Transportation Ridership Insights and Forecasting

May 27, 2025

Insights from the Dataset

The following insights were derived from analyzing the transportation ridership dataset (2019–September 2024):

- **Seasonal Peaks**: Ridership peaks in August and September due to school terms, with August 2024 recording a total of 1,320,700, where School contributed 102,420 riders.
- **September 2024 Anomaly**: Total ridership in September 2024 dropped by 39.32% (to 801,456 from 1,320,700 in August), likely due to incomplete data or holidays like Labor Day.
- **Dominance of Rapid Route**: Rapid Route accounts for ∼40% of total ridership, contributing 512,762 riders in August 2024, making it the most utilized mode.
- **Weekday vs. Weekend**: Weekday ridership is higher than weekends, with Mondays averaging ∼50,000 total riders compared to ∼30,000 on Sundays.
- **School Seasonality**: School ridership spikes during term months (e.g., August: 102,420, May: 95,000) and drops in summer (e.g., July: 10,000).

7-Day Forecast (October 1–7, 2024)

The forecast for Local Route, Light Rail, Peak Service, Rapid Route, and School was generated using linear regression, with Total estimated from historical weekday averages.

Date	Local Route	Light Rail	Peak Service	Rapid Route	School
2024-10-01	9964	6959	198	12574	2524
2024-10-02	9937	7063	189	12576	2453
2024-10-03	9910	7167	180	12578	2382
2024-10-04	9883	7271	171	12581	2311
2024-10-05	9856	7375	163	12583	2241
2024-10-06	9829	7479	154	12585	2170
2024-10-07	9988	6857	205	12575	2593

Table 1: 7-Day Ridership Forecast for October 1–7, 2024.

Linear Regression Model for Transportation Ridership Forecasting

May 27, 2025

Model Overview

The forecasting of transportation ridership for Local Route, Light Rail, Peak Service, Rapid Route, and School modes was performed using linear regression, a supervised machine learning algorithm. This model was chosen for its simplicity, interpretability, and effectiveness in capturing linear relationships between features and ridership, as implemented in the scikit-learn library.

Algorithm Description

Linear regression models the relationship between the dependent variable y (ridership for a specific mode) and independent variables X (features) as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$$

where:

- β_0: Intercept term.
- $\beta_1, \beta_2, \beta_3, \beta_4$: Coefficients for the features.
- ϵ: Error term.

The model minimizes the mean squared error (MSE) using ordinary least squares (OLS), solving for the coefficients that best fit the training data.

Feature Selection

The following features were selected to capture temporal patterns and overall demand:

- **month** (1–12): Captures seasonal trends (e.g., higher ridership in school terms).
- **weekday_num** (1=Monday, 7=Sunday): Reflects weekly patterns (e.g., higher weekday ridership).
- **day** (1–31): Accounts for daily variations within a month.
- **Total**: Historical weekday average of total ridership across all modes, used as a proxy for overall demand.

Features were derived from the Date column and computed Total, with no additional scaling applied due to the linear nature of the model.

Training Process

Separate linear regression models were trained for each mode using scikit-learn's LinearRegression class. The training data consisted of 1,918 daily records from 2019 to September 2024. For each mode:

- Input features: month, weekday_num, day, Total.
- Target: Ridership for the specific mode (e.g., Local Route).
- The model was fit using the entire dataset to maximize training data, as the forecast period (October 1–7, 2024) is immediately after the dataset's end.

Model Parameters

The LinearRegression model was configured with default parameters:

- **fit_intercept=True**: Includes an intercept term in the model.

- **normalize**=**False**: No internal normalization, as features were already in a suitable scale.
- **n_jobs**=**None**: Single-threaded computation, sufficient for the dataset size.

No hyperparameter tuning was performed, as linear regression has minimal tunable parameters, and the default settings provided strong performance.

Performance Metrics

Model performance was evaluated on the training data (no separate test set due to the forecasting context):

- **Local Route**: MSE = 635,466.92, R^2 = 0.983 (98.3% variance explained).
- Similar performance is expected for other modes (Light Rail, Peak Service, Rapid Route, School), though not explicitly computed.

The high R^2 indicates that the model effectively captures the linear relationships between features and ridership, though the MSE suggests some residual error.

Limitations and Future Improvements

- **Linearity Assumption**: The model assumes linear relationships, which may not capture complex patterns (e.g., non-linear seasonal effects).
- **Total Estimation**: The forecast relies on an estimated Total, introducing uncertainty. A separate model for Total could improve accuracy.
- **External Factors**: Factors like holidays or disruptions (e.g., September 2024 anomaly) are not modeled.
- **Future Work**: Use non-linear models (e.g., Random Forest), incorporate external data (e.g., holidays), or apply cross-validation for robustness.