**Summary Paper:**

**Predicting Toronto Public Transit Delays**

**DS4 - Group 2:** Daniel Cebula, Melissa Hartwick, Allan Sales, McKinleigh Needham, Aravind Kakarala, Athithian Selvadurai

# **Introduction and Objective**

As part of re-opening of the economy, we felt it was timely to review TTC data to determine if we could classify and predict transit delays, in hopes of being more prepared for our organizations’ return to work plans, and provide one more data point to help us make the right decision for our commute.

The objective of our analysis is to predict the length of Toronto Transit Commission (TTC) subway, bus and streetcar delays for the period 2014 - 2019. Our hypothesis is that the publicly available TTC delay data, when combined with weather data, is sufficient enough to predict the length of a delay within +/- 10 minutes.

# **Data Preparation**

Two datasets were used for our analysis:

* TTC Delay Data: Subway delay data, Bus delay data, Streetcar delay data

Subway Delay Data: <https://open.toronto.ca/dataset/ttc-subway-delay-data/>

Bus Delay Data: <https://open.toronto.ca/dataset/ttc-bus-delay-data/>

Streetcar Delay Data: <https://open.toronto.ca/dataset/ttc-streetcar-delay-data/>

* Canadian Historical Weather

<https://climate.weather.gc.ca/historical_data/search_historic_data_e.html>

<https://drive.google.com/drive/folders/1WJCDEU34c60IfOnG4rv5EPZ4IhhW9vZH>

The data was retrieved through a series of HTTP requests and downloaded as a comma-separated values (csv) file. The data was then imported into a pandas dataframe in Jupyter Notebook for data cleaning and preparation.

After data cleaning and preparation, we had one final data set that had 653,736 observations and 8 columns with ‘Min Gap’ to be used as the target variable for predictions.



# **Model Design**

Principal Component Analysis was used to reduce dimensions while preserving 95% of variance. We then set-up pipelines to fit RandomForestClassifier, GradientBoostedClassifier and LinearSuportVectorClassfier. GridSearchCV was then done on selected hyperparameters to find the highest accuracy. The following hyperparameters were chosen for the models, as these yielded optimal accuracy while ensuring an acceptable run-time:

1. Random Forest Classifier (n\_estimators=10)
2. Gradient Boosted Classifier (n\_estimators=10)
3. Support Vector Classifier (tol=0.1)

Once these models were developed, we also ran an ensemble model in an attempt to reduce the generalization error of the predictions from each individual model. The VotingClassifier ensemble was built using these three classifiers using a hard voting scheme.

# **Model Evaluation**

Of the models used, their accuracy in predicting ‘Min Gap’ was as follows:

| **Model Name** | **Accuracy Score** |
| --- | --- |
| Random Forest Classifier | 46.6% |
| Gradient Boosted Classifier | 59.7% |
| Support Vector Classifier | 60.2% |
| Ensemble | 59.8% |

# **Conclusion**

We realized that predicting delays is actually quite difficult due to the additional factors that are not typically present in a dataset (e.g. knowledge level of the mechanic).

For future iterations of the model, we would recommend keeping the mode of transportation data separate and looking to create a prediction model of each of the three modes of transportation (i.e. subway, streetcar, and bus). We hypothesize that by combining the modes of transportation, we assume that the nature of these delays are somewhat similar which isn’t necessarily the case.