Executive Summary

Predicting TTC Delays

Rachael Lam (rachael.lam@mail.utoronto.ca) & Hamid Yuksel (h.yuksel@mail.utoronto.ca)

I. INTRODUCTION

The Transit Toronto Commission (TTC) is a public transportation agency that operates streetcar services in the City of Toronto. Additionally, it controls bus, subway and paratransit services in the Greater Toronto Area (GTA). The TTC serves 5.5 million people in the GTA with roughly 1.7 million customer journeys on a typical weekday.

II. BUSINESS OBJECTIVES

Many customers rely on the TTC to travel to their destinations safely and promptly. Customer dissatisfaction could not only lead to massive repercussions for the TTC including decreased ridership, loyalty and profitability, but also consequences for customers in their daily lives, such as job loss, childcare fees and healthcare access. With this understanding, it's important to know where delays occur and how delays are affected.

The assumption is that with this knowledge, the TTC will be able to make adjustments to their services to better assist customers. The objective is the following:

1) *Investigate to which extent external events are related to TTC streetcar delays.* The intuition is that external events such as rush hour traffic or emergency incidents cause more or fewer extreme delays. This information could help inform how to adjust services if our intuition is correct.
2) *Predict the time of day that a delay will occur based on selected features.* With this prediction, it will be easier for the TTC to better handle delays based on the time of day.

III. RESULTS

A. Investigate to which extent external events cause TTC streetcar delays.

We investigated the relationship between location, time, incident and length of delay using a k-means clustering algorithm. Figure 1 shows the silhouette score for every observation and the average silhouette score (dashed red line) with eight clusters.

Only one cluster is less than the average, thus insinuating it is close to other clusters. While no cluster has a silhouette score of less than 0, clusters are still nearer than expected, proving a somewhat weak relationship between location, time, incident and length of delay. Even so, location, time and incident still have an impact on the length of delay and this information can help optimize TTC's services to better account for the length of delay within each cluster.

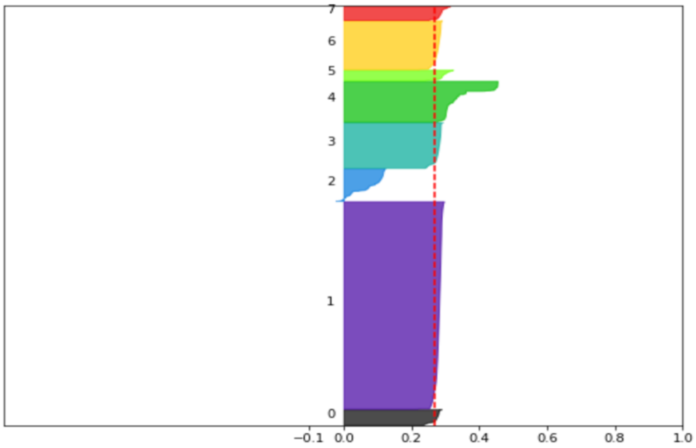

Fig. 1. Silhouette plot for $k = 8$

B. Predict the time of day that a delay will occur based on selected features.

Our machine learning model can predict with low error the min delay for a TTC bus given n features. Predicting the min delay has a Root Mean Square Error (RMSE) of ~1.4. We implemented and compared two forms of regression models: one using CatBoost and one using a simple decision tree from SciKit. While both performed much better than our baseline estimates of ~20.5, CatBoost noticeably had a much higher RMSE of ~7.5 than that of the simple decision tree.

CatBoost can train using the categorical features right away, whereas the simple decision tree required the categorical features to be encoded. We found that a depth of 6 to be optimal for the CatBoost model, based on automated depth testing. However, CatBoost's ease of use shows it is not always going to offer the best results. In this case, the less complicated simple decision tree regressor prevails. This model can help TTC anticipate delay lengths more accurately so that they may better keep the public informed with more accurate estimates. In doing so, TTC patrons will get less annoyed as delay times extend past estimates.

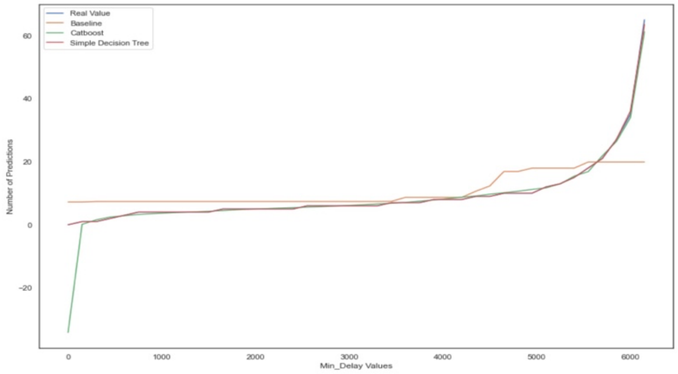

Fig. 2. Modelling and predictions