Toronto Transit Commission (TTC) Delayed Bus Analysis

Supervised by.: Prof. Graham Wall
Submitted by.:

Bahauddin Kalyani (C0886857) | Dev Makwana (C0885064) | Krina Patel (C0886861) | Mahaveersinh Chauhan (C0884854) | Trushna Patel (C0886910)

1. Introduction

The TTC (Toronto Transit Commission) is the public transportation provider in Toronto, Canada, responsible for operating buses, streetcars, and subways throughout the city. As with any transportation system, delays can impact the service's efficiency and passengers' satisfaction. An analysis report has been prepared and presented to understand better the frequency and causes of delays on TTC buses. It provides a comprehensive overview of the uncertainties that occurred during a specific period and the factors contributing to them. It also includes delay prediction for the year 2023. By identifying the causes of delays, the TTC can develop strategies to minimize them and improve the overall quality of its bus service.

2. Data Collection

The dataset used in this report is from the City of Toronto's Open Data Portal published by the TTC. (*TTC, Toronto – Bus Times, Routes & Updates*, n.d.) The dataset contains the data from January 1, 2014, to December 31, 2022, in the form of an Excel file.

Table 1 describes features and their meaning for analyzing and predicting.

3. Data Validation and Cleaning

In terms of data cleaning, an effort was made to eliminate missing and null values from the dataset. Significant invalid, inconsistent, and missing entries in the Route and Direction columns were eliminated.

Also, the column Vehicle was removed because it only contained unique values. The heatmap in Figure 1. Shows the missing values in each column.

Field Name	Description
Report Date	The date when the incident occurred.
Route	The number of the bus route.
Time	The time when the incident occurred.
Day	The name of the day of the week.
Location	The location/station of the incident.
Incident	The reason behind the incident delay.
Min Delay	The delay, in minutes, to the schedule for the following bus.
Min Gap	The total scheduled time, in minutes, from the bus ahead of the following bus. (Minutes between two buses)
Direction	The direction of the bus route, where B, b, or BW indicates both ways.
Vehicle	Vehicle Number

Table 1. Dataset features and descriptions

Figure 1. Null value counts for each column

4. Data Aggregation and Representation

Date Aggregation involves combining and summarizing data to gain insights and communicate the findings effectively. Each year data had a separate excel file aggregated for further analysis. Figure 2 represents the 10 Locations that witnessed delays the most which state that maximum delay occurs at Kennedy and Scarborough center stations with more than 20000 bus delays. It is also evident from Figure 2 that a delay of 10 minutes occurred the most while a range of 10-20 minutes is an average delay. The locations having the most delays are the main intersections of the Greater Toronto Area.

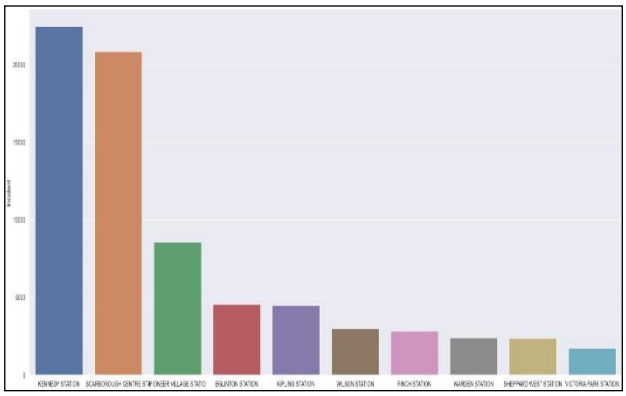

Figure 2. Locations facing the most delays

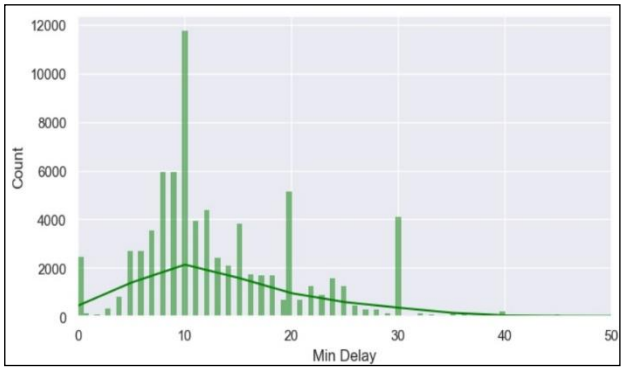

Figure 3. Occurrence of Delay in minutes

It is observed from Figure 3 that the maximum delays are caused on weekends than on weekdays since people tend to move out on weekends.

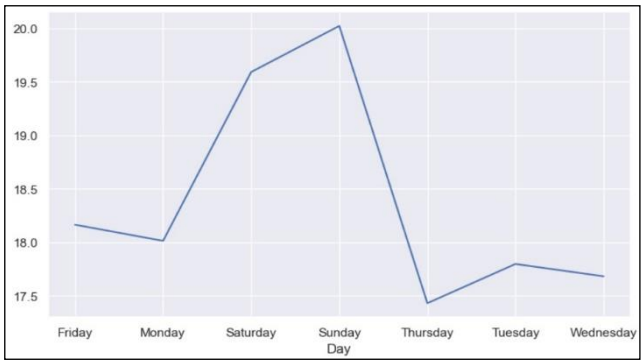

Figure 4. Delays on days of the week for existing data

The entire dataset was represented with a combination of front-end and back-end to view and analyze. Figure 5 depicts the dtale library in Python to view datasets, correlate, visualize missing values and outliers, summarize data, conduct feature analysis, time-series analysis, and predictive sources.

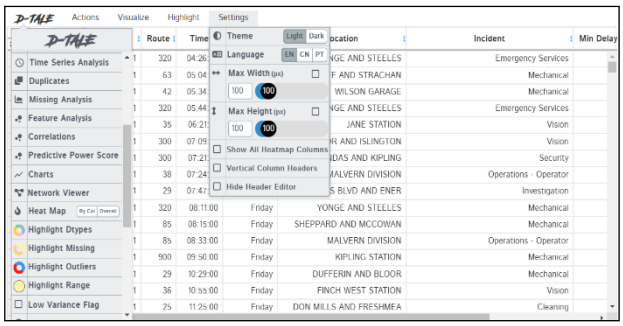

Figure 5. Visualizing and analyzing the dataset

5. Data Analysis and Visualization

5.1 Types of Analysis

There are four types of Data Analysis: Descriptive Analysis, Diagnostic Analysis, Predictive Analysis, and Prescriptive Analysis.

5.1.1 Descriptive Analysis

Descriptive Analysis answers questions about the events that occurred in the past to generate information. This report conceptualizes the average delay and the gap between two consecutive buses in minutes.

	Route	Min Delay	Min Gap
count	78616.000000	78616.000000	78616.000000
mean	195.603185	18.238310	31.119874
std	3217.833302	40.053344	41.550283
min	1.000000	0.000000	0.000000
25%	37.000000	9.000000	18.000000
50%	71.000000	11.000000	22.000000
75%	120.000000	20.000000	36.000000

Figure 6. Average Delay and Average Gap in minutes

5.1.2 Diagnostic Analysis

Diagnostic Analysis determines the cause of a particular event that occurred in the past. This report involves an analysis of the major incidents that caused a delay in the schedule, along with analyzing what time of the day witnessed the maximum delay. Figure 2 represents the word cloud to quickly identify incidents responsible for most delays, while Figure 3 shows the maximum delay between 02:00 and 06:00 pm.

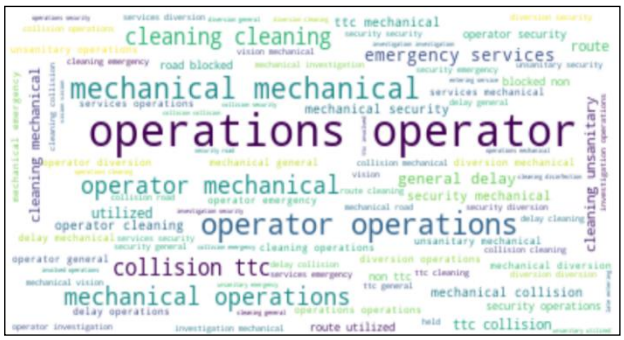

Figure 7. Significant incidents that cause most delays

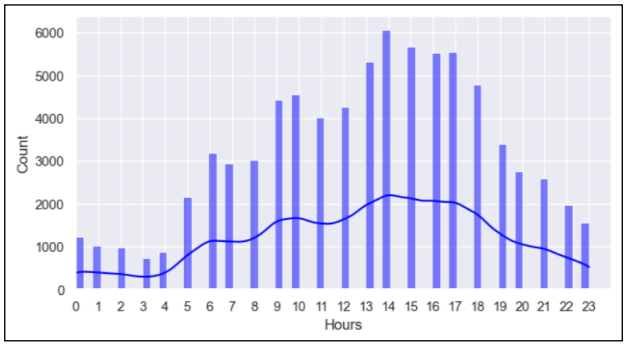

Figure 8. Time of the day witnessing most delays

5.1.3 Predictive Analysis

Predictive Analysis determines what will happen in the future based on the events that occurred in the past. This report includes predicting delays in 2023 using Prophet for time series analysis depicted in Figure 4.

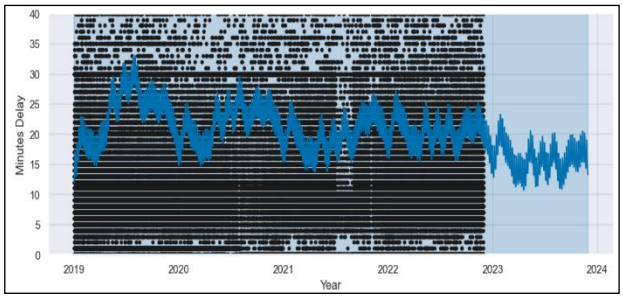

Figure 9. Prediction of delay in 2023

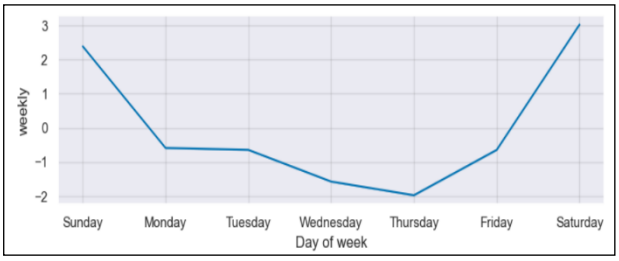

Figure 10. Predicting delay for days of the week

It shows the decline in wait for the upcoming year compared to the past few years. According to Figure 5, the maximum delay will likely happen during weekends in 2023.

5.2 Machine Learning Regression Algorithm

There are various Machine Learning algorithms for Regression problems, like, Linear Regression, Logistic Regression, Support Vector Machine, Decision Tree Regression, Random Forest Regression, and many more. This report emphasizes trying various Regression algorithms and exploring the one most likely to give better results. A target label should be well-defined to apply the Regression technique to the dataset, which will be predicted from the remaining features. In this case, the target label is expecting a delay in minutes. Moreover, the dataset should be divided into train and test data with a ratio of 80% and 20%, respectively. Regression algorithms such as Decision Tree, K-Nearest Neighbor, Random Forest, XGBoost, AdaBoost, and Support Vector Machine are used to analyze the performance of each algorithm on the same dataset. Table 2 shows the comparison of these algorithms that were implemented on the TTC dataset.

Algorithm	Score
Decision Tree	0.852
Support Vector Machine	0.018
Random Forest	0.932
K-Nearest Neighbor	0.878
AdaBoost	0.824
XGBoost	0.946

Table 2. Comparison of Regression Algorithms

As the Random Forest Regression technique is robust to outliers and reduces the risk of overfitting (Kho, 2019), we tuned its hyper-parameters to explore its working further, so Random Forest Regressor is the baseline algorithm for this project.

5.2.1 Random Forest Regressor

The Random Forest technique divides the dataset into random subsets and applies the Decision Tree algorithm to predict the label. Later, it averages all the predicted values to provide the final prediction for the entire dataset.

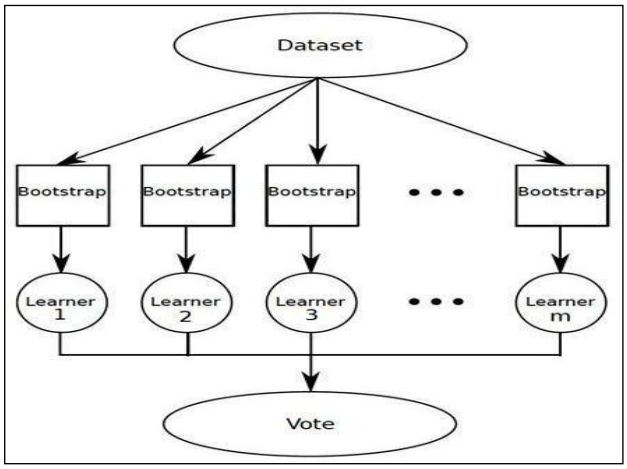

Figure 11. Random Forest Regressor Working (Kho, 2019)

There are several hyper-parameters for RandomForestRegressor, such as n_estimators, max_depth, min_samples_split, max_features, max_leaf_nodes, n_jobs, random_state, min_samples_leaf, and verbose. They are used to tune the model to give values for these parameters yielding the best result.

A hyper-parameter named n_estimators is used to specify how many decision trees the technique should produce. Also, max features determine how many features at most should be used during training. So, we built a function called RandomizedSearchCV from the sklearn library to set these hyper-parameters.

```
n_estimators = [int(x) for x in np.linspace(start = 100, stop = 1200, num = 12)]
max_features = ['auto', 'sqrt']
max_depth = [int(x) for x in np.linspace(5, 30, num = 6)]
min_samples_split = [2, 5, 10, 15, 100]
min_samples_leaf = [1, 2, 5, 10]

random_grid = {'n_estimators': n_estimators,
               'max_features': max_features,
               'max_depth': max_depth,
               'min_samples_split': min_samples_split,
               'min_samples_leaf': min_samples_leaf}
```

Figure 12. Hyper-parameter Tuning with Randomized Search

Since the parameter values are lists, a Randomized Search will try all possible combinations and find the best parameters. These best parameters are used to train the model to give optimized results.

```
rf_random.best_params_

{'n_estimators': 400,
 'min_samples_split': 5,
 'min_samples_leaf': 5,
 'max_features': 'auto',
 'max_depth': 15}
```

Figure 13. Best Parameters

5.2.2 Model Evaluation

This trained model is now analyzed for the performance using Mean Absolute Error which is the mean distance between actual and predicted values; Mean Squared Error which means the mean squared distance between actual and predicted values; and R-square, which checks how well the model fits. (Trevisan, 2022)

```
# check mean_absolute_error
mean_absolute_error(y_test, y_pred)
1.363203127534241
# check mean_squared_error
mean_squared_error(y_test, y_pred)
52.8262262168185
# check r2_score
r2_score(y_test, y_pred)
0.9503266795541878
```

Figure 14. MAE, MSE, and R-square Evaluation

5.2.3 Analyzing Results

The outcomes and learnings from the dataset analysis can be utilized to respond to inquiries about gaps, delays, and occurrences that led to a delay in historical data. Additionally, these findings can be used to forecast the delay for 2023

by day, week, and year to estimate how much of a delay an occurrence might bring about in particular areas.

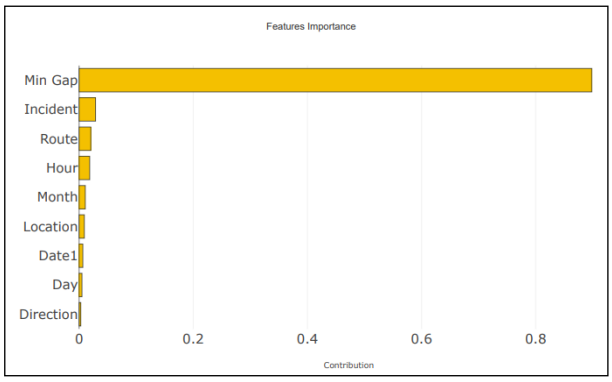

Figure 15. Feature Importance

Dataset | Dataset Filters | True Values Vs Predicted Values

_index	_predi	Route	Day	Location	Incident
6	12.65	300	0	2948	23
83	9.24	38	2	8616	0
248	111.8	97	5	4517	22
250	20.18	14	5	4321	3
270	12.01	985	5	8608	0
353	20.65	85	6	6529	0
369	7.41	53	6	7769	0
383	13.12	34	6	8616	0
450	29.98	97	4	1901	17
501	12	38	4	8616	0
644	15.3	134	0	8616	17
780	20.08	48	1	8537	0

Figure 16. Dataset, Predicted Values, True vs. Predicted Values

6. Utilization of Results

Questions based on gaps, delays, and incidents that caused a delay in previous data can be answered using the findings and insights from the dataset analysis. Also, these findings can be used to forecast the amount of delay that will occur daily, weekly, and annually in 2023 to assess the potential impact of incidents in specific regions.

7. Summary

The delayed TTC bus dataset is analyzed in the paper to uncover different unreported elements. Although there are other regression techniques, Random Forest is the project's baseline technique since it handles outliers well. To sum up, bus drivers are the leading cause of delays. Although there will be fewer delays overall in 2023, the weekend tendency will continue to follow the same pattern.

References

Kho, J. (2019, March 12). *Why Random Forest is My Favorite Machine Learning Model*. Medium. https://towardsdatascience.com/why-random-forest-is-my-favorite-machine-learning-model-b97651fa3706

Trevisan, V. (2022, March 25). *Comparing the robustness of MAE, MSE and RMSE*. Medium. https://towardsdatascience.com/comparing-robustness-of-mae-mse-and-rmse-6d69da870828

TTC, Toronto – bus Times, Routes & Updates. (n.d.). Retrieved January 23, 2023, from https://moovitapp.com/index/en/public_transit-lines-Toronto_ON-143-434