Lesson 27: Classification Metrics

Data Science with Python – BSc Course

45 Minutes

Learning Objectives

The Problem: Our classifier has 95% accuracy – is that good? What if 95% of samples are one class? How do we properly evaluate classification models?

After this lesson, you will be able to:

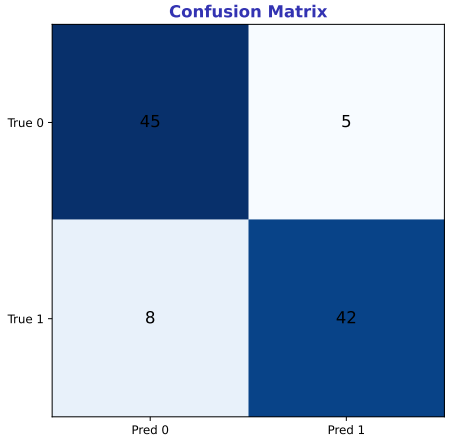
- Build and interpret confusion matrices
- Calculate precision, recall, and F1 score
- Plot and interpret ROC curves and AUC
- Choose metrics appropriate for your problem

Finance Application: Evaluating fraud detection and default prediction models

Confusion Matrix

The Foundation of Classification Metrics

- 2x2 table: TP, TN, FP, FN (True/False Positive/Negative)
- All other metrics derive from these four numbers

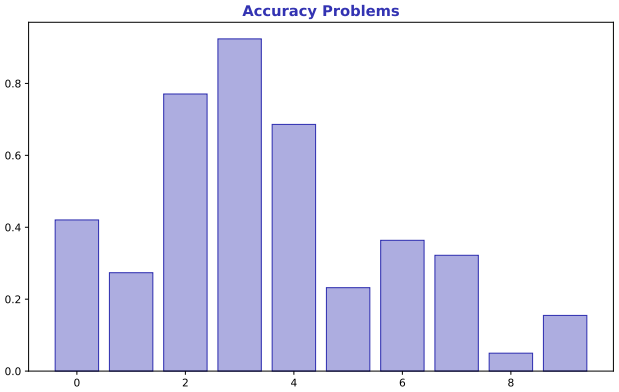


TP = correct positive, FP = false alarm (Type I), FN = missed positive (Type II)

The Accuracy Trap

When Accuracy Misleads

- 1% fraud rate: predicting "no fraud" gives 99% accuracy
- But you catch zero frauds – useless model

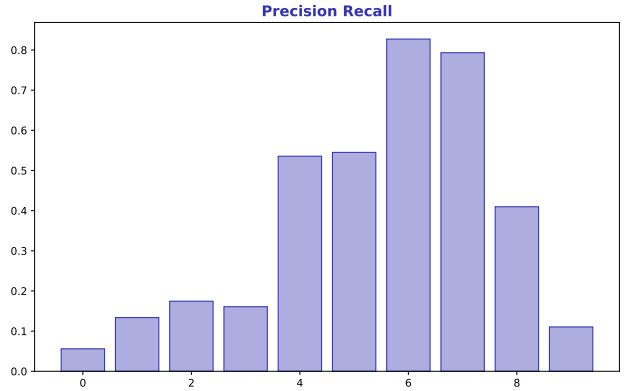


Rule: Never use accuracy alone when classes are imbalanced

Precision and Recall

Two Perspectives on Model Quality

- Precision = TP / (TP + FP) – "Of predicted positives, how many correct?"
- Recall = TP / (TP + FN) – "Of actual positives, how many found?"

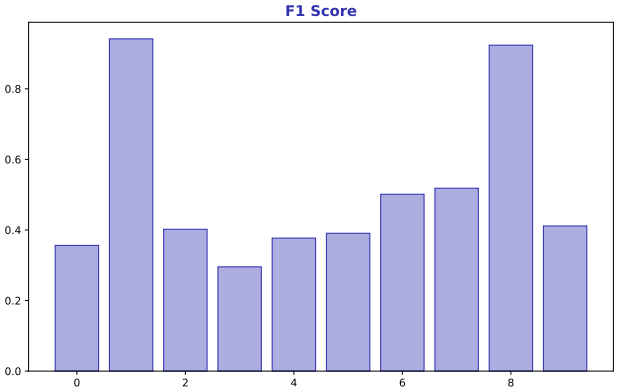


Trade-off: High threshold = high precision, low recall. Low threshold = opposite.

F1 Score

Balancing Precision and Recall

- $F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$ – harmonic mean
- Single number when you need both precision and recall

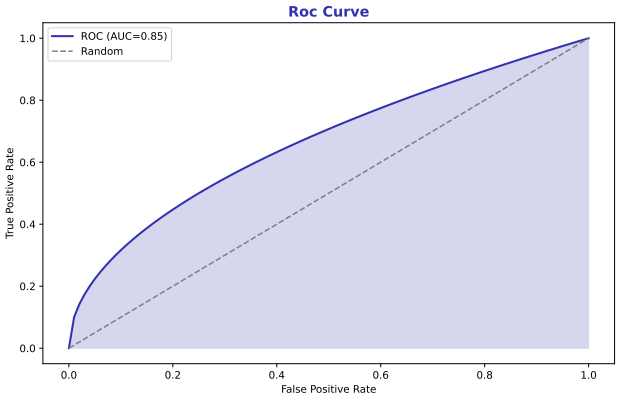


F1 = 0 if either precision or recall is 0. Max F1 = 1 (perfect).

ROC Curve

Performance Across All Thresholds

- X-axis: False Positive Rate (FPR). Y-axis: True Positive Rate (Recall)
- Each point is a different decision threshold

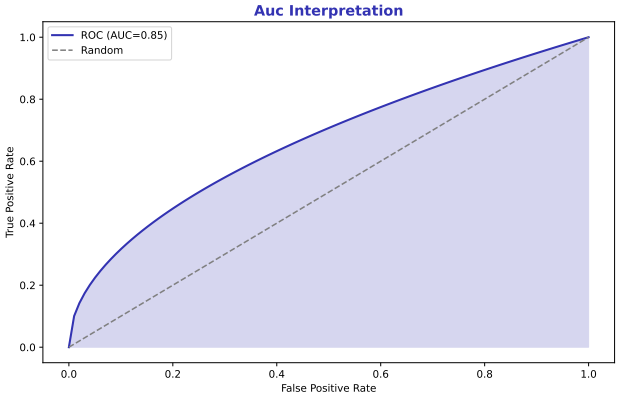


Perfect classifier: curve goes to top-left corner (TPR=1, FPR=0)

AUC Interpretation

Area Under the ROC Curve

- AUC = 0.5: random guessing. AUC = 1.0: perfect separation
- Interpretation: probability that model ranks positive higher than negative

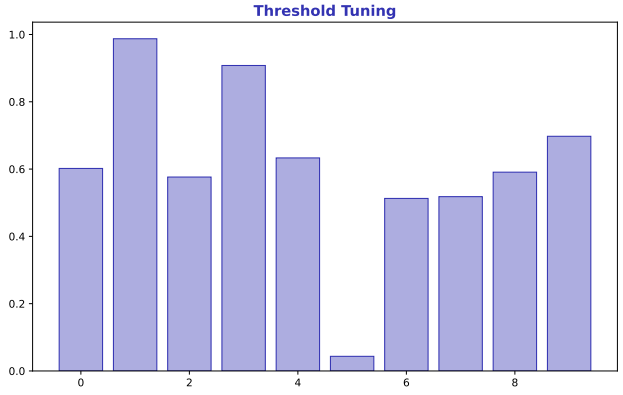


Industry benchmarks: AUC $>$ 0.7 acceptable, $>$ 0.8 good, $>$ 0.9 excellent

Threshold Tuning

Choosing the Right Operating Point

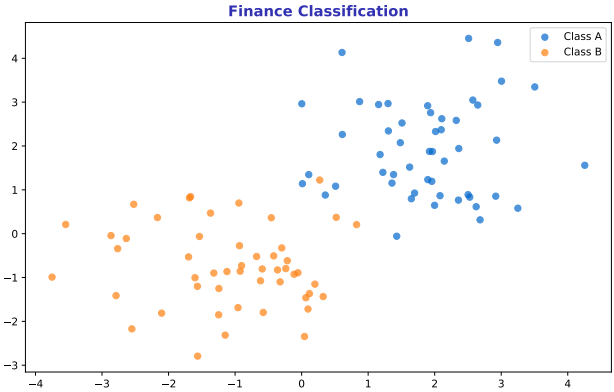
- Default threshold (0.5) is rarely optimal
- Tune based on business costs: cost of FP vs cost of FN



Example: Fraud costs $1000, investigation costs $10 – lower threshold is better

Metrics for Financial Problems

- Default prediction: high recall (catch all defaults), accept lower precision
- Trading signals: high precision (avoid false signals), accept lower recall



Match metric to cost structure: what's worse, missing a default or false alarm?

Hands-On Exercise (25 min)

Task: Evaluate a Default Prediction Model

1. Fit logistic regression on credit data (target: default yes/no)
2. Generate confusion matrix with `confusion_matrix()`
3. Calculate precision, recall, F1 using `classification_report()`
4. Plot ROC curve and calculate AUC
5. Try thresholds 0.3 and 0.7 – how do metrics change?

Deliverable: Confusion matrix heatmap + ROC curve with AUC annotation.

Extension: Calculate expected cost at each threshold using custom cost matrix

Lesson Summary

Problem Solved: We now have a toolkit of metrics beyond accuracy for proper model evaluation.

Key Takeaways:

- Confusion matrix: TP, TN, FP, FN – foundation of all metrics
- Precision = quality of positives, Recall = coverage of positives
- ROC/AUC: threshold-independent performance measure
- Choose metrics based on business costs

Next Lesson: Class Imbalance (L28) – handling rare events

Memory: Precision = Positive predictions that are correct. Recall = Positives that we Recovered.