Forecasting MTA Subway Delay by Decision Tree and Random Forest

Ruilin Peng

2024-10-20

1. Introduction

As the largest public transit authority in North America, Metropolitan Transportation Authority carries over 11 million passengers on an average weekday system wide. However, delay is also a daily issue that causes inconvenience to the passengers and impacting the efficiency of the system. Subway delays, in particular, would increase passengers' travel time, disrupt train schedules, and lead to backlogs.

The ability to predict subway delays would potentially optimize service delivery, increase service efficiency, and improve passengers' experience. Thus in this report, I would try multiple algorithms to forecast the subway delays.

2. Data

2.1 Data Description

The dataset used in this study comes from New York State open data uploaded by MTA. It consists monthly records of subway delay-number of delays in each cause and in each line.

Attributes Description

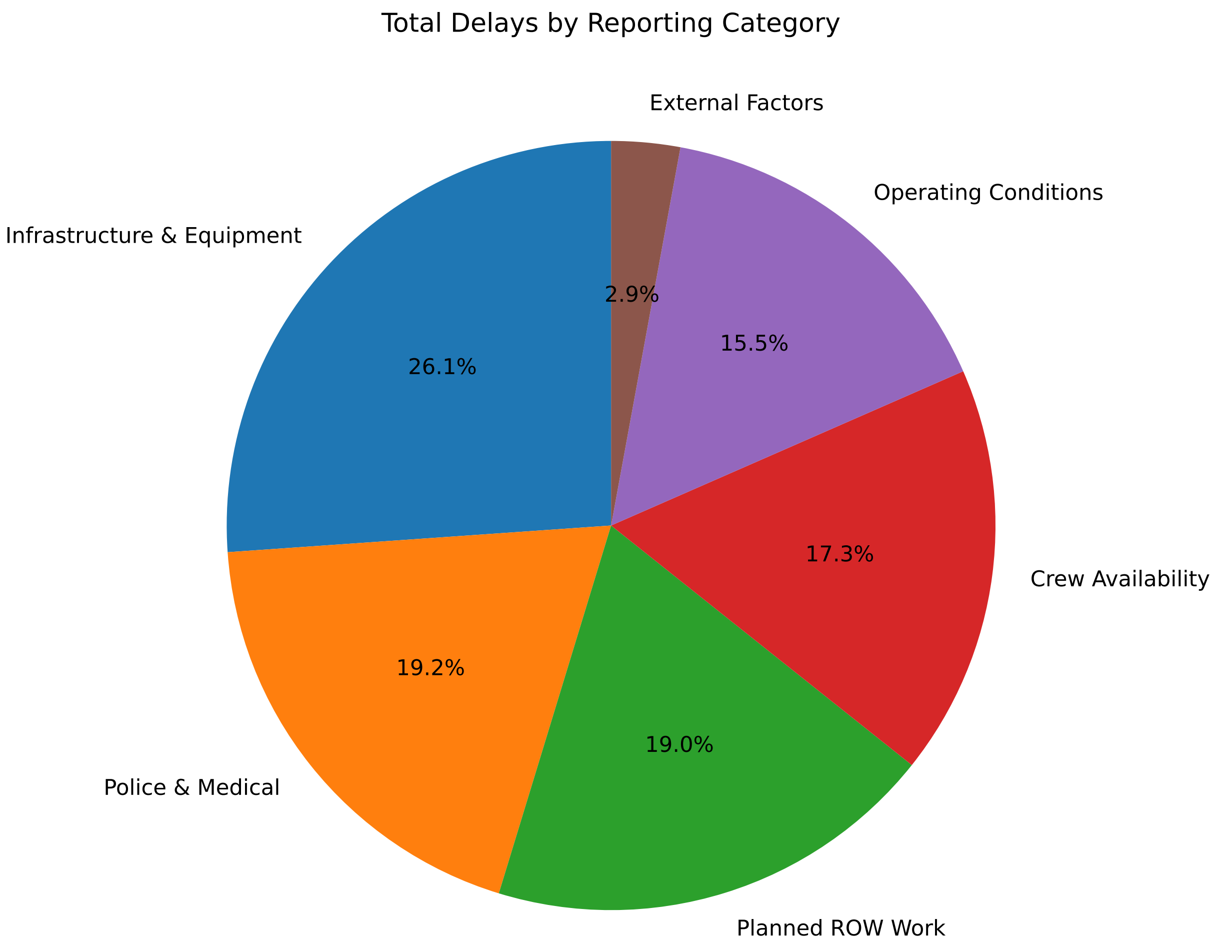
Attribute	Description	Data Type
month	The month in which subway trains delayed is being calculated (yyyy-mm-dd).	*Floating Timestamp*
division	The A Division (numbered subway lines), B Division (lettered subway lines) and systemwide.	*Text*
line	Each subway line (1, 2, 3, 4, 5, 6, 7, A, C, E, B, D, F, M, G, J, Z, L, N, Q, R, W, S 42nd, S Rock, S Fkln).	*Text*
day_type	Represents weekday as 1 and weekend as 2.	*Integer*
reporting_category	The main category under which the delay was reported (e.g., infrastructure, crew).	*Text*
subcategory	A more specific description of the cause of the delay (e.g., braking issues, weather).	*Text*
delays	The total number of delays reported for that particular instance.	*Integer*

2.2 Explanatory Data Analysis

2.2.1 Delays by Reporting Category

Table 1: Summary of total delays by category

Reporting Category	Total Delays
Infrastructure & Equipment	1013808
Police & Medical	744767
Planned ROW Work	738486
Crew Availability	670674
Operating Conditions	603564
External Factors	112218



From the summary and the plot, we can observe that the largest source of delay came from "Infrastructure & Equipment", making up over a quarter of all subway delays. This suggests that events such as track maintenance, signal failure turn out to be a major bottleneck for MTA subway system's efficiency.

As the second largest cause for delay, "Police & Medical" makes up almost 20% of all causes.

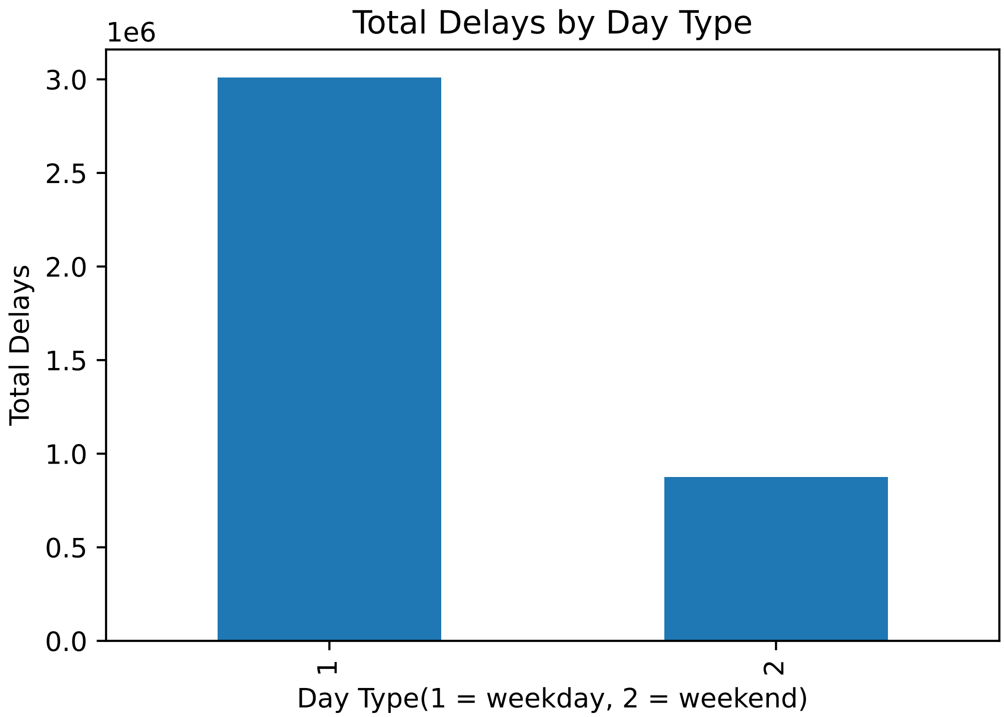
Roughly the same proportion as the second largest cause, "Planned ROW Work" accounts for 19% of the delays. Thus the schedules of such event could be optimized.

2.2.2 Delays by Day Type

Table 2: Summary of total delays by day type

Day Type(1 = weekday, 2 = weekend)	Total Delays
1	3009006

Day Type(1 = weekday, 2 = weekend)	Total Delays
2	874511



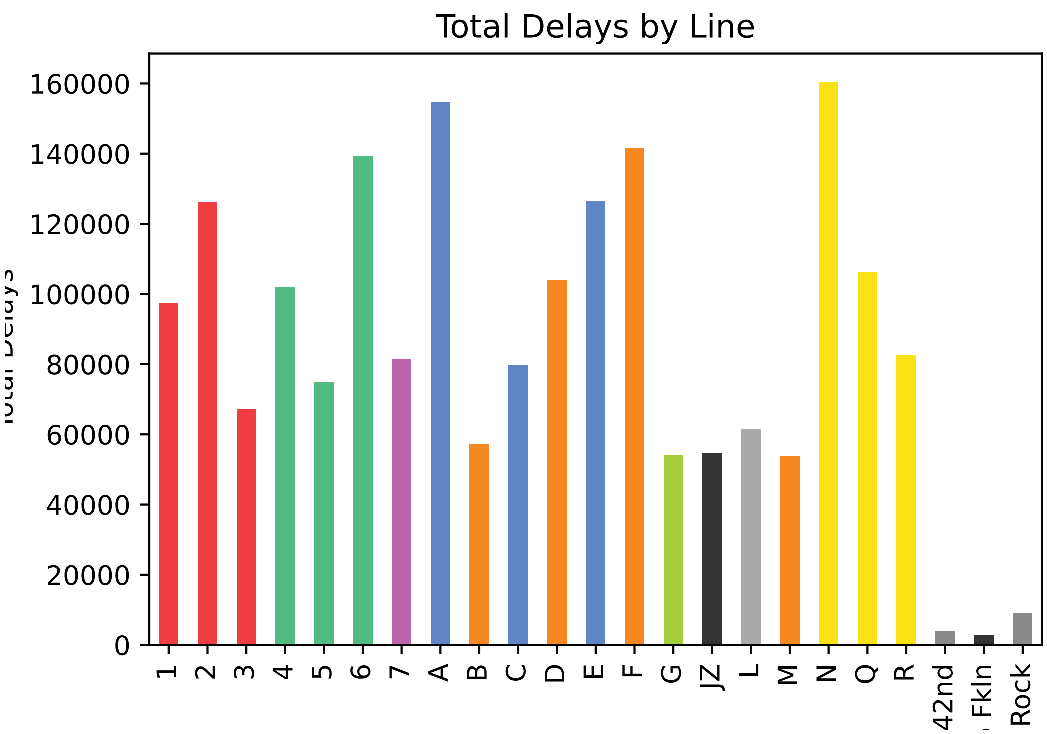
Therefore, we can observe that most delays occur on weekdays when, theoretically, more passengers need to transit to work.

2.2.3 Delays by Line

Table 3: Summary of total delays by line

Line	Total Delays
1	97502
2	126173
3	67100
4	101942
5	74961
6	139341
7	81421
A	154733
B	57196
C	79692
D	104078
E	126512
F	141530
G	54186
JZ	54596
L	61574

Line	Total Delays
M	53705
N	160507
Q	106112
R	82642
S 42nd	3898
S Fkln	2720
S Rock	9061



The lines with the most total delays such as N, A, F, and 6, are the lines that run through upper, midtown, lower Manhattan, which are parts of the city with the highest population densities.

2.3 Data Preprocessing

2.3.1 Preprocessing of month

A new column is introduced called "year", derived from the original "month" value which is a timestamp. Same is done to the month column which is transformed to a column that only contains the numerical value of the month. The reason for doing so is that most statistical models cannot directly handle datetime object.

2.3.2 Preprocessing of other features

Table 4: Mapping for Subway Divisions

Original Value	Encoded Value
A DIVISION	0
B DIVISION	1

Table 5: Mapping for Subway Lines

Original Value	Encoded Value
1	0
2	1
3	2
4	3
5	4
6	5
7	6
A	7
B	8
C	9
D	10
E	11
F	12
G	13
JZ	14
L	15
M	16
N	17
Q	18
R	19
S 42nd	20
S Fkln	21
S Rock	22

Table 6: Mapping for Reporting Category

Original Value	Encoded Value
Crew Availability	0
External Factors	1
Infrastructure & Equipment	2
Operating Conditions	3
Planned ROW Work	4
Police & Medical	5

Table 7: Mapping for Subcategories

Original Value	Encoded Value
Braking	0
Capital Work - Other Planned ROW	1
Crew Availability	2
Door-Related	3
External Agency or Utility	4
External Debris on Roadbed	5
Fire, Smoke, Debris	6
Inclement Weather	7
Insufficient Supplement Schedule	8
Other - CE	9

Original Value	Encoded Value
Other - Sig	10
Other Infrastructure	11
Other Internal Disruptions	12
Other Planned ROW Work	13
Persons on Roadbed	14
Propulsion	15
Public Conduct, Crime, Police Response	16
Rail and Roadbed	17
Service Delivery	18
Sick/Injured Customer	19
Signal Modernization Capital Project	20
Subways Maintenance	21
Train Brake Activation - Cause Unknown	22
Work Equipment	23
NA	24

First, as I noticed there were quite a few records having input mistakes for the division column, having values such as "2020-06-01" or "Systemwide", these lines were removed. Also, for line, the records with value "systemwide" was also removed as it does not help with forecasting.

As for features other than month, which consist of categorical values, since statistical models require numerical inputs, these categorical values are transformed into numerical values using label encoder. For example, for division, "A DIVISION" would become 1 and "B DIVISION" would become 2. For lines, "1" which stands for line 1 would become 0 and similarly line 2 would become 1.

2.3.3 Train Test Split

Using train_test_split, 70% of the dataframe are selected for fitting the model and 30% are used as test data.

3. Forecasting Methods

3.1 Decision Tree

Decision Tree Regression observes the feature of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output.

In our problem, there doesn't seem to have any linear relationships between factors causing the delay and the number of delays. However, a decision tree would be able to model non-linear relationships by making splits in the data based on different conditions.

3.2 Random Forest

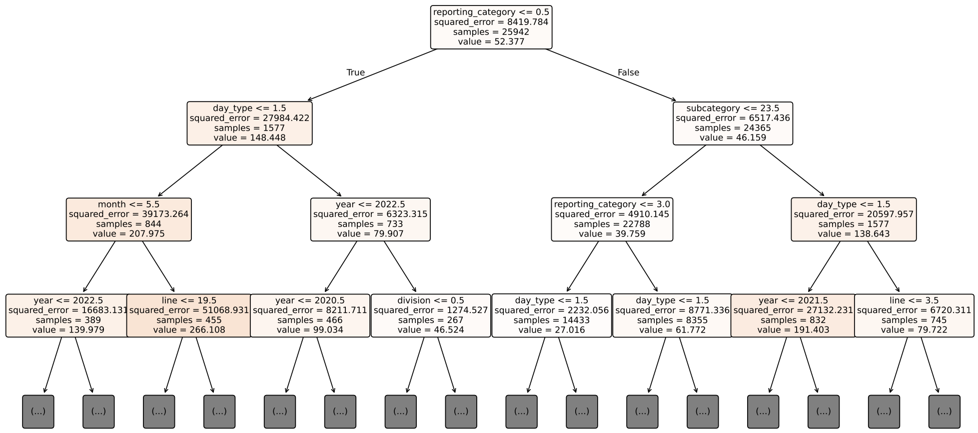
Random Forest is an algorithm that creates a number of decision trees during the training where each tree is fit using a random subset of the training data. The randomness introduces variability among individual trees, reducing the risk of overfitting and improving overall prediction performance.

Thus applying Random Forest Regressor on the MTA forecasting problem might be able to reduce overfitting and create better forecasting than one single decision tree.

4. Results

4.1 Decision Tree Result

4.1.1 Decision Tree Representation



Above is a visualization of Decision Tree regression model fitted using the DecisionTreeRegressor() from sklearn with the max-depth of 20.

4.1.2 Decision Tree Prediction

Using the model fit and tables 4 to 7 for encoded values, we can predict the delay of given conditions. For example, the forecasted number of delays for September 2025, division B, line a on a weekday, caused by Police & Medical, sub-category fire, smoke & debris would be 13.

4.1.3 Decision Tree Metrics

Table 8: Model Evaluation Metrics

Metric	Value
Mean Squared Error	2632.3453480
R-squared	0.6724606

Table 9: Feature Importance

Feature	Importance
line	0.2982052
subcategory	0.2156180
year	0.1446413
month	0.1319372
reporting_category	0.1012043
day_type	0.0852511

Feature	Importance
division	0.0231429

By plugging test data in the result random forest and comparing the expected test result with the actual result, we can get the above metrics.

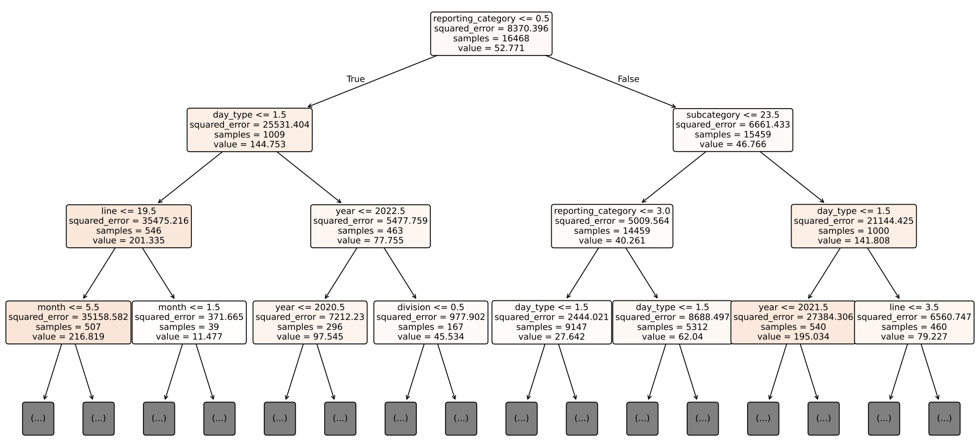
The Mean Squared Error, the average squared difference between the actual and predicted results is 2632.34.

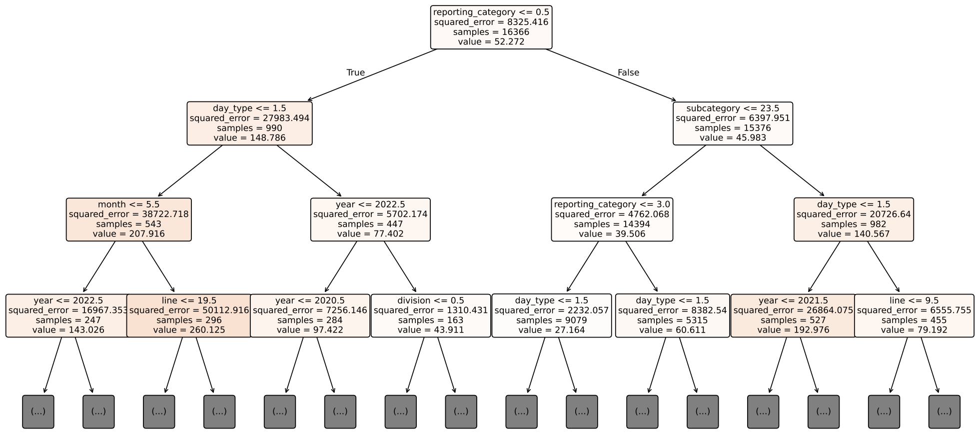
The R-squared value, which means the proportion of variance in the response variable(delay) that can be explained by the model features, is 0.672. This means the decision tree we have obtains 67.2% of the variance of subway delays

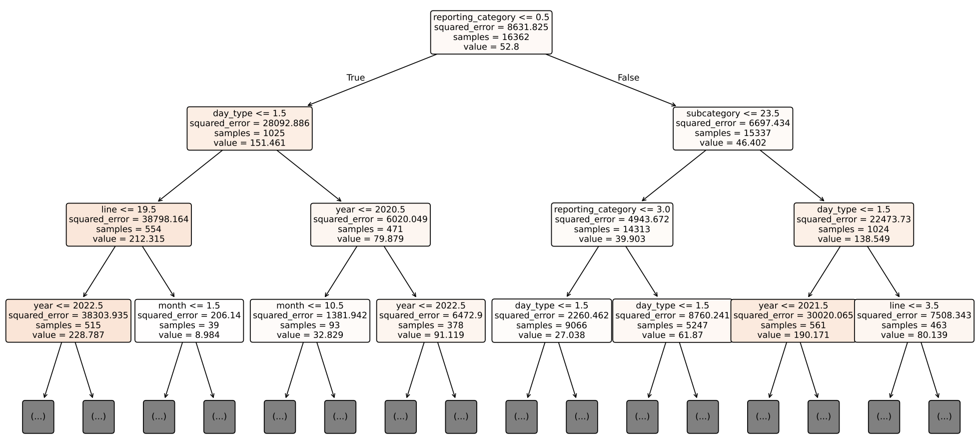
For feature importance, line and subcategory are the most important features, in this decision tree model.

4.2 Random Forest Result

4.2.1 Random Forest Representation







Above are three of the decision trees the Random Forest fitted.

4.2.2 Random Forest Prediction

Using the same conditions for decision tree, September 2025, division B, line a on a weekday, caused by Police & Medical, sub-category fire, smoke & debris, the expected number of delay given by the Random Forest Algorithm would be 37.75. The reason that the Random Forest is not giving an integer for the number of delays is that it takes the average from multiple decision tree forests.

4.2.3 Random Forest Metrics

Table 10: Model Evaluation Metrics

Metric	Value
Mean Squared Error	2412.9813513
R-squared	0.6997558

Table 11: Feature Importance

Feature	Importance
subcategory	0.2701671
line	0.2472030
reporting_category	0.1449106
year	0.1421856
day_type	0.1114333
month	0.0684022
division	0.0156981

By plugging test data in the result random forest and comparing the expected test result with the actual result, we can get the above metrics.

The Mean Squared Error, the average squared difference between the actual and predicted results is 2412.98, which is lower than that of a single decision tree.

The R-squared value, which means the proportion of variance in the response variable(delay) that can be explained by the model features, is 0.699. This means the decision tree we have obtains 69.9% of the variance of subway delays, which is a higher percentage than that of a single decision tree.

For feature importance, line and subcategory are the most important features, in this random forest model, same as in the previous Decision Tree.

5. Conclusion

In this report, we fit and evaluated Decision Tree and Random Forest to forecast MTA subway delays based on division, line, and operational factors. In terms of performance, the Random Forest Regressor outperformed the Decision Tree as suggested by the lower Mean Squared Error and Higher R-squared value. This is because a lower MSE suggests the model has smaller errors in its prediction and a higher R-squared value means the model is explaining more of the variance in the response variable.

Both models suggested that subcategories and lines were the most influential features.

Bibliography

GeeksforGeeks. (2023, April 18). Label encoding in Python. https://www.geeksforgeeks.org/ml-label-encoding-of-datasets-in-python/

GeeksforGeeks. (2023, January 11). Python: Decision tree regression using sklearn. https://www.geeksforgeeks.org/python-decision-tree-regression-using-sklearn/

GeeksforGeeks. (2024, July 12). Random Forest algorithm in machine learning. https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/?ref=header_outind

Interface to python. Interface to Python • reticulate. (n.d.). https://rstudio.github.io/reticulate/

Plot_tree. scikit. (n.d.). https://scikit-learn.org/stable/modules/generated/sklearn.tree.plot_tree.html

Appendix

Generative AI Statement

I used the following generative artificial intelligence(AI) tool: Chat GPT 4o. I used the suggestions such as how to install python package in r's reticulate environment and how to assign python variables to R variables.

How is this report generated

This report is generated using R markdown with the pdf as the output option. The code chunks that do the plotting, model fitting, and metrics such as MSE, R2 computation were Python except the tables were using R's knitr kable.