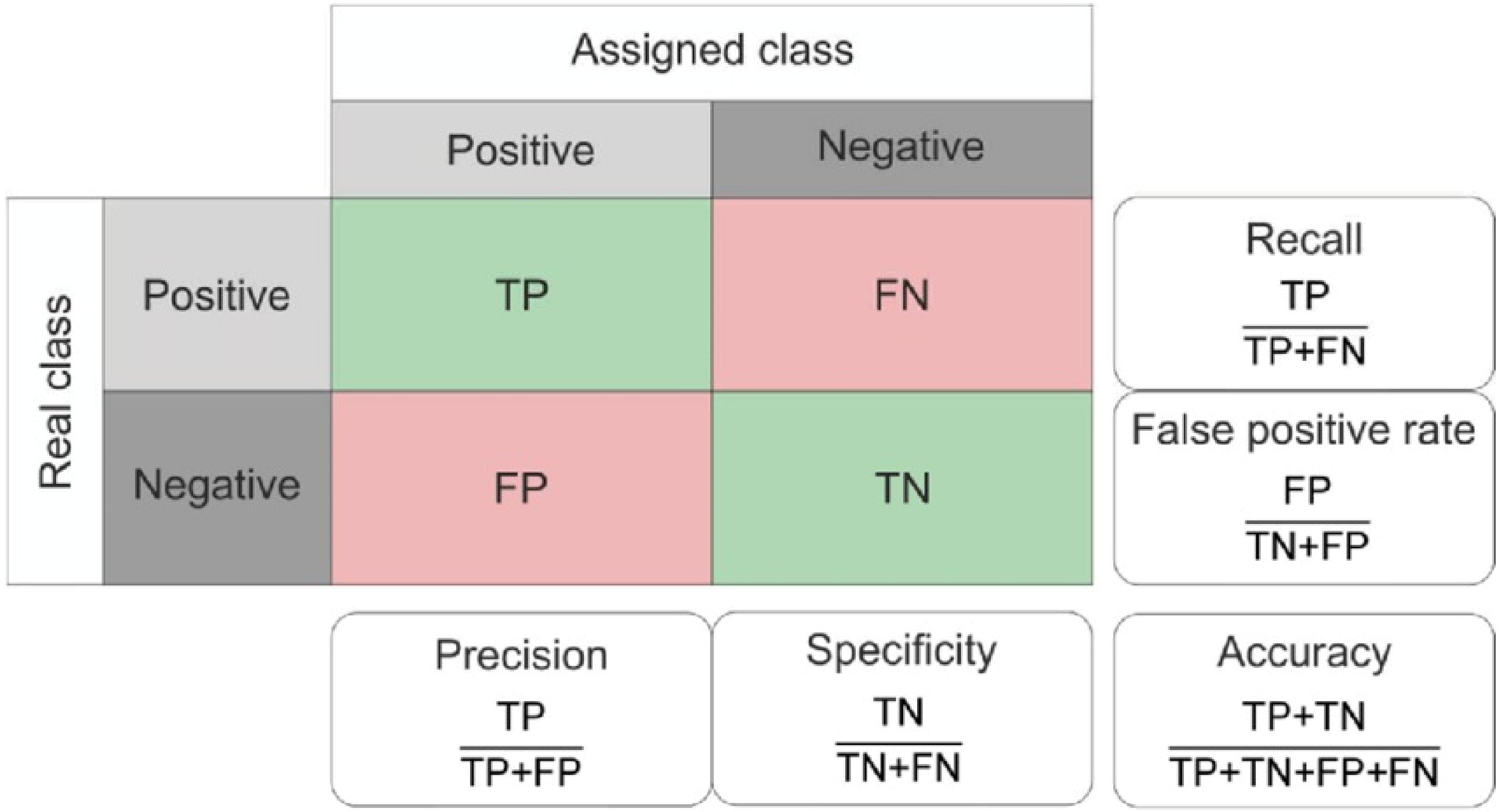
Specificity $\frac{TN}{TN+FN}$

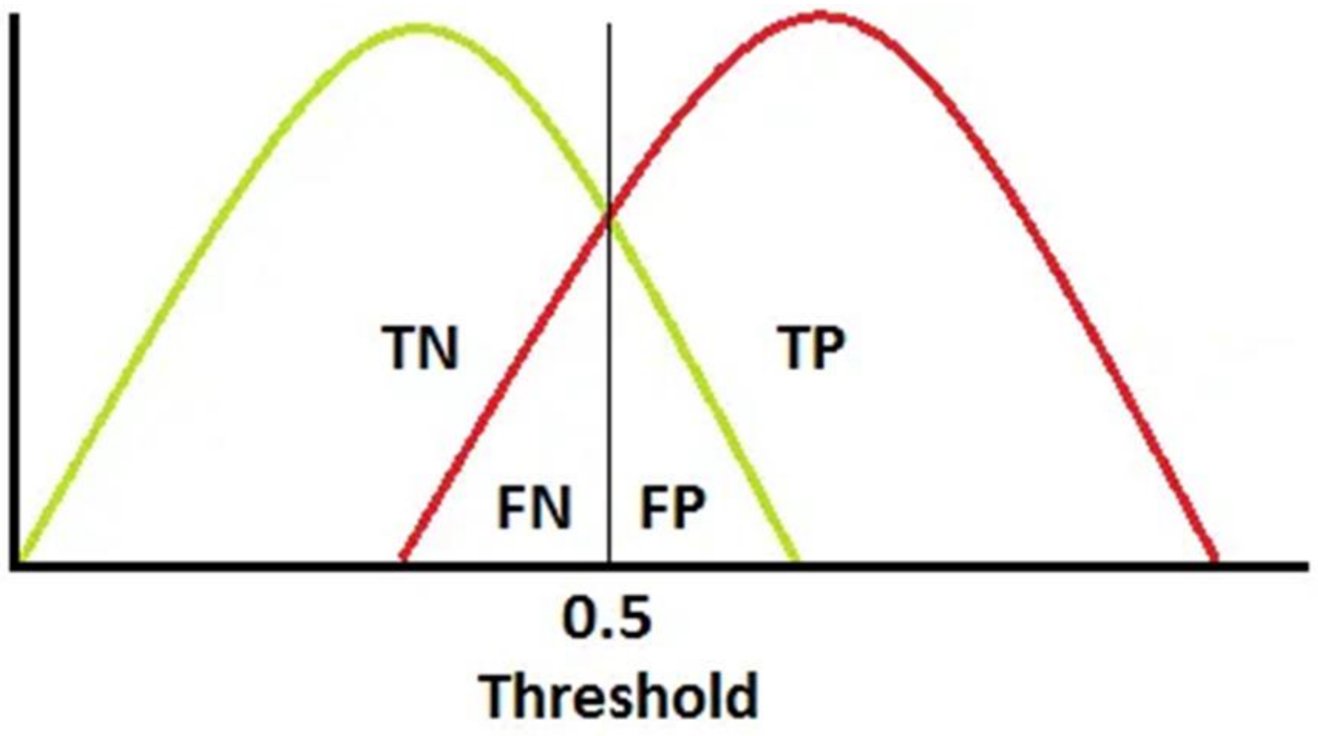
Accuracy $\frac{TP+TN}{TP+TN+FP+FN}$



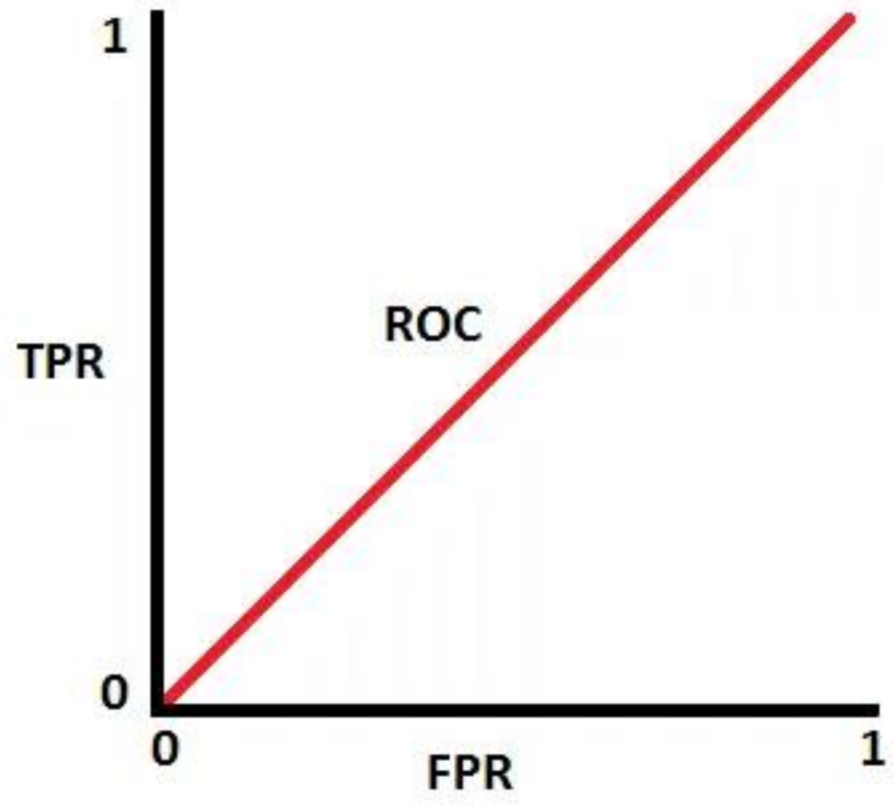
False positive rate versus threshold

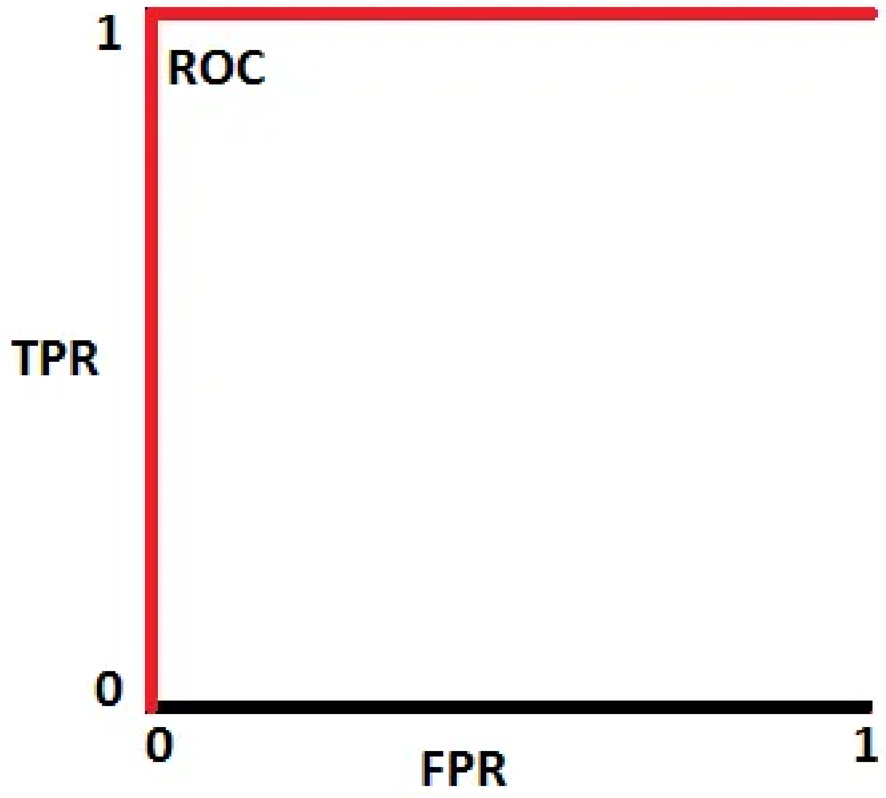


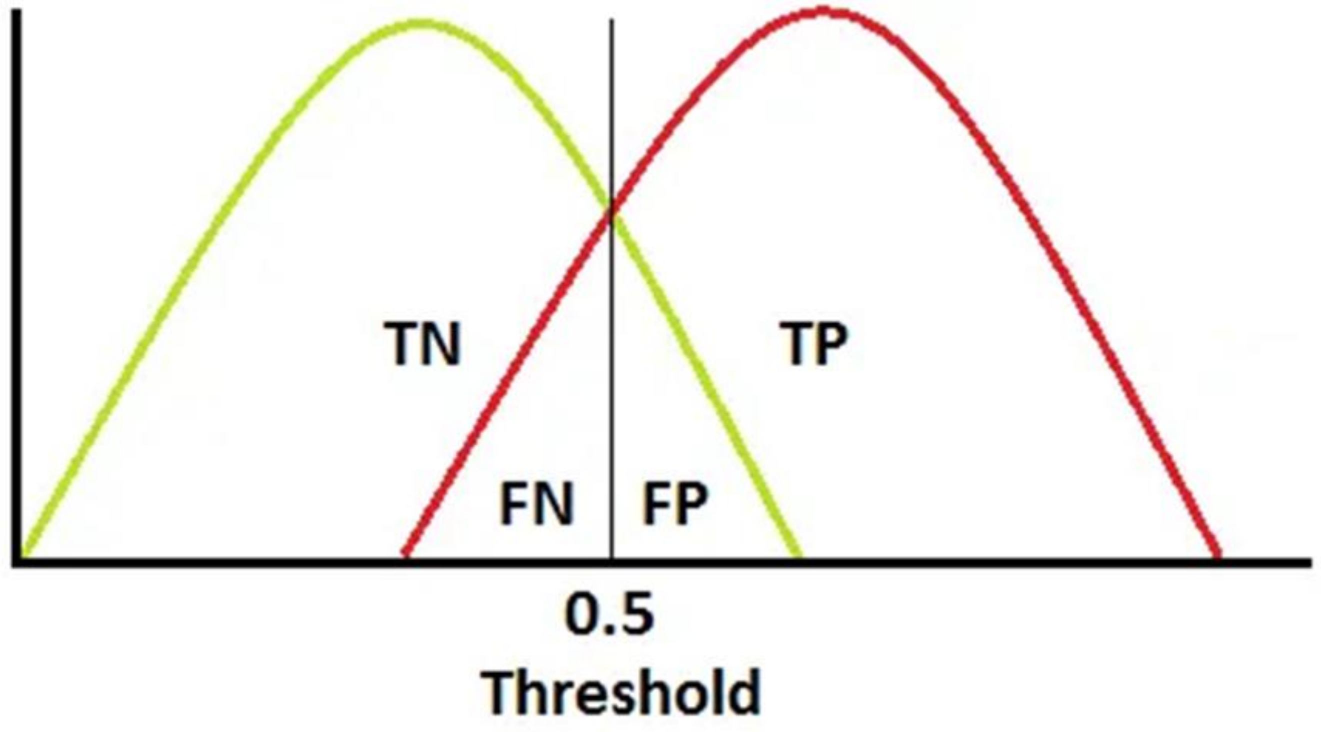


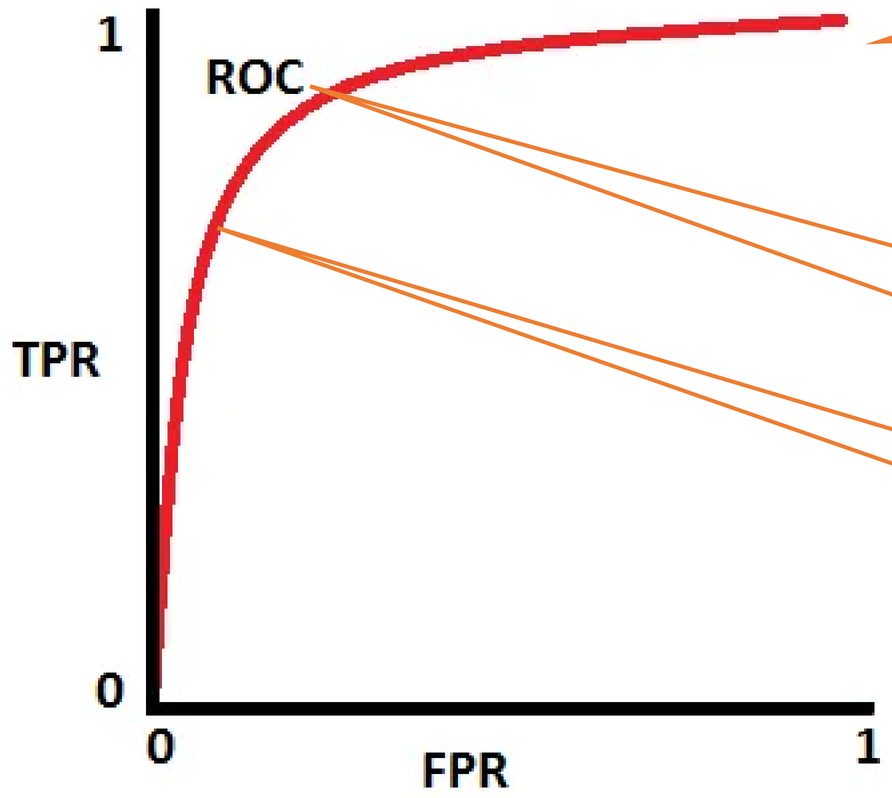


AU ROC





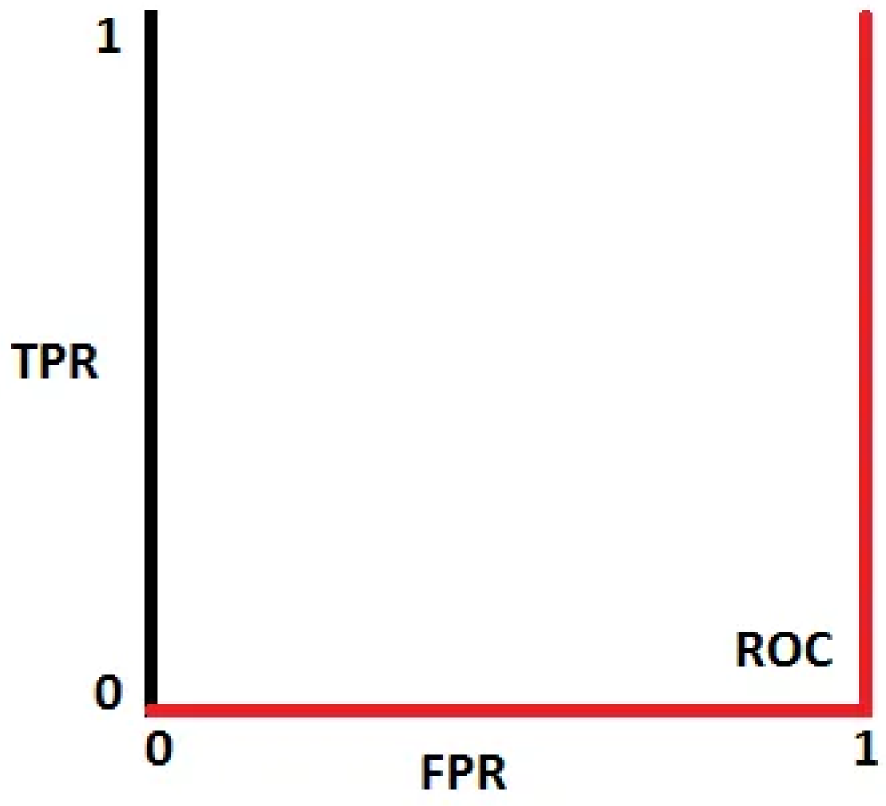




When data is balanced

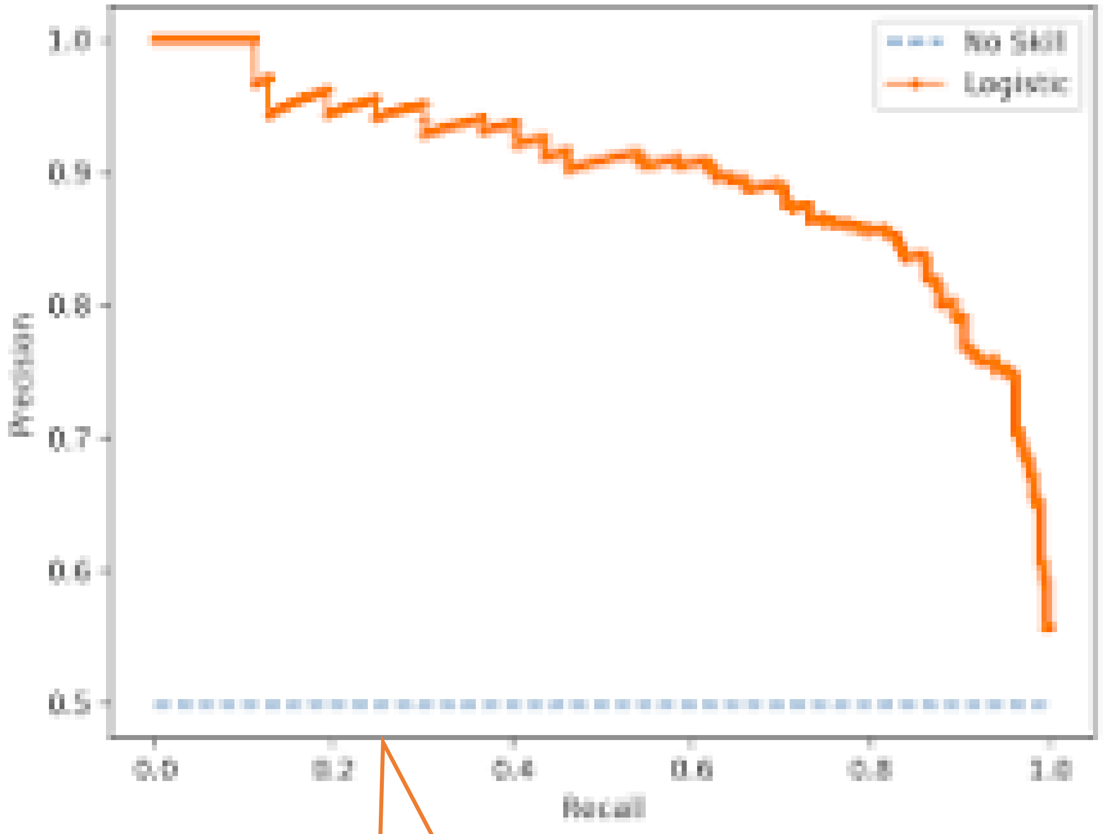
Measure of robustness

Left is better, Up is better

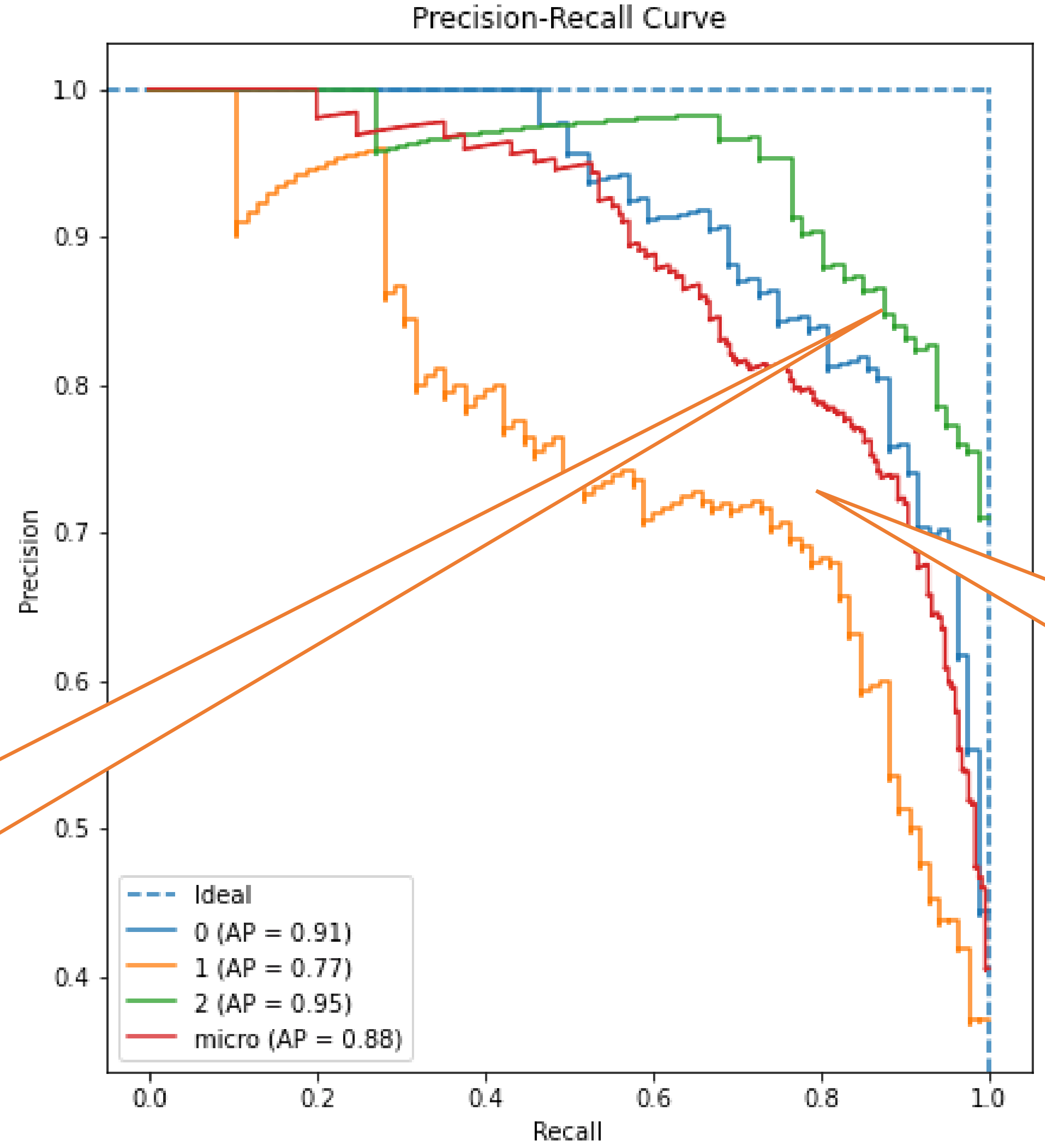


AU PRC

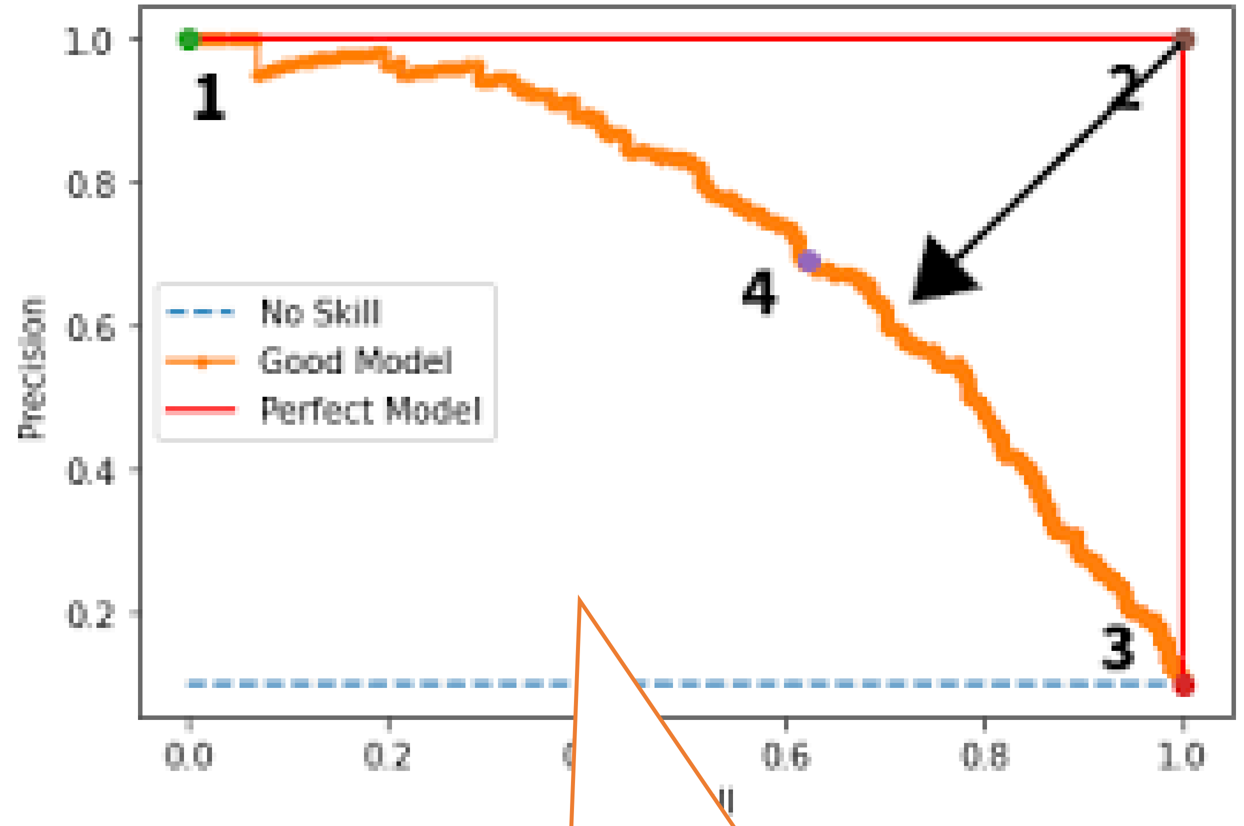
Use when data is imbalanced



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

Label	Per-Class F1 Score	Support	Support Proportion	Weighted Average F1 Score
Airplane	0.67	3	0.3	$(0.67 * 0.3) + (0.40 * 0.1) + (0.67 * 0.6)$ **= 0.64**
Boat	0.40	1	0.1	
Car	0.67	6	0.6	
Total	-	10	1.0	


QUESTIONS