Advanced Evaluation Metrics: A Visual Guide for Beginners

Confusion Matrix

Predicted Class

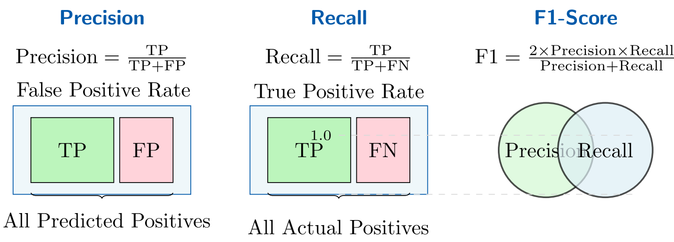
Actual Class

True Positive (TP)	**False Positive (FP)**
False Negative (FN)	**True Negative (TN)**

All other metrics are derived from these four values!

A confusion matrix shows how your model is performing by comparing predicted vs. actual classes.

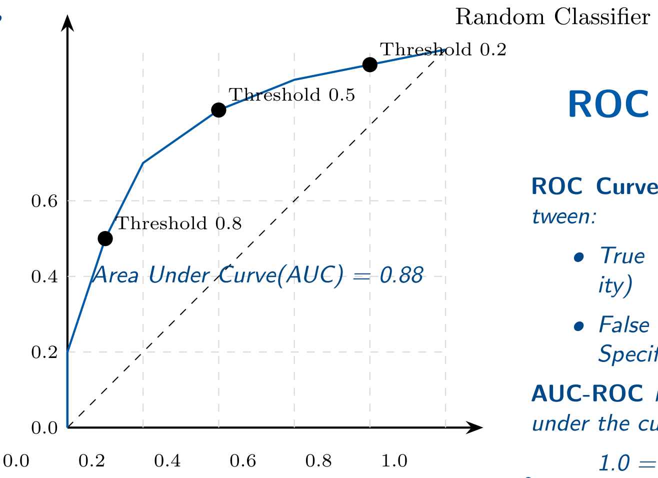
- *TP: Correctly predicted positive*
- *TN: Correctly predicted negative*
- *FP: Incorrectly predicted positive*
- *FN: Incorrectly predicted negative*



Harmonic mean of precision and recall(balances both concerns)

How many actual positivesdid we correctly identify?

How many of our positivepredictions were correct?



ROC Curve and AUC-ROC

ROC Curve *shows the tradeoff between:*

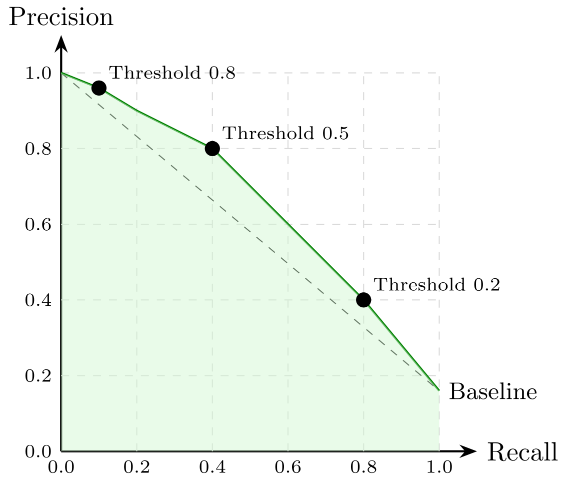
- *True Positive Rate (Sensitivity)*
- *False Positive Rate (1-Specificity)*

AUC-ROC *measures the entire area under the curve:*

- *1.0 = Perfect classifier*
- *0.5 = Random guessing*
- *¡ 0.5 = Worse than random*

Use when: *Evaluating how well a model can distinguish between classes, especially with balanced datasets.*

Precision-Recall Curve and AUC-PR



Area Under Curve(AUC-PR) = 0.75

PR Curve *shows the tradeoff between:*

- *Precision (focusing on reducing false positives)*
- *Recall (focusing on reducing false negatives)*

AUC-PR *measures the entire area under the PR curve.*

Use when: *Working with imbalanced datasets where the positive class is rare and more important.*

Key difference from ROC: *Focuses on performance on the positive class only.*

Per-Class Evaluation and Multi-Class Metrics

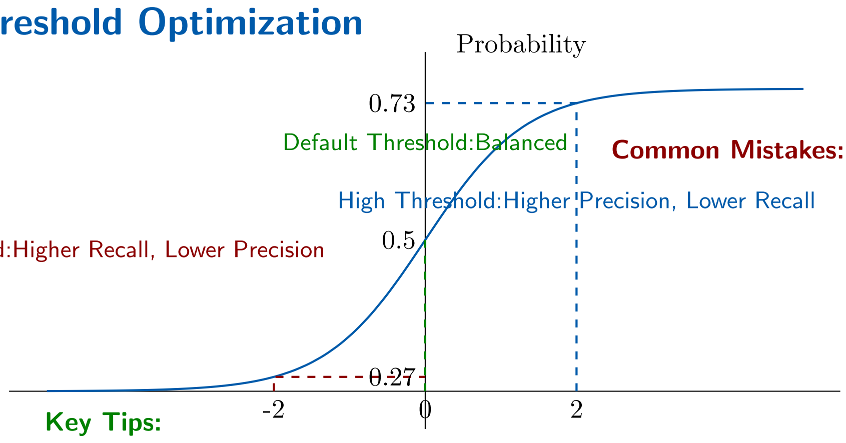
Class-Specific Metrics:

Class	Precision	Recall	F1-Score
Cat	0.85	0.92	0.88
Dog	0.91	0.78	0.84
Bird	0.72	0.83	0.77
Average	0.83	0.84	0.83

Averaging Methods:

Macro Average: Simple average across classes
Weighted Average: Weighted by class frequency
Micro Average: Calculated from summed TP, FP, FN

Threshold Optimization



Common Mistakes:

Focusing only on overall accuracy
Ignoring performance on minority classes
Using macro-averaging for highly imbalanced data

Key Tips:

Always examine per-class performance in multi-class problems
Choose appropriate averaging method based on class distribution
Pay extra attention to minority classes

Different applications have different needs:
Medical Testing: Higher recall to catch all cases (low threshold)
Spam Detection: Higher precision to avoid false alarms (high threshold)
Balanced Use Case: Default threshold (typically 0.5)

Best Practices & Common Pitfalls

Optimize threshold based onyour application's needs!

DOs

Compute precision, recall, F1-score for each class
Plot ROC and PR curves, analyze both
Use AUC-ROC for balanced data; AUC-PR for imbalanced data
Present confusion matrix to understand misclassifications
Optimize thresholds based on your specific use case

DON'Ts

Don't rely solely on accuracy, especially with imbalanced data
Don't report metrics on validation data post-hoc—use a dedicated test set
Don't ignore per-class performance in multi-label problems
Don't skip threshold optimization if using sigmoid outputs
Don't focus on a single metric without context