Predictive Maintenance for NYC Subway Lines

Atmika Pai
Cornell Tech, New York, NY
MS in Information Systems,
Concentration in Urban Tech
aap253@cornell.edu

Bhoomika Mehta
Cornell Tech, New York, NY
MS in Information Systems,
Concentration in Connective Media
bm726@cornell.edu

Nikhil Jain
Cornell Tech, New York, NY
MS in Information Systems,
Concentration in Urban Tech
nj289@cornell.edu

Abstract

New York City's Metropolitan Transit Authority (MTA) boasts the largest subway system in the United States, serving 3.6 million daily riders annually, but it is not without its issues. It is in dire need of infrastructure investments to maintain safe and timely operations. The MTA has published a 20-year Needs Assessment [2] that highlights aging infrastructure, evolving climate change, and increasing ridership as problems that will affect the scale and speed at which the subway needs targeted maintenance. With that in mind, our research investigates novel machine learning techniques to assess the probability of a service disruption type and its expected impact on passenger travel times based on subway line and weather-related factors. After building the respective supervised learning models, we cluster subway lines based on their contribution to network vulnerability using unsupervised learning techniques. This holistic approach to service disruptions will help MTA take more concerted maintenance measures and optimize resources allocation.

1. Motivation

Disruptions in public transport significantly impact passengers by increasing travel times and burdening transit authorities with higher repair and maintenance costs. While MTA reports a strong on-time performance of over 80%, defined as the percentage of subway trips running on schedule, these service disruptions still delay more than 30,000 trains each month, primarily driven by internal factors like infrastructure or equipment breakdowns, crew shortages, and operational challenges.[3] Over time, such disruptions have significant financial consequences, so it is in the interest of transit authorities to prioritize maintenance efforts to deliver outstanding service to its constituents.

Two key studies guide our analysis. In *Artificial Intelligence-Aided Rail Transit Infrastructure Data Mining*, Liu and Dai propose an ML-based method for predicting urban rail transit signal failures up to one month in advance.[1] Motivated by their findings, we integrate weather characteristics and adopt XGBoost as one of our primary ML models. The second study, *Predicting Disruptions and Their Passenger Delay Impacts for Public Transport Stops* by Yap and Cats[4], provides a replicable framework for modeling disruption frequencies and passenger delay impacts across public transport networks, which forms the crux of our analysis. While their study focuses on the Washington D.C. Metro at the station level, we adapt their framework to the NYC Subway System at the line level due to the lack of granular station-level data.

2. Model Framework

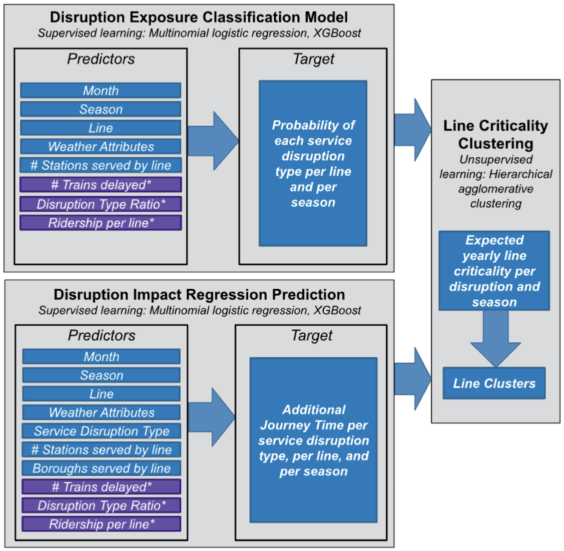
We develop a machine learning framework consisting of three models: *Disruption Exposure Classification Model* predicts the frequency of service disruption types across subway lines; *Disruption Impact Regression Model* evaluates their impact on passenger journey times, and *Line Criticality Clustering* clusters lines by seasonal line criticality, which measures a line's importance in maintaining system efficiency and reliability. Figure 1 highlights each model's target variable, feature vectors, and their role within the overall research framework.

2.1. Disruption Exposure Classification Model

The objective of this model is to forecast the probability $\hat{p}_{d,l,s}$ of each service disruption d and season s per subway line l. Once trained, this model enumerates the degree to which a particular line is exposed to disruptions.

2.2. Disruption Impact Regression Model

The objective of this model is to evaluate the impact of service disruptions on passenger journey times, segmented by service disruption type d, subway line l, and season s. Leveraging MTA's Customer Journey-Focused Metrics, the model utilizes Additional Journey Time (AJT) — the estimated extra time one passenger spends on their journey compared to the scheduled time — as the target variable.


Disruption Exposure Classification Model
Supervised learning: Multinomial logistic regression, XGBoost
Predictors
Month
Season
Line
Weather Attributes
# Stations served by line
# Trains delayed*
Disruption Type Ratio*
Ridership per line*
Target
Probability of each service disruption type per line and per season
Line Criticality Clustering
Unsupervised learning: Hierarchical agglomerative clustering
Expected yearly line criticality per disruption and season
Line Clusters
Disruption Impact Regression Prediction
Supervised learning: Multinomial logistic regression, XGBoost
Predictors
Month
Season
Line
Weather Attributes
Service Disruption Type
# Stations served by line
Boroughs served by line
# Trains delayed*
Disruption Type Ratio*
Ridership per line*
Target
Additional Journey Time per service disruption type, per line, and per season

* Purple indicates previous month historicals were used in model

Figure 1. Modelling Framework (Adapted from Yap and Cats Research)[4]

As such, this model gives us estimated AJT, $\hat{AJT}_{d,l,s}$, per service disruption d, season s, and line l, enumerating the impact of a disruption on passengers' travel times once a service disruption has occurred.

2.3. Line Criticality Clustering

The objective of this model is to categorize lines based on their expected susceptibility to disruptions. By predicting disruption exposure and impact, the clustering identifies stations that contribute most to the overall vulnerability of the subway network. For each disruption type d at line l and season s, the criticality score is computed by multiplying the disruption probability $\hat{p}_{d,l,s}$ by the disruption impact $\hat{w}_{d,l,s}$, summed across all disruption types and seasons:

$$\hat{c}_l = \sum_{s \in S} \sum_{d \in D} \hat{p}_{d,l,s} \times \hat{AJT}_{d,l,s} \tag{1}$$

The criticality score captures both a line's vulnerability to disruptions (exposure) and the severity of their impact (importance). This holistic approach enables transport agencies to prioritize mitigation strategies by distinguishing between lines that are highly susceptible to frequent and severe disruptions and those with lower susceptibility.

3. Data

We primarily use data provided by MTA through NYC's Open Data Portal, with variables dating from 2020 onwards. Our analysis indicates that the number of train delays due to service disruptions stabilized in mid-2021 as seen in Figure 2. As such, we limit the study period to July 2021, which is also when the pandemic lockdown was lifted in NYC, to October 2024. We also limit our dataset to weekdays and peak hours, defined by the MTA as 7 a.m. to 10 a.m. and 4 p.m. to 7 p.m., to ensure consistency when merging various datasets. In total, our dataset consists of 5,152 observations spanning three years and 23 subway lines. For ease of analysis, the S lines, which include shuttle services (42nd Street, Rockaway Park, and Franklin Avenue), were aggregated.

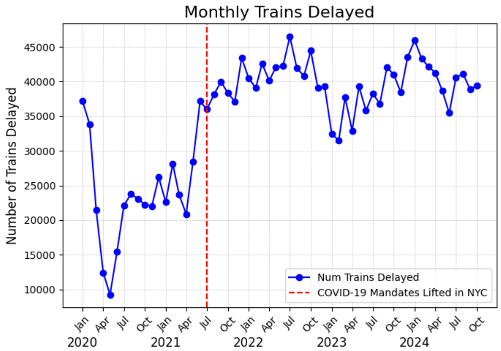


Figure 2. Research Time Period

3.1. MTA Datasets

The MTA Subway Trains Delayed Dataset records the number of subway trains delayed from scheduled service, categorized into six distinct service disruption types: Crew Availability, Infrastructure & Equipment, Operating Conditions, Planned ROW Work, Police & Medical, and External Factors. The dataset includes a subcategory detailing the causes of service disruptions. We create the variable `Disruption_Subcategory_Count` to capture the number of potential causes for each disruption type. We also introduce the variable `Disruption_Type_Ratio`, which represents the ratio of disruptions of a specific type to the total disruptions for the month.

The MTA Customer Journey-Focused Metrics dataset tracks monthly subway performance and ridership by line, focusing on delays, reliability, and service quality. We primarily use the passenger volume per subway line and Additional Journey Time (AJT) metric, which represents the average extra time each passenger spends compared to the scheduled time. It is the sum of the average additional time spent waiting on platforms (APT) and the average additional time spent onboard a train (ATT) by one passenger.

We use the MTA Subway Entrances & Exits dataset to collect subway line attributes such as number of stops and the boroughs served by each line.

3.2. Weather Characteristics

We incorporate weather characteristics from Climate Data Online by the National Oceanic and Atmospheric Administration. Using the *Global Summary of the Month* dataset, we focus primarily on minimum and maximum temperatures, precipitation and snow levels, as well as extreme temperatures expressed as binary indicators.

3.3. Feature Engineering

We employ the previous month's number of trains delayed (`Num_Trains_Delayed`), the newly added Disruption Type Ratio (`Disruption_Type_Ratio`), and Passenger Volume (`Ridership`) as lagged variables to account for temporal patterns in the data and how past events influence current disruptions. By including these terms, the model can better assess recurring trends or hidden signals, which are particularly useful for detecting rare disruptions.

We also one-hot encode categorical variables such as `Service_Disruption_Type`, `Season`, and `Borough_Serviced`, allowing the model to effectively process these non-numeric attributes. We standardize numerical variables like `Num_Trains_Delayed`, `Ridership` and all weather attributes to ensure they are on the same scale, preventing any single variable from dominating model results.

4. Methodology

4.1. Disruption Exposure Classification Model

We utilize a supervised learning approach for this model, experimenting with two different machine learning techniques: Multinomial Logistic Regression and XGBoost classification as recommended by our research literature.

To evaluate the performance and generalizability of our models, we partition the dataset into an 80% training set and a 20% test set. We employ performance metrics like the F1 score, logarithmic loss, and precision, with a particular focus on the model's ability to handle class imbalance. Logarithmic loss was used to compute the negative log-likelihood of the true label, based on the predicted probability of a sample belonging to that label. The F1 score is used as a primary accuracy metric, calculated globally by combining the total true positives, false negatives, and false positives. This approach ensures that the metric accounts for the class imbalance present in our dataset, particularly in the case of rarer service disruption types like *External Factors*. Finally, we analyze these metrics for both the training and test sets to identify potential overfitting.

4.2. Disruption Impact Regression Model

For the prediction of passenger impacts of disruptions, we also apply a supervised learning approach. To quantify passenger delays, we use MTA's Additional Journey Time (AJT) metric as our target variable.

We experiment with simple linear regression model as a baseline and an XGBoost model for comparison. To evaluate model generalizability of our models, we again partition the dataset into an 80% training set and a 20% test set. The RMSE (Root Mean Squared Error) and R² are used as the primary performance metrics for the regression models.

4.3. Line Criticality Clustering

The final models for predicting disruption exposure and impact are used to cluster lines based on their expected criticality. The disruption exposure model predicts disruption probabilities for each type d and season s, while the disruption impact model estimates the impact for each disruption type on each line l for every season s. These predictions are combined (e.g., using Equation 1) to calculate the expected criticality for each disruption type, line, and season, expressed in average passenger delay in minutes.

An unsupervised learning method is then applied to cluster lines based on this criticality, revealing differences in susceptibility to various disruption types and grouping lines with similar exposure and impact patterns. The criticality values are normalized to ensure comparability across lines, and the optimal number of clusters is determined using the elbow method.

5. Results and Discussion

5.1. Disruption Exposure Classification Model

The logistic regression model achieved a test F1-score of 0.803, performing well for high-frequency service disruption categories like *Infrastructure & Equipment* but struggling with minority classes like *Planned ROW Work* and *Police & Medical*. The model generalized well, as evidenced by similar F1 scores on the training and test sets.

Table 1. Model 1 Performance Comparison

Metrics	**Logistic**	**XGBoost**
Test F1 Score	0.803	0.938
Train F1 Score	0.829	1.00
Test Log-Loss	0.428	0.0142
Train Log-Loss	0.417	0.0100

The XGBoost classifier achieved a significantly higher test F1-score of 0.94 and lower test log-loss of 0.0142, likely due to its ability to capture non-linear relationships and complex feature interactions. XGBoost may be overfitting, as indicated by the perfect train F1 score of 1.000. To address this, we removed features with low predictive power and applied cross-validation and grid search to tune regularization parameters and reduce the maximum tree depth.

5.1.1 Observations

1. *Planned ROW Work* and *Crew Availability* had the highest probability of service disruption at 0.171 independent of line and season.

2. The G, M, and Q lines show elevated disruption probabilities due to *External Factors* and *Operating Conditions* service disruption types. The G line runs through the boroughs of Brooklyn and Queens, so it may be particularly vulnerable due to its elevated tracks and intersections with areas undergoing rapid urban development and street-level construction, which is also the case with M and Q lines.
3. The 3, 7, and L, which lines serve major commercial corridors in Manhattan and Queens, experience more disruptions, likely due to infrastructure strain and heavy ridership.
4. The 1, 2, 3, 4, and 5 lines, despite not having the highest disruption probabilities in specific categories, face disruptions related to *Crew Availability* and *Operating Conditions.*
5. Evaluating feature importance, attributes directly related to disruptions (i.e. unique count of service disruption subcategories, number of trains delayed, and Disruption Type Ratio) demonstrated the highest predictive power, while weather-related features (i.e. season) had very low predictive power.

5.2. Disruption Impact Regression Model

For disruption impact prediction, linear regression served as a baseline, achieving an R² score of 0.771 but struggling to capture complex interactions between features such as ridership and weather conditions. In contrast, the XGBoost regression model substantially improved prediction accuracy with an R² score of 0.94 and a mean absolute error (MAE) of 443,471, nearly halving the linear model's MAE of 858,924 .The models demonstrated strong generalization, as indicated by the similar R² and RMSE scores across the train and test sets. While XGBoost showed slightly lower RMSE and higher R² on the training set compared to the test set, this minor gap suggests a small degree of overfitting that remains within an acceptable range.

Table 2. Model 2 Performance Comparison

Metrics	Linear	XGBoost
Test RMSE Score	1,207,522	622,771
Train RMSE Score	1,221,522	567,472
Test R²	0.771	0.939
Train R²	0.764	0.949

5.2.1 Observations

1. AJT did not vary too much by service disruption type.
2. Higher AJT values are linked to lines with high ridership, aging infrastructure, and external disruptions, such as B, C, and Q, which result in significant delays.

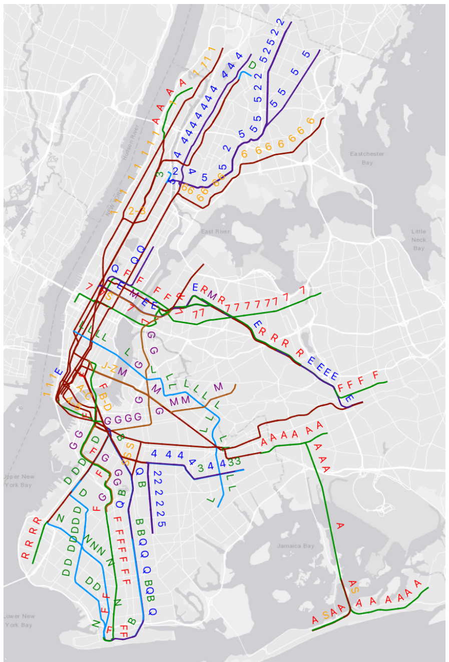


Figure 3. Mapped NYC Subway Lines with Associated Cluster

3. *Crew Availability* is a consistent factor contributing to disruptions across various lines, especially those with moderate to high disruption impact, such as Q and M.
4. Lines serving less dense or peripheral areas, such as L, F, and 1, tend to exhibit lower disruption impacts, possibly due to more efficient operations, lower ridership, or newer infrastructure.
5. Ridership and weather attributes emerged as the most critical features. These findings underscore the need for integrating weather and temporal data into subway maintenance predictive frameworks. In contrast, boroughs serviced by the line (represented as indicators) had low predictive power.

5.3. Line Criticality Clustering

The elbow method analysis identified five optimal clusters among NYC's subway lines, reflecting susceptibility of each subway line and its overall vulnerability to the greater subway network. Table 3 presents the clusters and their respective subway lines and Figure 3 visualizes the results on a map (legend is available in Figure 3).

5.3.1 Observations

1. Cluster 4 and 0 lines have the highest normalized criticality scores and disruption exposure probabilities, reflecting their vulnerability in the subway network.

Table 3. Subway Lines by Cluster

Cluster	Subway Lines	Color
0	R, 7, 7X, A, F	Green
1	3, N, B, D, L	Blue
2	2, 4, 5, Q, E	Purple
3	M, C, G	Orange
4	6, 6X, 1	Red

2. Lines in Cluster 3 play vital role in linking residential areas to commercial hubs, resulting in moderate criticality even though they have high AJT.

3. These findings highlight the need to prioritize reliability improvements for Clusters 4 and 0, as disruptions in these clusters have the most severe impact on passengers.

4. Clusters 0 and 4 show high disruption exposure but median or low AJT, suggesting they may represent subway lines with robust operational procedures that manage disruptions effectively. This could present a learning opportunity for other subway lines to adopt similar strategies for better disruption management.

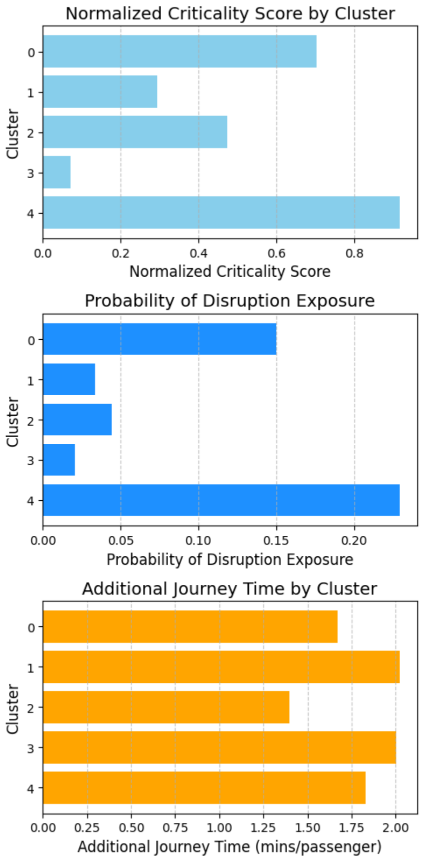


Figure 4. Normalized Criticality Score by Cluster

6. Conclusion

This research evaluated service disruptions across NYC subway lines by modeling both disruption exposure and its impact on passenger journey times to asses vulnerability of each subway line within the NYC subway system.

The *Disruption Exposure Classification Model* highlighted that factors such as *Crew Availability* were major contributors to disruption probabilities, with lines like G, M, and Q being particularly vulnerable due to construction and operational issues. The *Disruption Impact Regression Model* showed that AJT values were significantly higher on lines with high ridership, aging infrastructure, and exposure to external disruptions like weather. Our clustering analysis revealed that Clusters 4 and 0, which include high-traffic lines such as the 6, 7, A, and F, are the most critical, with higher disruption exposure probabilities. Targeting the most vulnerable lines could optimize resource allocation, reduce passenger delays, and minimize financial losses, ensuring a more resilient and efficient transit system.

As for model performance, XGBoost outperformed simpler linear models by effectively capturing non-linear relationships and complex interactions within the data. Another interesting insight from our models was that while seasonality had low predictive power in the *Disruption Exposure Classification Model*, it was a crucial factor in predicting AJT in the *Disruption Impact Regression Model*. This suggests that while seasonality may not directly affect the frequency of disruptions, it significantly influences their severity, likely due to extreme weather exacerbating delays.

Building on prior studies, this research advances disruption management by integrating machine learning predictions, weather, and ridership data to improve operational decision-making. Future work could explore the inverse relationship identified between AJT and disruption exposure, specifically in Cluster 4 and 0 or adding real-time data to refine these models.

References

[1] X. Liu and J. Dai. Artificial intelligence-aided rail transit infrastructure data mining. Technical Report CAIT-UTC-REG 43, Rutgers University. Center for Advanced Infrastructure and Transportation, March 2022. 1

[2] Metropolitan Transportation Authority. Future mta, 2024. Accessed: December 13, 2024. 1

[3] Metropolitan Transportation Authority. Mta subway operational metrics dashboard, 2024. Accessed: December 13, 2024. 1

[4] M. Yap and O. Cats. Predicting disruptions and their passenger delay impacts for public transport stops. *Transportation*, 48:1703–1731, 2021. 1, 2