

# Predicting Ridership of New York City Subway

Jia Bloom • Noah Clark • Chris Enslin • Kshitij Gurung • Minh Duc Pham



## Introduction

Weather conditions can significantly influence people’s decisions to use public transportation, yet the extent of this impact is not fully understood. In a city like New York, where millions depend on the subway, understanding these patterns can support more informed decisions by both commuters and transit planners.

## Datasets & Preprocessing

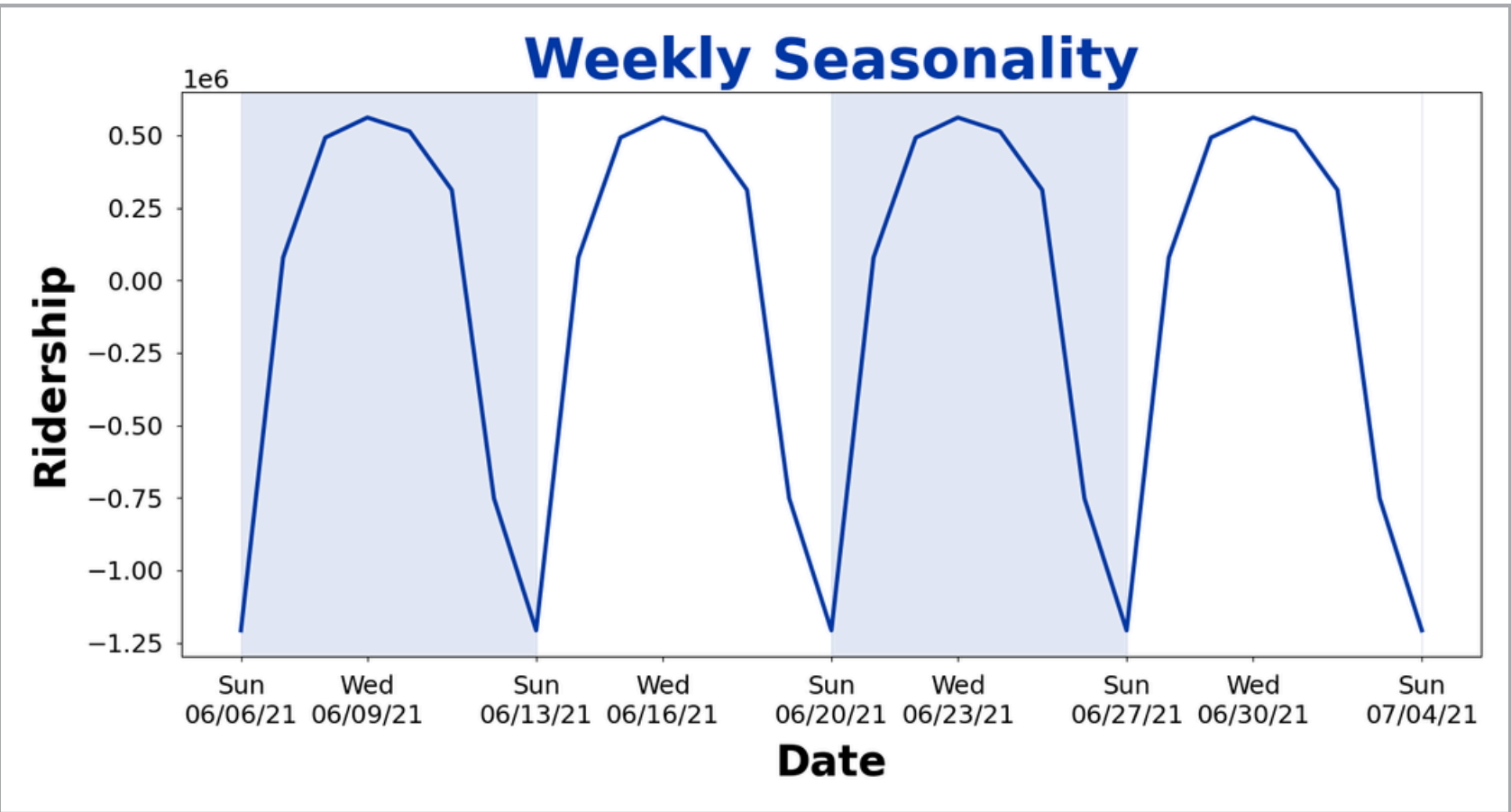
Our main dataset was the **MTA Subway Hourly Ridership: 2020-2024**<sup>1</sup> from data.ny.gov, which contains ~111 million rows and 12 columns. As the title suggests, it includes time series data on hourly ridership at each station complex, originally broken down by payment type. By building a normalized SQLite database, we were able to reduce the dataset size from a >16 GB CSV file to a 7 GB database, aggregated by station and hour. The resulting hourly ridership values served as the dependent variable in our analysis.

In addition, we also incorporated the following datasets:

- **Weather**<sup>2</sup> → accessed historical weather data from Open-meteo API (1,325 rows x 26 cols)
- **Federal Holidays**<sup>3</sup> → sourced from the U.S. Office of Personnel Management (11 rows x 2 cols)
- **School Holidays**<sup>4</sup> → sourced from NYC Public Schools website (39 rows x 3 cols)

## Time Series Analysis

As part of exploratory data analysis, we first conducted basic time series analysis. Our autoregressive (AR) models can then be used as a comparative benchmark for more complex models augmented with additional features, which are expected to yield superior predictive performance. **Decomposition** revealed a weak but slightly positive trend in ridership, while both the seasonal component and **autocorrelation** analysis showed strong weekly cyclicity. We then tested predictive models, including **Holt-Winters Exponential Smoothing** and **SARIMA**, finding that both over-predicted during holidays—highlighting the importance of incorporating holiday data to better predict ridership.



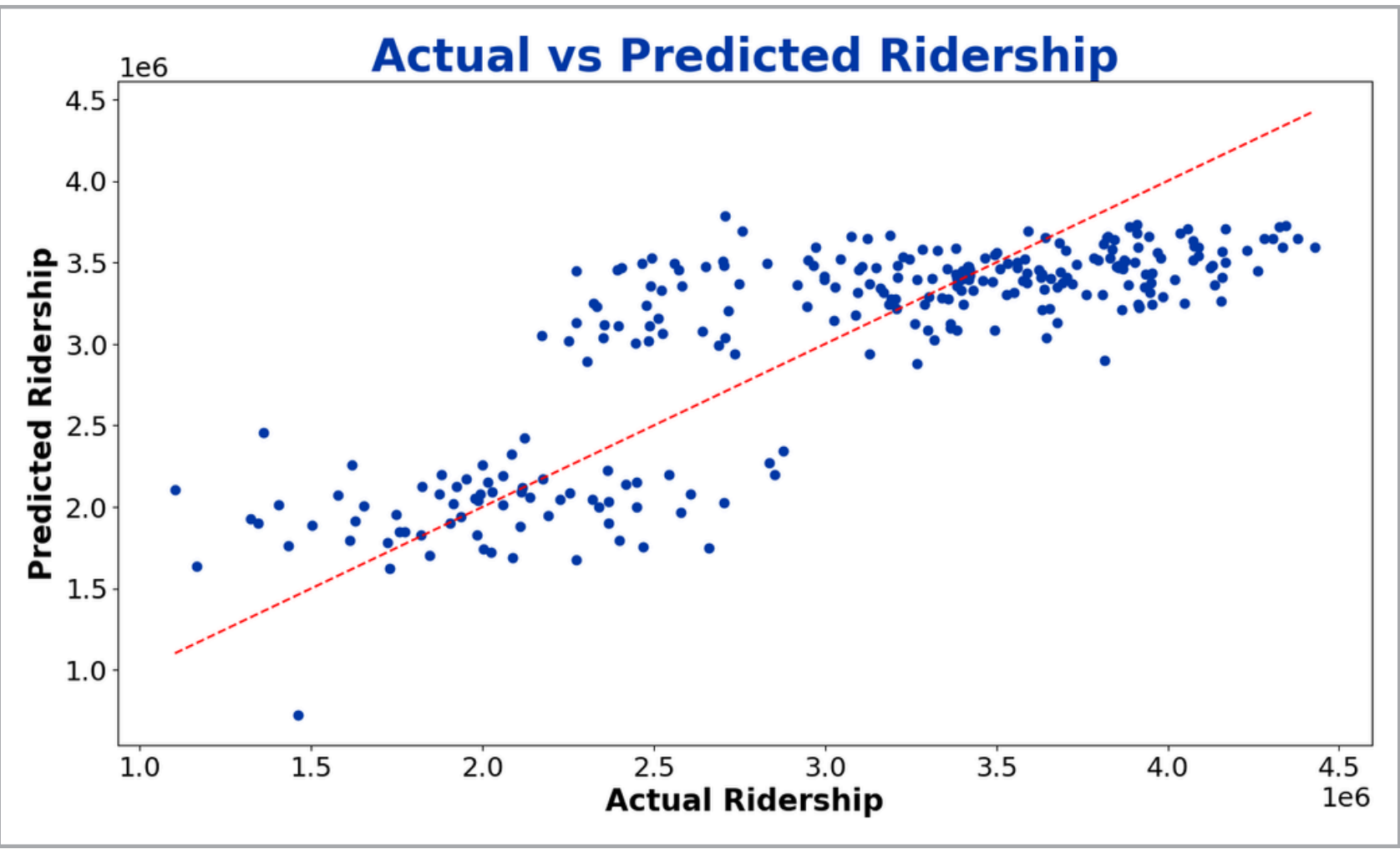
## Feature Engineering

After joining the daily ridership data with the historical weather data, we removed six redundant weather predictors identified through a **correlation matrix** analysis of all variables. Next, we applied feature scaling, including **log transformations** for right-skewed variables such as precipitation and snow.

## Multiple Linear Regression (OLS)

Next, we fit an ordinary least squares multiple linear regression on the reduced and scaled set of weather variables.

## Multiple Linear Regression Results

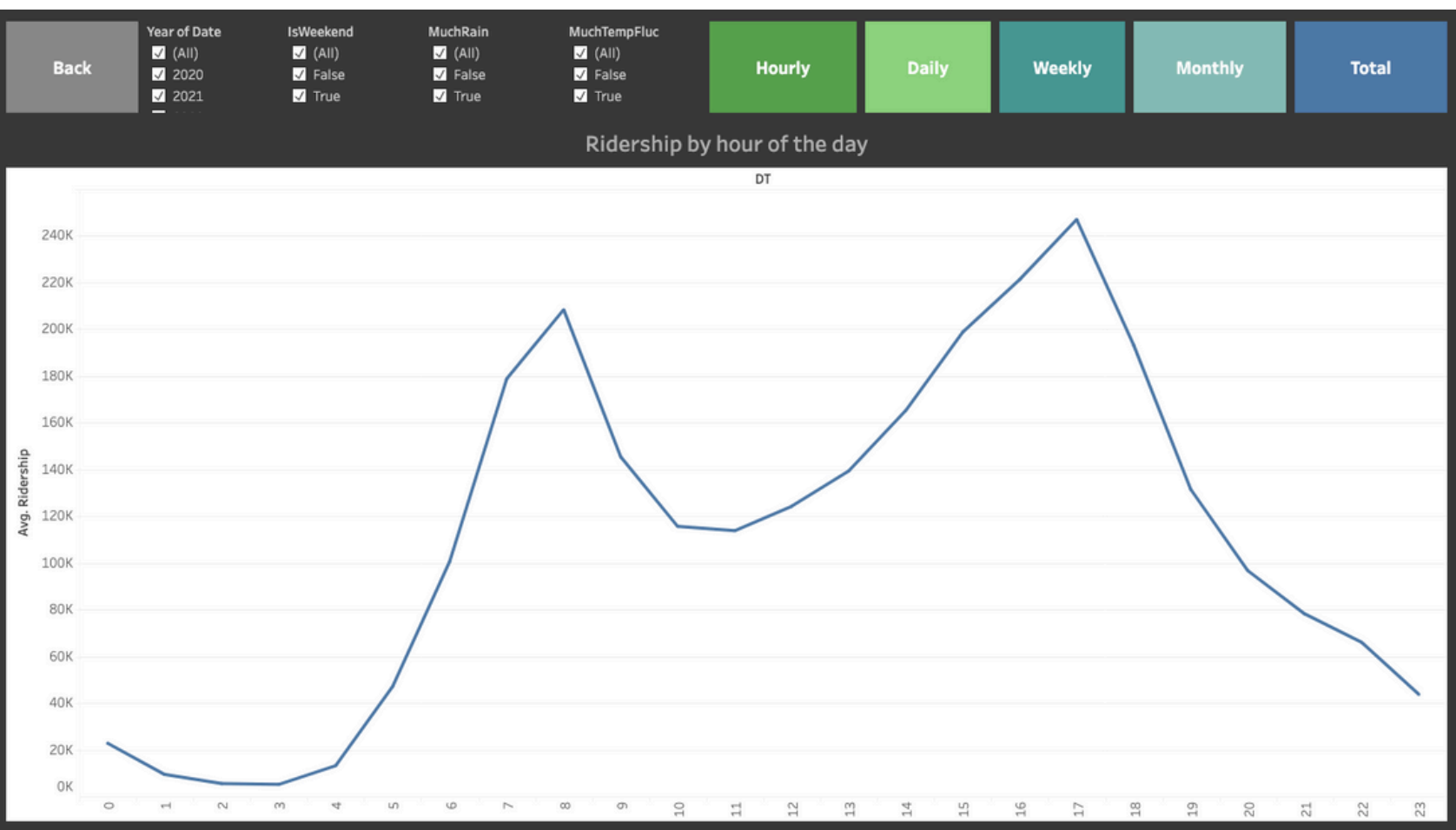


## Random Forest

Given the nonlinear relationships and complex feature interactions observed in the data, we implemented a **Random Forest** model due to its ability to handle high-dimensional data and capture non-additive effects. To assess robustness, we developed several variants with a model incorporating **PCA** for dimensionality reduction and a model reintroducing temporal features (e.g., day of the week). Notably, the model confirmed exploratory insights, attributing significant predictive importance to weekday effects, precipitation (negative), and daily temperature variability (positive).

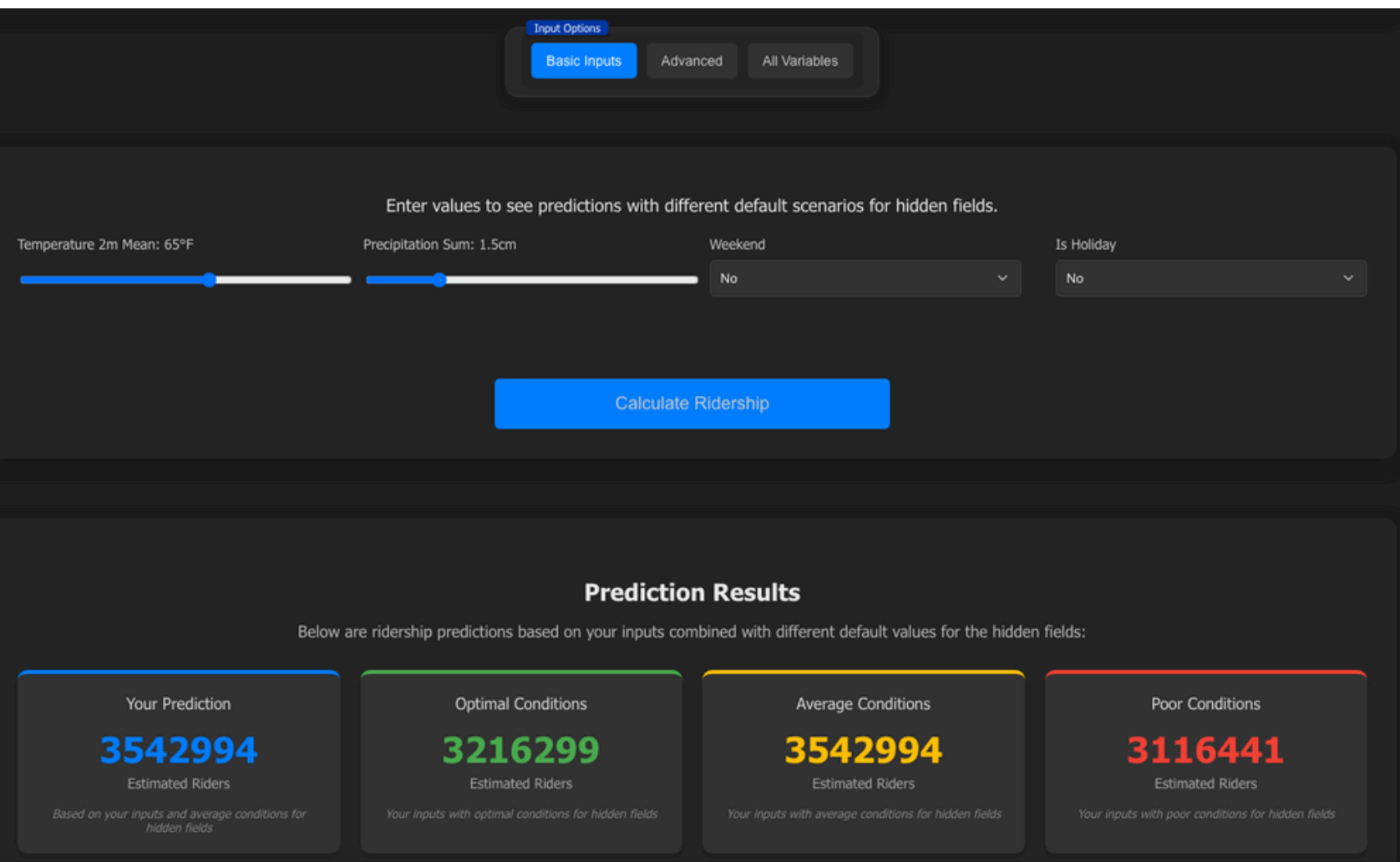
## Interactive Visualizations: Historical

To enhance our data analysis and facilitate independent exploration of the datasets, we developed a **React**-based web application featuring **D3.js** interactive visualizations which analyze how weather and holidays influence subway ridership. Furthermore, we integrated a **Tableau** dashboard, allowing dynamic data filtering to investigate patterns under various weather conditions, holidays, and day-of-week effects over user-selected time horizons.



## Interactive Visualizations: Predictive

A dedicated tab in our web app enables users to input **custom parameters** into the **random forest model**, generating predicted subway ridership values. This feature enhances interactivity by quantifying the influence of individual variables on ridership and improving accessibility and interpretability.



## Other Models



















We explored several alternative methodologies to enhance predictive performance, such as regularized regression techniques—including **Ridge**, **Lasso**, and **Elastic Net regression**. We also implemented a **Recurrent Neural Network (RNN)** with **LSTM** to capture sequential dependencies in ridership data, leveraging the data’s weekly seasonality for better forecasting.

## Methodological Comparisons

Our project introduces several key innovations relative to prior research on weather and transit ridership.

1. **NYC** → Prior studies have primarily examined transit systems in China and other Asian cities.
2. **Expanded Modeling** → Unlike many existing studies which rely solely on linear regression, we evaluate a wider range of models to improve predictive power and capture nonlinear relationships.
3. **Integrated Weather & Holiday Data** → Previous research often focuses solely on weather effects, whereas our approach incorporates holidays.
4. **Interactive Public Tool** → Our development of an interactive platform enhances the practical utility of our research compared to conventional static studies.

## Results Summary

Model	Predictors	RMSE	R <sup>2</sup>
Random Forest	  	398,865	0.748
CATBoost Regressor	  	409,241	0.735
Recurrent Neural Network (RNN) + LSTM	  	429,260	0.729
Ridge Regression	  	482,812	0.631
Random Forest w/ PCA	 	483,999	0.629
Multiple Linear Regression	 	515,576	0.580
Holt-Winters Exponential Smoothing		554,598	n/a
SARIMA		617,654	n/a

 temporal/AR data •  weather data •  holiday data

## Conclusions

The **Random Forest** model, leveraging weather, holidays, and temporal features, proved the most accurate predictor of subway ridership, outperforming linear regression and other machine learning approaches. Notably, ridership peaks mid-week—especially on Tuesdays and Wednesdays with longer daylight duration—but declines on weekends and holidays. The model also reveals how environmental factors influence demand: higher temperatures, wider temperature swings, increased precipitation, and stronger winds all negatively impacts ridership. By capturing these complex patterns, the framework offers transit planners and commuters actionable insights, supported by interactive tools, to optimize decision-making in NYC’s dynamic public transportation system.

## References

1. MTA Subway Hourly Ridership: 2020-2024 [https://www.data.ny.gov/d/wuig-7c2s]
2. OpenWeather API [https://www.open-meteo.com]
3. Federal Holidays [https://www.opm.gov/policy-data-oversight/pay-leave/federal-holidays/#url=Overview]
4. NYC Public Schools Calendar [https://www.schools.nyc.gov/calendar]