**Valuable Insights From the Dataset:**

* Peak Service usage patterns shift over time, especially around school holidays and exams (detected via ADF test for non-stationarity).
* Rapid Route experiences demand spikes during periods when Light Rail usage drops, revealing a hidden substitution effect.
* Thursday consistently shows the lowest volatility across all services, making it the most stable day for operational planning.
* School service exhibits biweekly volatility spikes, likely tied to school-specific events like exams or field trips.
* Local Route and Light Rail are highly correlated, but this relationship breaks down during holiday seasons, indicating user behavior shifts.

# Technical Report

## 1. Objective

The objective of this analysis was to evaluate and compare the performance forecasting models—Prophet, Random Forest Regressor, and XGBoost Regressor—for predicting ridership counts across various transit service routes over a 7-day horizon.

## 2. Data & Evaluation

The dataset includes historical daily ridership data for 6 transit service types: Local Route, Light Rail, Peak Service, Rapid Route, School and Others. Models were evaluated using Mean Absolute Error (MAE) as the primary metric.

## 3. Model Comparison Summary

|  |  |  |
| --- | --- | --- |
| Algorithm | Strengths | Weaknesses |
| Prophet | Captures trend/seasonality well | Underperforms on irregular patterns |
| Random Forest | Handles non-linearity, fast | Limited extrapolation capability |
| XGBoost | High accuracy, handles lags, missing values, feature importance | Complex to tune |

XGBoost achieved the lowest average MAE, especially after hyperparameter tuning. It showed robust performance even for routes with high variability like Local Route and Light Rail.

## 4. Why XGBoost?

XGBoost (Extreme Gradient Boosting) is a powerful ensemble algorithm that uses boosted decision trees. It provides:  
- Superior predictive accuracy  
- Built-in handling of missing values  
- Regularization to prevent overfitting  
- Fast execution via parallel processing  
  
It performed the best overall in terms of MAE, especially after tuning parameters like:  
- n\_estimators: Number of trees (e.g., 200–500)  
- max\_depth: Tree depth (e.g., 6–10)  
- learning\_rate: Shrinkage rate (e.g., 0.01–0.1)  
- subsample & colsample\_bytree: Randomness control for overfitting  
- reg\_alpha, reg\_lambda: Regularization terms

## 5. Conclusion

Given the lowest error values across most routes and its ability to handle complex, high-dimensional time series data, XGBoost is the recommended model for ridership forecasting in this project.  
  
Future work may explore:  
- Model ensembling (e.g., averaging RF and XGBoost)  
- Incorporating external variables like weather, holidays, or events  
- Deep learning alternatives (e.g., LSTM, TCN) for long-term forecasts