

Exercise walk-through

13.7

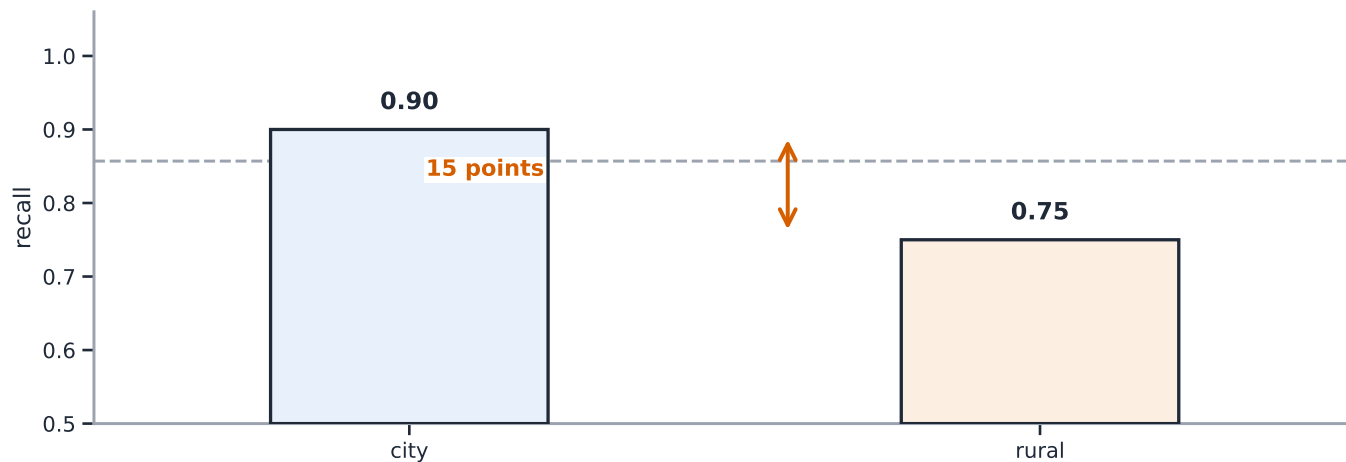
Fairness and limitations

University Computer Science

Fairness and limitations: audit per group

- 1 **Gap** City **recall** $27/30 = 0.90$; rural $9/12 = 0.75$ - 15 points: the model misses 1 in 4 rural lates but 1 in 10 city lates.
- 2 **Cause** Rural lates are 13% of the region vs city 29% - the thinner class is under-served; fix records or weights, not the metric.
- 3 **Card** Overnight is synthetic, 200 records, two regions, one season: the card states intended use, the 15-point gap, out-of-scope uses.

Error disparity chart



city TP 27 FN 3 (30 lates); rural TP 9 FN 3 (12); rural lates 13% vs city 29%

Recall by region: city 0.90, rural 0.75.

Check yourself

- Course graph A* expansions before the goal...? (Four.)
- $P(\text{Storm} \mid \text{Complaint})$ with the course numbers...? (0.667.)
- The course line...? ($y = 0.9x + 1.3$.)
- City vs rural **recall**...? (0.90 vs 0.75.)

Exercises 2.1-2.3

2.1 Why does overall accuracy hide the gap?

[3]

2.2 Where does the disparity most likely start?

[3]

2.3 Name two fairness criteria that can conflict.

[3]

Exercises 2.1·2.3

2.1 Why does overall accuracy hide the gap? [3]

2.1

- The majority region dominates the pooled counts – the minority's mistakes are diluted to decimals ✓.

2.2 Where does the disparity most likely start? [3]

2.2

- In the data: the thinner minority class under-serves the boundary ✓.

2.3 Name two fairness criteria that can conflict. [3]

2.3

- Equal-error-rate parity vs demographic parity (any two of the named criteria that conflict) ✓.

Exercise 3.1

3.1 A second audit adds a third region: mountain, with 6 later of 40 records, recall $3/6 = 0.50$.

Recompute the per-region recall table and the worst disparity in points. State the cause-level explanation for mountain and the data-side fix. Give the class-weighting alternative and its risk. State what the model card must now record. [8]

Exercise 3.1

3.1 A second audit adds a third region: mountain, with 6 later of 40 records, recall $3/6 = 0.50$.

Recompute the per-region recall table and the worst disparity in points. State the cause-level explanation for mountain and the data-side fix. Give the class-weighting alternative and its risk. State what the model card must now record. [8]

3.1

- Table: city 0.90, rural 0.75, mountain 0.50 – worst disparity 40 points (city vs mountain) ✓.
- Cause: mountain later are $6/40 = 15\%$ of records – thinnest class; the boundary is fitted mostly on city data ✓.
- Data-side fix: collect more mountain records (or rebalance the sampling) so the class is not 15% ✓.
- Class weighting: weight mountain later more in the loss – risk: precision in mountain may collapse (more false later) ✓.
- Card: record the three-region recall table and the 40-point gap as a measured disparity ✓.

Exercise 3.2

3.2 The owner proposes fixing all disparity by moving each region's threshold until recalls match.

State what parity criterion this implements. Give one reason it can be the right call and one reason it can mislead. Propose the evaluation that would justify (or veto) the move. State the limitation-statement duty the owner retains either way. [7]

Exercise 3.2

3.2 The owner proposes fixing all disparity by moving each region's threshold until recalls match.

State what parity criterion this implements. Give one reason it can be the right call and one reason it can mislead. Propose the evaluation that would justify (or veto) the move. State the limitation-statement duty the owner retains either way. [7]

3.2

- It implements equal-error-rate (recall) parity by separated thresholds ✓.
- Right call: the cost of a missed late is the same per customer – parity matches the harm ✓.
- Mislead: thresholds tuned on small validation counts can overfit; and parity here can break precision parity, trading one harm for another ✓.
- Evaluation: per-region precision AND recall with confidence intervals on held-out data, before and after the move ✓.
- Duty: the limitation statement must disclose the separated thresholds and the measured residuals either way ✓.