LogiFlow Railway Pipeline & Delay ML

Document version: June 2026. Describes the Indian Railways cargo decision pipeline, scraped delay corpus training, and the three public accuracy quantifiers.

1. Rail pipeline overview

LogiFlow rail optimization chains: (1) station resolution from city names, (2) route discovery from the 2017 IR schedule CSV plus live ConfirmTkt/RailRadar enrichment, (3) feature engineering (tariff, weather, live delay APIs), (4) ML delay prediction adjusting ETA and risk, (5) ranking by cost/time/risk priority. Map geometry uses full A-to-B schedule slices with offline station coordinates.

2. Training data

Primary ML corpus: ir_train_delays.csv scraped from runningstatus.in (15,650 labeled train-days). Each row is a station on a train run_date with scheduled/actual times and delay text. Labels are parsed from arrival_delay_min, delay_text (e.g. '4 hrs 48 mins'), or actual minus scheduled clock times.

3. Model & validation

Algorithm: GradientBoostingRegressor (gbm). Features: stations_on_run, route_distance_km, train_type, scheduled_hour, day_of_week. Validation: 5-fold GroupKFold grouped by train_number (no train leakage). Temporal backtest: leave-one-run_date-out on scraped history days.

4. Quantifier 1 - +/-15 min accuracy

Value: 59.3%. Definition: fraction of cross-validated predictions where |predicted - actual| <= 15 minutes. Derivation: sklearn cross_val_predict with GroupKFold splits; counted over all held-out folds.

4. Quantifier 2 - +/-30 min accuracy

Value: 80.9%. Same CV protocol as Quantifier 1 with a 30-minute tolerance band. Used for operational planning where wider slack is acceptable.

4. Quantifier 3 - Past-date backtest (+/-30 min)

Value: 80.1% (average across 3 held-out run_dates). For each scraped date D: train on all other dates, predict every train on D, measure +/-30 min hit rate. Simulates forecasting delay on a past day using prior history.

5. Error metrics

CV MAE: 22.7 min. CV RMSE: 53.6 min. R-squared: 0.171. MAE is mean absolute error in minutes; RMSE penalises large outliers more heavily.

6. Retraining

Run: make train-delay-ml or python backend/scripts/train_delay_ml.py. Updates scraped_delay_model.pkl and scraped_delay_metrics.json. Redeploy backend so /railway/model-info serves fresh quantifiers.