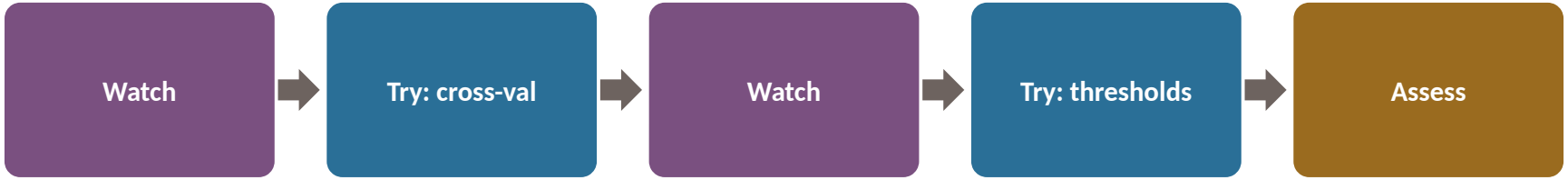
MODULE 5

Rigorous model evaluation

Training error flatters. This module is about the number you can actually defend, and its uncertainty.

Watch a little, then do a little



Each recorded segment is short. You pause, practice the concept hands-on, then return for the next piece.

Where this module sits in the lifecycle

Evaluation sits inside the model layer but governs everything: it decides whether a result is real. The through-line is that every preprocessing step which learns from data must be fit inside the training folds only.

PART 1 OF 2

Holdout, cross-validation, and leakage

About 12 minutes on camera

The holdout ladder

Training error is an optimistic, biased estimate of future performance. The ladder buys a more stable, honest estimate as you climb it, at the cost of compute.

How leakage sneaks in

Fit-before-split

A scaler, imputer, selector, or target encoder fit on the full data before splitting.

Temporal leakage

Using future information to predict the past because rows were shuffled across time.

Grouped records

Duplicate or grouped rows straddling the train/test boundary so the model has seen its test.

The tell

A held-out score that looks too good is usually leakage, not skill.

NOW TRY IT

MLU-Explain Cross-Validation + Bias-Variance

- Step through how folds are formed and how the averaged estimate stabilizes as k changes.
- In Bias-Variance, see how flexibility moves you between under- and over-fitting and where test error is minimized.

About 15 minutes, then come back for Part 2.

PART 2 OF 2

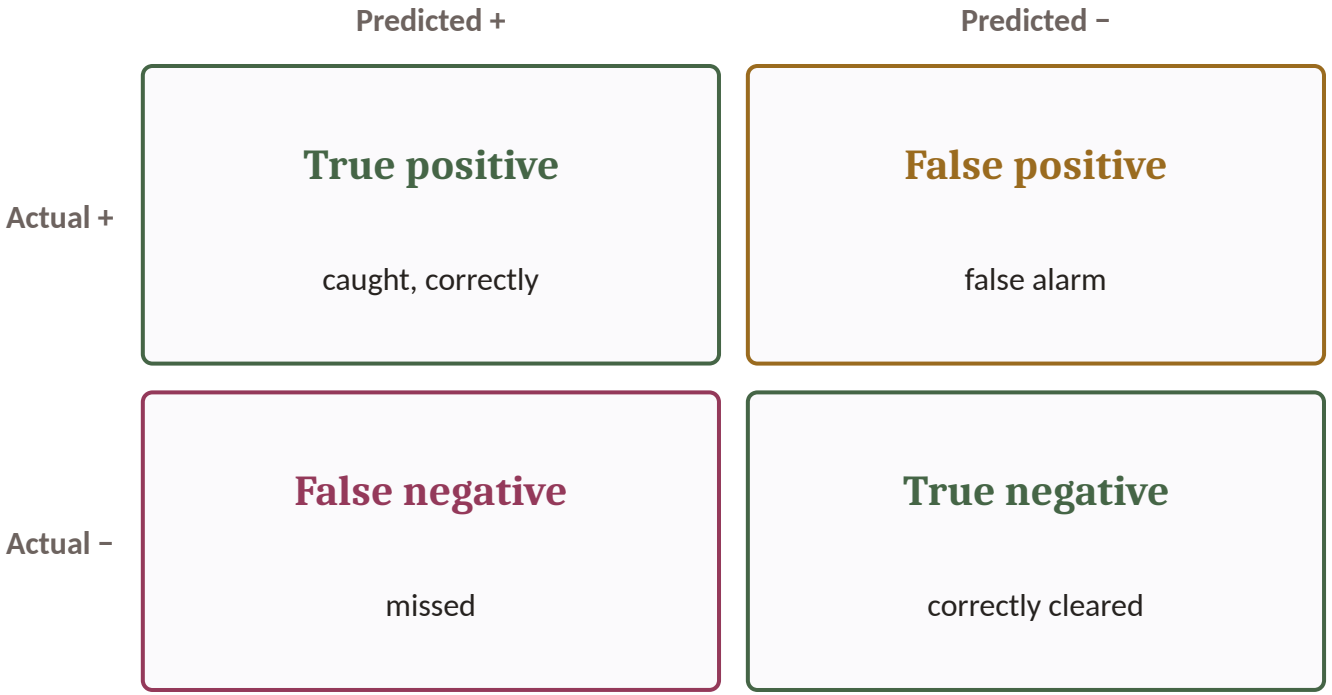
Honest metrics and uncertainty

About 13 minutes on camera

97% accuracy can be worse than useless

- Under class imbalance a high accuracy can hide total failure on the rare positive class.
- Read the confusion matrix, then precision, recall, and F-scores, to see what the headline number conceals.
- A classifier outputs a score; the decision threshold turns scores into labels, and ROC and PR curves summarize every threshold at once.

Read the confusion matrix, not the headline



A metric is a sample, not a fact

- A point estimate is one draw from a distribution. Report it with a bootstrap or cross-validated confidence interval.
- Interpret the interval correctly: it is about the variability of the estimate, not a probability about the truth.
- Disaggregate: an aggregate score that looks strong can hide a subgroup where the model collapses.

NOW TRY IT

Thresholding and confidence intervals

- In ROC & AUC, move the threshold and watch the confusion matrix update; note how one model traces a whole curve.
- In Google MLCC Thresholding, use the imbalance slider to feel precision and recall respond as the positive class becomes rare.
- In Interpreting Confidence Intervals, watch how intervals from repeated samples behave.

About 20 minutes, then build the formative.

Formative deliverable (completion points)

- Take any model you have trained (or a provided baseline) and pick one metric appropriate to its target.
- Report that metric with either a bootstrap or a cross-validated confidence interval.
- Add one sentence explaining why the point estimate alone would mislead a decision-maker.

How it counts. Submit the work with a short reflection or evidence you ran it. Credited on completion, not scored for correctness. This is the practice that makes the graded simulation go well.

GRADED ASSESSMENT

The Ridgeline Renewables model audit

- A departing contractor left a “production-ready” failure-prediction model scoring 0.97 AUC, with a note recommending immediate deployment.
- The handoff hides planted defects: target leakage, train/test contamination, and a metric chosen to flatter under imbalance.
- Your job is to find them, quantify the honest performance, and recommend for or against deployment with evidence.