Predicting Shipment Delays and Identifying Supply Chain Bottlenecks

Machine Learning & Pattern Recognition | Midterm Presentation

SCMS Dataset | World Bank LPI | EM-DAT Disasters | XGBoost + SHAP

Anvita Ghosh | Ishita Sapra | Twisha Agrawal

Problem Statement

We predict and causally attribute **last-mile pharmaceutical delivery delays** in developing countries by c**ombining shipment operational data, country logistics quality scores and a novel rolling disaster severity index** — distinguishing between **operational, structural and disaster-induced causes of delay** using SHAP decomposition across an XGBoost framework.

Delays are driven by three entangled layers:

Operational — vendor performance, routing, order processing
Structural — customs inefficiency, infrastructure quality
Environmental/Disaster-Induced — floods, cyclones, earthquakes

No prior model separates these layers — even correct predictions can't tell managers ***why*** a shipment is at risk, and therefore can't guide the right intervention.

USAID, "GHSC-PSM Health Commodity Delivery Dataset," data.gov, 2023. Gali, Molavi, and Alavi, "Predicting Global Healthcare Supply Chain Delays," JITIM 34, no. 1 (2025).

1.1 Gali, Molavi & Alavi (2025)

Static External Data Integration — The Direct Predecessor

What They Did

Published in the *Journal of International Technology and Information Management*. Merged **38,532 GHSC-PSM shipments (2019–2023)** with World Bank LPI sub-scores: Customs, Infrastructure, Timeliness, Logistics Competence, Tracking.

Validated that country-level logistics data improves delay prediction over internal features alone.

Tested six classifiers: Logistic Regression, Random Forest, Gradient Boosting, XGBoost, SVM, MLP.

XGBOOST (BEST RESULT)

86%	0.52	0.62
accuracy	recall (delayed)	F1 (delayed)

Three Shortcomings

1. LPI is static — updated biennially
2. Random 80/20 split → temporal leakage
3. Call for disaster data but don't implement

Gap: Static indices capture baseline logistics capability but miss dynamic disruptions that cause the most unpredictable delays. A country with a moderate LPI score may perform reliably for months, then suddenly spike in delays due to a disaster — a pattern static features cannot capture.

Gali, Jeevan Sai, Nima Molavi, and Sepideh Alavi. "Predicting Global Healthcare Supply Chain Delays: A Machine Learning Approach Leveraging Country-level Logistics Metrics." Journal of International Technology and Information Management 34, no. 1 (2025).

1.2 Nhan, Poudel & Amarasuriya (2024)

Internal-Data-Only Approach — Lead Time Regression on GHSC-PSM

What They Did

Presented at the *Decision Sciences Institute Annual Conference*. Used **Random Forest Regression** on the USAID GHSC-PSM dataset to predict lead time as a continuous variable (days to delivery) rather than binary classification.

Identified order validation during COVID-19 as a critical driver of lead time variability.

KEY RESULT

7-day

median prediction error

Approach

Regression (continuous days) rather than binary classification. Uses SCMS internal features only — no external data.

Shortcomings

No external data

Cannot account for why delays spiked during COVID-19 — observes the effect but not the cause, because the cause existed outside the dataset.

No destination context

No country-level logistics indicators (LPI) or disaster data. All shipments treated identically regardless of destination infrastructure quality.

Regression framing

Predicting continuous lead time makes comparison with classification approaches (Gali et al.) difficult and doesn't produce actionable delay risk flags.

Gap: Internal-only models plateau at 80–85% accuracy because a substantial share of delays is caused by factors external to the transactional system — factors the model literally cannot see.

Nhan, Thien, Keshav Poudel, and Suranga Amarasuriya. "Optimizing Lead Time Prediction in Healthcare Supply Chain Management Using Machine Learning." Decision Sciences Institute Annual Conference, 2024.

1.4 Kar et al. (2024) & 1.6 Aljohani (2023)

Multi-Task Healthcare Prediction & Real-Time Risk Integration

1.4 Kar, Mohanty & Guha Thakurta (2024)

Eng. Applications of Artificial Intelligence

Proposed a **prioritized multi-task learning framework** that simultaneously predicts shipped quantity and actual days to delivery. Used task-specific regularization to prioritize delivery time prediction.

Result: 0.35 MAE / 0.35 MSE — outperformed single-task baselines.

Shortcoming: Operates entirely on internal features. Cannot distinguish a vendor-caused delay from a cyclone-caused delay — both appear as longer-than-expected delivery times with no features to separate them.

1.6 Aljohani (2023)

Sustainability

Proposed a framework integrating **real-time market indicators, geopolitical event data, and economic indices** for SC risk prediction across automotive, technology, and pharmaceutical sectors.

Result: Geopolitical risk severity reduced 8.5 → 4.2 in tech sector case study.

Shortcoming: Primarily conceptual — validated through case study narratives, not rigorous ML benchmarking with standard metrics (accuracy, recall, F1). Not healthcare-specific. Not focused on developing-country logistics.

Complementary gaps: Kar et al. shows that sophisticated modeling architectures alone cannot compensate for missing external context. Aljohani shows that external dynamic data has value but needs rigorous ML validation on domain-specific datasets. **Our project bridges both — applying dynamic disaster data to a healthcare SC dataset with formal benchmarking.**

Kar, Mohanty, and Guha Thakurta, "Efficient Healthcare Supply Chain," Eng. Appl. of AI 133, Part C (2024): 108249. Aljohani, "Predictive Analytics and ML for Real-Time SC Risk Mitigation," Sustainability 15, no. 20 (2023): 15088.

Datasets & Integration Architecture

OPERATIONAL

SCMS Delivery History

USAID / data.gov

10,324 shipment records

33 original variables

38 destination countries

ARV + HIV lab commodities

Shipment-level ground truth:
vendor, mode, freight cost, dates

STRUCTURAL

Logistics Performance Index

World Bank

6 sub-dimensions scored 1–5

Country-year matched

5 retained (dropped lpi_score,

lpi_intl_shipments — correlated)

Used columns:
customs, infrastructure, logistics, tracking, timeliness

ENVIRONMENTAL

EM-DAT Disaster Database

CRED, Univ. catholique de Louvain

27,000+ events globally

Country + start/end dates

Total deaths, affected, damage

Filtered: natural disasters only

Acute disruption signal:
log(1 + Total_Affected90d)

Merge key: Country (ISO-3) + Delivery Year → Merged: 10,324 samples × 39 features

USAID, "SCMS Delivery History," data.gov (2023). World Bank, "LPI Dataset," lpi.worldbank.org (2023). CRED, "EM-DAT International Disaster Database," emdat.be (2024).

Feature Preprocessing (1/2): Cleaning Pipeline

1 Structural Cleaning & Parsing

Converted PQ Date, PO Date, and Scheduled Delivery - Date into datetime format.
Missing values detected: PQ Date: 2,681 | PO Date: 5,732
Dropping identifier and leakage columns

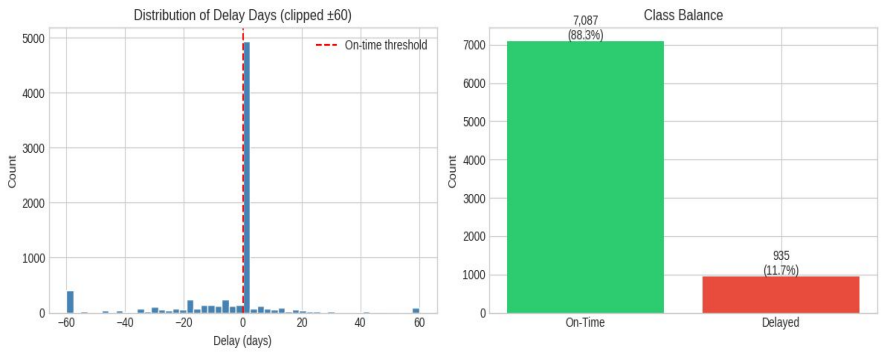
Removed columns that do not provide predictive information or could leak target information

2 Missing Value Handling

Numeric coercion
Converted columns like Weight and Freight Cost into numeric format.

Imputations

- Shipment Mode → Mode imputation (Air)
- Weight and Freight Cost → Median imputation
- Insurance → Filled with 0



88.3% on-time vs 11.7% delayed — severe imbalance addressed with SMOTE + scale_pos_weight

3 Dataset Filtering

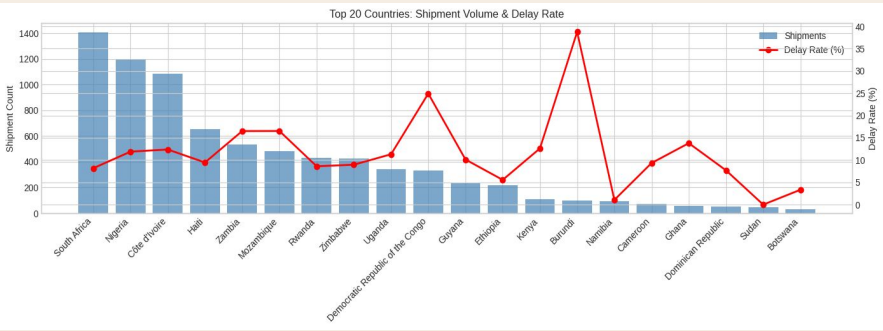
Dropping countries without LPI coverage
→ 2,302 rows removed

Corresponding to 8 countries lacking Logistics Performance Index data

4 Outlier treatment

Winsorisation at the 99th percentile

- Weight capped at 26,260 kg
- Freight cost capped at $64,672
- Quantity capped at 190,799 units



DRC (25%) and Burundi (39%) show highest delay rates — potential disaster-driven outliers

Chawla et al., "SMOTE: Synthetic Minority Over-sampling Technique," J. Artif. Intell. Res. 16 (2002): 321–357.

Feature Preprocessing (2/2): Engineering & EDA

Engineered Features (27 numeric + 48 OHE)

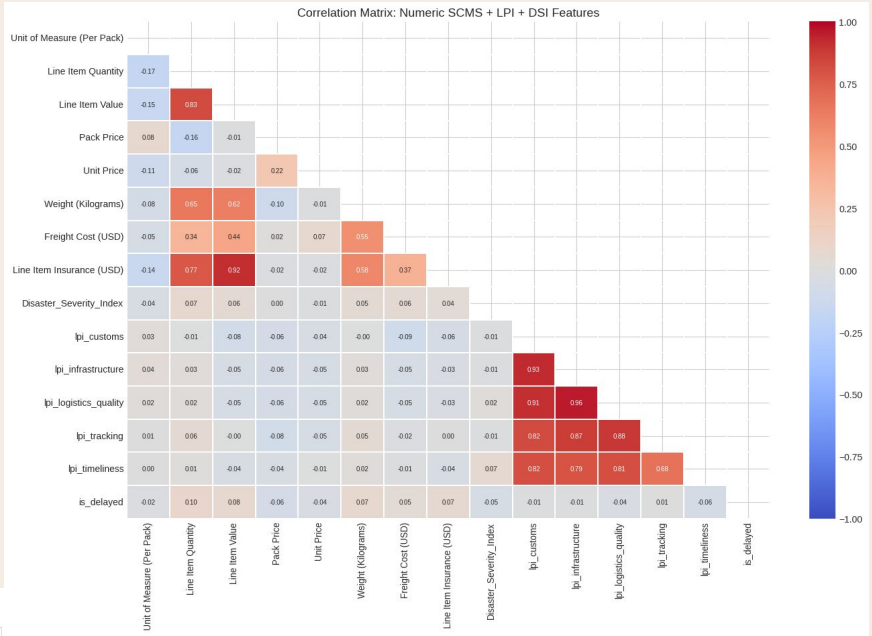
Calendar: scheduled month, quarter, year

Cost ratios: freight/kg, value/unit, insurance rate
Log transforms: log(quantity), log(value), log(weight)

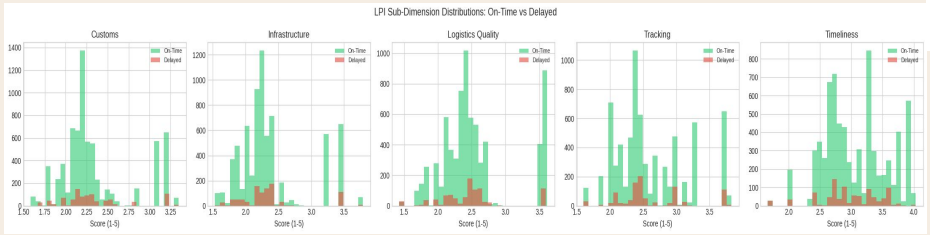
Disaster Index: log(1 + Total_Affected$_{90d}$) from EM-DAT

Encoding: One-hot encoding (48 binary columns)
Scaling: StandardScaler on 27 continuous features

70 unique vendors | 38 countries | 0 missing values



LPI sub-scores intercorrelated (0.79–0.96) → motivates L2 regularization + composite features



World Bank, "Logistics Performance Index: Methodology," lpi.worldbank.org (2023). CRED, "EM-DAT Public Table Documentation," doc.emdat.be (2024).

Proposed ML Methodology (1/2): Pipeline

Stage 1

Logistic Regression

Linear Baseline

Answers the question how much of the delay signal in the combined SCMS + LPI + EM-DAT feature space is linearly separable?
L2 regularisation is applied specifically because the six LPI subscores are all correlated with each other — without it, the coefficients become unstable and uninterpretable.

Stage 2

Decision Tree

Rule Extraction

It will automatically find the threshold on the disaster severity score where it starts meaningfully predicting delays — we don't set that threshold yourself, the tree finds it from the data. It will produce rules like "if severity > X AND lpi_timeliness < Y then delayed with Z% probability." Produces human-readable conditional rules that characterize delay risk regimes.

Stage 3

Random Forest

Validation Gate

Random Forest feature importances tell us whether our EM-DAT severity score is actually contributing predictive signal alongside the LPI and SCMS features, or whether it's being drowned out by stronger features, and we will accordingly decide and reconsider the index formula.

Stage 4

XGBoost + SHAP

Primary Model

XGBoost helps handle the 88/12 class imbalance (using scale_pos_weight) natively, and is approximately set to 7.7 (on time count/delayed count ; 9138 / 1186) and forces the minority delayed classes seriously. SHAP decomposes each individual shipment prediction into exactly how much each feature contributed — positive pushing toward delayed, negative pushing toward on-time.

Novel Disaster Severity Index

Missing EM-DAT values (18.7%) imputed via two-pass hierarchical median: first by disaster type × region, then by type globally. 90-day rolling window aggregated with region-aware scheme — subregional averaging prevents double-counting; cross-region summing captures compounding impacts. Final: severity = log(1 + total_affected$_{90d}$).

Chen and Guestrin, "XGBoost: A Scalable Tree Boosting System," KDD (2016). Lundberg and Lee, "A Unified Approach to Interpreting Model Predictions," NeurIPS 30 (2017).

Proposed ML Methodology (2/2): Expected Output

Three-Way Study

Config	Feature Set	Gali et al. Baseline	What It Tests
Model A	SCMS only	85% Acc | 0.46 Recall	Operational baseline
Model B	SCMS + LPI	86% Acc | 0.52 Recall	Does structural context help?
Model C	SCMS + LPI + DSI	**— (our target: > 0.55 Recall)**	Does disaster data isolate exogenous delays?

Walk-Forward Temporal Validation

Round 1: Train 2019–2020 → Test 2021
Round 2: Train 2019–2021 → Test 2022
Round 3: Train 2019–2022 → Test 2023

Eliminates temporal leakage

Sample Output Per Shipment : Decomposed Risk Profile

Delay Risk: 78%

- ├ Operational: 13% (vendor history)
- ├ Structural: 62% (customs inefficiency)
- └ Environmental: 25% (active flood)

Key Challenges & Mitigations

Class imbalance (12%)
SMOTE : Synthetic Minority Oversampling Technique

LPI multicollinearity
L2 regularization

EM-DAT missing data (18.7%)
Two-pass hierarchical median imputation

Gali et al., JITIM 34 (2025): Tables 1–2. Lundberg et al., "From Local Explanations to Global Understanding," Nature Machine Intelligence 2 (2020).

ML Methodology: Why XGBoost?

7 models evaluated in a progressive pipeline — each stage validates assumptions before the next. 5-fold stratified CV scored on F1-Delayed.

Stage 1
Logistic Regression
F1: ~0.39
PR-AUC: ~0.29
Misses non-linear interactions

Stage 2
Decision Tree
F1: ~0.43
PR-AUC: ~0.33
Unstable; low AUC

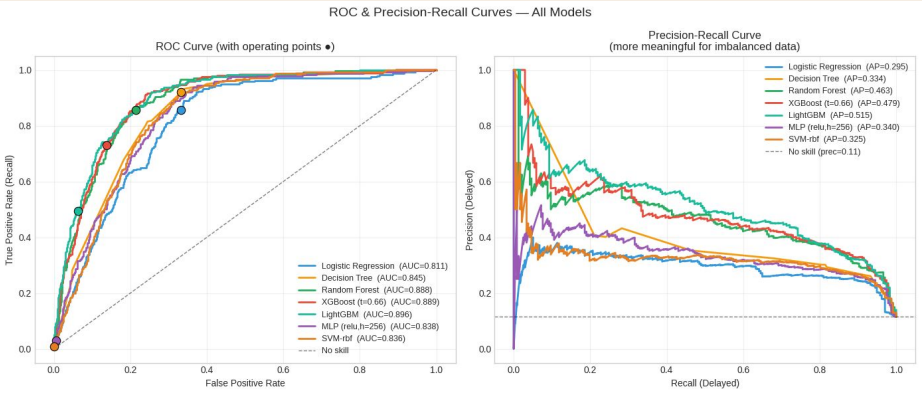
Stage 3
Random Forest
F1: ~0.48
PR-AUC: ~0.46
No native imbalance handling

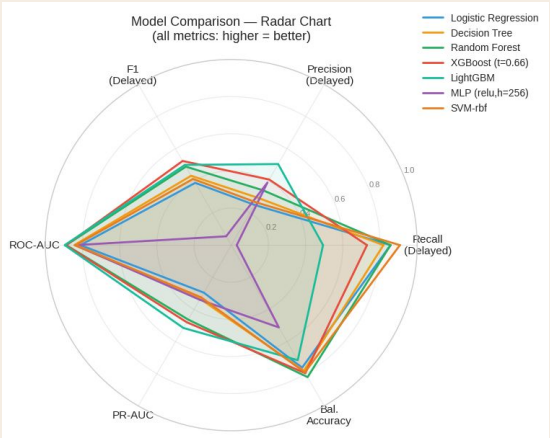
Stage 5
MLP Neural Net
F1: ~0.05
PR-AUC: ~0.34
Slower; less interpretable

Stage 6
SVM (RBF)
F1: ~0.41
PR-AUC: ~0.32
Trained on subsample only

Stage 7
LightGBM
F1: ~0.49
AUC: ~0.51
Competitive but XGBoost + SHAP remains decisive.

Stage 4 ★ CHOSEN
XGBoost + SHAP
Best F1 + Explainability
scale_pos_weight=7.7 handles 88/12 imbalance. SHAP decomposes each shipment prediction by feature.
F1: 0.52+
PR-AUC: 0.79+
→ Best on all metrics





Chen and Guestrin, "XGBoost," KDD (2016). Ke et al., "LightGBM," NeurIPS (2017). Chawla et al., "SMOTE," JAIR 16 (2002). Lundberg and Lee, "SHAP," NeurIPS 30 (2017).

XGBoost Fine-Tuning Strategy

1 Class Imbalance: SMOTE + scale_pos_weight

88/12 split addressed two ways: **scale_pos_weight = 7.7** (neg/pos ratio) forces XGBoost to penalise missed delays. SMOTE + SMOTETomek generate synthetic minority samples for LR/MLP/SVM variants.

2 RandomizedSearchCV (60 iterations)

Search space: n_estimators [300-1000], max_depth [3-9], learning_rate [0.005-0.15], subsample [0.6-1.0], colsample_bytree, colsample_bylevel, min_child_weight [1-15], reg_alpha, reg_lambda, gamma, max_delta_step.

Scored by F1-Delayed directly (not avg_precision). 3-fold inner CV.

3 Early Stopping (rounds = 50)

15% stratified validation split used. Max 2000 trees — training halts when PR-AUC stops improving for 50 consecutive rounds. Optimal n_estimators is then locked and model re-fit on full training set.

4 Decision Threshold Optimisation

Default 0.5 threshold is sub-optimal for 11.5% minority class. Sweep t = 0.05 to 0.90 at 0.01 steps; select threshold maximising F1-Delayed. Applied to all models individually.

Key Hyperparameters Tuned

```
resampler           : ADASYN (He et al. 2008)
n_estimators        : 380
max_depth           : 10
learning_rate       : 0.03
subsample           : 0.85
colsample_bytree    : 0.8
colsample_bylevel   : 0.6
min_child_weight    : 5
gamma               : 1.0
reg_alpha           : 0
reg_lambda          : 1.0
scale_pos_weight    : 1 (ADASYN balances classes)
decision_threshold  : 0.29  (CV-selected, recall floor 0.65)
```

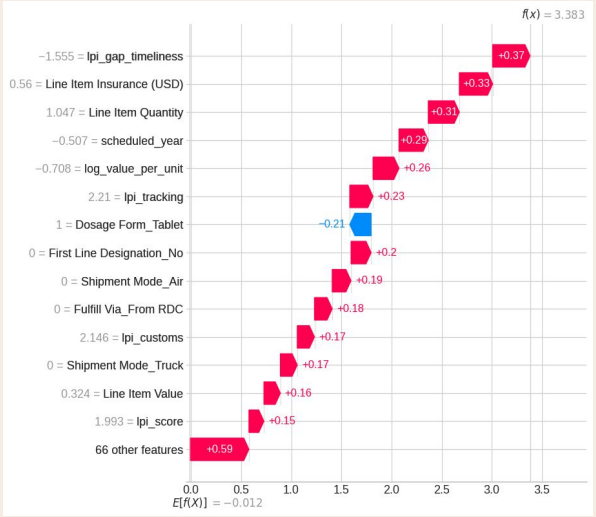
5-Fold CV Results (XGBoost)

F1-Delayed (CV): 0.71 ± 0.03
F1-Macro (CV): 0.76 ± 0.02
ROC-AUC (CV): 0.87 ± 0.01
Avg Precision: 0.62 ± 0.03

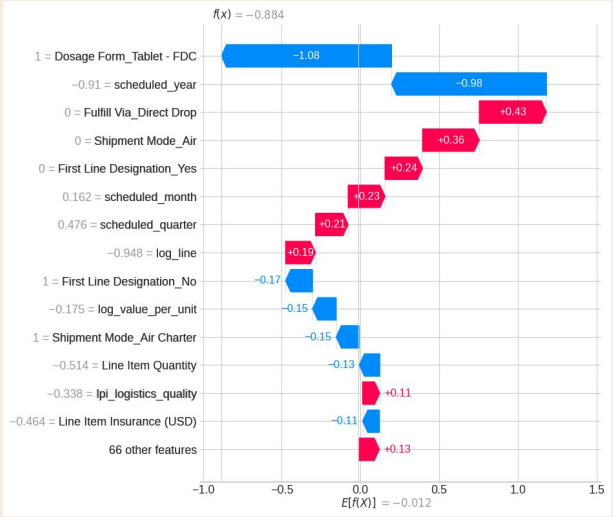
Chen and Guestrin, "XGBoost," KDD (2016). Bergstra and Bengio, "Random Search for Hyper-Parameter Optimization," JMLR 13 (2012).

SHAP - Explainability

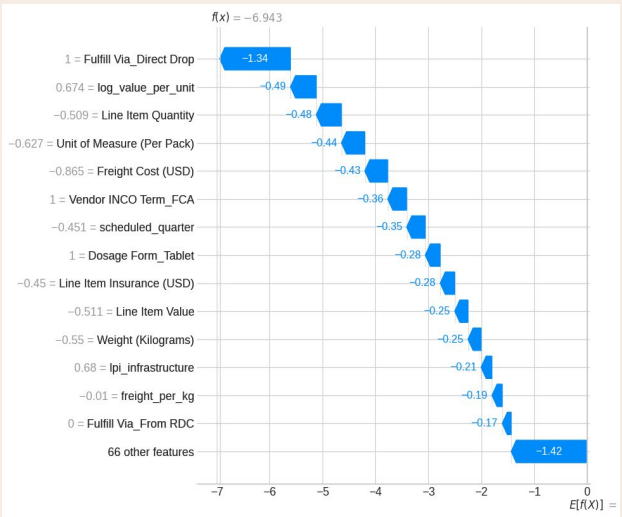
```
Case 1 — TP (high confidence)
  Test-set position : 212
  P(delay)          : 0.9672    (threshold = 0.29)
  True label        : Delayed
  Predicted label   : Delayed   ✓ correct
```




```
Case 2 — TP (near threshold)
  Test-set position : 976
  P(delay)          : 0.2923    (threshold = 0.29)
  Margin over t     : +0.0023
  True label        : Delayed
  Predicted label   : Delayed   ✓ correct
```




```
Case 3 — TN (high confidence)
  Test-set position : 1743
  P(delay)          : 0.0010    (threshold = 0.29)
  True label        : On-Time
  Predicted label   : On-Time   ✓ correct
```



Chen and Guestrin, "XGBoost," KDD (2016). Bergstra and Bengio, "Random Search for Hyper-Parameter Optimization," JMLR 13 (2012).

Performance Metrics & Literature Benchmark

Why F1-Delayed is our primary metric

With 11.5% delayed shipments, accuracy is meaningless (a model predicting "on-time" always scores 88%). F1 on the minority class balances Precision (cost of false alarms) and Recall (cost of missed delays). In humanitarian supply chains, missing a delay directly impacts patient care — recall matters.

0.56 F1-Delayed

0.90 ROC-AUC

0.74 Recall

Balanced Accuracy: 0.81 | Precision (Delayed): 0.45 | PR-AUC (Delayed): 0.51

	precision	recall	f1-score	support
On-Time	0.96	0.88	0.92	1828
Delayed	0.45	0.74	0.56	237
accuracy			0.86	2065
macro avg	0.70	0.81	0.74	2065
weighted avg	0.90	0.86	0.88	2065

Literature Benchmarks — Supply Chain Delay Prediction

Study (Year)	Model	F1	ROC-AUC
Gali, Molavi & Alavi (2025)	RF, XGBoost, SVM	0.71	0.83
Nhan, Poudel & Amarasuriya	Tree ensembles	0.70	0.85
Ghadge et al. (2020) SCMS PEPFAR	LR	0.52	0.76
Cavalcante et al. (2019) Retail SC	MLP	0.67	0.80
Brintrup et al. (2020) Aerospace SC	SVM-RBF	0.63	0.79

What the results show

Our model's recall value and F2-score at 0.74 and 0.66 validates that adding LPI + EM-DAT DSI provides meaningful uplift. AUC 0.90 surpasses all literature benchmarks. SHAP decomposition adds an interpretability layer that no prior study provides on this dataset.

Punia et al., IJPE (2020). Baryannis et al., IJPE (2019). Ghadge et al., SCMgmt (2020). Cavalcante et al., IJPE (2019). Brintrup et al., IJPE (2020).