Analyzation of Traffic Patterns in Relation to Delays in Boston Public Transit

Cristian Mendivil
crme7282@colorado.edu

Lucas Lyon
luly2738@colorado.edu

Vamshi Arugonda
vaar2387@colorado.edu

Theodore Freeman
thfr5770@colorado.edu

Abstract

Do delays in public transit times correlate with increased travel times for uber? What are the most influential weather patterns affecting bus travel time? Based on the delay from weather patterns, how much extra travel time should a Boston commuter account for?

Generally we found that a fair number of trips in the uber data were delayed similarly to bus routes in the mtba data set, but there were also plenty of trips that were actually faster than average suggesting only a slightly positive correlation. Attempts to accurately predict percentage of daily buses on time from weather conditions were unsuccessful.

1. Introduction

The authors of this paper will layout the methods and sources of data for an analysis of traffic, ride sharing, and public transportation in the Boston metro area and how weather plays into the delays . Data will be sources from 2016 to present, and most of the data will serve to provide a baseline for normal travel times between various sectors of the city. By examining the delays in bus transit times we will provide the Massachusetts Bay Transportation Authority (MBTA) with useful information on where to account for extra travel time and wariness of traffic. The city of Boston may also be able to use this information to prioritize or de-prioritize road maintenance in bad weather. This paper will go into the application of knowledge gained, previous work that has been done on the data, details about the data sets and where to download them, methods of evaluation, and data mining tools.

2. Related Work

There has been quite a bit of prior work about traffic prediction and analysis. One such study was done by a group of researchers at the University of Southern California with a goal of accurately predicting and quantifying impact of traffic incidents. This is a pretty good study, as in their conclusion they claim that their model can increase "prediction accuracy of baseline approaches by up to 45% [1]" for the impact of traffic incidents on road networks. Another group did some work on developing a support system for using real time bus location data to accurately estimate arrival times. This study may be useful considering that all the data we intend on using is public transit data or Uber. Perhaps it can give us some ideas of how to use our public transit data in a cleverer way. A way in which our project will be different from the described research above is in a couple of ways. First, the most recent of these projects was done in 2016 so there is potential at least to have more currently relevant results. Second our work is going to try to learn how individual traffic delays affect city-wide transit rather than just providing time estimates for when the next bus will arrive or route prediction for obstruction avoidance. Both projects may be useful to us though by providing different ideas for how to use and view our data as well as what we might avoid. If we find that we are getting stuck in a corner though and neither of these are able to help get us out it seems there is plenty of other research out there which if we searched for we may be able to find our answers. [2]

There is also some work done on analyzing the impact of weather on traffic. We found an article in which writers were researching the relationship between inclement weather and traffic flow in Istanbul. Researchers used Remote Traffic Microwave Sensor (RTMS) and weather data from two highway roads in the Istanbul metropolitan area. The research found that traffic went slower when it was raining, but traffic went a bit faster if the road was wet and it was not raining. Researches also found that snow decrease the total number of cars in the roads by a significant amount. [3]

3. Data Sets

Our team has utilized three data sets: one from Uber, and one from MBTA, and a third from NOAA. The Uber dataset must be downloaded in quarter-year increments from Uber's exclusive data tool, Movement. It can be found at *https://movement.uber.com/*. The MBTA dataset must similarly be downloaded in a few chunks rather than in one file from their dashboard at *http://www.mbtabackOntrack.com/performace/index.html\#/download*. You must select the radio box "Reliability". Our team had difficulty downloading the entire dataset at once, and had to split the download into three time frames: January 1st 2016 - January 1st 2017, January 2nd 2017 - January 1st 2018, and January 2nd 2018 - March 5th 2018.

The MBTA Reliability dataset has 395,130 rows and 9 attributes. The attributes include service date and time, whether the row is for Off-Peak service or Peak service times, the type of transport (including rail, commuter, and bus), the route line, stop station, metric measured (including Passenger Wait Time and Schedule Adherence), and varying numerators and denominators for those metrics.

We primarily used the service date and time, and numerator and denominator attributes. The Peak vs Off-Peak hours attribute was not very helpful for establishing a link between delays in public transit and traffic, and similarly was not effective when training regression models.

The Uber Movement dataset is split into 7 distinct .csv files, each containing 3 months worth of travel times between every Uber-defined source and destination in Boston. Each file has 7 attributes: sourceid, dstid, hod, meanTravelTime, standardDeviationTravelTime, geometricMeanTravelTime, and geometricStandardDeviationTravelTime. We decided to use only the geometric mean travel time attribute in our analysis as it was most relatable to the bus data. This data set turned out to be more difficult to use effectively as we lacked dates for each entry which would have allowed us to more directly compared bus data. This also would have allowed us to do similar or more complex attempts at predicting delays based on weather.

Our third data set was daily aggregated weather values for the boston area from the National Oceanic and Atmospheric Administration (NOAA). This data set included 44 attributes, the majority of which were empty. In order to use this data we trimmed those 44 down to 5 which were inches of precipitation, inches of snow, average temperature, max temperature, min temperature. Initial analysis was performed with all of these but later average temperature was removed. Unfortunately it's removal didn't increase performance of the models.

4. Main Techniques Applied

4.1 Correlation Analysis

4.2 Regression

At some point late in our semester we were informed that we were a little off track and still needed to do some data mining. In order to achieve this goal we decided to download the weather data and add another question to answer. We downloaded the NOAA data set and decided that the attributes concerning precipitation, snow and temperature values were the only useful ones in the set that would also not confuse the regression models. The question we decided to answer was "to what degree do different weather conditions affect the percentage of buses on time?" In order to do this we first had to combine the MTBA bus data with the newly acquired NOAA weather data. This was done by formatting their date columns to be the same and then using the pandas join method to apply the daily weather entries to the appropriate days in the bus set.

After doing this, we reorganized the newly joined data in Excel so that the columns would include the weather stats, if the entry was during peak hours or off peak hours, and what bus route the entry was for with percent of buses on time as our "score". Dates were removed because they might mudle our results. In order to make the bus route usable we would have had to use one-hot encoding to transform the bus route numbers into categorical data and would have added around 500 new attributes to our data set. Therefore, we decided not to use these in order to preserve good performance. The peak or off peak attribute was binary in nature so it was transformed from string entries into a binary (one or zero) format for simplicity.

Initial attempts for regression were to use this data set above split nearly half and half for training and testing and attempt to train a number of different regression models on them. The regression algorithms we tried were a Multi-Layer Perceptron Neural Network, a Random Forrest Regressor, Logistic Regressor, TheilSen Regressor, and a couple others. Initially it appeared that the regressors weren't performing well and that it was because the peak or off peak attribute was muddling the performance of the models. When this feature was then removed and training re-attempted and evaluated it seemed that suddenly all of the models were performing quite well but these positive results turned out to be from bugs in code. Where the data was being partitioned for training and testing sets the column indices were off by one and so the models were actually being fit to predict what the minimum temperature would be based on the other weather data. When the bug was fixed, it turned out that the models were unable to accurately predict

5. Key Results

[TODO]

6. Applications

Given a full set of results, there are numerous applications

7. Further Research

8. Figures

9. References

[1] B. Pan, U. Demiryurek, C. Shahabi, and C. Gupta, “Forecasting Spatiotemporal Impact of Traffic Incidents on Road Networks,” in 2013 IEEE 13th International Conference on Data Mining, 2013, pp. 587–596.

[2] F. Sun, Y. Pan, J. White, and A. Dubey, “Real-Time and Predictive Analytics for Smart Public Transportation Decision Support System,” in 2016 IEEE International Conference on Smart Computing (SMARTCOMP), 2016, pp. 1–8

[3] Akin, D., Sisiopiku, V.P., Skabardonis, A., 2011. “Impacts of weather on traffic flow characteristics of urban freeways in istanbul.” Procedia: Soc. Behav. Sci. 16.

Routes v Percentage on time

Date v Difference (Average-Actual)

Date v Ratio

Weather Data

MBTA reliability