“Toronto’s TTC Subway Delays”

MGT 6203 GROUP PROJECT FINAL REPORT TEAM 49: Tadeus Rossetti Marchesi, Chetan Tewari, Meaghan Wright

Introduction

With climate change initiatives such as the European Green Deal and Resolution 109 in the United States recommending providing incentives for public transportation in the coming years; promoting commuters to use city transit and maximize efficiency is top of mind [1] [6]. Delays cost the economy time and money, with 800,000 delays on the British national rail network costing approximately £1 billion in lost time [2] [3]. New York City subway delays could cost as high as 389 million US dollars annually [4].

The Toronto subway is a rapid transit system serving Toronto and the neighboring city of Vaughan in Ontario, Canada, operated by the Toronto Transit Commission (TTC) [10]. It is a multimodal rail network consisting of three heavy-capacity rail lines operating underground, and one elevated medium-capacity rail line [10]. Two light rail lines, which will operate both at-grade and underground, are under construction [10]. The TTC had 386,443,400 passengers in 2021, with 1,790,800 per weekday as of the second quarter of 2022 and these figures should soar over the next years [1]. The objective of this study is three-fold:

a. What can we learn about train delays?
b. Can we predict them?
c. Can we prevent them?

Data Collection and Preprocessing

The first stage included collecting the publicly available datasets from the TTC website, as well as merging and cleaning the resulting data. We accessed seven train delay files in .csv format between 2014 to 2022 from the Toronto city’s TTC database, which updates monthly. These were loaded and merged into a single data frame in a Jupyter notebook, with the raw data containing 167,958 rows of data between 2014-2022 and 10 unique feature columns, as seen in the first table:

	Date	Time	Day	Station	Code	Min Delay	Min Gap	Bound	Line	Vehicle
0	2017-05-01	00:18	Monday	KIPLING STATION (ENTER	TUSC	0	0	W	BD	5251
1	2017-05-01	00:58	Monday	ISLINGTON STATION	MUSC	0	0	W	BD	5182
2	2017-05-01	01:07	Monday	UNION STATION (DOWNSVI	MUPAA	3	7	N	YU	5811
3	2017-05-01	01:18	Monday	MCCOWAN STATION	MRTO	4	9	S	SRT	3006
4	2017-05-01	01:42	Monday	CASTLE FRANK STATION	MUO	7	11	W	BD	5296
...	...	...	...	...	...	...	...	...	...	...
14897	2022-09-30	00:07	Friday	KIPLING STATION	SUDP	11	15	E	BD	5331
14898	2022-09-30	00:13	Friday	MUSEUM STATION	MUI	16	21	S	YU	5521
14899	2022-09-30	01:15	Friday	CASTLE FRANK STATION	PUSTC	0	0	E	BD	5072
14900	2022-09-30	10:25	Friday	LESLIE STATION	EUSC	3	8	W	SHP	6161
14901	2022-09-30	17:41	Friday	SHEPPARD-YONGE STATION	EUBO	5	10	E	SHP	5161

167958 rows × 10 columns

Data cleaning was a bit more involved, as we were trying to understand the meaning of some of the cateogrical columns and identify useful features for our model. In the raw data, there is a column called “Code”, which originally has 200 delay codes indicating why a delay happened, such as human or mechanical issues. First, we decided to simplify these codes into clusters of broader groupings to reduce the modeling complexity and minimize having too many categorical features. This could increase the efficiency when training various models. The original “SUB RMENU CODE” groups from the TTC were used to reduce to 11 groupings based on the “CODE DESCRIPTION”, a detailed description of what each “Code” meant and provided from the TTC website.

The new Code Groups are Security, Human, Electrical, Mechanical, Fire, Logistical, Radio, Weather, Miscellaneous, Other, and Unavailable. An example of the newly engineered “Code Group” feature columns is in the following table:

	SUB RMENU CODE	CODE DESCRIPTION	Code Group
0	EUAC	Air Conditioning	mechanical
1	EUAL	Alternating Current	electrical
2	EUATC	ATC RC&S Equipment	mechanical
3	EUBK	Brakes	mechanical
4	EUBO	Body	mechanical
...	...	...	...
195	TRNOA	No Operator Immediately Available	human
196	TRO	Transportation Department - Other	miscellaneous
197	TRSET	Train Controls Improperly Shut Down	mechanical
198	TRST	Storm Trains	weather
199	TRTC	Transit Control Related Problems	human

200 rows × 3 columns

Also, a train that was not delayed will not provide useful information for our model. Since a train that is not delayed will have a “Min Delay” value of 0, these points have been filtered out from the dataset. Two columns were identified as having missing values: “Bound” and “Line”. These were investigated in detail to understand which datapoints were missing and if a probable reason could be found. These two columns were ruled out for our modelling purposes to limit the amount of geospatial complexity in the modelling. Finally, originally the column “Min Gap” was originally poorly understood and excluded. However, since the Progress Report more detailed information on the TTC website was found and the “Min Gap” is the time length between trains in minutes and has since become an important feature for predicting Min Delays.

All other columns have complete and non-null values. This results in 57,267 rows of data remaining for exploratory data analysis and modelling.

Exploratory Data Analysis

To evaluate the final model’s performance and reduce overfitting the model, we have decided to keep a test dataset separate for evaluating model performance. The 2014-2021 data will be used for training and is the data used for the following exploratory data analysis; it contains most of the dataset (50,517 rows). The 2022 data was held out for testing the model’s performance (7,250 rows) later. With the cleaned and Code grouped test dataset, initial plots for exploratory data analysis were conducted to see if there were any identifiable trends available with the remaining data [7]. Our initial hypotheses included:

1) Train delays would be most significant during rush hour traffic and

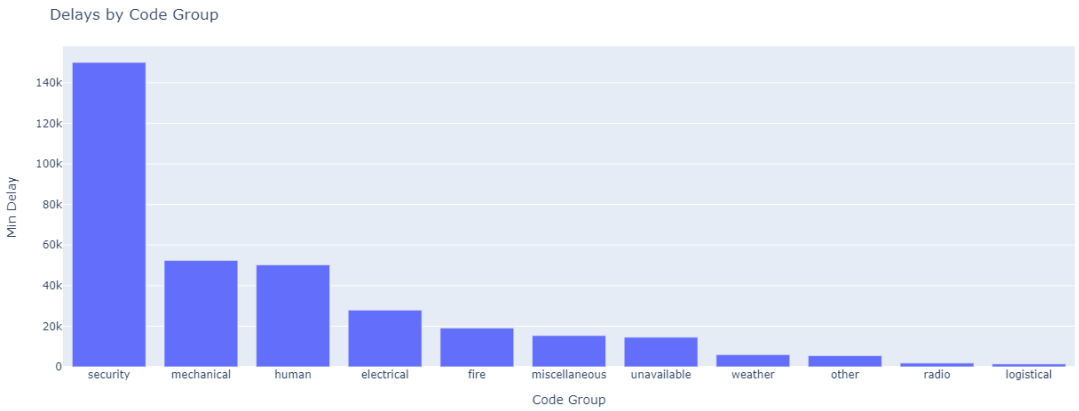
2) Train delays could be impacted by inclement weather at times in winter months.

A quick check of the Station column shows the “Vaughan MC Station” in 2021 appears to have the highest “Min Delay” values. This station is a rapid transit station in Vaughan, Ontario and is the north terminus of the western section of the Toronto subway’s Line 1 Yonge-University [10]. This could be related to high commuter volumes coming from Vaughan into Toronto.

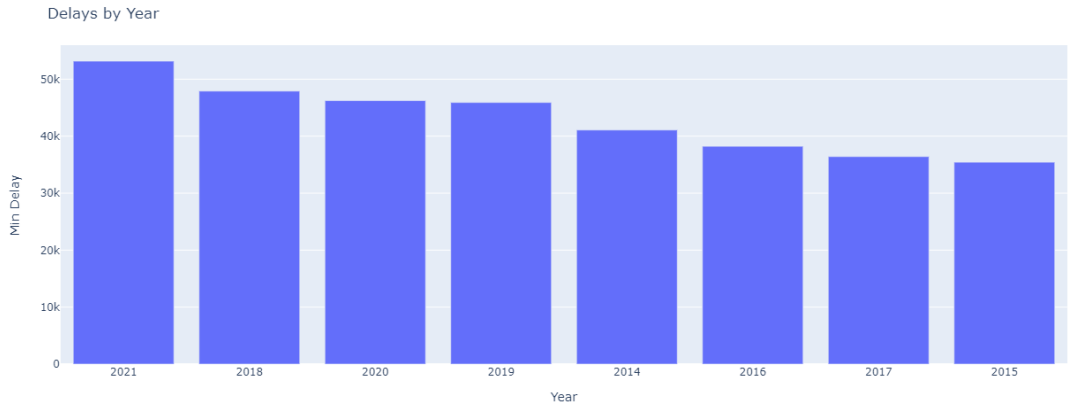
	Year	Station	sum
942	2021	VAUGHAN MC STATION	3079
485	2018	SHEPPARD WEST STATION	2935
810	2020	VAUGHAN MC STATION	2873
858	2021	EGLINTON STATION	2160
752	2020	MCCOWAN STATION	1998
...	...	...	...
567	2019	DUPONT MIGRATION POINT	3
846	2021	DAVISVILLE BUILD-UP	3
345	2017	MUSEUM STATION (APPROA	3
654	2019	UNION TO KING STATION	3
598	2019	KIPLING TO HIGH PARK	2

988 rows × 3 columns

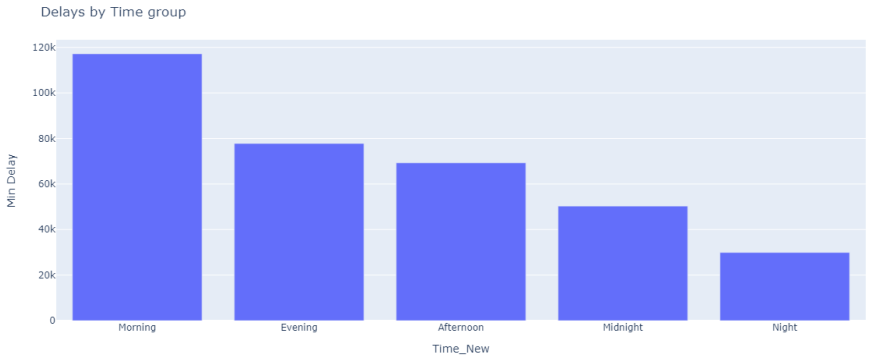
Next, by grouping the Min Delays by Code Group we see that security, mechanical and human caused delays are the most prevalent causes of delay in the test dataset:



There does not seem to be a strong visual trend by grouping the Min Delays by Year as seen in the following figure, although more recent years (2018, 2020-2021) show a slight increasing trend in Min Delays:

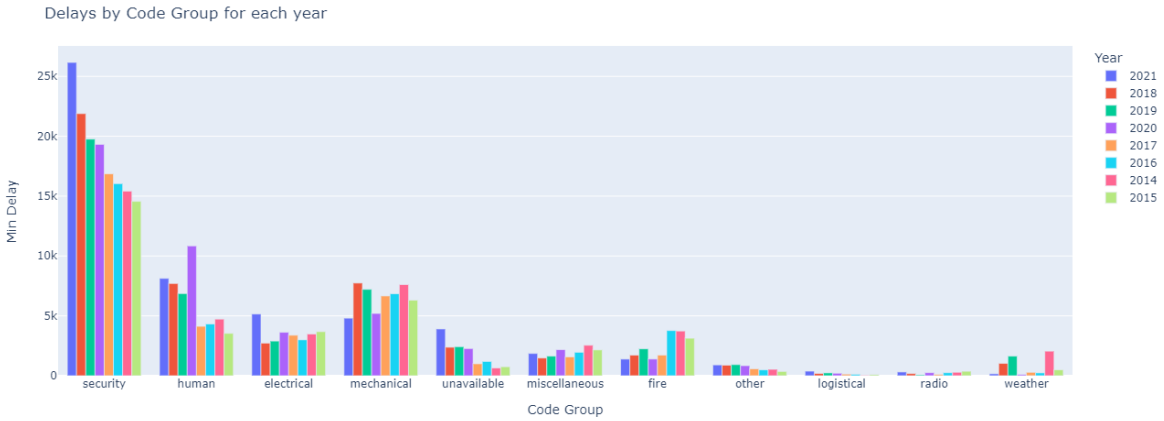


An interesting trend was discovered by grouping the Min Delays by the Time of Day (morning, evening, afternoon, midnight, and night). The team took a novel approach by grouping the times of day into a separate engineered feature column, since we hypothesized that rush hour commutes would potentially increase delays [7]. The most delays occurred during the morning, which appears consistent with our hypothesis and ultimate objective of optimizing city transportation.

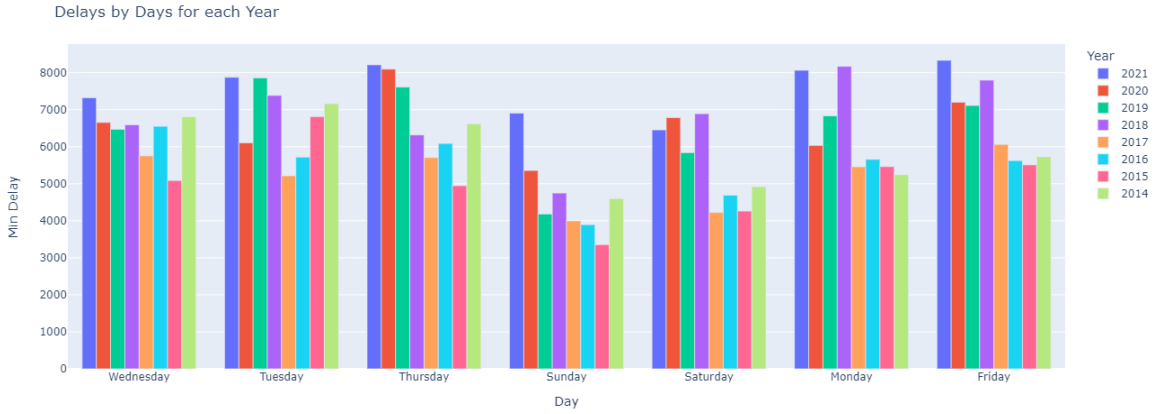


Next, looking at the Delays by Code Group for each Year, there is a strong trend of security causing the most delays across all years. In 2021 there are the highest incidents of delays caused by security, with a general increasing trend since 2016. Human caused delays, typically medical in nature, were also a significant source of delays. Mechanical failures tend to

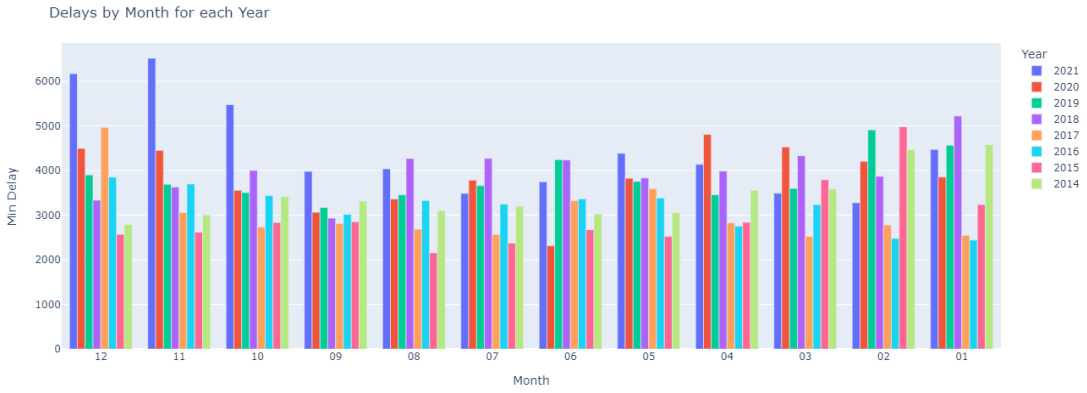
have a higher prevalence than electrical failures in causing delays since 2014. Weather, logistical and radio groups show limited delays in comparison.



There does not seem to be any strong trends by grouping the Delays by Days for each Year as shown in the following figure, although Sunday has slightly lower delay day of the week visually and may be a day of rest for commuters.



No significant trends are observed in the delays by month for each year, however there were higher values between October to December in 2021 as seen in the following figure. This is potentially related to increased transit use during the holiday season.



Finally, we were interested in understanding whether there were any seasonal trends in this data, realizing that in winter months there could potentially be more delays due to inclement weather or increased traffic due to the holiday season [5] [7]. The team created a “Season” column by extracting the Month from the date index of the data and creating a categorical column for Winter, Spring, Summer, and Fall. Winter appears to have the most delays as seen in the following figure.

Modeling Methodology and Analysis

From the exploratory data analysis, feature importance and key trends from the training dataset were identified. Since the progress report and with additional modeling analysis, the key independent variables have changed and been identified as “Min Gap” (time in minutes between trains), “Day”, “Station”, “Season”, “Code Group”, and “Time of Day”. Originally Min Gap was poorly understood and not expected to be a feature of importance. With further investigation on the TTC website, it was understood to be the time length (minutes) between trains and became a key feature in predictive modelling. The dependent feature remains the Min Delay (train delay length in minutes). There are three features that were engineered by the team, used to drill down further into the data analysis and included in the final modelling: “Code Group,” “Seasons” and “Time of Day”. The final features selected for the modeling purposes are as follows:

1. Min Gap (Time length in minutes between trains)
2. Day
3. Station
4. Season (Engineered feature)
5. Code Group (Engineered feature)
6. Time of Day (Engineered feature)

The Linear Regression Equation for modeling Min Delay is as follows:

$$\boldsymbol{Delay} = \boldsymbol{\beta_0} + \mathit{Time}.\boldsymbol{\beta_1} + \mathit{Station}.\boldsymbol{\beta_2} + \mathit{Code}.\boldsymbol{\beta_3} + \mathit{Season}.\boldsymbol{\beta_4} + \mathit{Day}.\boldsymbol{\beta_5} + \boldsymbol{MinGap}.\boldsymbol{\beta_6}$$

The goal for linear regression is to develop a model with strong predictive power that provides confidence in the accuracy of expected minimum delay of a subway train in Toronto’s TTC, given the features mentioned above. As a reminder, the models are trained on the 2014-2021 dataset (50,517 rows), with the 2022 (7,250 rows) providing a hold-out test dataset.

The primary approach was to use a multiple linear regression as a starting base model and use more complex models such as Ensemble models (Random Forest, AdaBoost and Xgboost) to determine if statistical performances can be improved upon. Parametric optimization was employed to optimize hyper-parameters in the models, minimizing the errors between predicted and actual delays, and along with EDA which helped inform the final predictors used in the modeling. Hyperparameter tuning allowed us to efficiently test various inputs to the model. Cross-validation, by resampling the data into groups of test and training sets to fit the model, was also implemented to minimize overfitting the models, allowing them to be trialed on different subsets of the data, and allow for comparisons between various models and their performance.

The model performances were compared using statistical measures of the model’s variance and error:

1. R Square
2. Mean Square Error (MSE)

The results for all the models are summarized below:

Model	R2	MSE
Linear Regression (**Base Model**)	0.869	25.37
Linear Regression (**Base Model-CV**)	0.862	-32.35
Random Forest	0.88	23.29
Random Forest (CV)	0.792	-49.48
Random Forest (Parameter Tuning)	0.85	29.14
AdaBoost	0.303	135.42
AdaBoost (CV)	-0.826	-259.49
XGBoost	0.89	21.13
XGBoost (CV)	0.85	-43.3

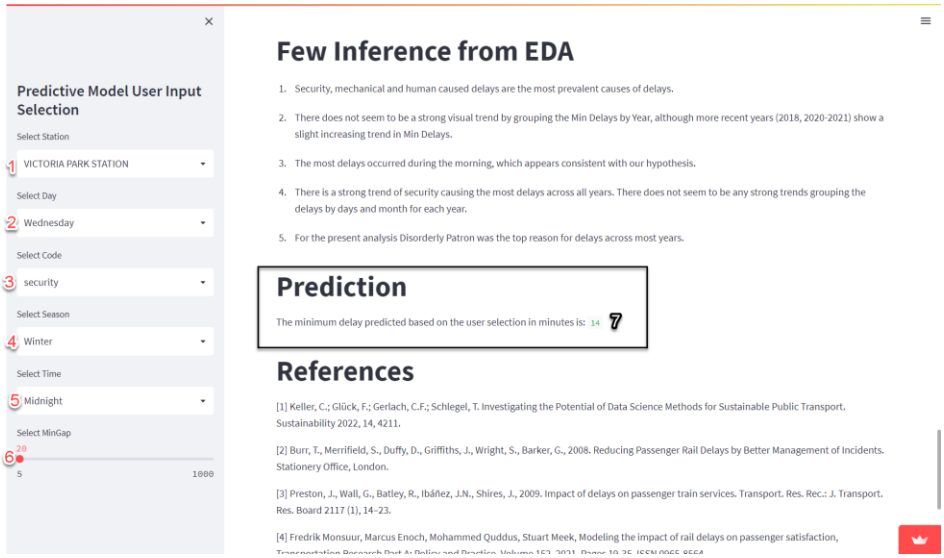
Overall, XGBoost had the strongest performance with an R Squared value of 89% and 85% (cross-validation); the poorest performing model was the cross-validated AdaBoost (30%) and Random Forest model (79.2%, cross-validation). From documentation, XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable [11]. It implements machine learning algorithms under the Gradient Boosting framework [11]. Gradient boosting leverages ensembles of weak predictive models, usually in the form of decision trees, improving modeling results by learning from prior, weaker nodes ("creating strong learners from weak learners") [11]. It was able to leverage these and produce a stronger R Squared, accounting for most of the variation known and unknown, while also reducing the MSE.

AdaBoost produced the weakest statistical performance, as well as having the longest run time to produce predictions; AdaBoost took hours, while XGBoost took minutes. The cross-validation AdaBoost model produced negative R2 and the largest MSE (also negative); this is a very poorly fit model to the data. Negative R2 values can also be indicative that something is wrong with the intercept value of the model; the AdaBoost CV model was quickly ruled out as a contender. Simple, multiple linear regression also produced a high R Squared value of 86.9% (86.2% cross-validation), also with a low MSE, giving confidence that in the case of Toronto's TTC subway delay predictions, a simple heuristic produces quite reasonable results without necessarily going into more complex models, which can also be more difficult to explain the key stakeholders. The XGBoost model was the final model selected for this project based on its performance.

Finally, the project is hosted in Streamlit which is an open-source app framework in a shareable web application. The XGBoost model for predicting TTC's Minimum Delay is used in the Streamlit app. Please allow the app a couple minutes to load due to size of input data and model:

Team049 Toronto's TTC Subway Delays Streamlit App

Example of final app:



In the app, the end user can adjust the controllable features shown in #1-6 (red, left-hand side), based on their preferred station, Day, Safety Code, Season, Time of Day, and Min Gap (time in minutes between trains). This input to the model results in a prediction of the Minimum Delay (in minutes) expected for that subway train, shown in the prediction highlighted at #7 in the figure above. Additional EDA (Exploratory Data Analysis) plots that informed the final model is also included for context to the end user on how the model was defined. Using Streamlit allowed for the most user-friendly way to share the resulting model.

Given additional time, next steps could include adding a geospatial component to the Streamlit application. Using the XGBoost model, a real-time prediction could be created and layered onto a GIS map of the transit station. This would empower transit users to plan their routes with better insights into the on-going delays and organize their travels accordingly. The model could be used going forward on monthly refreshes of the transit data, to test its prediction accuracy and help inform city planning for improving reliability of trains with higher cases of minimum delays. Train scheduling could be used with the model to improve transit efficiency and increase train services through high traffic stations and lines.

Conclusions

The Toronto TTC subway delay project was completed within the Fall 2022 semester timeline. The initial questions our team set out to investigate were: What can we learn about train delays? Can we predict them? Can we prevent them?

The team has been able to successfully gain insights about Toronto's TTC subway delays from the exploratory data analysis. Seasonality, especially winter months, has been identified as likely having higher transit usage and longer delays. Human-related incidents, rather than mechanical or electrical failures, are the most common causes for train delays in Toronto's subway system. Rush hour commutes are also a probable cause of delays, with mornings and evenings having the most minimum delays due to increased traffic. The "Min Gap" column provided key information for predicting Min Delay and improved modeling results with its inclusion. These findings confirmed our earlier hypotheses of when most train delays could be happening.

The final XGBoost model successfully predicted train Minimum Delays given the 6 input features: Min Gap (time in minutes between trains), Day, Station, Season, Code Group and Time. This model has a high degree of confidence, with an R Squared value at 85-89% and low MSE of 21.13, indicating most of the variance is accounted for. A multiple linear regression also produced reliable results, with an R Squared value of ~86% and MSE of 25.37.

Preventing Toronto's subway train delays remains a multi-faceted challenge. Since most train delays are related to human-caused issues, such as security and medical incidents, this could warrant a broader discussion with external

organizations and local city authorities. According to the TTC website, additional paramedics and security have already been implemented at high-risk stations [10]. Based on the exploratory data analysis and trends in data, there may be an opportunity to target additional personnel and resources better to seasonality, commute times and specific train stations that cause the most delays. Given the large economic cost of train delays, additional funds and resources in the short term should produce large returns in the long term [5]. With Toronto's population growing since 2017 and expected to continue that trajectory in future years, additional resources may be needed to help prevent or minimize these incidents and ensure transit usage remains a reliable form of transportation in alignment with on-going climate change initiatives [8][9][10].

Literature Citations

[1] Keller, C.; Glück, F.; Gerlach, C.F.; Schlegel, T. Investigating the Potential of Data Science Methods for Sustainable Public Transport. Sustainability 2022, 14, 4211.

[2] Burr, T., Merrifield, S., Duffy, D., Griffiths, J., Wright, S., Barker, G., 2008. Reducing Passenger Rail Delays by Better Management of Incidents. Stationery Office, London.

[3] Preston, J., Wall, G., Batley, R., Ibáñez, J.N., Shires, J., 2009. Impact of delays on passenger train services. Transport. Res. Rec.: J. Transport. Res. Board 2117 (1), 14–23.

[4] New York city comptroller Link

[5] Nils O.E. Olsson, Hans Haugland, Influencing factors on train punctuality—results from some Norwegian studies, Transport Policy, Volume 11, Issue 4, 2004, Pages 387-397, ISSN 0967-070X.

[6] Wiggenraad, P.B.L., 2001. Alighting and Boarding Times of Passengers at Dutch Railway Stations. Report, TRAIL Research School, Delft.

[7] Flier, H., Gelashvili, R., Graffagnino, T., Nunkesser, M. (2009). Mining Railway Delay Dependencies in Large-Scale Real-World Delay Data. In: Ahuja, R.K., Möhring, R.H., Zaroliagis, C.D. (eds) Robust and Online Large-Scale Optimization. Lecture Notes in Computer Science, vol 5868. Springer, Berlin, Heidelberg.

[8] Fredrik Monsuur, Marcus Enoch, Mohammed Quddus, Stuart Meek, Modeling the impact of rail delays on passenger satisfaction, Transportation Research Part A: Policy and Practice, Volume 152, 2021, Pages 19-35, ISSN 0965-8564.

[9] Tijs Huisman, Richard J. Boucherie, Nico M. van Dijk, A solvable queueing network model for railway networks and its validation and applications for the Netherlands, European Journal of Operational Research, Volume 142, Issue 1, 2002, Pages 30-51, ISSN 0377-2217.

[10] Toronto TTC Website Link

[11] XGBoost documentation Link