Predicting NHS RTT Breach Risk Using Open National Data

Saad Ahsan | MSc Data Science (CST4090) | Supervisor: Dr Giacomo Nalli
Middlesex University London

Poster No
02

Scan for Data/code

Open Data
Reproducible pipeline

Early Warning
Month-ahead screening

Explainable
SHAP-driven reasons

CONTEXT & GOAL (WHAT PROBLEM AM I SOLVING?)

RTT waiting times reflect whether elective care is flowing smoothly or getting stuck. During 2023–24, long-wait breaches remained widespread, but risk was uneven across providers and specialties. National reporting describes overall pressure, yet local teams still need short-term signals for prioritisation.

Aim: build an interpretable early-warning model (open data only) that flags provider–specialty services likely to breach next month.

Targets (provider × specialty × month)

- **PropOver18**: proportion of incomplete RTT pathways waiting >18 weeks
- **Exceeded18**: 1 if PropOver18 > 10% (screening threshold)

Why 10%?

It provides a practical "meaningful breach state" for early warning, avoids tiny fluctuations around the 92% constitutional expectation, and focuses attention on services with a growing long-wait tail.

Data (open national view)

Inputs (2023–24)

- **RTT incomplete pathways:** week-band distributions ⇒ long-wait tail
- **HES APC provider-level:** elective/emergency activity + LOS / waiting proxies

Strengths vs blind spots

- **Strength:** transparent and reproducible across providers
- **Blind spots:** staffing, cancellations, theatres, bed state (not in open data at needed granularity)

Quick context (what the data shows)

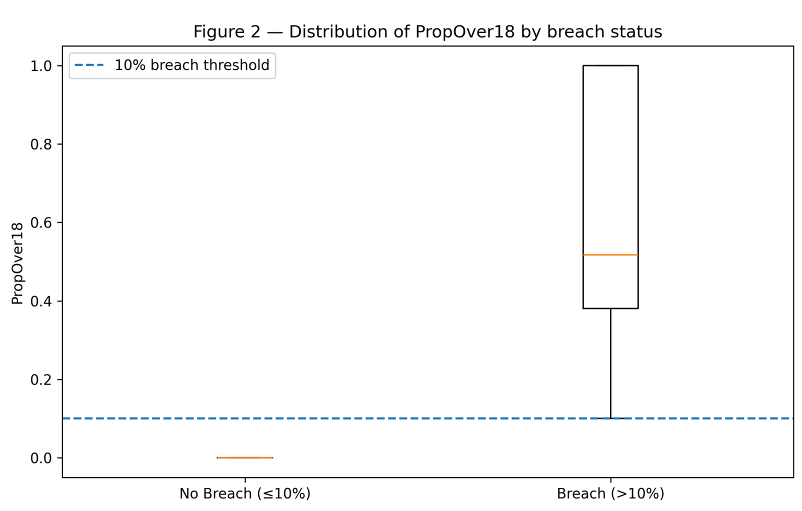


Figure 2 — Distribution of PropOver18 by breach status

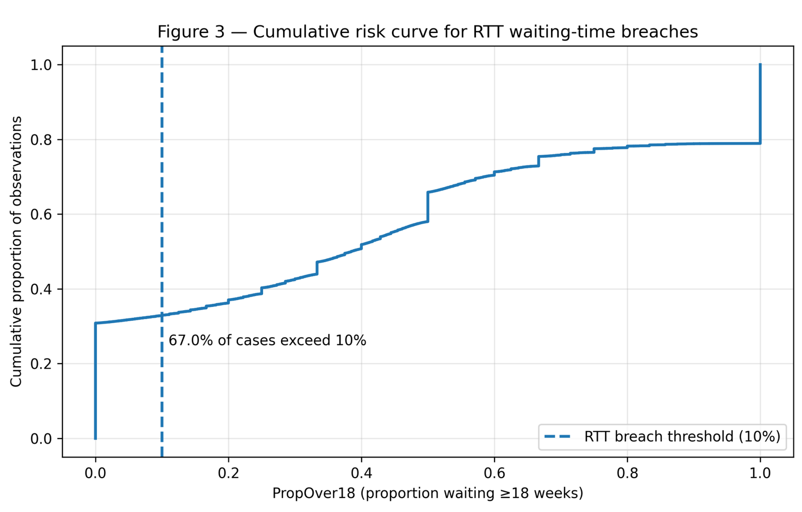


Figure 3 — Cumulative risk curve for RTT waiting-time breaches

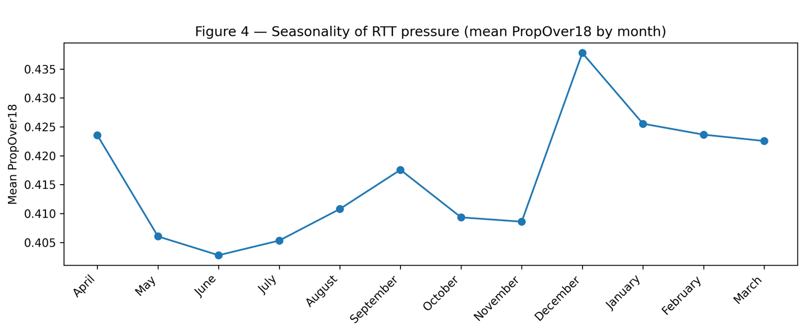


Figure 4 — Seasonality of RTT pressure (mean PropOver18 by month)

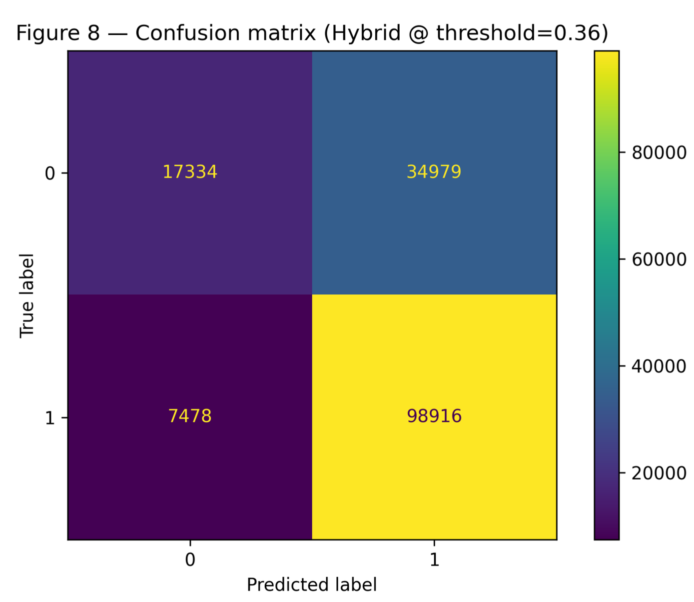


Figure 8 — Confusion matrix (Hybrid @ threshold=0.36)

The distribution shows clear separation plus a borderline band near 10%. Seasonality is visible, supporting explicit month/winter/holiday indicators.

Method overview (end-to-end pipeline)

Mini pipeline (fast to understand)

RTT monthly CSVs
Filter to *Incomplete Pathways*; build `Over18Weeks`, `PropOver18`

HES provider-level APC
Emergency/elective activity + LOS/wait proxies

Merge on provider code
Create provider–specialty–month panel

Feature engineering
Lagged breach + list size + seasonality + pressure proxies

Models
Classification Baseline; Global XGBoost followed by Hybrid (global + per-specialty) threshold optimisation.
Regression Baseline; Ridge Regressor followed by Random forest and XGBoost Regressors.

Explainability
SHAP global + local reasons for flags

Why two tasks? (Classification vs Regression)

They answer different operational questions. Classification supports **escalation triage** (breach vs non-breach). Regression supports **severity ranking** once a service is flagged.

What the model is learning?

The classifier mainly learns whether a service has escaped its long-wait tail before. If not, it assumes risk persists and then adjusts that risk based on list size, emergency pressure, and seasonal context. In effect, it behaves conservatively: breach states are assumed to continue unless there is strong evidence of recovery.

Evaluation

Metrics emphasise operational asymmetry: high recall is preferred (missing a deteriorating service can be costly). ROC–AUC summarises discrimination, while PR curves and the confusion matrix show the trade-off at the chosen threshold.

Model selection visuals

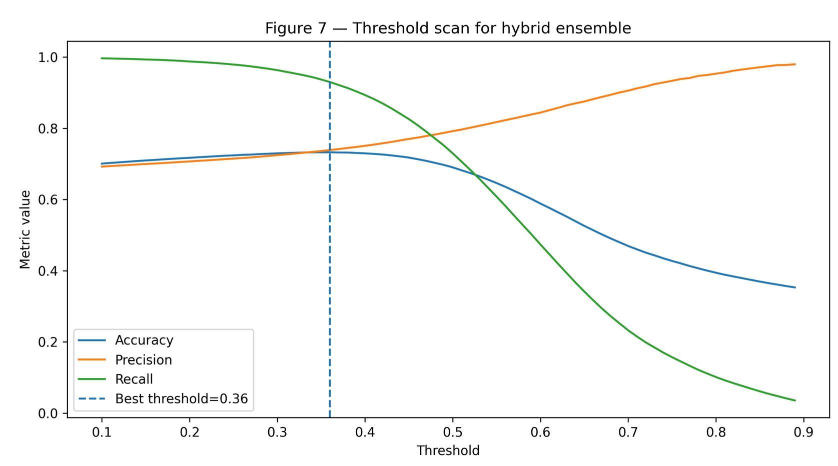


Figure 7 — Threshold scan for hybrid ensemble

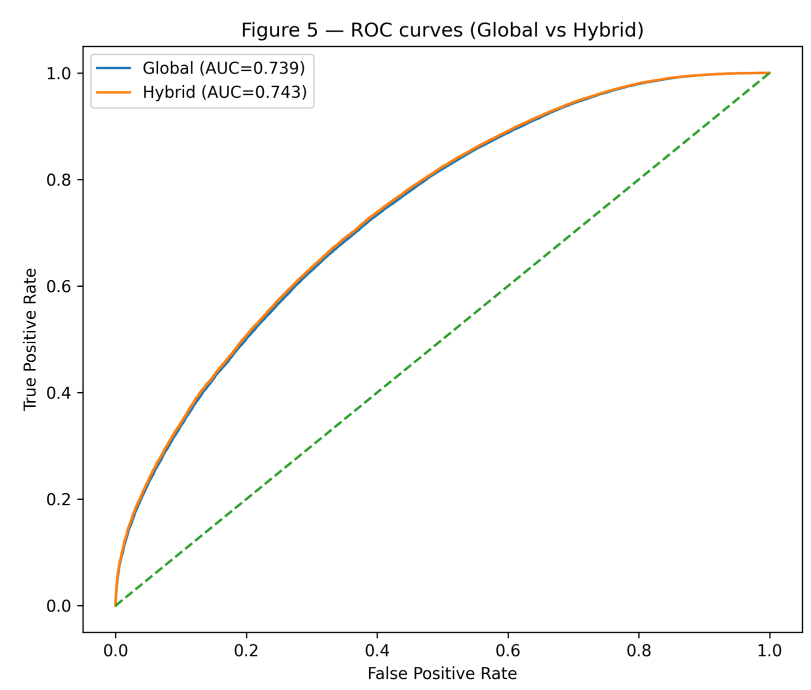


Figure 5 — ROC curves (Global vs Hybrid)

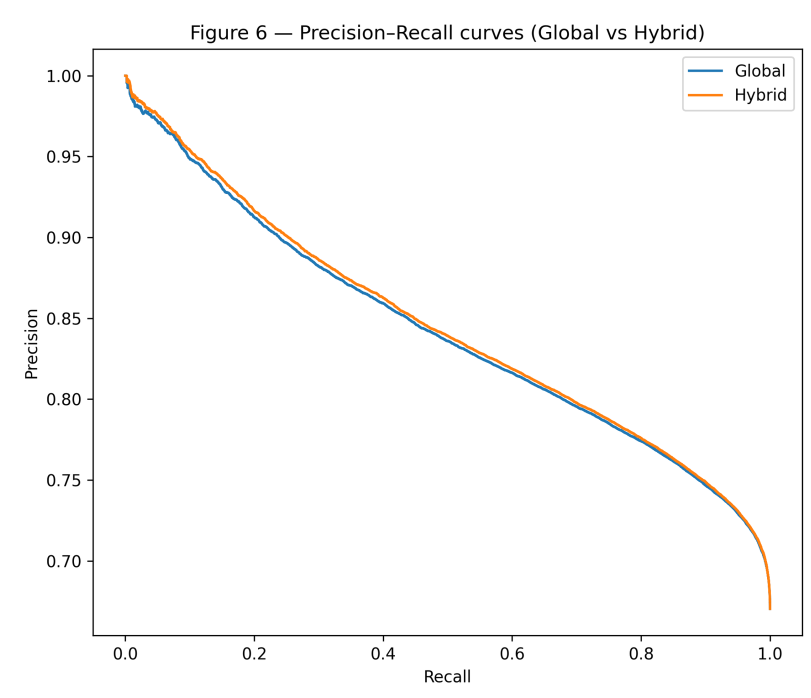


Figure 6 — Precision–Recall curves (Global vs Hybrid)

Threshold scanning selects an operating point aligned with screening (high recall). ROC/PR curves summarise discrimination and the precision trade-off.

Results (headline impact)

Best classifier (final)

Hybrid XGBoost + threshold optimisation
(operating threshold ≈ 0.36). This prioritises recall so genuinely at-risk services are rarely missed.

Accuracy: 0.73 **ROC–AUC: 0.74**
Recall: 0.93 **F1: 0.82**

How to interpret a risk score

A higher probability does not mean failure is certain. It means the service resembles past provider–specialty–months that tended to remain in breach. Scores work best for ranking and triage, not as an automated decision rule.

Why classification beats regression here

Predicting exact PropOver18 requires visibility of short-term operational actions (extra lists, cancellations, staffing changes). These are not observable in open data. By contrast, breach-state prediction relies more on persistence and scale, which are visible, producing stronger and more stable classification performance.

Regression— "How Severe is the breach?"

Target: PropOver18 (continuous proportion beyond 18 weeks)

Regression aims to estimate **breach intensity** (e.g., 12% vs 35% long waits). This is harder because precise proportions respond to local actions not visible in open data (extra lists, cancellations, staffing shifts). For that reason, regression is positioned as a **severity ranking tool**, not a perfect numeric forecast.

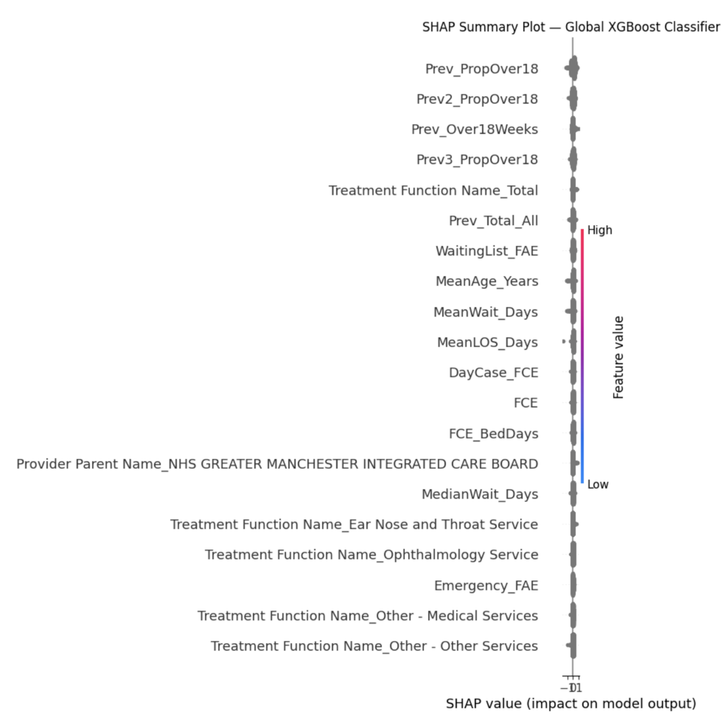
Models compared

- Ridge regression (strong baseline; linear)
- Random Forest regression (non-linear; robust)
- XGBoost regression (non-linear; strong on tabular data)

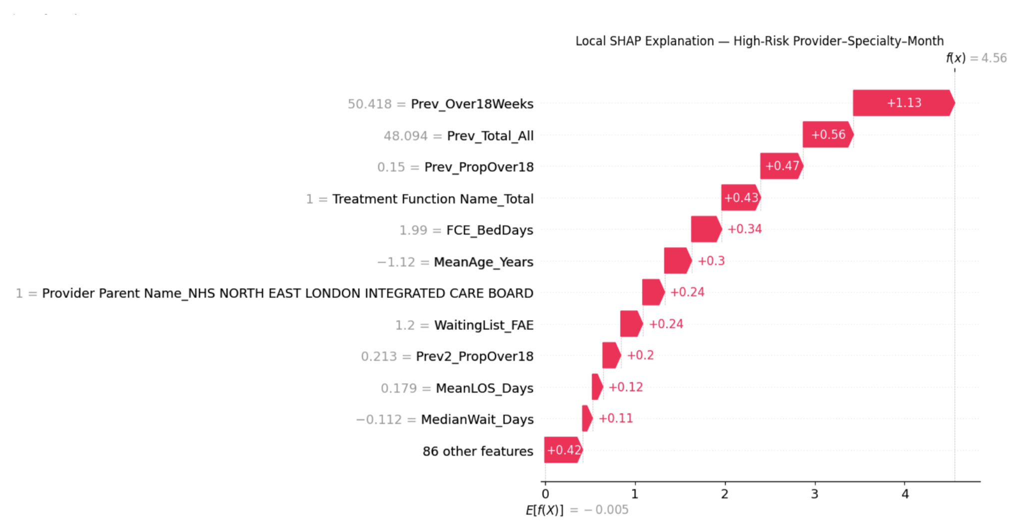
Explainability (SHAP: score → reason)

Consistent top drivers

- Lagged breach proportion (persistence)
- Waiting list size (scale)
- Emergency pressure proxies (context)
- Seasonality (winter + holiday effects)



SHAP Summary Plot — Global XGBoost Classifier



Local SHAP Explanation — High-Risk Provider-Specialty-Month

SHAP makes alerts defensible: "flagged because last month's long-wait tail stayed high, the list is large, and winter pressure increases risk."

Key message: with transparent features and explainable models, open national data can support proactive RTT management—even without privileged operational feeds.