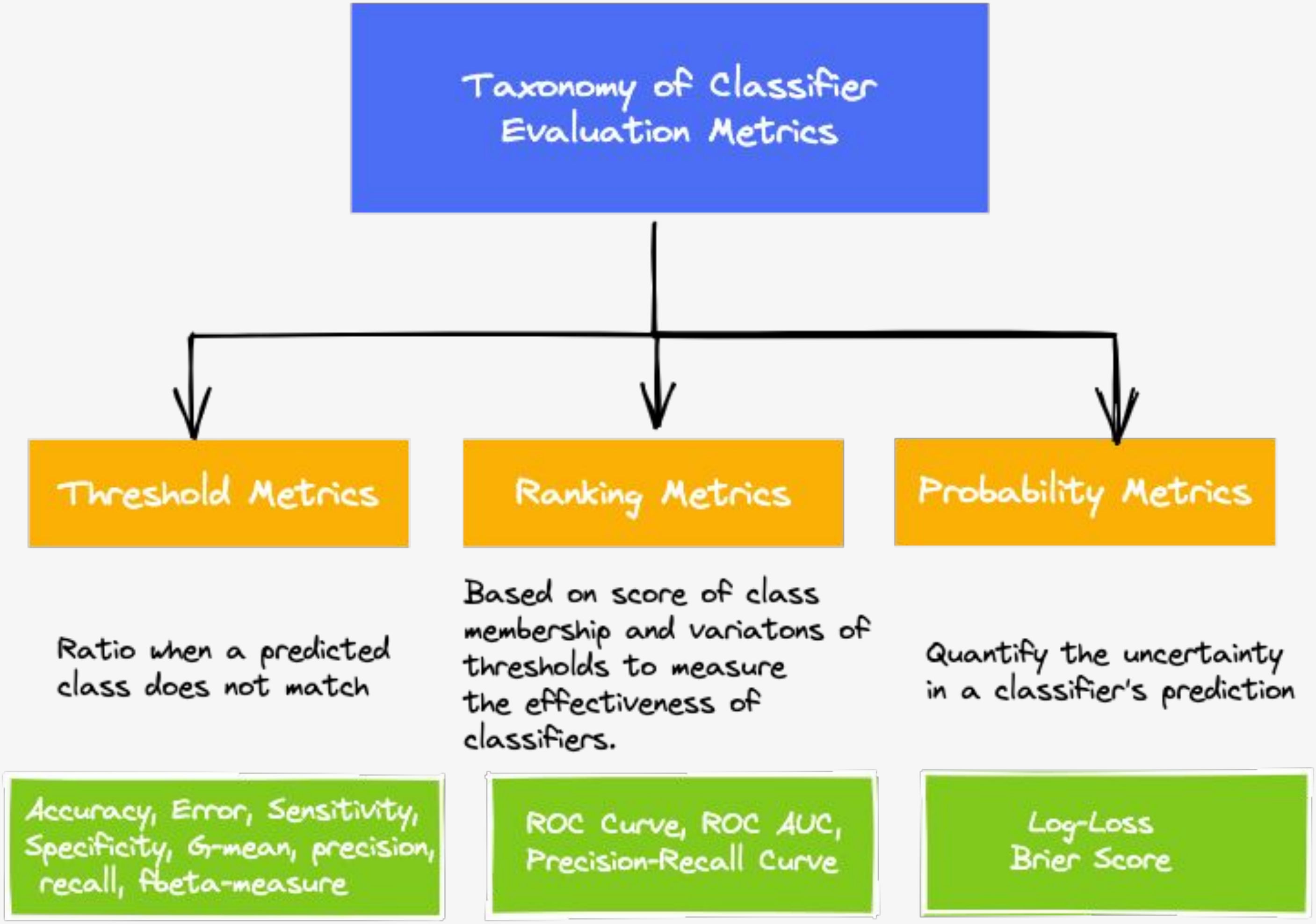
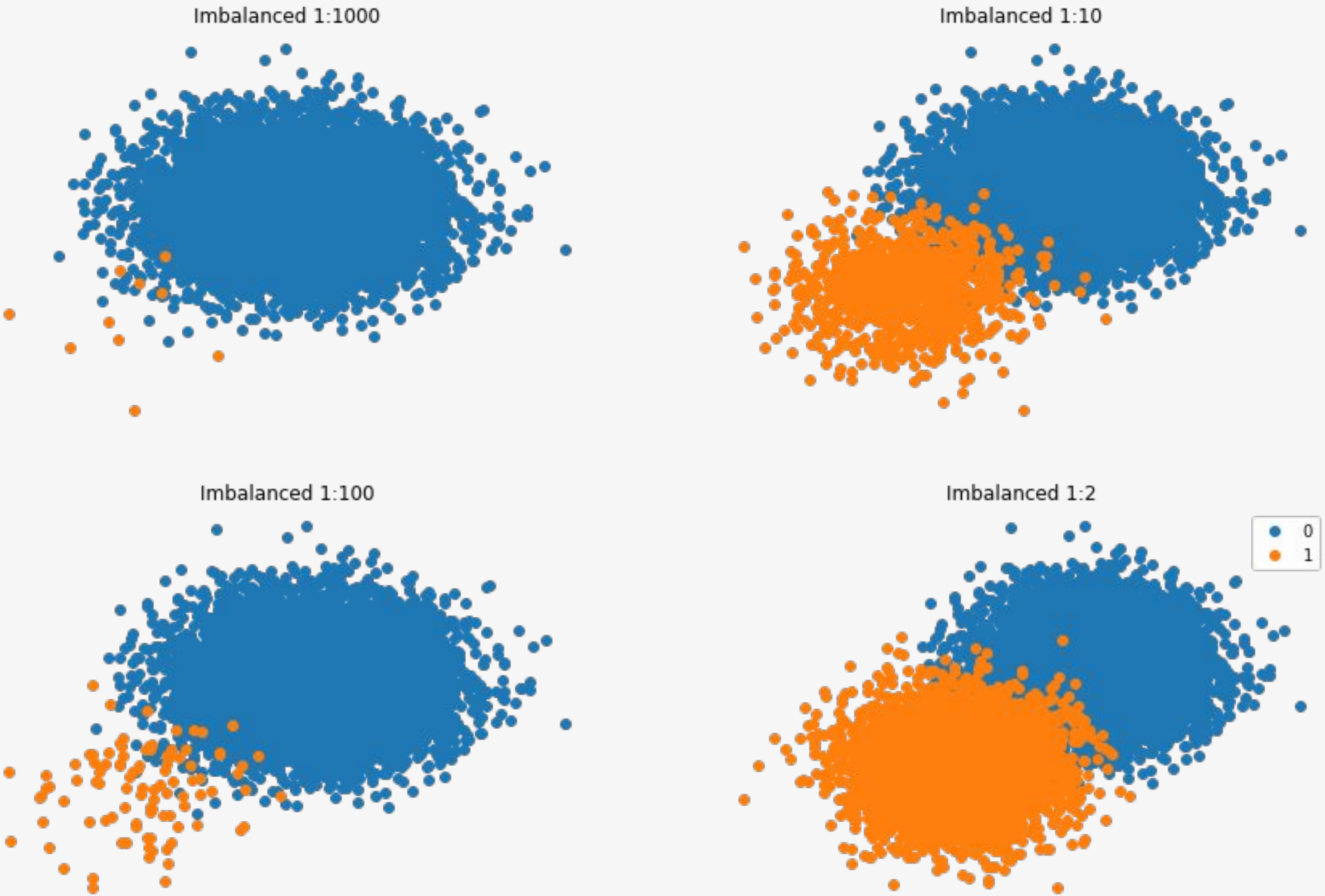
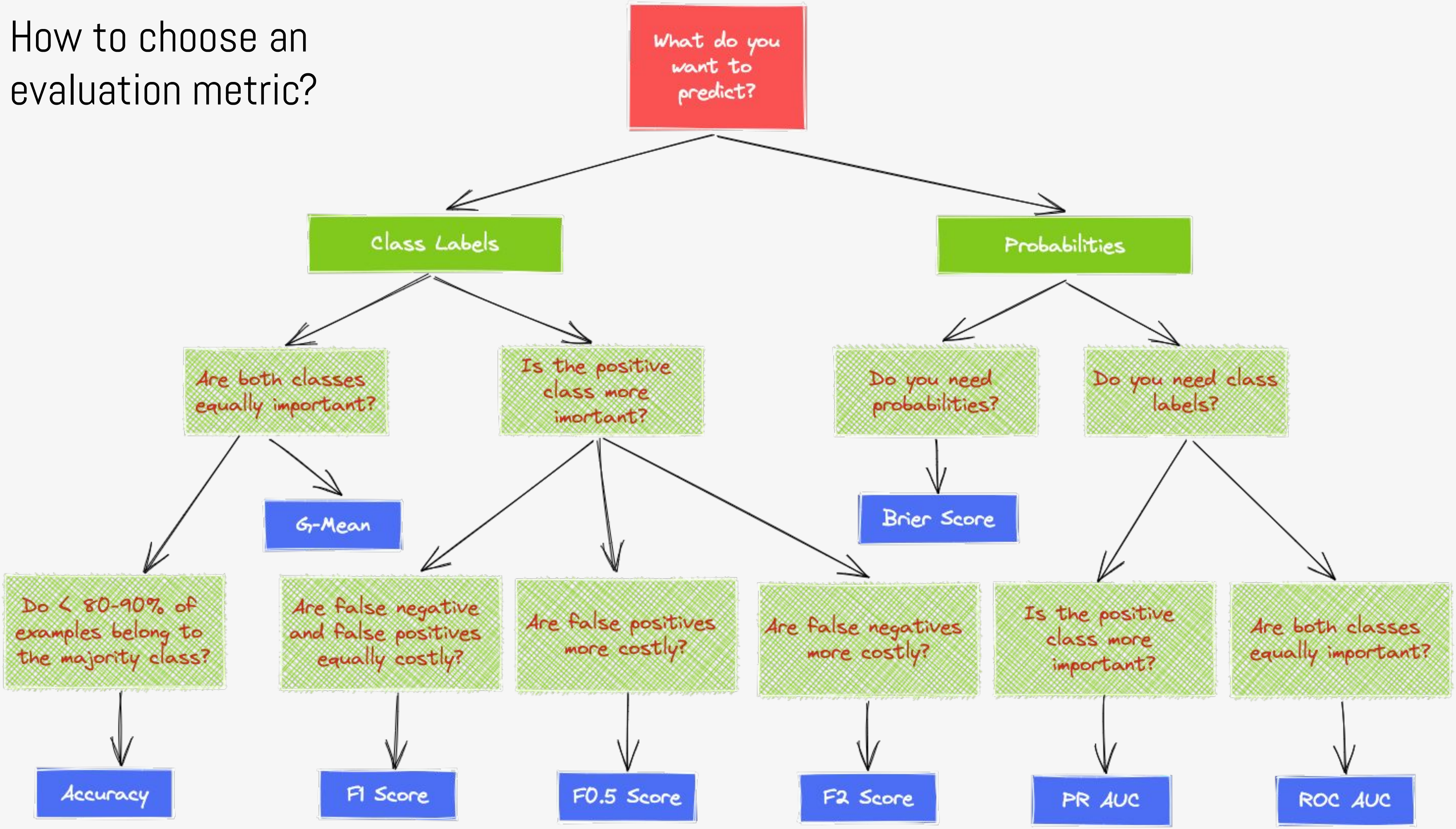
Machine Learning Evaluation Metrics


Taxonomy of Classifier Evaluation Metrics
Threshold Metrics
Ranking Metrics
Probability Metrics
Ratio when a predicted class does not match
Based on score of class membership and variatons of thresholds to measure the effectiveness of classifiers.
Quantify the uncertainty in a classifier's prediction
Accuracy, Error, Sensitivity, Specificity, G-mean, precision, recall, fbeta-measure
ROC Curve, ROC AUC, Precision-Recall Curve
Log-Loss
Brier Score

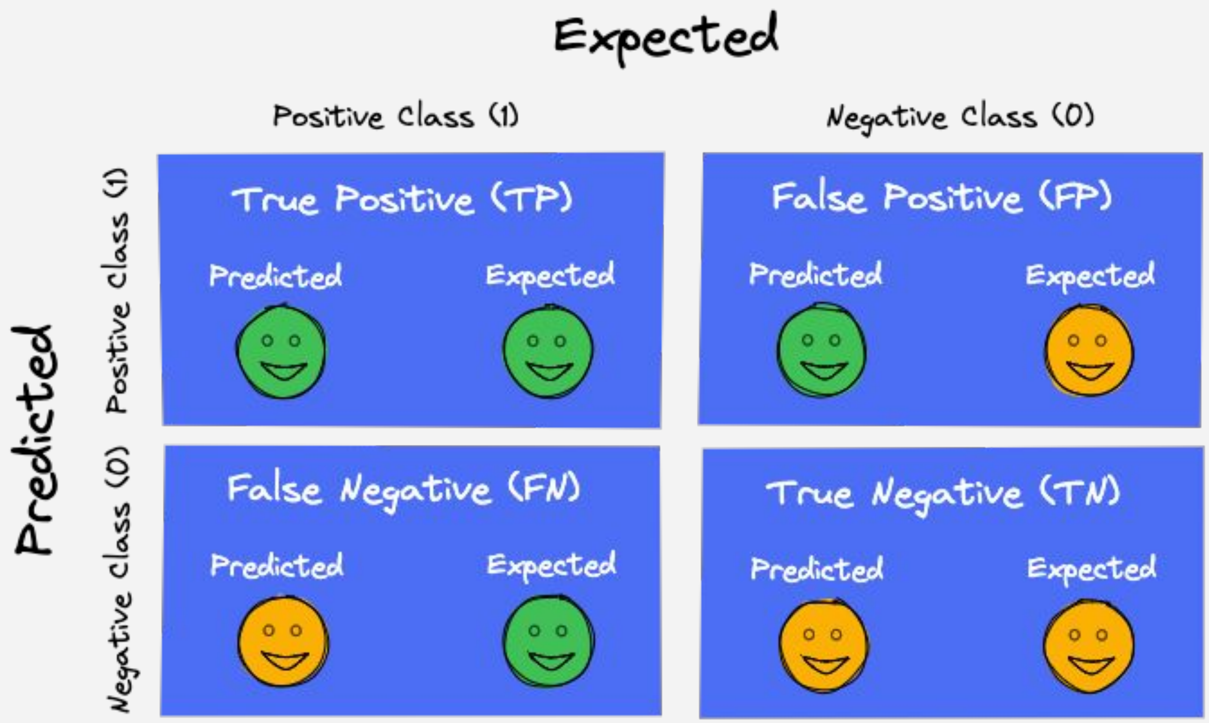

Imbalanced 1:1000
Imbalanced 1:10
Imbalanced 1:100
Imbalanced 1:2
0
1

How to choose an evaluation metric?



@Brownlee, Jason. Imbalanced classification with python.

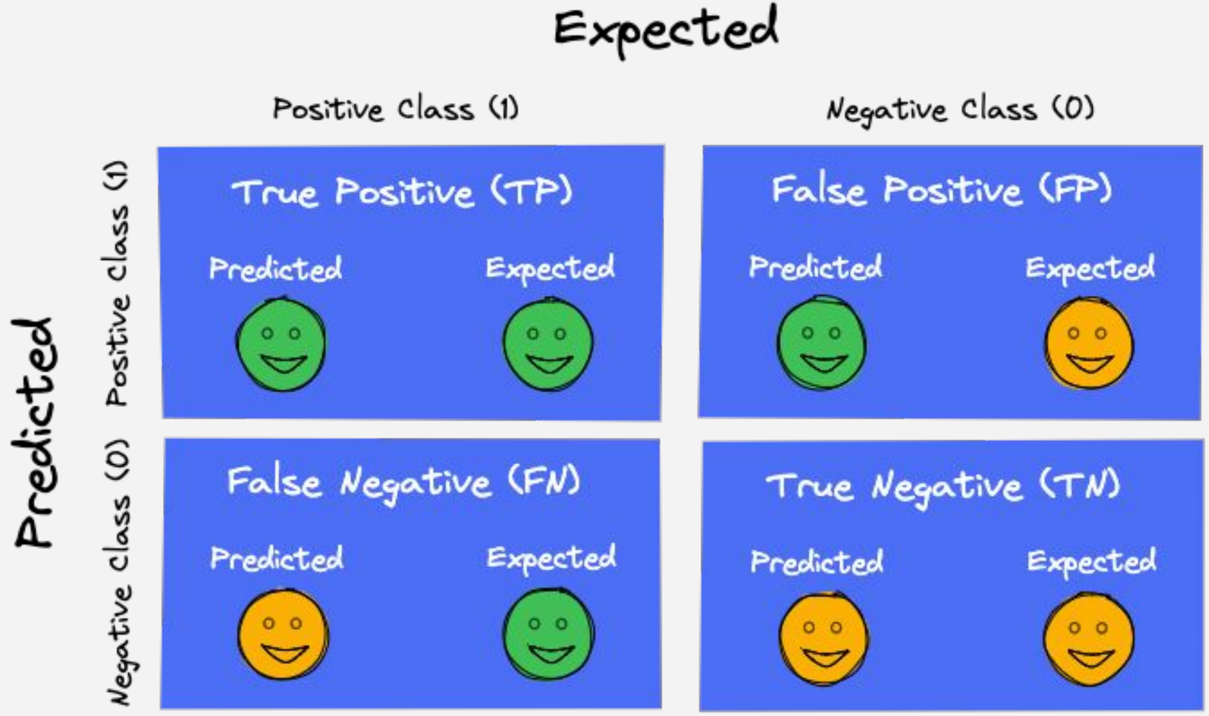
Confusion Matrix



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

$$\text{Error} = 1 - \text{Accuracy}$$

Confusion Matrix



$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{FP + TN}$$

$$\text{G-mean} = \sqrt{\text{Sensitivity} \times \text{Specificity}}$$

Confusion Matrix

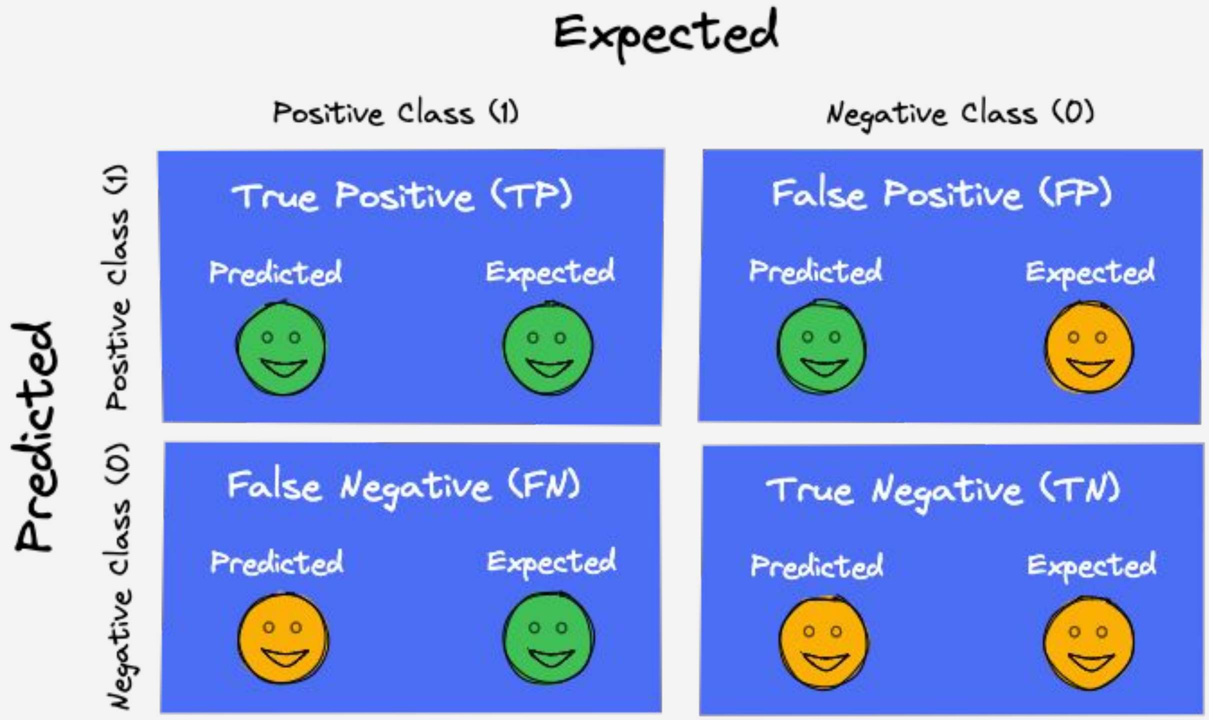
Predicted \ Expected	Positive Class (1)	Negative Class (0)
Positive class (1)	True Positive (TP) Predicted / Expected	False Positive (FP) Predicted / Expected
Negative class (0)	False Negative (FN) Predicted / Expected	True Negative (TN) Predicted / Expected

Precision = (positive predicte value - PPV) $\frac{TP}{TP + FP}$

Precision = (negative predicte value - NPV) $\frac{TN}{TN + FN}$

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

Confusion Matrix



$$\text{Fbeta-measure} = \frac{(1 + \beta^2) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}}$$

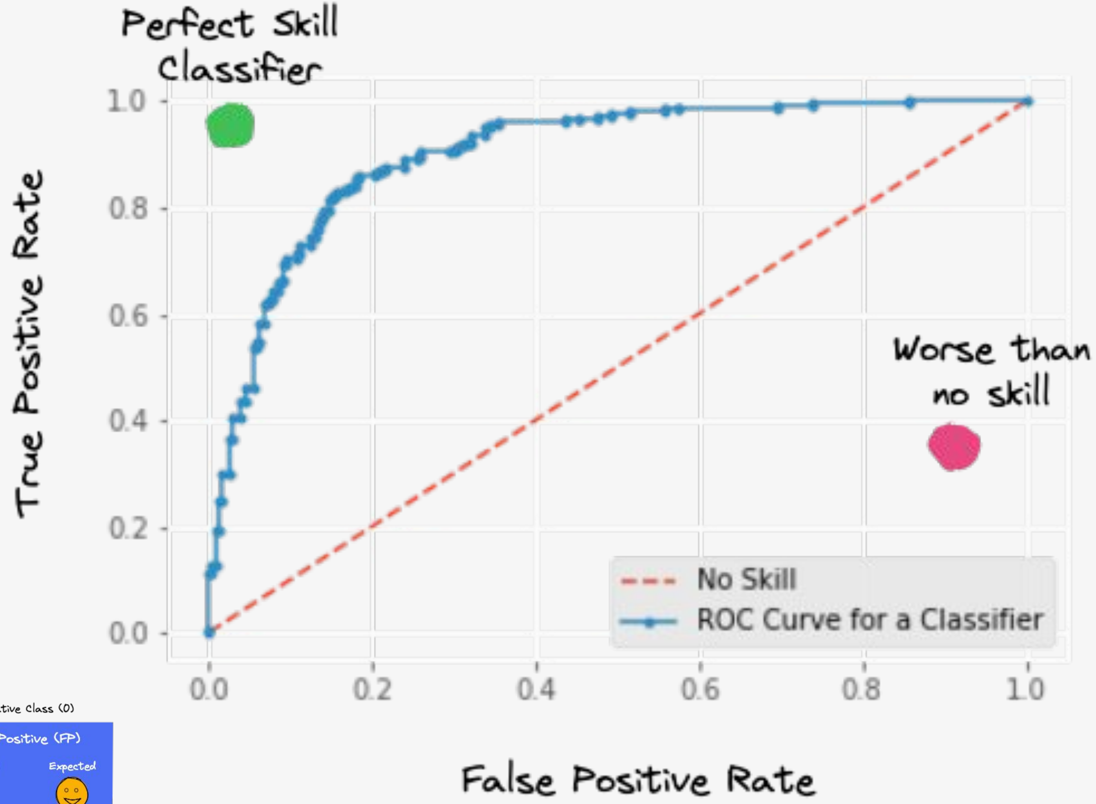
$$\beta == \begin{cases} 0.5, & \text{more weight on precision} \\ 1.0, & \text{balance on weight PR and RE} \\ 2.0, & \text{less weight on precision} \end{cases}$$

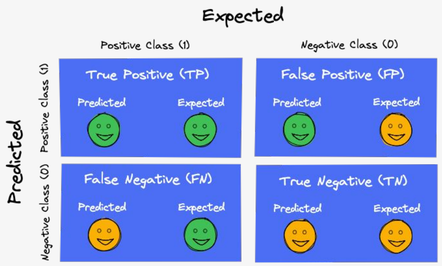
Rank metrics are more concerned with evaluating classifiers based on **how effective** they are at separating classes.

These metrics require that a **classifier predicts a score** or a probability of class membership. From this score, **different thresholds** can be applied to **test the effectiveness of classifiers**. Those models that maintain a good score across a range of thresholds will have good class separation and will be ranked higher.

Receiver Operating Characteristic (ROC)

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

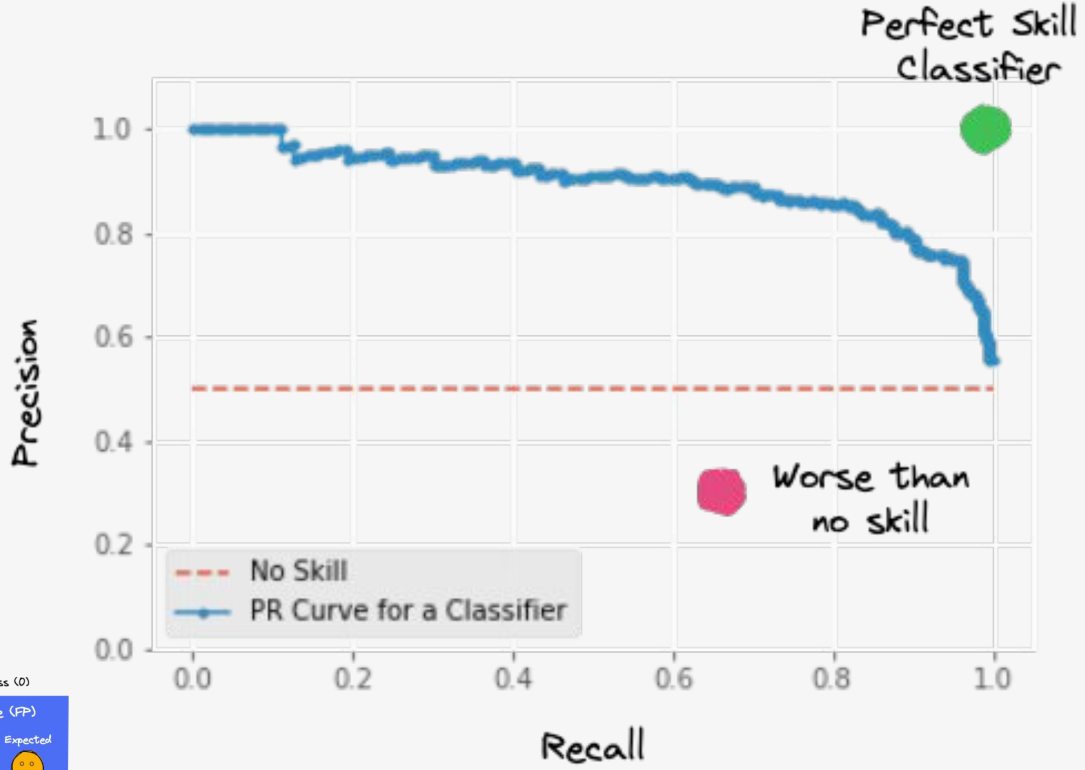




$$FPR = \frac{FP}{FP + TN}$$

Precision-Recall (PR) Curve

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



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