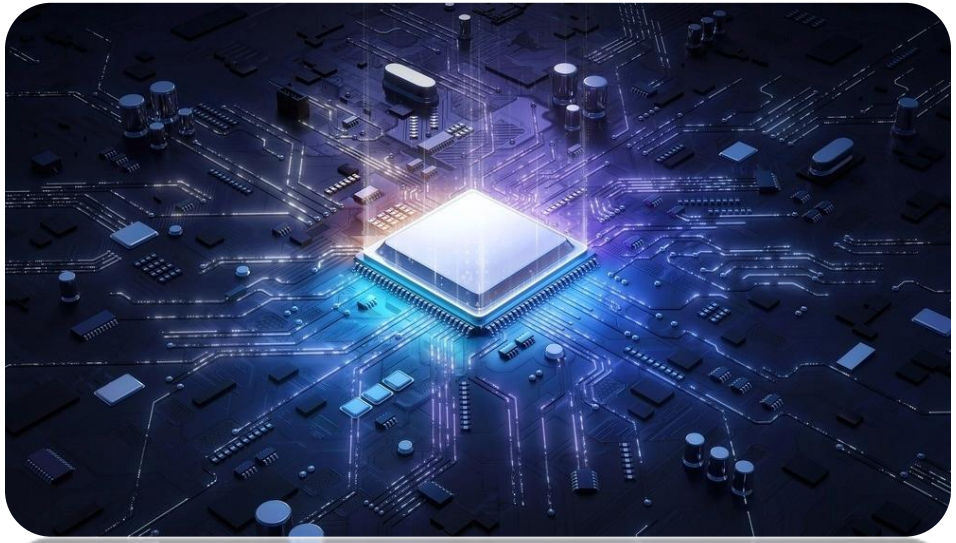
Semiconductor Wafer Yield Prediction

Leakage-Safe • Cost-Aware ML • Chrono Split • Calibration • SHAP

by Himanshu Saini



Outline

- Problem Statement
- Exploratory Data Analysis
- Leakage-Safe ETL Pipeline
- Split & Metrics
- Results
- Interpretability
- Conclusion

Problem Statement & Goal

Problem Statement: In a semiconductor fab, a small fraction of wafer runs (~6.6%) fail final quality control, yet each miss is costly; rule-based alarms either miss excursions or flood operators with false alerts. The telemetry is high-dimensional (590 sensors), partially missing, correlated, and time-ordered, so naive training leaks future information and accuracy metrics are misleading under severe class imbalance. The problem is to predict—before QC release—which runs are likely to fail, producing calibrated probabilities that enable cost-aware thresholds and a controlled alarm load.

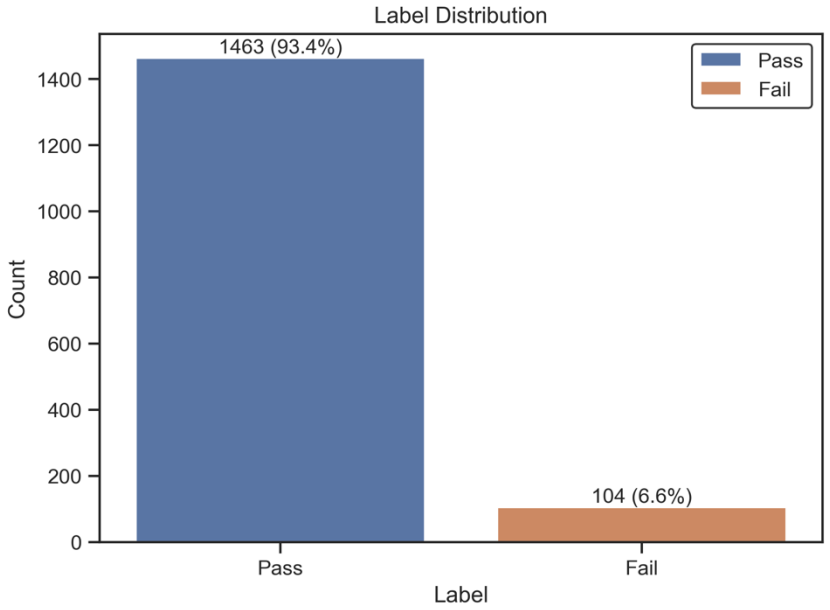
Goal: Deliver a leakage-safe, chronologically validated ML system that outputs calibrated per-run failure probabilities and selects the cost-minimizing threshold for the fab's chosen FN:FP ratio and alarm budget.

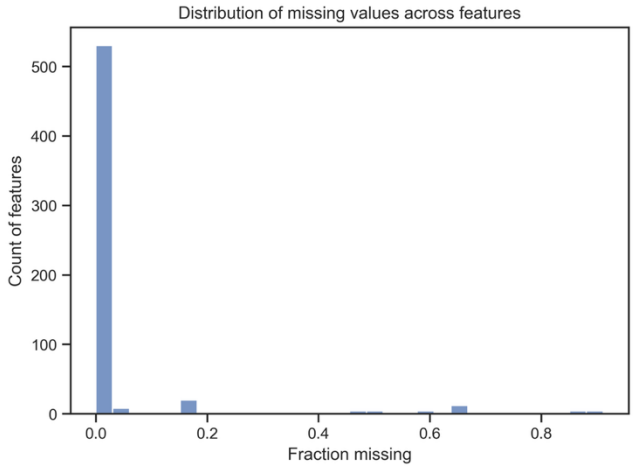
Motivation: Cut rework and downtime, focus engineering on high-risk runs, and manage alarm volume.

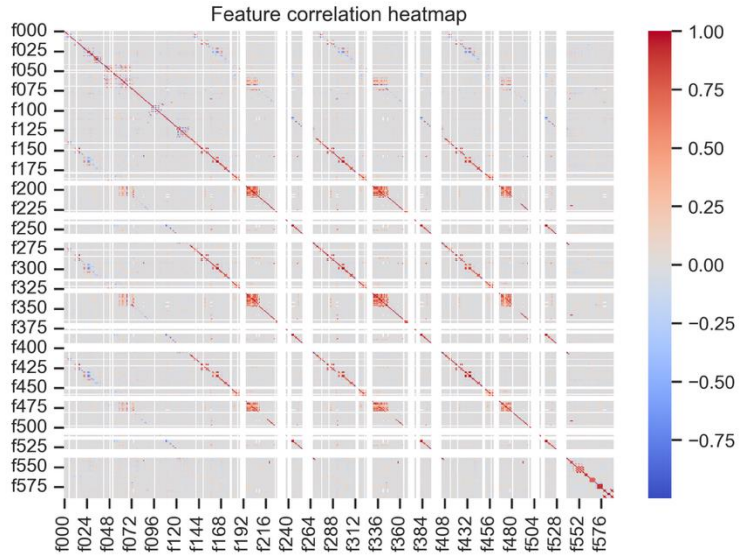
Exploratory Data Analysis (EDA)

EDA summary

- **Data shape:** 1,567 runs, 590 sensors.
- **Distributions:** highly imbalance with 6.6% fail prevalence.
- **Missingness:** median ≈0.3%; 8 sensors >70% dropped; 44 sensors 10–70% flagged.
- **Structure:** correlated sensor blocks → prune at $|r| \geq 0.98$.



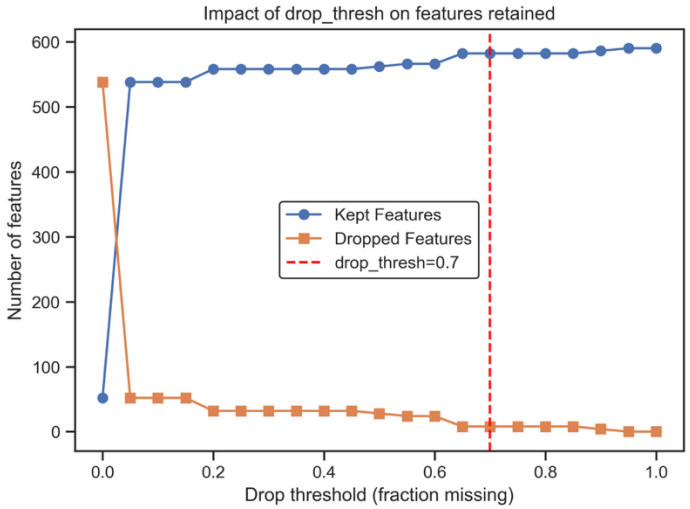


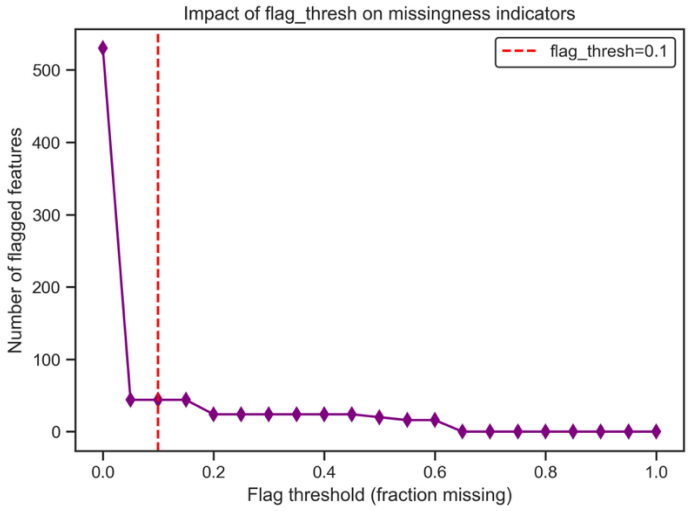


Leakage-Safe ETL Pipeline

ETL summary

- **Missing data:** drop >70% (8 sensors); flag 10–70% (44 indicators); median impute <10%.
- **Outliers:** log1p for skewed non-negative; winsorize 1–99% for heavy tails.
- **Split:** chronological split; timestamp not a feature.
- **Pruning (train data):** low-variance (<1e-8), exact duplicates (38), correlation $|r| \geq 0.98$.
- **Scaling:** Standardize for LR/MLP; trees unscaled.
- **Final feature set:** 375 features (6 indicators).

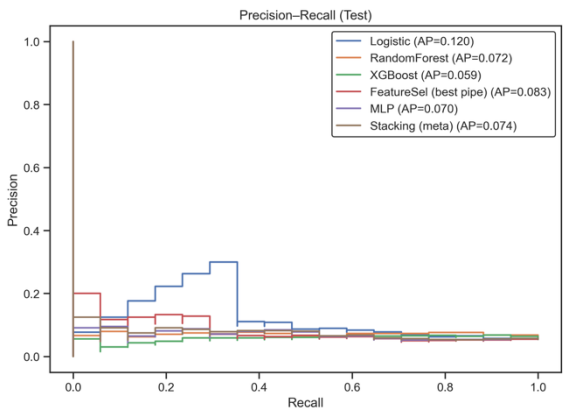


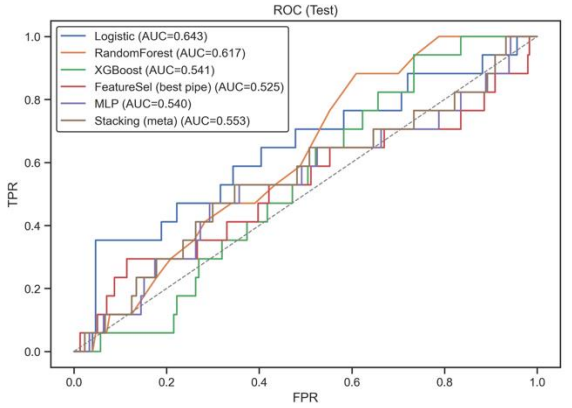


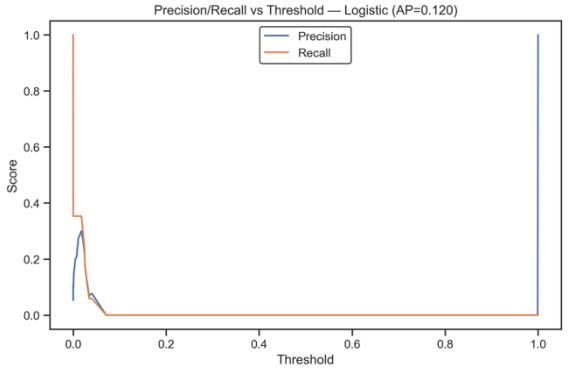
Split & Metrics

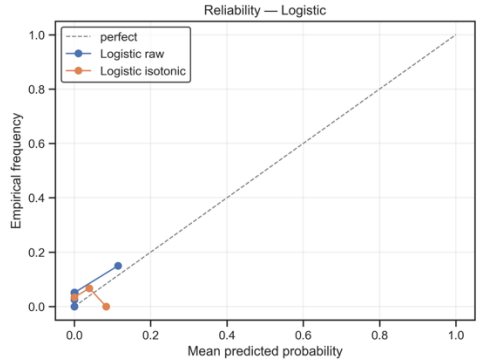
Evaluation design

- **Split:** chronological 60/20/20 by timestamp; fit transformers on train only.
- Thresholds tuned on validation; test used once for final reporting.
- **Primary metric:** PR-AUC (rare-event focus).
- **Secondary:** ROC-AUC, Balanced Accuracy.
- **Calibration on validation:** isotonic and sigmoid; thresholds re-picked post-calibration.





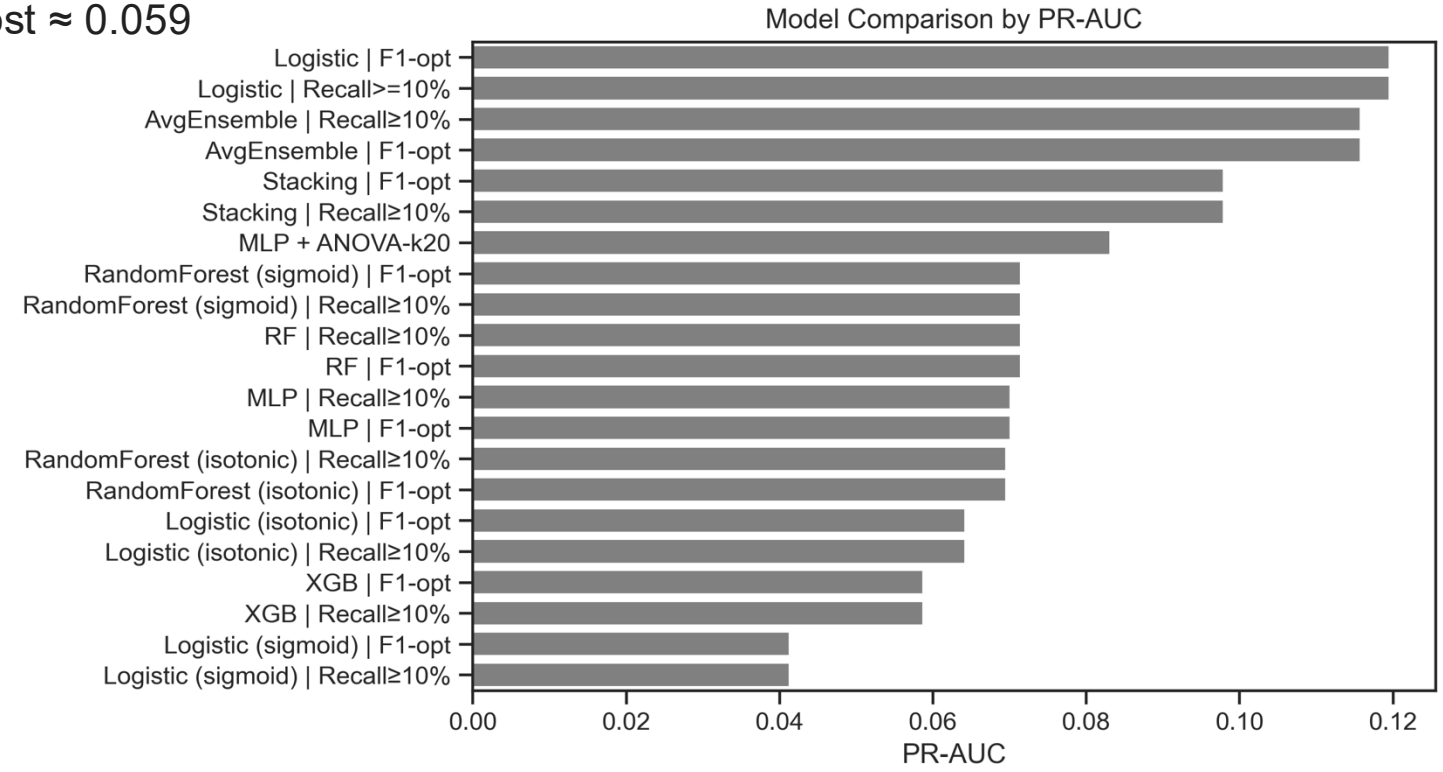




Results — PR-AUC

Key findings

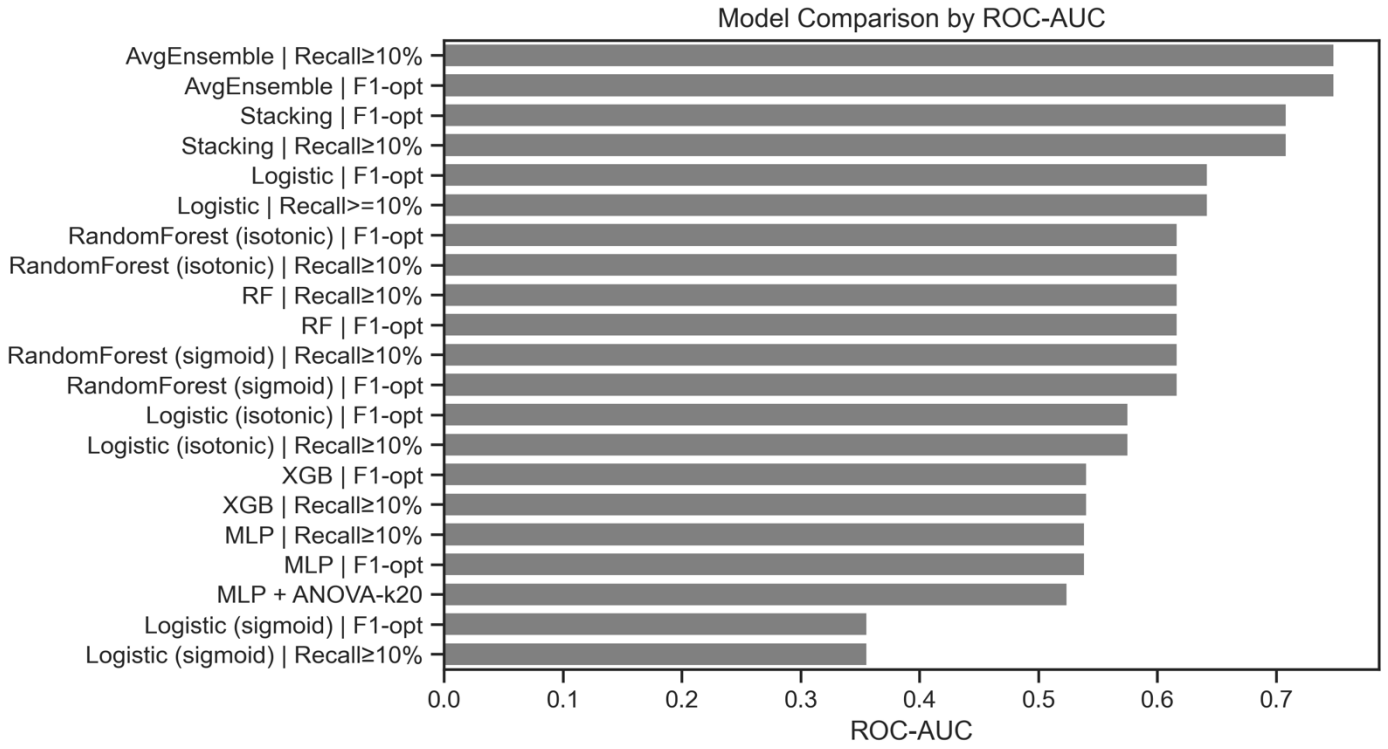
- Logistic Regression PR-AUC ≈ 0.120 (test)
- Simple averaging ≈ 0.116; Stacking ≈ 0.098
- Random Forest ≈ 0.072; MLP ≈ 0.070; XGBoost ≈ 0.059
- Absolute precision low in rare-event regimes



Results — ROC-AUC

Secondary check

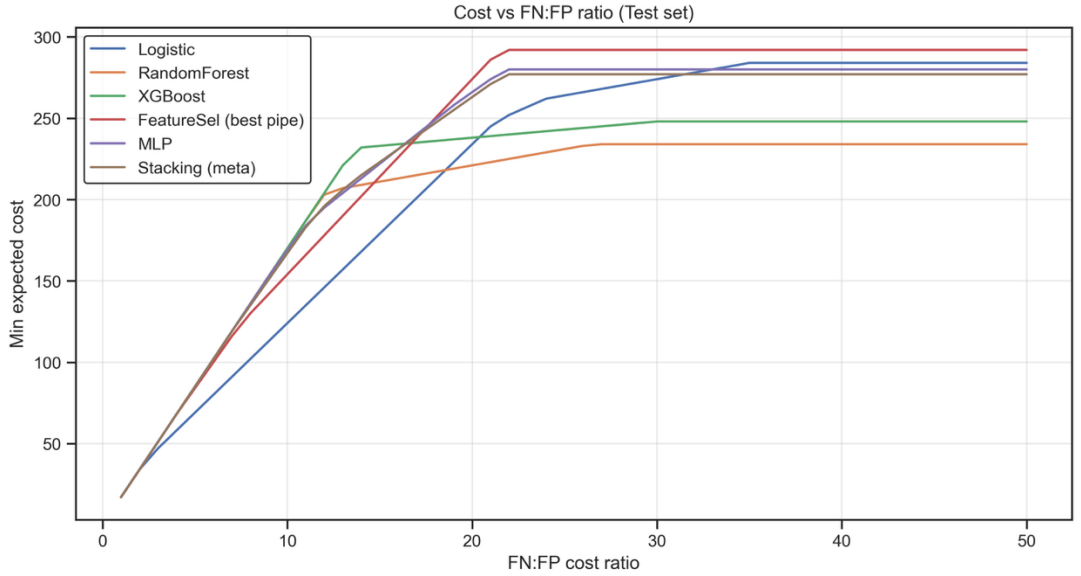
- Avg ensemble ROC-AUC ≈ 0.749; Stacking ≈ 0.709
- Logistic ≈ 0.643; Random Forest ≈ 0.617
- ROC is supportive. PR-AUC stays primary under 6.6% prevalence.

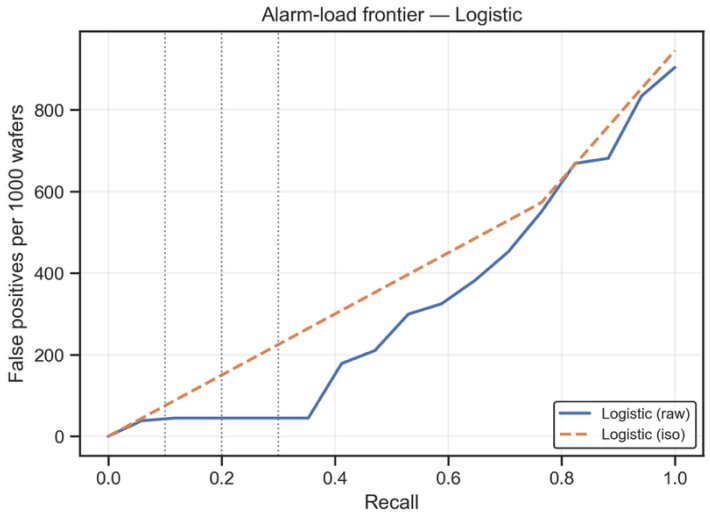


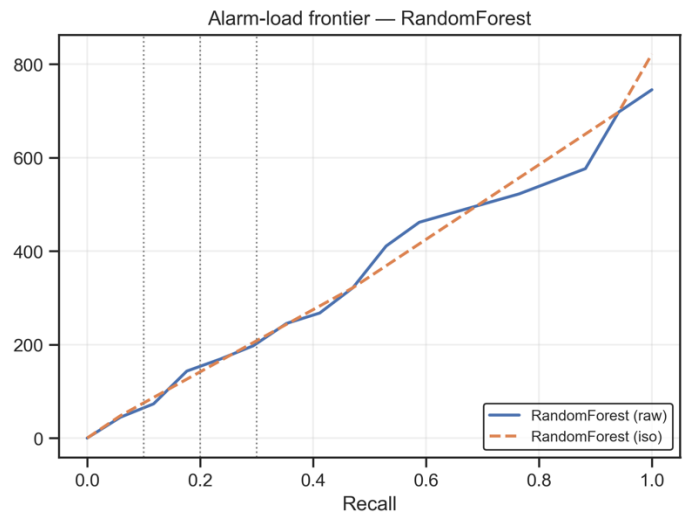
Cost-Sensitive Operations

Decision guidance

- **Expected cost per 1K wafers:**
  $$c_{1k} = 1000(FN/N \cdot c_{FN} + FP/N \cdot c_{FP})$$
- **Logistic cheaper** when FN:FP ≤ ≈18:1.
- **Random Forest cheaper** when FN:FP ≥ ≈19:1.
- Pick threshold on validation to meet recall and alarm-load targets.
- Use **calibrated probabilities** for stable thresholding.
- **Callout:** Crossover ≈ 19:1 (RF overtakes LR).



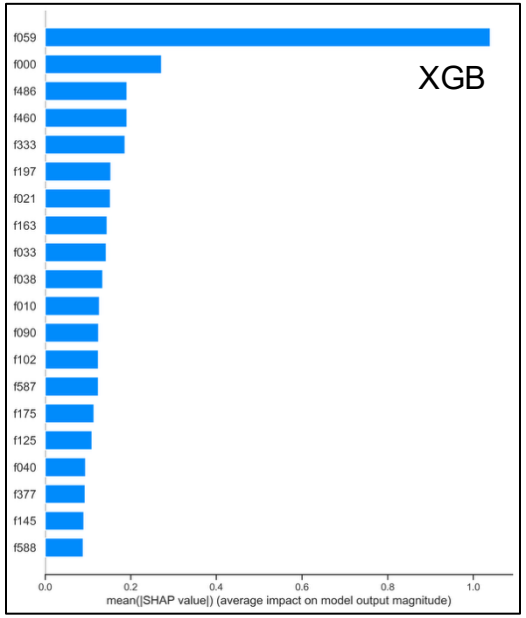


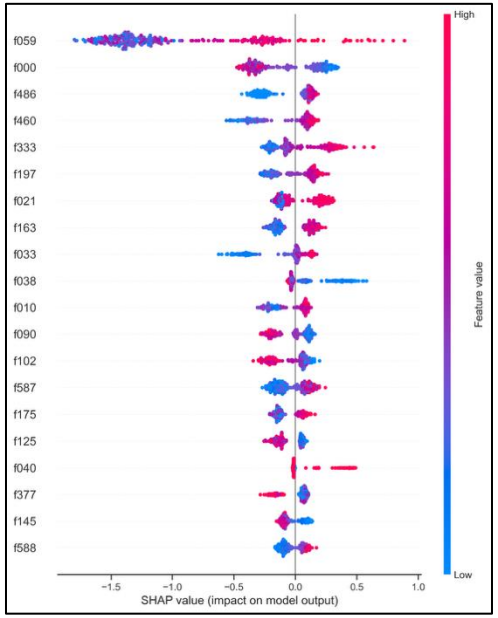


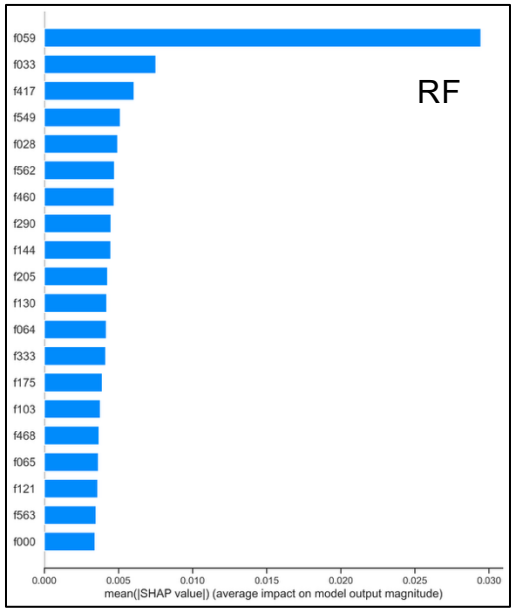
Interpretability — SHAP

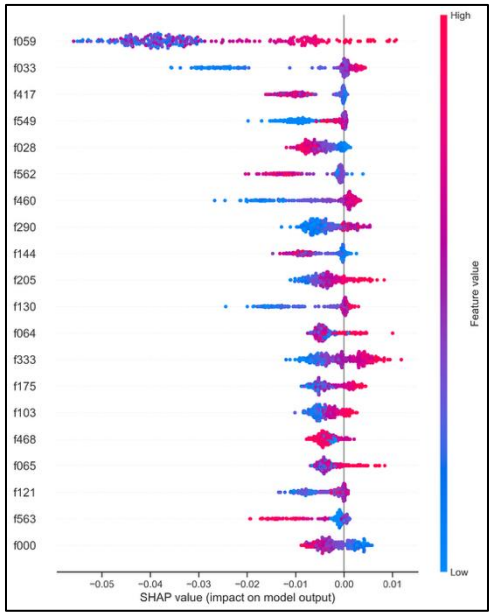
Stable signals

- TreeSHAP on RF/XGB; consensus top sensors: f033, f059, f175, f333, f460.
- High f059 pushes toward Fail; LR shows f059_missing as risk.
- Combine global bars, beeswarms, and dependence for monitoring.









Conclusion

- Leakage-safe pipeline across EDA → ETL → chrono split.
- PR-AUC as primary metric under 6.6% prevalence
- Model choice depends on FN:FP; LR ≤18:1, RF ≥19:1
- SHAP isolates stable sensor drivers (f033, f059, f460, …)
- Calibrated probabilities enable actionable thresholds

GitHub: https://github.com/himanshusaini11/DataScience/tree/master/Project3-SECOMSemiconductorYieldPrediction

Thank you!