Christian Downey

December 16th, 2024

Introduction

The COVID-19 pandemic brought unprecedented challenges to urban mobility systems, with New York City's subway network experiencing one of its most significant disruptions in history. In March 2020, as the city became an early epicenter of the pandemic, lockdown measures, social distancing guidelines, and widespread fear of virus transmission led to a steep decline in subway ridership. Daily ridership plummeted by over 90% during the pandemic. The subsequent recovery of subway ridership has been uneven and influenced by a complex interplay of factors, underscoring the importance of understanding these dynamics for urban mobility and economic recovery.

This final project seeks to expand on my midterm project, which predicted future ridership data using a Linear Regression model. Now, comprehensive COVID-19 case and vaccine data are included, as well as daily weather information.

Data description

A comprehensive analysis of New York City subway ridership recovery requires a robust dataset that captures the various factors influencing commuter behavior. This study integrates data from three primary sources: NYC Subway ridership data from data.ny.gov, weather data from NOAA's Climate Data Online (CDO) API, and COVID-19 vaccination/case data from the New York City Department of Health's GitHub repo.

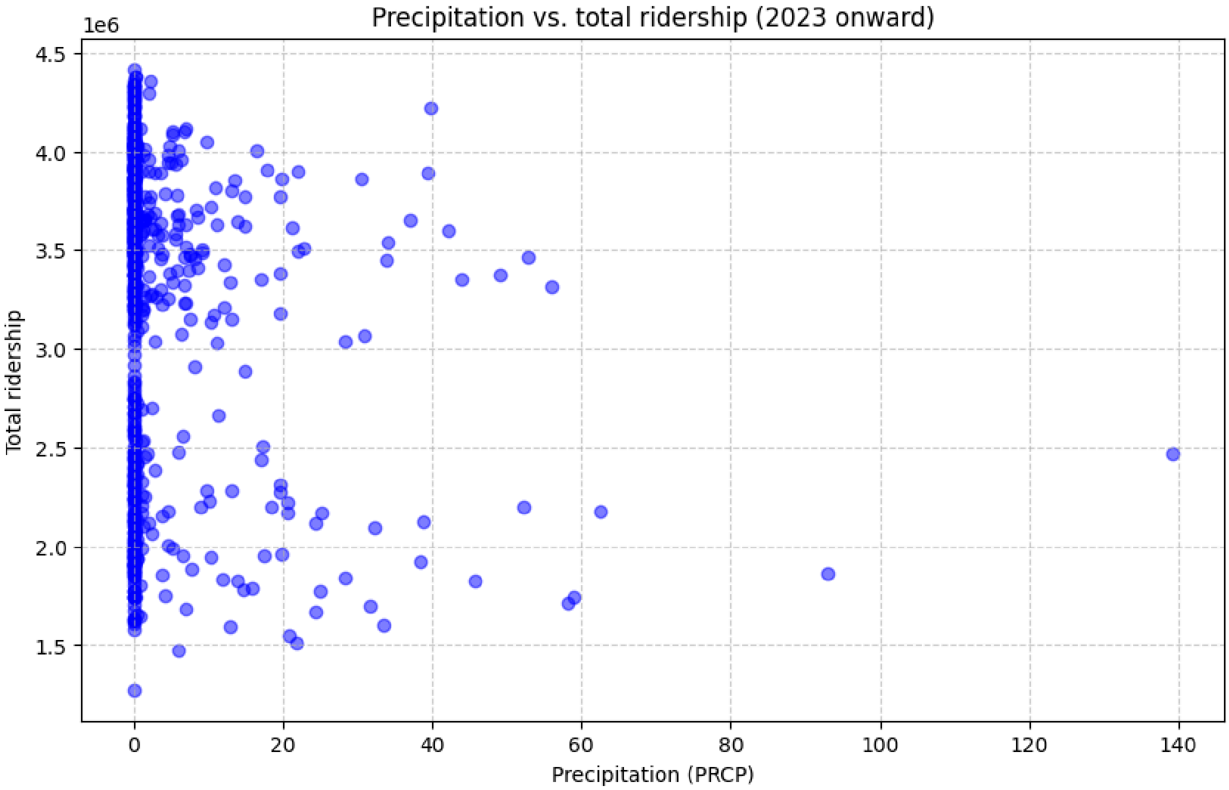
The primary dataset for this study is the NYC Subway Ridership Data from data.ny.gov. This dataset provides daily subway ridership counts for the entire system, as well as comparable pre-pandemic ridership figures. The subway ridership data serves as the dependent variable in the study's predictive models.

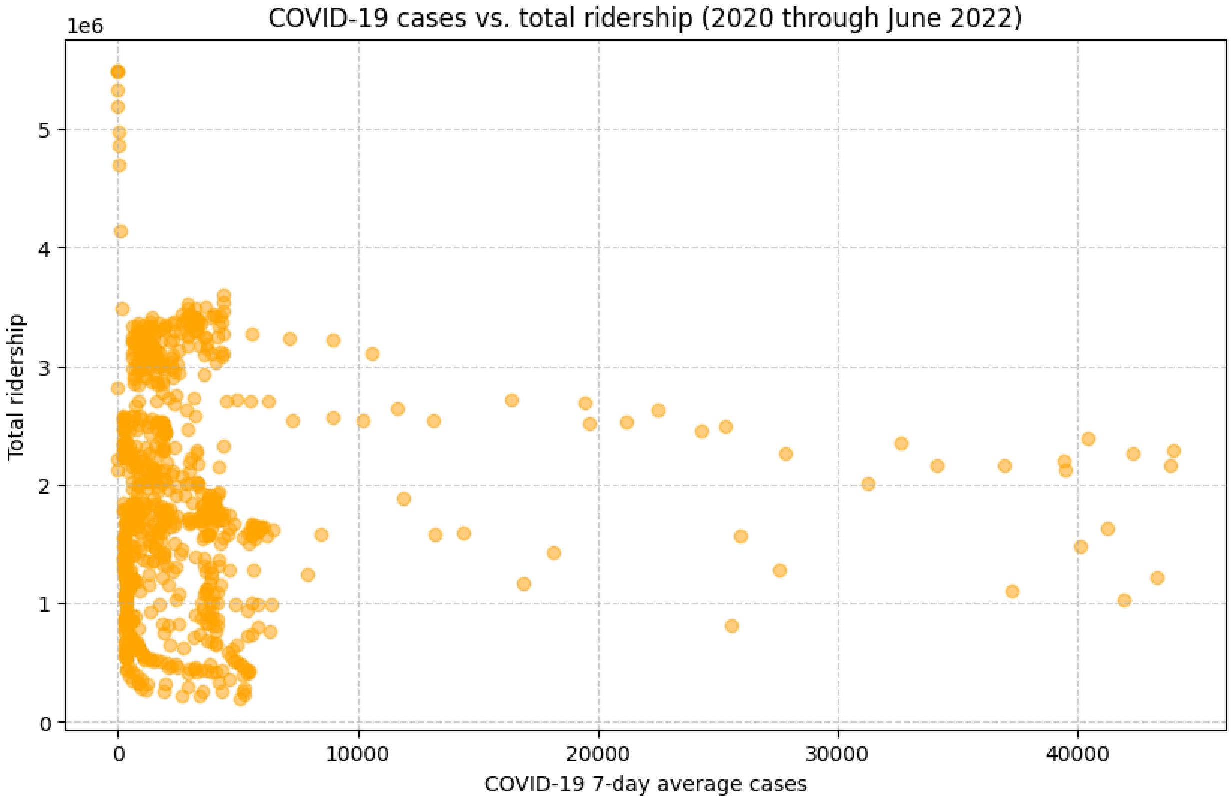
To account for external environmental factors, weather data is obtained from NOAA's Climate Data Online (CDO) API. Weather variables such as daily maximum temperature (TMAX), minimum temperature (TMIN), and precipitation (PRCP) are included. Weather conditions play a crucial role in commuter decision-making, as extreme weather events may discourage travel or disrupt service. On the other hand, inclement or cold weather may encourage subway ridership to avoid such conditions.

Health-related data is drawn from two critical sources: COVID-19 vaccination rates and COVID-19 case counts, both sourced from the NYC Department of Health's GitHub repo. The vaccination dataset tracks the daily cumulative percentage of New York City residents who have received at least one dose of a COVID-19 vaccine. The COVID-19 case data captures the seven-day average of daily new cases, reflecting the perceived health risk at any given time. As case counts rise, ridership may decrease due to heightened risk aversion, while declining case counts may have the opposite effect.

Exploratory data analysis

Before going into making predictions, I reviewed the data to see if I could visualize any relationships between precipitation and ridership as well as COVID-19 cases and ridership.





I tried filtering the dataset twice: data after 2023 for precipitation, and data between March 2020 and June 2022 for COVID-19 cases. I did this to avoid cases where one variable may not truly affect ridership data: for example, a very rainy day may cause one to think that ridership would be lower (due to delays or other behavioral changes), however, ridership is actually lower because COVID-19 cases are at an all-time high.

Even still, it’s hard to see a clear correlation between cases or precipitation on ridership.

Models and methods

Given the complex nature of ridership recovery, no single model is sufficient to account for all possible relationships and dependencies in the data. Thus, three predictive models were selected for analysis: Linear Regression, Random Forest Regressor, and ARIMA.

Since the COVID-19 vaccine data only extends to September 2023, the data before that date was used to train each model and make predictions to today. Afterward, each model was compared to actual ridership figures from September 2023 to December 2024.

Linear Regression

The first model used in this study is Linear Regression, one of the most fundamental and widely used predictive modeling techniques. The objective of this model is to identify and quantify the linear relationship between predictors (weather, vaccination rates, and COVID-19 case counts) and subway ridership recovery. Linear regression is especially useful for its interpretability, as the coefficients of the model directly indicate the magnitude and direction of the impact each predictor has on the target variable. To train and test the model, the data was split into two subsets: 80% of the data was used for training, and the remaining 20% was reserved for testing.

The features used in this model included daily maximum temperature (TMAX), daily minimum temperature (TMIN), precipitation (PRCP), vaccination coverage (CITY_PERC_1PLUS), and COVID-19 case counts. The response variable was subway ridership recovery, expressed as a proportion of pre-pandemic ridership levels.

The primary strength of linear regression lies in its simplicity and interpretability. Coefficients provide measurable insight into the direction and magnitude of each predictor's impact on ridership. However, linear regression assumes that relationships between predictors and the response variable are linear, which may oversimplify things. It also struggles to capture nonlinear interactions between variables.

Random Forest Regressor

To address the limitations of linear regression, the second model applied in this study is the Random Forest Regressor, a more flexible, non-parametric ensemble learning method. Random forest models are well-suited for handling non-linear relationships and feature interactions, which are often present in real-world data. Unlike linear regression, random forest does not assume a linear relationship between predictors and the target variable. Instead, it constructs multiple decision trees and aggregates their predictions to form a robust, stable model.

Similar to linear regression, the data was split into an 80-20 ratio, with 80% used for training and 20% reserved for testing. This approach prevents overfitting and ensures the generalization of model predictions.

The Random Forest's key strength is its ability to capture non-linear relationships and interactions between features. It also provides insight into feature importance, which can inform policymakers about which factors have the greatest influence on subway ridership. However, random forests require greater computational resources and offer lower interpretability compared to linear regression, as it is challenging to extract simple coefficients to explain individual feature effects.

ARIMA

While linear regression and random forest models rely on explanatory variables to make predictions, ARIMA (Auto-Regressive Integrated Moving Average) takes a different approach. It focuses on the time-series structure of subway ridership itself, accounting for trends, seasonality, and autocorrelation. This is especially useful for capturing daily, weekly, or seasonal fluctuations in subway use, which are not fully explained by predictors like weather or health indicators.

Unlike the linear regression and random forest models, which use multiple predictors, ARIMA requires only the ridership time series as input. The focus of ARIMA is to model the temporal dependencies within the ridership data itself.

The ARIMA model was configured with an order of (1, 1, 1), which specifies 1 autoregressive term (AR), 1 difference (I) to achieve stationarity, and 1 moving average term (MA). After playing around with different orders, (1, 1, 1) seemed to work best on this dataset.

ARIMA's primary strength lies in its ability to model time-series and seasonal patterns. This makes it well-suited for forecasting short-term changes in subway ridership. However, ARIMA has a major limitation: it does not incorporate exogenous predictors (like weather or health data) into its analysis. It also seemed to flatline, regardless of what order I specified. There may be some issue in the code, because after about a few weeks of predictions, the ARIMA predictions stay at one value.

Results and interpretation

Each model offers a unique perspective, with linear regression highlighting simple relationships, random forest capturing non-linear feature interactions, and ARIMA accounting for temporal dependencies. By comparing the models' performance using key evaluation metrics such as R-squared (R^2), mean squared error (MSE), and qualitative insights, the study identifies which model is best suited for predicting subway ridership.

Overall, the Random Forest Regressor emerged as the most effective model, outperforming linear regression in both R^2 and MSE while providing insight into the relative importance of different predictors. ARIMA demonstrated strong short-term forecasting capabilities, excelling in modeling time-based patterns in subway ridership. Each model made a distinct contribution to understanding the factors driving ridership recovery.

Model performance table

Model	R^2	MSE	AIC	BIC	Feature importance
Linear Regression	0.6467	0.0062	N/A	N/A	Vaccine coverage (0.4386)
Random Forest	0.6567	0.0060	N/A	N/A	Vaccine coverage (0.755)
ARIMA	N/A	N/A	-2238.681	-2223.952	N/A

Conclusion and next steps

The COVID-19 pandemic deeply impacted subway ridership in New York City, necessitating a comprehensive analysis of the factors driving its recovery. This study applied three distinct predictive models—Linear Regression, Random Forest Regressor, and ARIMA—to understand the key drivers of ridership and forecast future usage. Each model brought unique strengths to the analysis, capturing different aspects of the ridership recovery process.

The most effective model was the Random Forest Regressor, which demonstrated superior predictive power by capturing non-linear relationships and identifying critical predictors. Among the five predictors used in the analysis—vaccine coverage, precipitation, maximum and minimum temperature, and COVID-19 case counts—vaccine coverage emerged as the most influential variable, with a feature importance score of 0.755. Additionally, random forest captured the complex, non-linear interactions between predictors, improving prediction accuracy relative to linear regression.

The ARIMA model also played an important role, offering a time-series approach that captured temporal dependencies and short-term seasonality. While it did not account for external predictors like weather and health data, ARIMA excelled at forecasting ridership based solely on its past values. With its ability to model time-based fluctuations, ARIMA is particularly useful for short-term forecasts, providing transit agencies with crucial insight into daily and weekly ridership changes.

Finally, Linear Regression served as a useful baseline model, offering clear and interpretable coefficients that quantified the linear impact of each predictor. Vaccine coverage had the largest positive impact, but the model struggled to capture the non-linear effects of precipitation and COVID-19 case counts. Despite its simplicity, linear regression provided valuable insights into how certain predictors influence ridership.

Future work

While the models used in this analysis offer valuable insights, there is room for improvement:

1. Feature expansion
   To improve predictive accuracy, future models could incorporate additional predictors beyond weather, vaccination rates, and COVID-19 case counts. Two key features for future analysis are unemployment rates and Google Mobility data. Unemployment rates affect commuter behavior, as unemployed individuals are less likely to commute for work. Google Mobility data, which tracks human movement, could offer real-time insight into population movement trends.

2. Model enhancements
   While the Random Forest Regressor outperformed the other models, its predictive accuracy could be further improved through hyperparameter optimization. Tuning key hyperparameters, such as the number of decision trees, maximum tree depth, and feature selection criteria, could enhance model performance. Similarly, ARIMA could benefit from a grid search approach to find the optimal combination of AR, I, and MA terms. More sophisticated time-series models, such as Seasonal ARIMA (SARIMA) or ARIMAX (which incorporates exogenous predictors), could also be explored. It would be incredibly useful to have seen ARIMA make a full prediction of the ridership, instead of just flatlining.

Final thoughts

The recovery of New York City's subway system is critical to the city's economic resurgence and social connectivity. This analysis demonstrates that a multi-model approach – incorporating Linear Regression, Random Forest Regressor, and ARIMA – is essential for capturing the complexity of subway ridership recovery. Each model provides distinct but complementary insights. Linear regression identifies clear linear relationships, Random Forest Regressor captures non-linear feature interactions, and ARIMA addresses time-series dependencies, making it useful for short-term forecasting. Together, these models offer a comprehensive view of the key drivers of ridership recovery.

The findings highlight the role of vaccine coverage as the most influential driver of ridership, while weather conditions such as precipitation and temperature play important but secondary roles. The identification of these key factors offers a roadmap for policymakers to take targeted action. Public health interventions, weather-based service adjustments, and the implementation of real-time forecasting dashboards can all help accelerate ridership recovery.

Looking forward, the integration of unemployment data, Google Mobility trends, and real-time data feeds from weather and health authorities will further enhance the accuracy of predictive models. The COVID-19 pandemic provided a unique opportunity to understand how health crises affect urban mobility. The insights gained from this analysis can be applied to future public health events or natural disasters, ensuring that New York City's subway system remains resilient, adaptable, and equitable for all commuters.