MONASH University

WEEK 5 TUTORIAL ACTIVITY — FORECAST FORENSICS

AI for Prediction & Forecasting: Part 1 · LO2, LO5 · ~60 minutes

> **Goal.** Become a "forecast forensics" team: diagnose *why* four forecasts are untrustworthy — class imbalance (the accuracy lie), data leakage, spurious regression, and miscalibration — and say exactly how you'd fix each. This is the judgement you'll put to work in Part 2's build.

Teams of 3–4 (cross-campus via Google Meet). Project deck: `BEX3202_week_05_prediction_pt1_activity_slides.pptx`.

PART A — DIAGNOSE THE FOUR CASES (≈30 MIN)

Four forecasts are untrustworthy. For **each**: name the **methodological failure**, say **what test or diagnostic exposes it**, and give the **fix**.

	Case	The failure?	What exposes it?	The fix →
A	A fraud model reports **96% accuracy**; on inspection its **recall is 12%** and the fraud rate is 3%.			
B	A team shuffled the monthly series into a **random 80/20 split** and got a near-perfect RMSE.			
C	An analyst regresses monthly revenue on a **trending macro index**, gets **$R^2 = 0.95$**, and declares a relationship.			
D	A forecast was shown as **a single line**; the team stocked to it and were caught out when demand fell in the (unshown) lower tail.			

PART B — COMPARE & GENERALISE (≈20 MIN)

- **Compare across teams.** Where did teams disagree on the failure or the fix? Argue the close calls.
- **Name the one check** that would have caught the most failures. (Hint: honest evaluation on *held-out time*, with a baseline and an interval.)
- **Turn it into a rule you'll use in Part 2:** every model must (1) **beat a baseline** on rolling-origin cross-validation, and (2) **report a calibrated interval**, not a single line — and you **never shuffle time**.

DEBRIEF (≈10 MIN, WHOLE CLASS)

The four diagnoses:

- **A = class imbalance / the accuracy lie** → look at recall and PR-AUC, use precision–recall, not raw accuracy.

- **B = data leakage** → use rolling-origin CV; the test set is always strictly *later* than training.
- **C = spurious regression** on non-stationary (I(1)) series → test with ADF/KPSS, then difference or check cointegration.
- **D = miscalibration / no interval** → report a prediction interval and check its coverage (PICP).

A defensible forecast **beats a benchmark on held-out time** and **shows its uncertainty** — that's the LO5 thinking Assessment 3 rewards, and the discipline you'll apply hands-on in **Week 6 (Part 2)**.

FACILITATOR NOTES (DO NOT DISTRIBUTE)

- **Push for precise vocabulary** — the *name* of the failure and the *diagnostic* (recall/PR-AUC, rolling-origin CV, ADF/KPSS, PICP), not just "it's wrong."
- **The single most common real error is B (leakage / shuffling time)** — make sure every team can articulate why a random split of a time series is invalid.
- **Cross-campus:** one shared forensics doc per team; coaches join the Meet breakout rooms. This week is pure diagnosis — the build happens in Part 2.