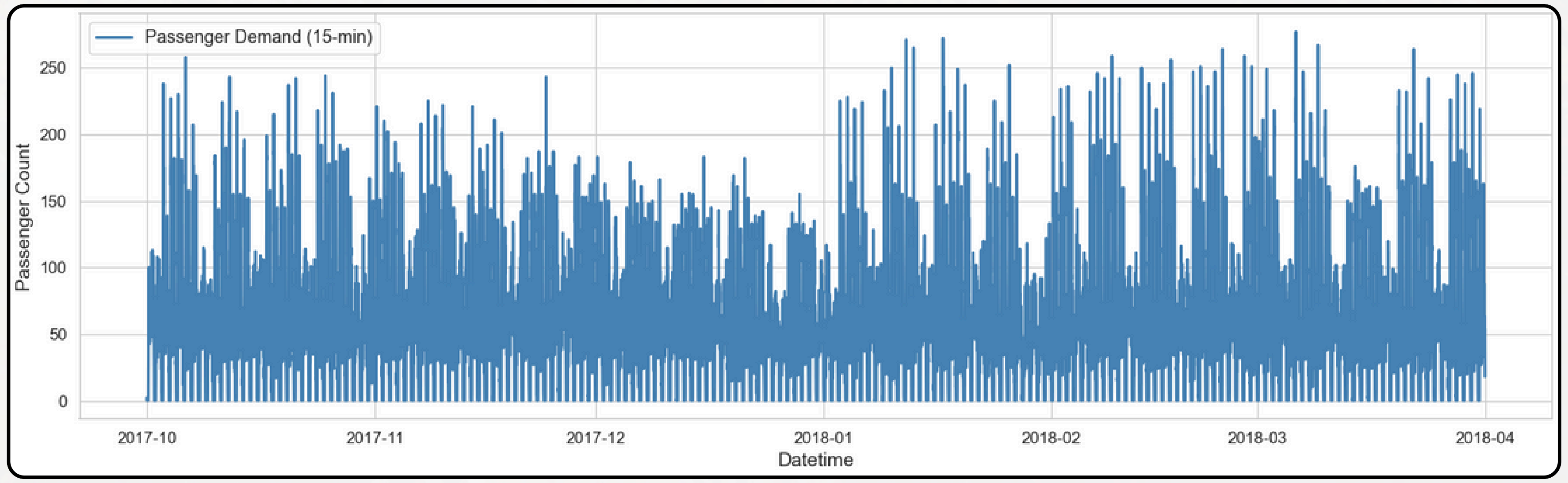
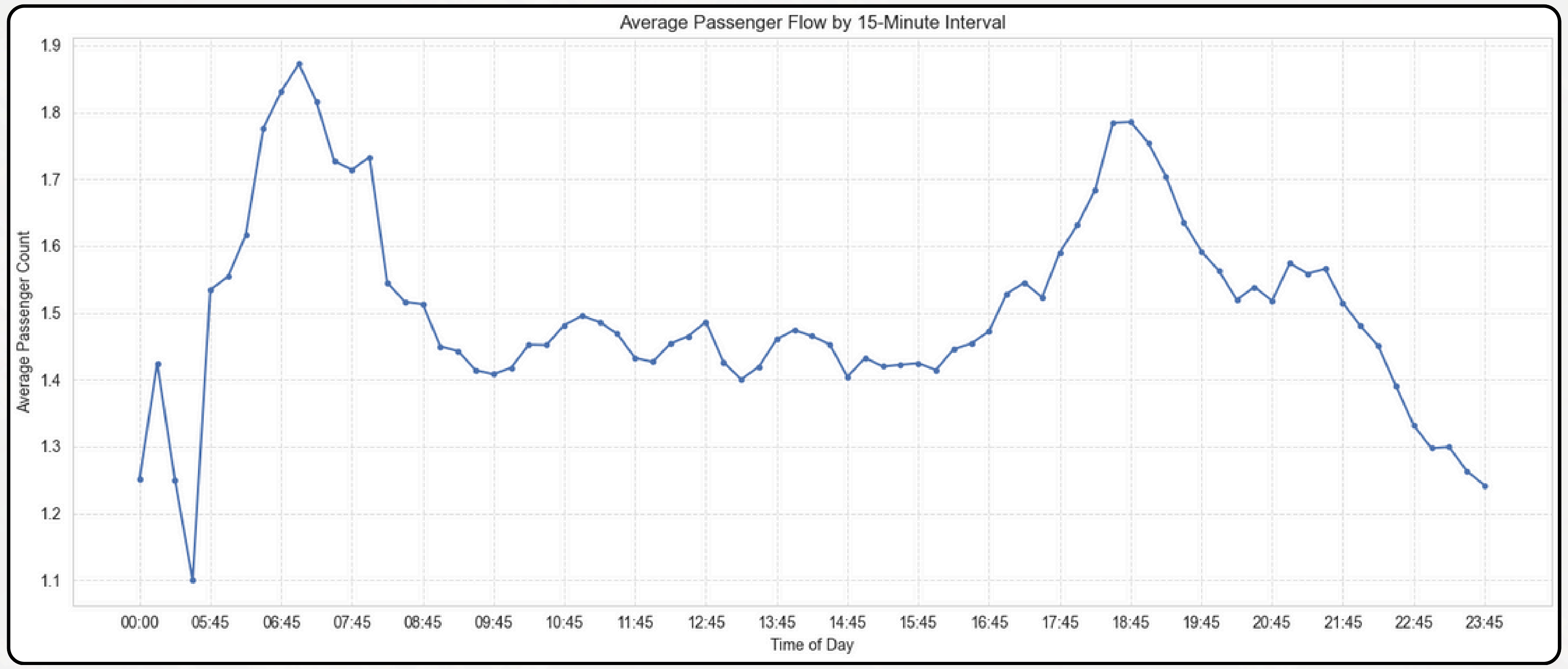
Urbanbus Passenger Demand Forecasting and Conformal Predictions

L Manoj Sai | Roll No: 22BT30019
Department of Bioscience and Biotechnology
Supervised By: Prof. Sushil Punia

Table of Contents

Business Problem & Challenges





1. Urban bus systems face growing demand and rising operational complexity
2. Agencies struggle to match supply with rapidly changing passenger volumes
3. Traditional planning relies on static timetables, leading to inefficiencies
4. High peak-hour variability causes overcrowding and service delays
5. Under-utilization during off-peak periods wastes fleet resources

Solution

How Forecasting Addresses Operational Challenges

- Provides route-level visibility into near-future passenger load
- Enables proactive matching of supply to demand instead of reactive adjustments
- Helps mitigate overcrowding, long wait times, and bus bunching

Operational Improvements Enabled

- Dynamic bus allocation during emerging peaks
- Frequency optimization aligned with expected short-term demand
- Better planning for high-demand corridors and transfer hubs
- Reduces resource waste in low-demand periods by preventing over-supply

Why 15-Minute Forecasts?

- UrbanBus data is aggregated into 15-minute intervals during regular operating hours
- Captures fine-grained ridership fluctuations observed across the dataset (e.g., sharp commuting peaks, weekday–weekend differences)
- Supports operational decisions such as dispatching extra buses or adjusting headways

OBJECTIVES

Develop a robust bus passenger demand forecasting framework combining optimized **LightGBM** models with **Conformalized Quantile Regression** to produce accurate predictions and reliable uncertainty intervals.The goal is to enable data-driven, risk-aware operational decision-making for route-level planning and dynamic resource allocation.

01

ANALYZE HISTORICAL BUS RIDERSHIP DATA AND IDENTIFY TEMPORAL PATTERNS

02

DEVELOP FORECASTING MODELS USING TIME SERIES AND MACHINE LEARNING TECHNIQUES

03

IMPLEMENT CONFORMAL PREDICTION METHODS FOR UNCERTAINTY QUANTIFICATION

Data overview

01 Dataset overview

- ~727 million AFC transactions (Oct 2017–Mar 2018)
- 5144 bus stops, 307 routes, second-level timestamp granularity
- Includes boarding/alighting stops, card types, route identifiers, and trip metadata

02

Stop-level:

High sparsity (27.6%), low mean flow (≈11)

Highly volatile, localized patterns

Route-level:

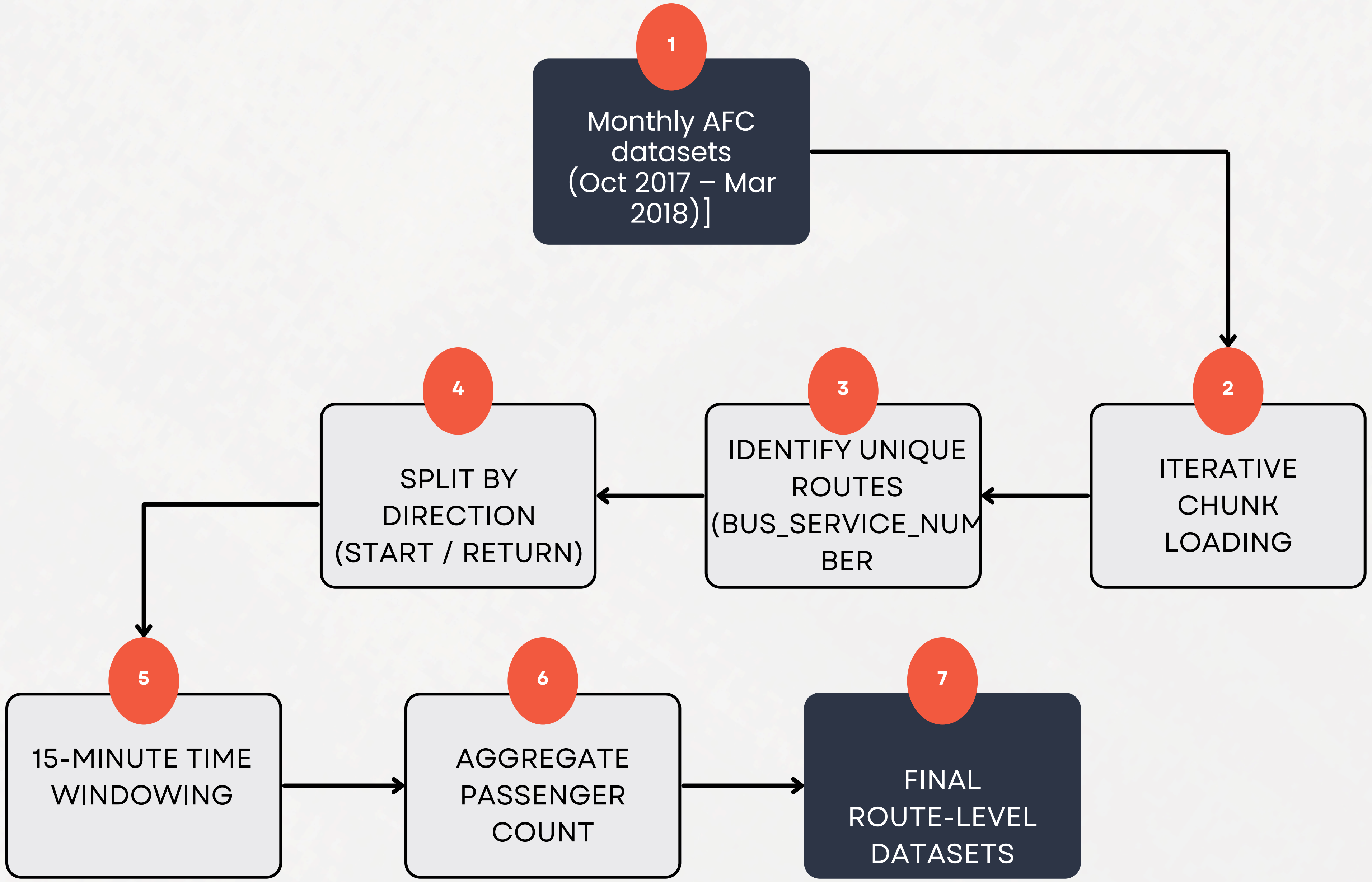
Higher mean flow (≈121), lower sparsity (3.1%)

Smoother aggregated demand—more suitable for forecasting

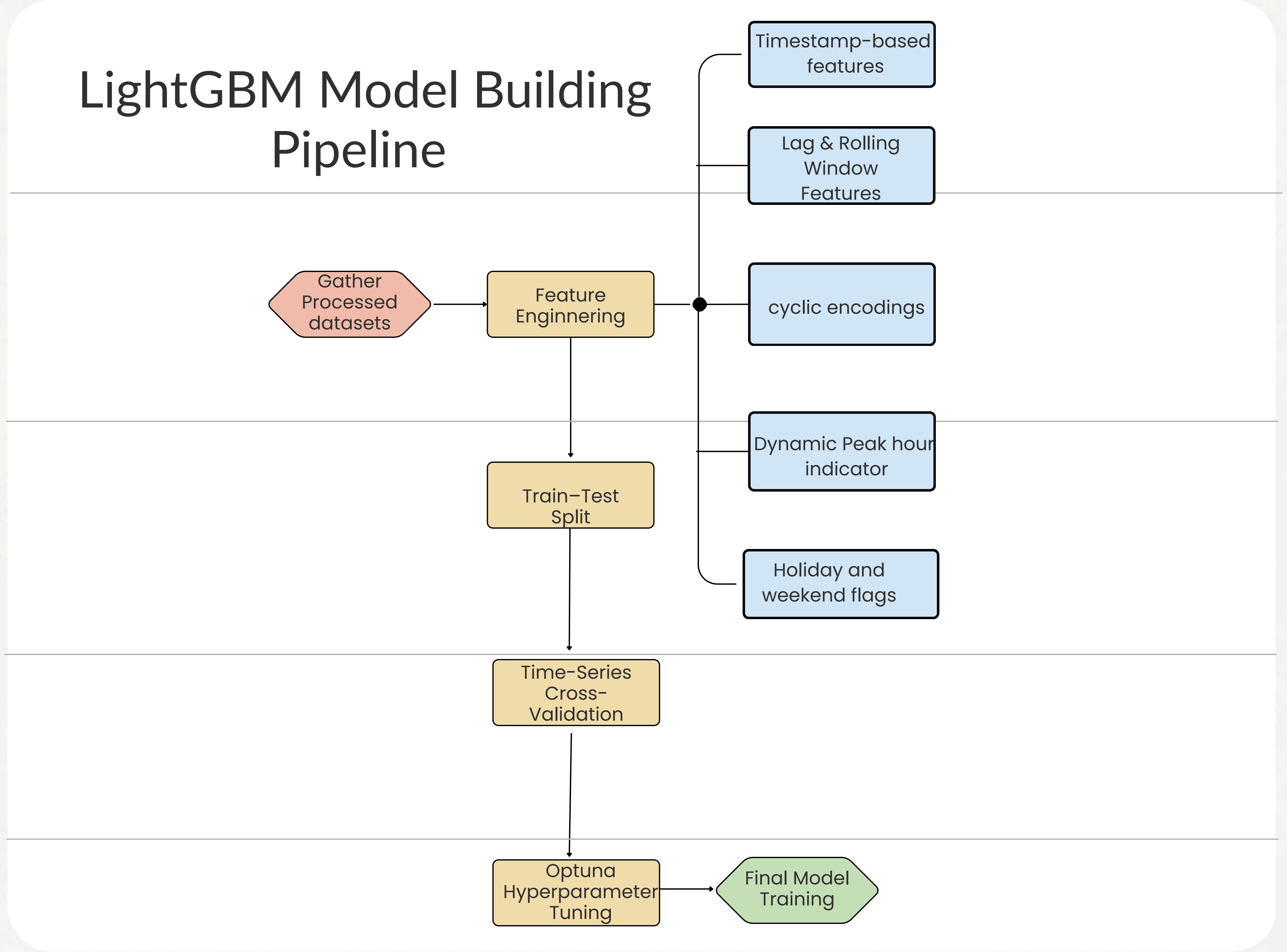
Type	File Name	Description
AFC Transaction	BUS_DATA_MAR_2018.csv	Detailed Bus data for March 2018.
AFC Transaction	BUS_DATA_FEB_2018.csv	Detailed Bus data for February 2018.
AFC Transaction	BUS_DATA_JAN_2018.csv	Detailed Bus data for January 2018.
AFC Transaction	BUS_DATA_DEC_2017.csv	Detailed Bus data for December 2017.
AFC Transaction	BUS_DATA_NOV_2017.csv	Detailed Bus data for November 2017.
AFC Transaction	BUS_DATA_OCT_2017.csv	Detailed Bus data for October 2017.
Bus Stop Related Information	BusStopList.csv	Names and IDs of all bus stops.
Bus Stop Related Information	DistanceMatrix.csv	Distance matrix between any two points.
Bus Stop Related Information	BusRoutes.pickle	Dictionary data of bus routes.

DATA PREPROCESSING

Card_Number	Unique identifier for the smart card used in the transaction
Card_Type	Category of the card
Travel_Mode	Mode of travel
Bus_Service_Number	Unique service number representing the bus route
Direction	Indicates the route direction (Start or Return)
Bus_Trip_Num	Unique trip identifier for the specific bus run
Bus_Reg_Num	Registration number of the bus vehicle
Boarding_stop_stn	Boarding stop station ID
Alighting_stop_stn	Alighting stop station ID
Ride_start_date, Ride_start_time	Date and time of boarding
Ride_end_date, Ride_end_time	Date and time of alighting



TRAINING AND INFERENCE PIPELINE

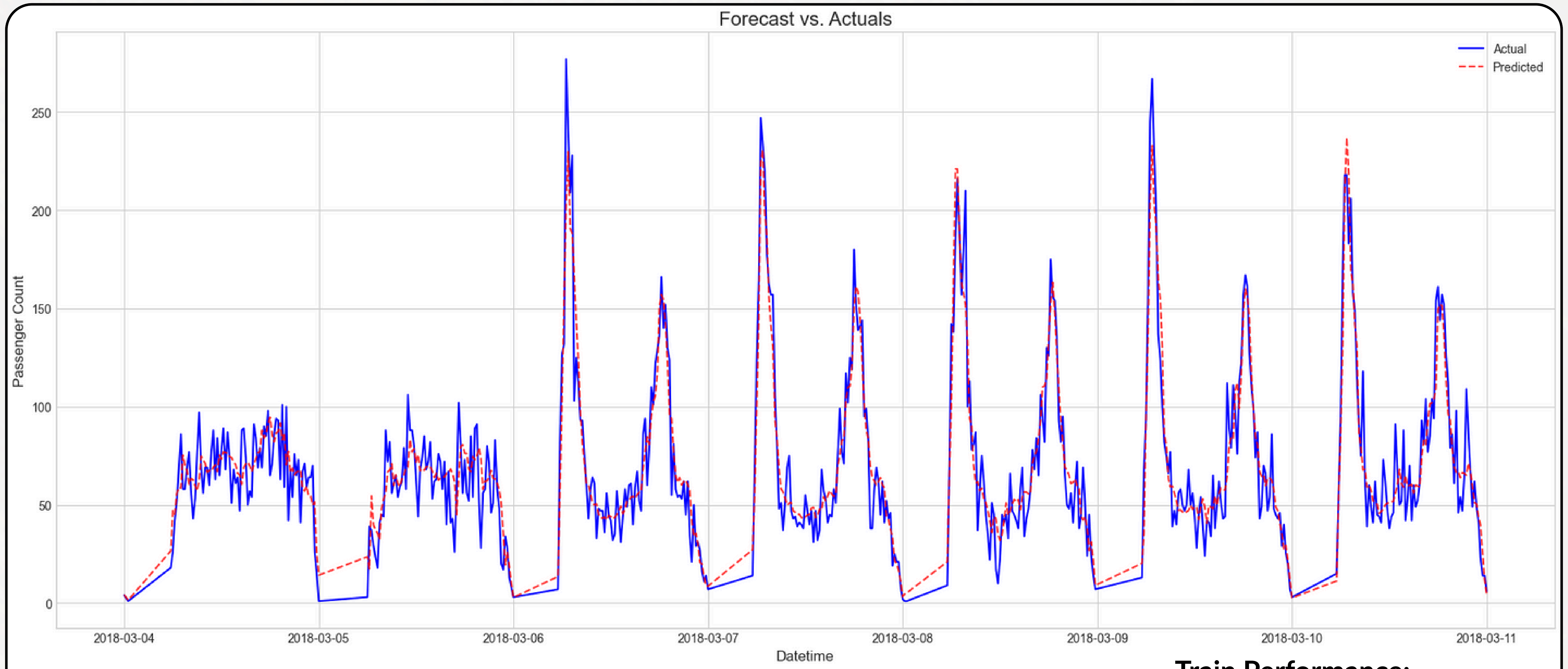


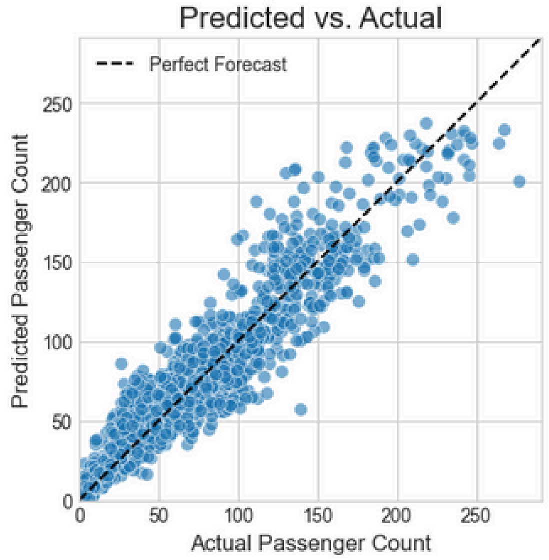
MODEL PREDICTION WORKFLOW:

- Constructed clean 15-minute route-level time series from six months of AFC data, with full temporal ordering and aggregation.
- Built feature engineering pipeline: temporal & cyclic features, weekend/holiday flags, dynamic peak-hour indicator, lag terms, and rolling statistics.
- Adopted a 28-day test split to prevent temporal leakage.
- Applied **TimeSeriesSplit** (3-fold) for chronological cross-validation to evaluate model stability.
- Used **Optuna** hyperparameter tuning to optimize **LightGBM** parameters by minimizing MAE.
- Trained final route-specific LightGBM models and evaluated them using MAE, MAPE, and sMAPE for comprehensive performance assessment.

Results Overview

Passenger demand forecasts on a random route (SER0b91)



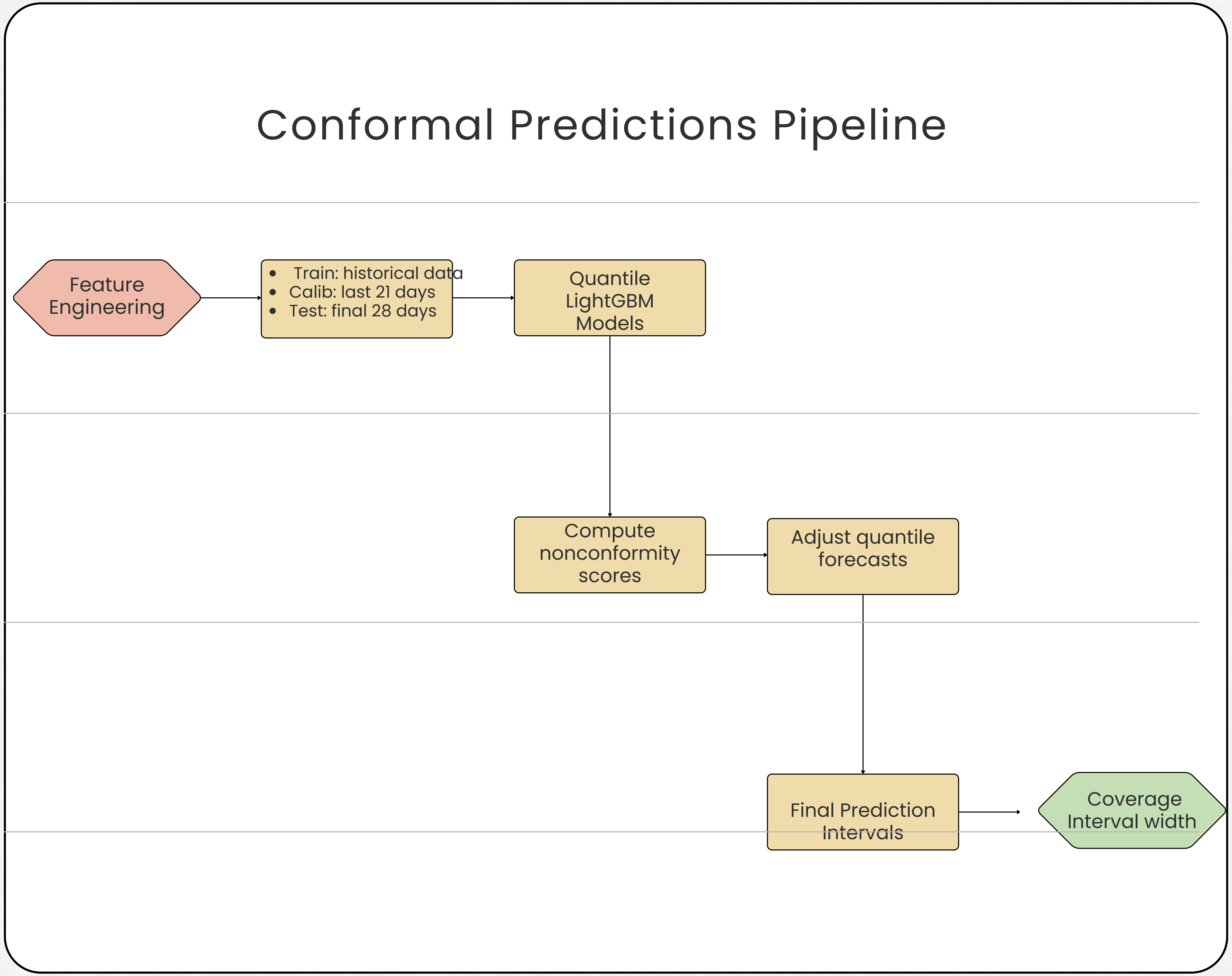


Train Performance:
MAE: 10.884
R^2: 0.8692
MAPE: 21.62%
sMAPE: 18.80%

Test Performance:
MAE: 12.3010
R^2: 0.8514
MAPE: 25.55%
sMAPE: 20.84%

	test_MAE	test_MAPE	test_sMAPE
mean	14.922828	50.8197	27.210507
std	7.057621	51.322	14.880648
min	1.358356	13.9777	12.09138
max	61.851253	358.5648	115.011525

- LightGBM effectively learns nonlinear daily/weekly patterns
- Rolling statistical features improve peak prediction smoothness
- Lag features strengthen short-horizon predictive power
- Performance varies slightly by route due to demand heterogeneity

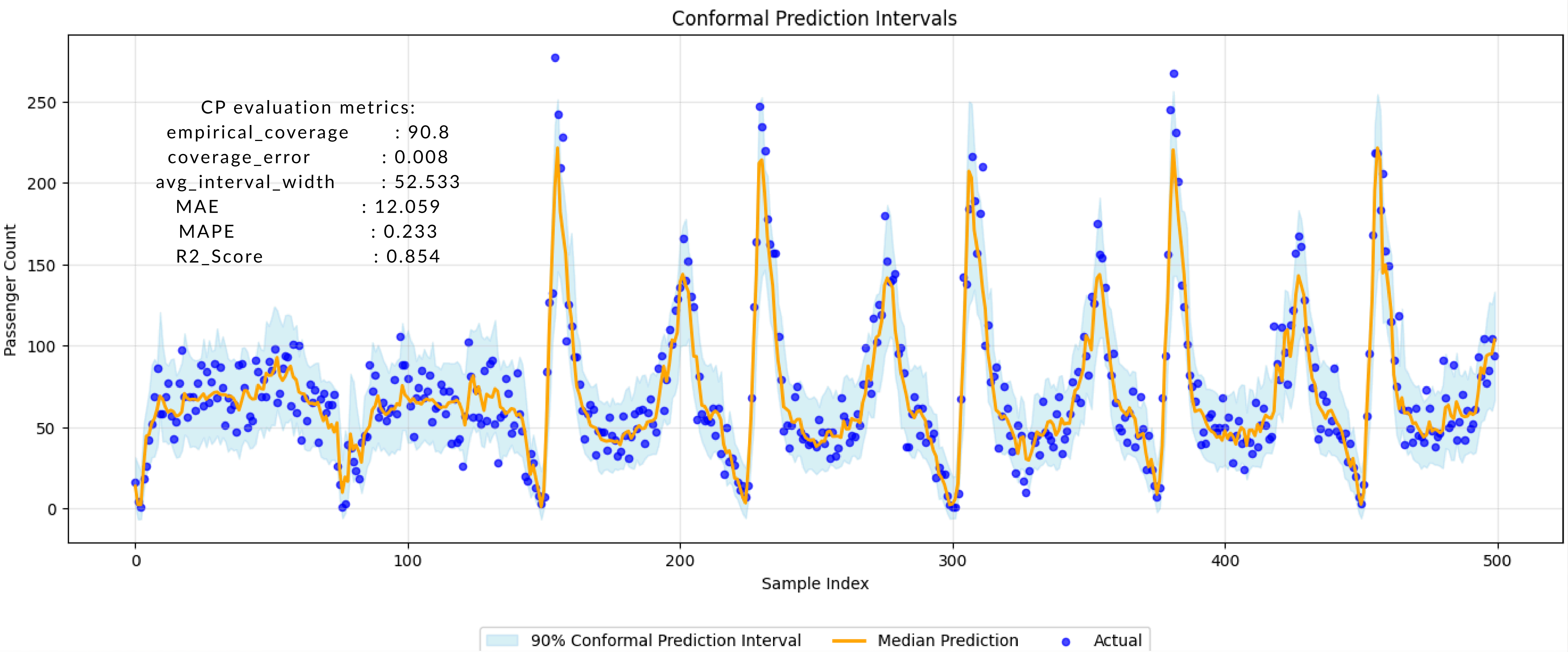


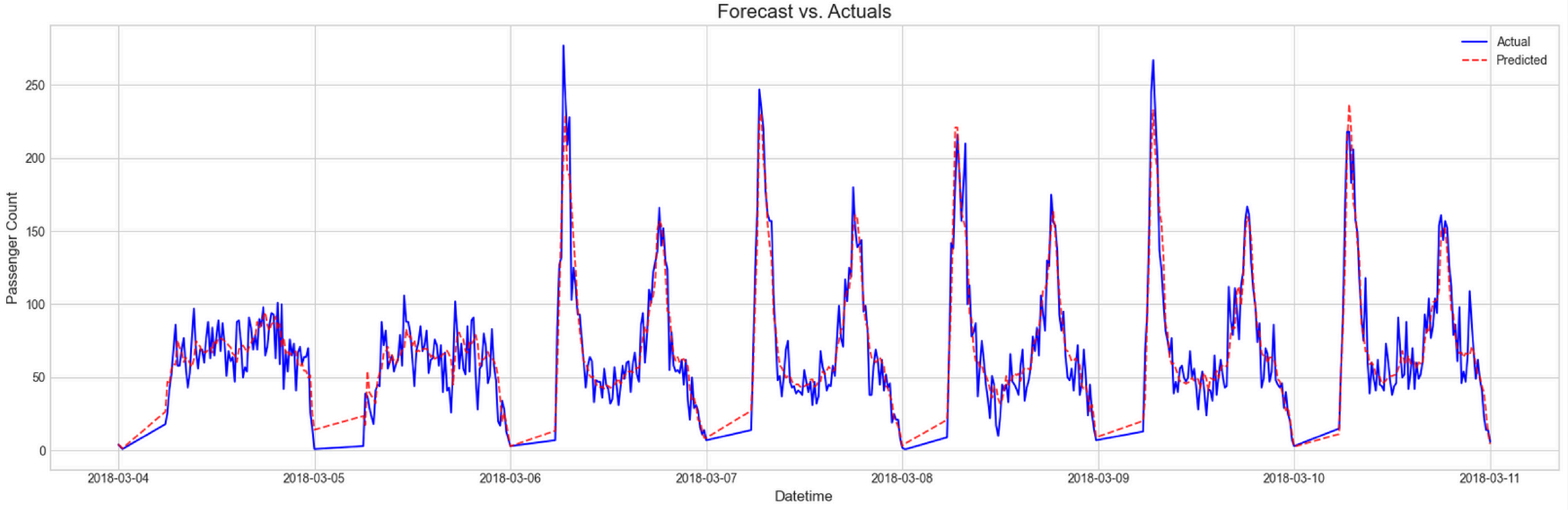
- Point forecasts do not capture uncertainty
- Passenger demand is highly stochastic (weather, holidays, events, disruptions)
- UrbanBus dataset shows large variance, sharp peaks, and strong route/stop correlations
- Need reliable prediction intervals, not just single-value prediction

What Conformal Prediction Provides

- Distribution-free prediction intervals (no assumptions about data distribution)
- Finite-sample validity → guaranteed target coverage (e.g., 90%)
- Model-agnostic → works on top of any ML model
- Produces route-specific uncertainty bands that reflect real variability

CONFORMAL PREDICTIONS





FUTURE OUTLOOK

Operational Deployment

- Integrate 15-min forecasts into on-demand bus allocation
- Enable dynamic dispatching and real-time schedule adjustments
- Use uncertainty intervals to trigger buffer fleet activation and adaptive headway control

Advancing Uncertainty Quantification

- Explore alternative conformal prediction methods beyond CQR:
- Ensemble Conformal Prediction (ECP)
- Jackknife+ and CV+ methods for robustness
- Adaptive Conformal Prediction (ACP) for time-varying uncertainty
- Quantile Regression Averaging (QRA) + Conformal hybrids

Modeling Enhancements

- Incorporate spatiotemporal deep learning models (AGCRN, MTGNN, STGCN)
- Add external factors: weather, events, disruptions, real-time traffic
- Combine multiple routes into a network-wide prediction system
- Explore multistep forecasting pipelines for longer planning horizons