EECE5644
Introduction to Machine Learning
Final Project Proposal

Abstract

We plan to explore the MBTA [Bus] Ridership dataset, which contains information about the number of passengers entering and leaving at each stop. The MBTA publishes publicly accessible data on their websites which we have included below for organizational purposes. Some examples of data which they provide include season and year in which a particular ride took place, the direction of the transportation vehicle, whether or not a specific trip took place on a weekday or a weekend, where the vehicle stopped, the average number of people who got on the vehicle as well as the average number of people who got off at any given stop. With the data of average ons and offs the number of people on the vehicle at any given time is also calculable and is provided under the data column titled: "average load". Perhaps the most important piece of data included in these published reports is the time it takes to get from one stop to another. Therein lies the focus for this project: the calculation of the time it takes to get from a specified starting point to a desired destination. We hope to achieve an accurate estimation for any given voyage utilizing the various factors outlined in the MBTA's published data. Some obvious correlations can be drawn from the data and the result we hope to obtain. If there are 100 passengers on the train at one time it can be expected that it may take longer for each of them to get off resulting in a longer commute time. Inversely if there is a specific stop that very few passengers get on or off at it can be assumed that this stop will not be the reason for delays in the commute. Through the in depth analysis of this provided data as well as alternate external factors we hope to include in our research we hope accurate commute estimations are possible.

As far as who this data can benefit, we believe our findings will be beneficial to anybody living in the Massachusetts area whether it is those who regularly use public transportation for their daily commute or those who take a train once a month to catch a sporting event. In our data collection we ensured that our data sets were region dependent that way we could come up with a conclusion that would be most beneficial to those in our area. A similar project could have been pursued on a nationwide scale, but we believed the greatest benefit would be a more niche take on the subject. As far as how we will handle this data we have decided to use decision trees as a way of looking at the data in terms of factors and results to see if any clear connections can be drawn. We intend to use a multitude of different factors from varying data sets to see if there are any non-obvious factors one might miss on a first take.

Bus Data:
https://mbta-massdot.opendata.arcgis.com/datasets/MassDOT::mbta-bus-ridership-by-trip-season-route-line-and-stop/explore
Commuter Rail
https://mbta-massdot.opendata.arcgis.com/datasets/MassDOT::mbta-commuter-rail-ridership-by-trip-season-route-line-and-stop/explore