ML_report

Omina Nematova

April 30, 2025

Project GitHub: https://github.com/omina26/TTC-Subway-Delay-Analysis

Project Website: https://omina26.github.io/TTC-Subway-Delay-Analysis/

Introduction

As businesses increasingly encourage employees to return to in-office work, reliable transporta- tion is becoming more important than ever for commuters. Many workers rely on subway systems for their daily commutes. Subway delays can disrupt schedules, increase travel time uncertainty, and even influence long-term decisions about where people choose to live or work. While some delays may be unavoidable, understanding when and where these delays occur, as well as how external factors like weather conditions and population density contribute to service disruptions can help both commuters and policymakers make informed decisions.

Motivated by the question, "What factors affect subway delays on Toronto's TTC?", this study builds on prior exploratory analysis by using machine learning to predict the duration of subway delays. A merged dataset was constructed using multiple sources. Subway delay reports (see reference) were obtained from the City of Toronto Open Data Catalogue, providing details on location, time, and cause of delays across the TTC subway network. Service population estimates were derived from Toronto Neighbourhood Profiles (see reference), also sourced from the Open Data Catalogue, while spatial data for neighbourhoods and station locations was incorporated using the Toronto Neighbourhoods shapefile (see reference), TTC Subway and Streetcar Map (see reference) and OpenStreetMap API (see reference). Additionally, hourly weather data from 2024 was retrieved using the Open-meteo API (see reference), allowing for an assessment of temperature, precipitation, and other environmental factors at the time of each recorded delay. To address the research question, five predictive models were developed and compared to determine which could most accurately estimate delay duration. Model performance was evaluated using root mean squared error (RMSE), mean absolute error (MAE), and R^2, and key factors influencing delay length were identified using variable importance plots.

Methods

This analysis used the merged dataset created during the exploratory phase, which combined subway delay records, hourly weather data, and population estimates for areas surrounding each TTC station. The TTC Subway Delay Data (2024) from the City of Toronto Open Data Catalogue was cleaned and filtered to retain relevant columns and delay durations between 1 and 30 minutes. Service population estimates were derived from the 2021 Toronto Neighbourhood Profiles and linked to stations using spatial joins based on the Toronto Neighbourhoods shapefile. Subway station coordinates were via the OpenStreetMap API. Hourly weather data from 2024 was retrieved using the Open-Meteo API and included a range of environmental variables. These datasets were merged by station name and hour to produce a unified dataframe used for modeling. All data cleaning and data wrangling steps were conducted in Python. Additional exploratory tables and plots used to analyze the data are included in the EDA report.

As part of the exploratory phase, three interactive visualizations were created using R to better understand subway delay patterns. First, a delay frequency map was generated using the `leaflet` package, displaying the total number of delays at each subway station using scaled circle markers, with popups showing station-level service population. Second, a scatterplot was created with `plotly` to visualize the relationship between service population and average delay duration, grouped by station and subway line. This allowed

identification of outliers and variability in delay severity across different areas. Third, a heatmap of delay frequency by hour of day and day of week was constructed using `plotly`, highlighting temporal patterns in delays, such as peaks during rush hours or weekends. These visualizations provided insight into when and where delays were most frequent, and guided feature selection for the modeling stage.

To predict TTC subway delay durations, five supervised learning models were trained and compared using a 70/30 train-test split. All modeling was conducted in R, and performance was evaluated using three key metrics:

- Root Mean Squared Error (RMSE): Measures the square root of the average squared differences between predicted and actual values. It penalizes larger errors more heavily and is expressed in the same units as the target variable (minutes).
- Mean Absolute Error (MAE): Calculates the average absolute difference between predictions and actual outcomes. It provides a straightforward interpretation of the average error in minutes.
- R-squared (R^2): Represents the proportion of variance in the response variable explained by the model. Higher R^2 values indicate better explanatory power.

Thirteen predictors were used in all models, including weather conditions, station, subway line, direction, service population, and time of day.

The Generalized Linear Model (GLM) with a Gamma family and log link was selected to model delay duration as a positively right-skewed continuous variable. Categorical variables and continuous predictors were used. The Generalized Additive Model (GAM) was fit using the same predictors, with smooth functions applied to continuous variables to capture non-linear effects. In both cases, predictor significance was assessed using model summary statistics.

The Random Forest model was implemented using the `ranger` method. This ensemble method builds multiple decision trees on bootstrapped subsets and averages predictions to reduce variance. It handles high-cardinality categorical features like station effectively. Tuning was performed over number of predictors at each split and minimum node size. The Gradient Boosting (GBM) model was trained using the `gbm` method, which builds trees sequentially, each trying to correct the residuals of the previous model. Its tuning involved number of trees, tree depth, learning rate, and minimum observations per node. XGBoost, an optimized implementation of gradient boosting, was trained using one-hot encoded predictors. It was tuned over a wide range of hyperparameters including number of rounds, maximum tree depth, learning rate, regularization term, column subsample ratio, minimum child weight, and row subsample ratio. For these three models, hyperparameters were tuned using 5-fold cross-validation, and varaible importance plots were generated. These plots helped identify the most significant predictors of delay duration based on impurity reduction and model contribution. The three plots were combined into a single visualization to highlight consistently important features across models.

Lastly, after fitting, predictions were generated for both training and test sets, and a comparison table was created to summarize each model's performance. The best model was defined as the one with the lowest RMSE and MAE, and the highest R^2 on the test set.

Results

Exploratory Analysis

Most exploratory tables and plots, along with their corresponding interpretations, are presented in the EDA report. Additional components of the exploratory analysis, including interactive visualizations, are available on the project website under the EDA tab with accompanying conclusions.

Model Performance

Table 1 presents the traing and testing RMSE, MAE, and R2 for all five models. The best model can be obtained by looking for the model with the lowest RMSE and MAE, and the highest R2 on the test set.

Table 1: Table 1: Model Performance Comparison

Model	RMSE Train	MAE Train	R2 Train	RMSE Test	MAE Test	R2 Test
GLM (Gamma)	4.763	3.325	0.008	4.782	3.301	0.004
GAM	4.699	3.256	0.034	4.761	3.273	0.013
Random Forest (Tuned)	2.899	1.998	0.632	4.761	3.316	0.013
Gradient Boosting (Tuned)	4.508	3.141	0.111	4.752	3.273	0.017
XGBoost (Tuned)	4.370	3.024	0.165	4.759	3.273	0.014

As shown in Table 1, the Gradient Boosting model achieved the best overall performance, with a test RMSE of 4.752, a test MAE of 3.273, and a test R2 of 0.017. However, the improvement over the other models is marginal, as the test RMSE, MAE, and R2 values for the remaining models are similarly close. Notably, all models exhibit R2 values near zero, indicating that they explain very little of the variance in delay duration. This suggests that, despite tuning and a diverse set of predictors, none of the models are very effective at accurately predicting subway delay durations.

Variable Importance

Table 2 presents the GLM coefficient summary, including an additional column indicating whether each predictor is statistically significant ($p < 0.05$). Similarly, Table 3 and Table 4 display the GAM summary, with Table 3 showing the parametric terms and Table 4 presenting the smooth terms. Together, these tables help identify which factors have a statistically significant impact on subway delay duration.

Table 2: Table 2: GLM Coefficients Summary

Term	Estimate	Std. Error	t value	p value	Significant
boundN	0.006	0.132	0.044	0.965	No
boundS	-0.047	0.132	-0.353	0.724	No
boundW	-0.028	0.029	-0.947	0.344	No
lineSHP	**0.182**	**0.056**	**3.243**	**0.001**	**Yes**
lineYU	0.045	0.131	0.345	0.730	No
apparent temperature	0.002	0.001	1.609	0.108	No
relative humidity	-0.001	0.001	-1.629	0.103	No
precipitation	0.001	0.017	0.086	0.931	No
snowfall	0.060	0.149	0.401	0.688	No
snow depth	-0.360	0.677	-0.531	0.596	No
cloud cover	0.000	0.000	-0.871	0.384	No
wind speed	0.004	0.005	0.850	0.395	No
wind gusts	-0.002	0.003	-0.927	0.354	No
hour	**0.004**	**0.002**	**2.477**	**0.013**	**Yes**
service population	0.000	0.000	-1.747	0.081	No

Table 3: Table 3: GAM Parametric Coefficients Summary

Term	Estimate	Std. Error	t value	p value	Significant
boundN	0.015	0.130	0.112	0.911	No
boundS	-0.021	0.130	-0.161	0.872	No
boundW	-0.024	0.029	-0.835	0.404	No
lineSHP	**0.149**	**0.057**	**2.640**	**0.008**	**Yes**

lineYU	0.050	0.128	0.386	0.699	No

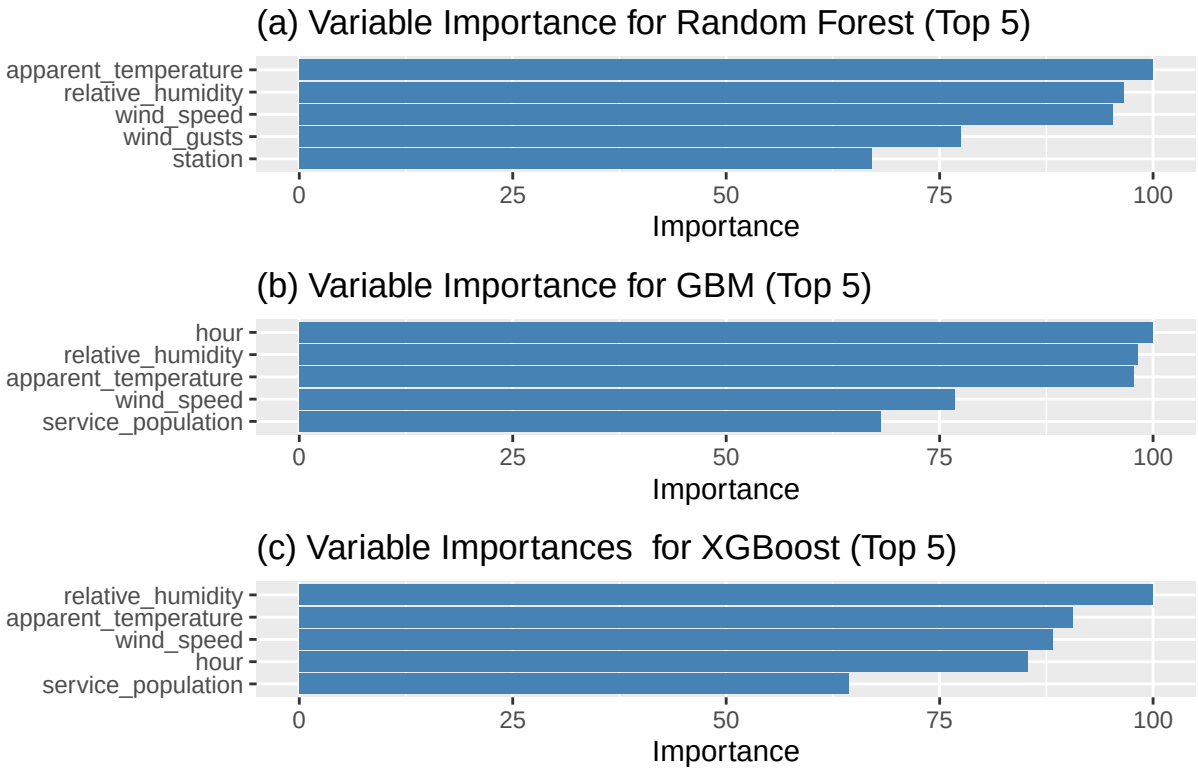
Table 4: Table 4: GAM Smooth Coefficients Summary

Term	Effective DF	Reference DF	F value	p value	Significant
s(apparent temperature)	6.110	7.317	1.074	0.413	No
s(relative humidity)	3.819	4.776	1.059	0.376	No
s(precipitation)	3.257	3.970	0.761	0.534	No
s(snowfall)	3.457	4.233	0.603	0.653	No
s(snow depth)	8.369	8.837	1.425	0.175	No
s(cloud cover)	1.844	2.255	1.534	0.193	No
s(wind speed)	1.001	1.002	2.728	0.099	No
s(wind gusts)	1.001	1.003	3.727	0.054	No
s(hour)	**6.825**	**7.874**	**9.531**	**0.000**	**Yes**
s(service population)	**8.802**	**8.986**	**6.635**	**0.000**	**Yes**

From Table 2, the variables 'lineSHP' and 'hour' are identified as statistically significant predictors in the GLM model, indicating that both the Sheppard subway line and the time of day have a notable impact on delay duration. In the GAM model, Table 3 shows that 'lineSHP' remains a significant parametric term, while Table 4 reveals that the smooth terms 'hour' and 'service population' are also statistically significant. This suggests that delay durations vary nonlinearly over the course of the day and with the population density near stations.

Figure 1 presents the variable importance plots for the Random Forest, Gradient Boosting, and XGBoost models, each displaying the top 5 most influential variables based on their respective importance metrics.

Figure 1: Variable Importance Across Tuned Tree−Based Models



From Figure 1, the most important variable in the Random Forest model was apparent temperature, while hour was most important in the Gradient Boosting model, and relative humidity ranked highest in the XGBoost model. Notably, all three models identified apparent temperature, relative humidity, and wind speed as among the most influential predictors.

When considering both the significant predictors from the GLM and GAM models and the variable importance results from the tree-based models, several variables stand out as consistently influential across models. Specifically, whether the subway operates on the Sheppard line (lineSHP), the hour of the day, service population, apparent temperature, and relative humidity appear to have the greatest impact on delay duration across multiple modeling approaches.

Conclusion

This study aimed to identify the key factors influencing subway delay durations on Toronto's TTC system by building predictive models using a combination of operational, environmental, and demographic data. Although none of the five models, GLM, GAM, Random Forest, Gradient Boosting, and XGBoost, demonstrated strong predictive performance, with R2 values near zero, they did reveal features most associated with delay severity. Among them, the Gradient Boosting model performed best, achieving the lowest prediction error on the test set.

Several findings from the modeling phase aligned with earlier exploratory data analysis (EDA). For example, lineSHP was identified as a significant predictor in both the GLM and GAM models, as well as in variable importance plots from tree-based models. This supports the EDA finding that the Sheppard Line experienced the longest delays among all TTC lines. Additionally, hour emerged as a statistically significant smooth term in the GAM and as a top predictor in ensemble models, reinforcing EDA insights that delays were most frequent during morning and afternoon rush hours. Interestingly, although peak-hour delays were more frequent, their durations tended to be shorter, suggesting a nonlinear relationship that was appropriately

captured by modeling hour as a smooth term.

Weather-related variables such as apparent temperature and relative humidity also ranked highly in the variable importance plots, even though EDA showed weak linear correlation between weather and delay duration. This suggests that nonlinear models may be capturing more complex interactions between environmental conditions and service disruptions.

However, the study has several limitations. First, the available data lacked operational details such as train frequency, staffing levels, or mechanical issues, which are likely to have a stronger and more direct effect on delay durations. Second, ridership data, another potentially informative feature, was not available for 2024 and could not be included in the analysis.

Despite these limitations, the study provides valuable insights into the types of factors associated with TTC subway delay durations. Enhancing model inputs with more detailed real-time data could yield more actionable results for transit planners and city officials.

References

- Open-meteo documentation: https://open-meteo.com/en/docs/historical-weather-api)
- OpenStreetMap API documentation: https://wiki.openstreetmap.org/wiki/API_v0.6
- Toronto Neighbourhood Profiles: https://open.toronto.ca/dataset/neighbourhood-profiles/
- Toronto Neighbourhoods: https://open.toronto.ca/dataset/neighbourhoods/
- TTC Subway and Streetcar Map: https://cdn.ttc.ca/-/media/Project/TTC/DevProto/Images/Home/Routes-and-Schedules/Landing-page-pdfsTTC_SubwayStreetcarMap_2021-11.pdf?rev=909317034177450b8b09ba5b247e24bf
- TTC Subway Delay Data: https://open.toronto.ca/dataset/ttc-subway-delay-data/