DS3000 Final Project: An Analysis of MBTA Green Line Traffic Congestion

Max Correia, Aayush Dhiman, Gavin Perkins

Northeastern University Khoury College of Computer Sciences

Abstract

The Massachusetts Bay Transportation Authority (MBTA) is the public transportation service that operates in the Greater Boston area. The goal of this project is to **determine which temporal features most affect system traffic** through our analysis of the gated fare data of the MBTA's Green Line in 2022. We ran a hyper-tuned a Random Forest Regression model over the data set to predict the factors that cause the greatest system congestion based on our generated decision trees. We also evaluate the performance of our model at predicting said parameters.

Introduction

The MBTA serviced more than 790,000 people in September 2023 [4]. It supports the livelihoods of workers and other commuters in the Greater Boston area. However, the MBTA has gained notoriety for inconsistencies due to poor management and constant delays among other factors. We aim to create a model that allows riders to determine which factors most affect commutes on the MBTA subway.

Our motivation for choosing this topic lies in our daily use of the MBTA. Several Northeastern students, ourselves included, have relied on the MBTA to get to classes, co-ops, or anywhere around the greater Boston area. Processing this data would provide a significant boon for all students, as we would be able to better plan our commute or communicate our needs regarding public transportation to the officials at the MBTA to further improve the service. Our goal for this project is to share this information with the general public, as it would make the information surrounding the MBTA much more digestible for the average MBTA user, as well as create a liaison between the people of Boston and the officials at the MBTA.

Our main goal is to determine which factors cause the most congestion using gated fare entries over 2022's MBTA subway fare data. Our end result would be to find the most impactful values that affect commutes using the MBTA subway system and potentially predict ridership trends in the future. We will clean our data and run an exploratory data analysis and determine which algorithm works best for our aforementioned goals.

Related Work

During our analysis, we took note of similar projects and sources compiled in the past that are closely related to our project and desired outcome:
Dwell Time Model and Analysis for the MBTA Red Line documents a more formulaic approach to determining the root causes of congestion at Kendall and South Station in 1999. This is similar to our work, and could be an inspiration for further analysis on our data
The **MBTA Blue Book** documents past ridership data for all of its services including its subway system. Notably, it also lists system shutdowns and events that may have impacts on ridership data during those time periods. It is similarly compiled to our modern dataset but does not go into analysis over said data.

Methodology

Data Acquisition

The data source we chose is a data set of gated fares over 2022, coming directly from the MBTA's Blue Book Open Data Portal. There were datasets spanning the past 10 years, but we chose to focus on the closest full year in order to get more relevant results for use in the present day.

Data Preparation

We first loaded our data set into a Pandas dataframe and isolated data points that relate to the Green Line. We first noted that the column stop_id in the dataset is redundant with the column station_name. Both of these columns indicate the same thing, and many stops do not have stop_ids, thus making it more difficult to use it to reference stops. As a result, we decided it would be best for the column to be removed entirely, as it was not providing additional information to the dataset. We then cleaned the data and searched for rows with missing or outlying data, but we failed to find any with the data that we had chosen. We also feature engineered columns for month, day, hour and minute through the time data in each point.

Model Selection

For our model, we chose to proceed with a Random Forest Regression model. The model is intended for use with continuous data, and can be used with large data sets like ours while being built to be resistant against overfitting. This is compared to algorithms like K-Nearest Neighbors (tuned based on the "n_neighbors" parameter) and Support Vector Regression (tuned based on the "C" regularization parameter); these algorithms were inaccurate and slow due to the density of our data. The best values for those models (n_neighbors=3, C=2) returned MSE values of 1985.87 and 7770.79 and r^2 values of 0.72 and -0.08 respectively.

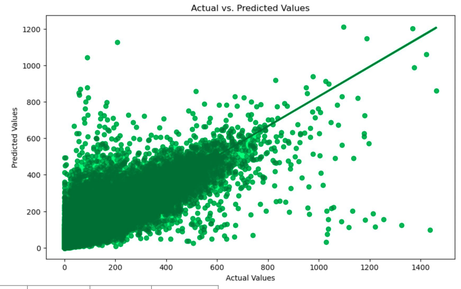
Model Explanation

The data was partitioned such that 80% of the data was used for the training set and 20% of the data was used for the testing set. After partitioning the data, we hypertuned the "n_estimators" parameter of an unseeded Random Forest Regressor through GridSearchCV, fitting the grid with the X and y training data and displaying the best "n_estimators" value with its corresponding MSE. We evaluate the performance of our model through the mean squared error and r^2 values; the smallest MSE value will give us the best performance model. We also compared the mean and standard deviation for our test and training sets in order to minimize bias and variance.

Results & Analysis

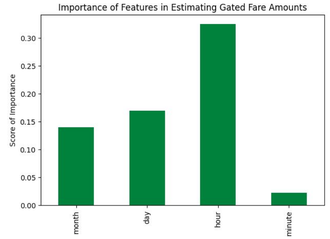
Based on the results from our Random Forest Regression model, we determined that **n_estimators=6** was the best value we could obtain through our model. The model's mean squared error, which measures the difference between the actual and the predicted value, was **1786.56**, which is exceptionally high. This means that the model is, on average, further away from the actual value and is failing to capture the true pattern in the data. The root mean squared error is not as high at **42.27**, which means that the magnitude of errors is not that high compared to the mean squared error. We also note that our model is **72%** accurate; while this may be lower than we had hoped for, it is the highest out of the models we tested.

Furthermore, the r^2 value is **0.75**; while this is also not as high as we would like, it was the largest out of all of our models, and means that our model does fairly well at correlating actual and predicted values. The model also states that the predicted average number of fares at any given time is **71**, which is consistent with MBTA data from 2022.



	Mean Train	STD Train	Mean Test	STD Test
6	0.945387	0.002221	0.726670	0.006566
5	0.934155	0.001596	0.724098	0.002782
4	0.942395	0.001657	0.719715	0.008327
3	0.922750	0.000666	0.701399	0.005732
2	0.898334	0.004442	0.660475	0.014263

We note that as n_estimators gets larger, the mean train and test scores consistently increase with them.



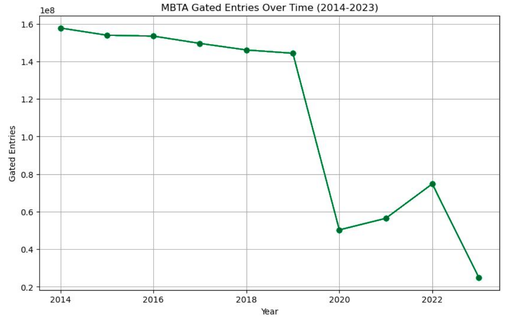
Notably, the hour of day is the most important out of our features in determining the amount of gated fares at a given time. This makes sense due to high usage during morning and evening commutes, while also factoring lulus in usage during midday hours. The day also makes a difference, as most holidays happen at specific times of the month, meaning that travellers are more likely to use the T during these specific days.

Impacts

The impact of this study is twofold: it would allow maps and transit apps to suggest the multiple alternate paths from one destination to another in order to provide the user with the option with the least amount of public transit congestion, and it serves as a demonstration of which times the MBTA is flooded; this data could be presented to the MBTA board in order to lobby for more public transit vehicles at those peak times. This data-driven advocacy approach allows for a more targeted and efficient allocation of resources, specifically during rush hours and times of high demand. Through showing that the time of day is the most important feature in estimating the number of fares at a given time, we can confidently state that the MBTA should prioritize the improvement of their systems during rush hour conditions, as opposed to any specific time of year. Especially during closures, the MBTA subway system suffers from large influxes of riders on their other lines; as such running more trains during those times would work towards reducing congestion within their system.

Conclusion

Given more computational resources, this project could be further expanded upon by looking at a larger period of time to obtain a more historical analysis of MBTA traffic data, or through other or all lines of the MBTA at once. Analyzing the other models could also be a potential avenue for further research, giving us a more holistic view of our data and allowing us to better compare our models for accuracy. It's important to note that the state of transportation had changed significantly over the last few years. Due to the COVID-19 virus and the lockdown that had occurred as a result, many of the trains, buses, and other public transportation was used significantly less. As such, working with those past results would potentially lead us to create an inaccurate model for predicting current and potentially future trends regarding the MBTA.



As such, adding this data to the dataset may introduce unintentional skew the model, resulting in more variability in the data that is increasingly unaccounted for by the model. However, a deeper analysis with the data from 2014 to 2023 rather than just 2022 data would be significantly more adept in predicting the number of gated entries.

Overall, our Random Forest Regression model was not as accurate as we had hoped, yet the data we did obtain indicated some interesting trends that should be further explored upon in future experiments.

References

1. Our data set: https://mbta-massdot.opendata.arcgis.com/datasets/mbta-gated-station-entries-historical/about
2. Puong, Andre. "Dwell time model and analysis for the MBTA red line." Massachusetts Institute of Technology Research Memo (2000): 02139-4307.
3. Authority, Massachusetts Bay Transportation. Ridership and service statistics. MBTA, 2010.
4. MBTA. "Ridership on the T." MBTA, www.mbta.com/performance-metrics/ridership-the-t. Accessed 3 Dec. 2023.