PREDICTIVE MAINTENANCE · FIELD CARD 08

The five evaluation denominators

Take an asset that fails twice a year and call it sick for the two hours before each failure: a model that says “healthy” every time scores 99.95% accuracy and catches nothing. The number moves with the bookkeeping, never with the model. Five questions replace it.

RECORD

Five questions, answered against the plant’s own ledger

01 **When it speaks, is it right?** Confirmed finds over *resolved* alerts, with the treatment of alerts still on watch declared before the count. Nine of twenty-eight resolved is 32%; counting three watches as false makes it 29%.

02 **When it matters, does it speak?** Flagged events over every event anyone eventually found, including the misses an RCA surfaced. Failures nobody ever attributed appear in no denominator, so this figure is an upper bound and is labelled one.

03 **Is it worth the attention?** False alarms over assets × months. Nineteen across four assets over six months is 0.8 per asset-month; the denominator, not the count, is what makes two quarters comparable.

04 **Did it buy usable time?** Lead time from first flag to the act-by the planner used, net of the response path — the median with its spread beside it, every claimed window logged against the as-found stage.

05 **Do its confidences mean anything?** Of the calls labelled high-confidence, what fraction held — over the count of calls in that band. At six such calls, a correct ordering is an encouraging hint, and the report prints that sentence beside the fraction.

ASK

Before the number goes on a slide

- **Which alerts were counted, and which were left out?**
  The convention moves the number more than the model does. Write it beside the figure, or the figure is not comparable to anything — including itself, last quarter.
- **Does the difference clear the noise?**
  32% on 28 resolved runs about 18–51%; 58% on 12 runs about 32–81%. They overlap from 32 to 51. Reporting that as “32% → 58%” with the intervals left off is how a programme believes a change that may be noise.
- **Was it graded on what was knowable at the time?**
  The events used to tune a threshold cannot be the events used to report it. Hold out by time *and* by whole assets: rows from one machine share its mounting and duty.
- **Costs, or counts?**
  A missed airend failure and a false ultrasound flag differ by orders of magnitude in consequence. Priced, a model with worse precision and better recall on the expensive mode is the better model.
- **What is n?**
  At nine confirmed events, quote the interval and rank the judges rather than crowning one. Where the count cannot carry the claim, the honest entry is **NOT DEMONSTRATED**, not a rounder number.

THE BOUNDARY THIS CARD CARRIES

No single accuracy figure appears here, and none should appear in the report this card feeds. The intervals quoted are Wilson score intervals at 95%, worth nothing until recomputed on your own counts. This card sets no target precision, recall, false-alarm budget, lead time or pass mark, licenses no model, establishes no diagnosis, and authorises no change to an operating or maintenance requirement, inspection, task or protective function.