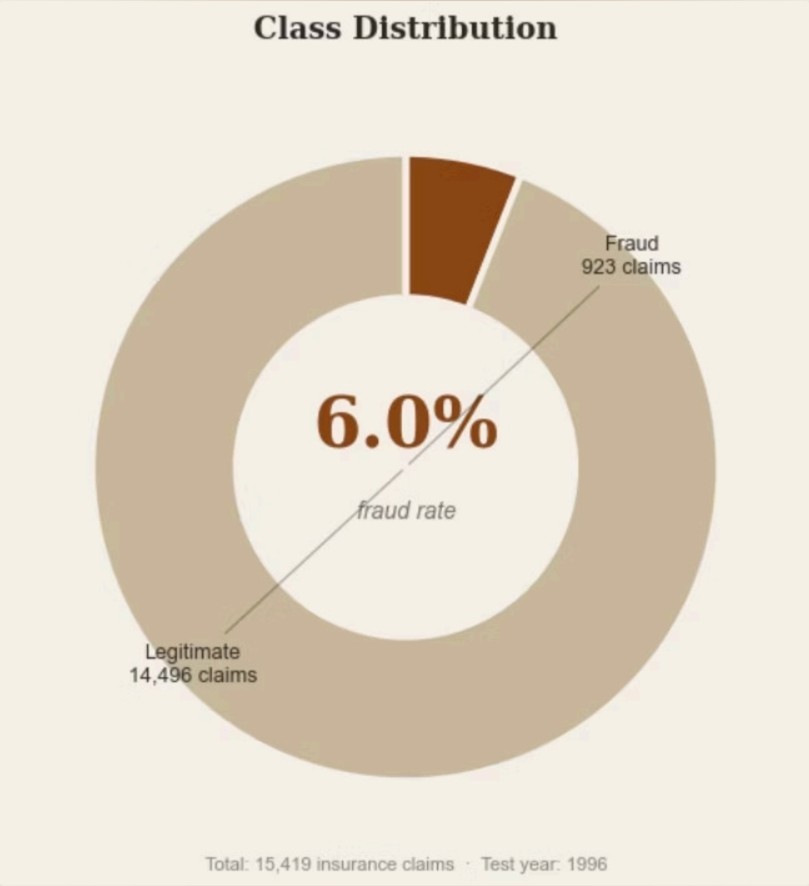
Fraud Risk Prioritisation for Motor Insurance Claims

Decision-support ranking that helps investigators prioritise which claims to review.



Executive summary in plain terms

We trained a model on claims from two years of data and tested it on most recent year. We use a year-by-year split to reflect real deployment on future claims and avoid overly optimistic results from mixing years. Fraud is rare (~6%), so the goal is to concentrate likely fraud into a small investigation queue. Investigating the top 10% of claims finds ~21% of that year's fraud (44 cases). Top 20% finds ~44% (94 cases). This means investigators see about twice the fraud they would if claims were chosen at random.

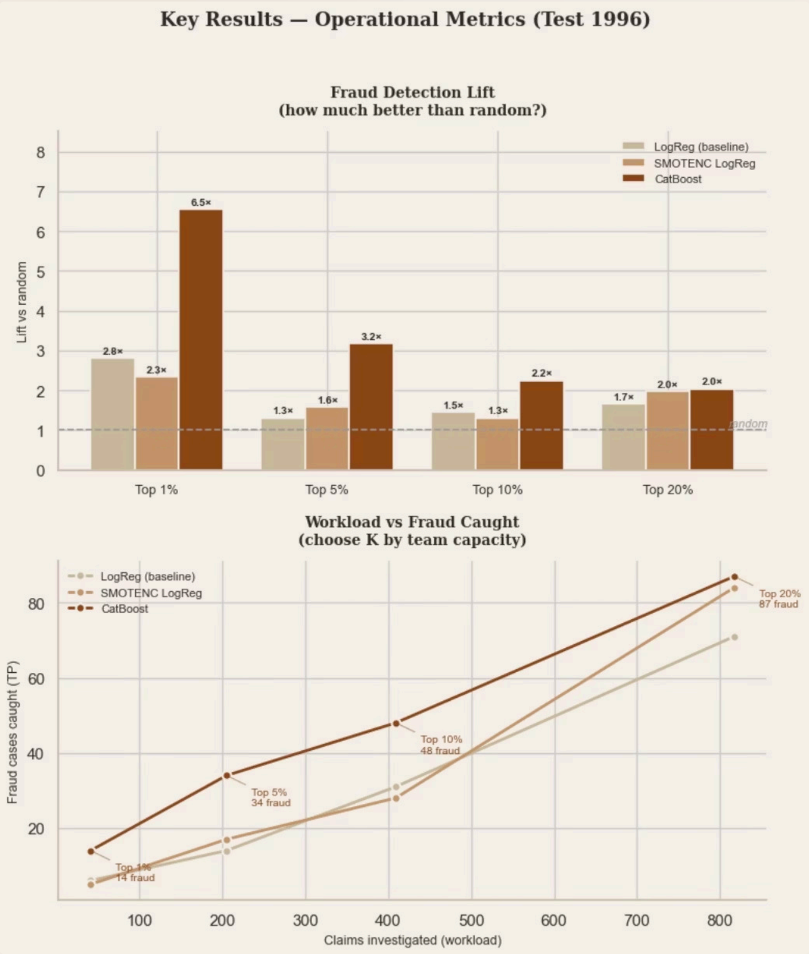
What the model does — and does not

What it does

- Scores and ranks claims by estimated fraud risk.
- Produces a Top-K queue for investigators to review.

What it does NOT do

- It does not prove fraud.
- It does not automatically deny claims or replace human judgement.



Key results — operational metrics (Test 1996)

Top 10% (409 claims)

- Fraud found: 44 (Precision ~10.8%)
- Lift vs random: 2.06×

Top 20% (817 claims)

- Fraud found: 94 (Precision ~11.5%)
- Lift vs random: 2.21×

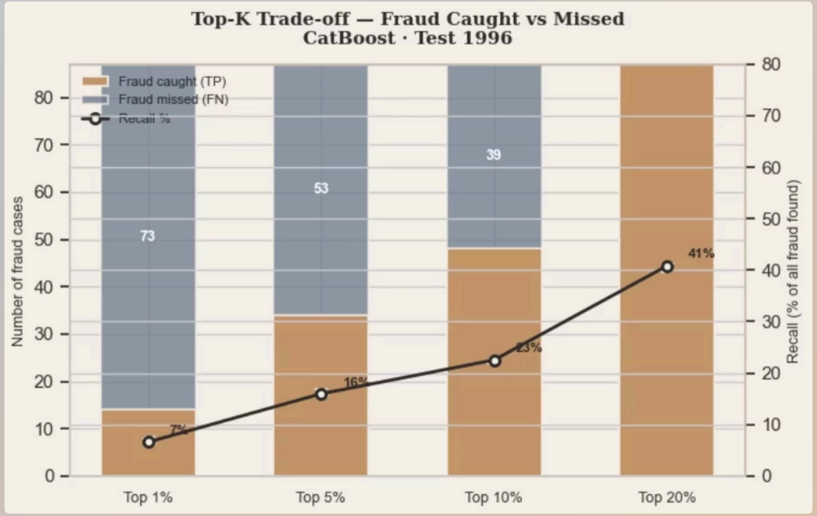
Top-20% captures 94 fraud cases and misses 119 fraud cases (out of 213 total fraud in 1996).

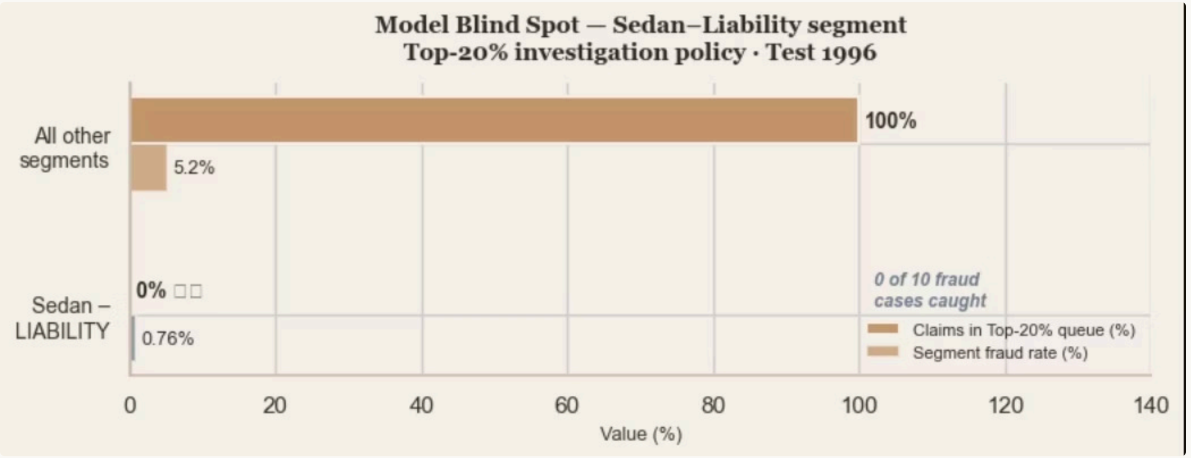
Interpretation: pick K based on investigator capacity and how many missed frauds you can tolerate. These numbers show trade-offs between workload and fraud caught.

Workload vs value — Top-K trade-off

Visual summary: as investigation capacity (K) increases, absolute fraud caught rises but marginal gains fall. Choose K where marginal benefit matches investigator cost.

Diminishing returns: doubling investigations from 10% to 20% increases fraud caught, but also doubles workload.





Model blind spot — segment coverage risk

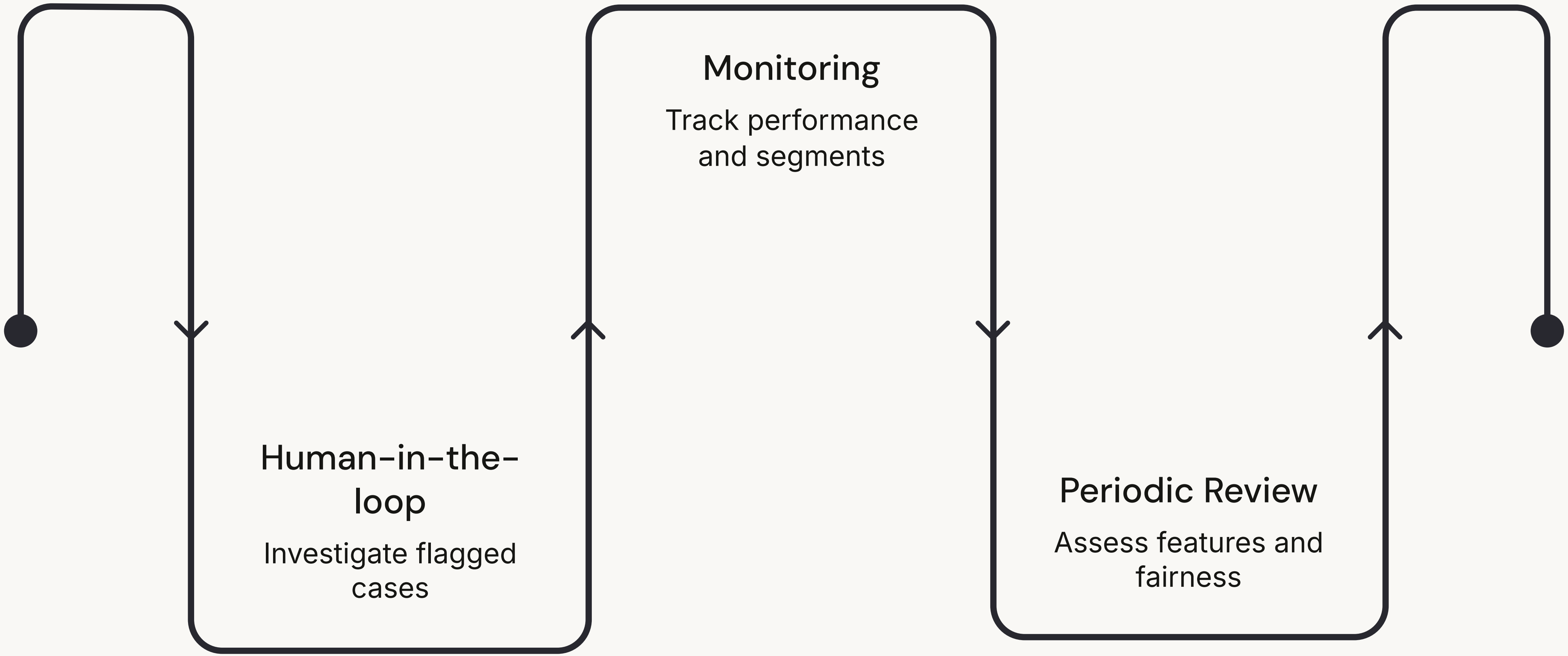
Under Top-20% the large segment "Sedan – Liability" (1,318 claims, fraud ~0.76%) had 0% of its claims included in the queue because its scores were all below the cutoff, so the queue provides 0% coverage.This boosts average efficiency but risks systematically missing rare fraud in that segment.

> Recommendation: reserve a small fraction of reviews for large low-risk segments (stratified sampling) or set a minimum coverage per segment for monitoring.

Explainability — what drives the score

SHAP analysis shows the model mostly uses product and claims process signals (Fault, CoverageType, VehicleCategory, and procedural indicators) rather than demographic attributes. This suggests decisions are driven by claim and product characteristics, not personal attributes — though SHAP shows associations, not causation. This does not guarantee the absence of bias; it only indicates these attributes are not primary drivers in this model.

Optional local check: review individual high-risk examples to validate the reasons before acting.



Before deployment: (1) keep humans in the loop, (2) monitor performance and segment coverage, (3) sample low-risk segments to detect drift and rare fraud.

Strengths, limitations, and proper use

- Strengths

  - Concentrates fraud for limited capacity reviews.
  - Interpretable with SHAP and Top-K metrics.

- Limitations

  - Labels may be noisy (insurer provenance).
  - Only three years of data → limited time generalization.
  - Risk of blind spots under strict Top-K rules.

- Appropriate use

  - Decision-support ranking for investigators with monitoring and governance.

Data growth & next steps: practical roadmap

1. Collect more labelled fraud examples from low-risk segments via targeted sampling.
2. Instrument monitoring: Top-K metrics, segment coverage, and feature drift dashboards.
3. Governance: clear policy that score ≠ guilt, human review mandatory, and periodic fairness checks.

With modest data growth and strong monitoring, the prioritisation system will become more reliable and safer for customers and investigators alike, especially in low-risk segments to reduce blind spots and improve coverage.