Comparing Trip Durations between CitiBikes and Taxis

Prathyush P Rao
Student ID: 1102225
Github repo with commit

November 4, 2022

1 Introduction

In a post-COVID world, a lot has changed. Work and work styles are different, healthier lifestyles have been adopted, and issues that weren't given much attention before are now the main focus of people's lives. One such trend that has changed is commuting to work. In New York City (NYC), there are now a plethora of commute options and recently, there has been a surge in biking and bike-sharing services [1]. The usage of bike-sharing services, such as CitiBike [2], has even surpassed levels seen in 2019 [3]. One particularly interesting insight seen post-COVID was that CitiBike usage had a rapid initial recovery after being struck by the Pandemic as compared to one of the most popular modes of transport in NYC - the Subway [4]! This gives a new perspective on how we are to look at commuting in the near future.

This report attempts to compare ride-share services such as taxis (yellow, green, and high volume for-hire vehicles (HVFHV)) and CitiBikes and predict the time taken to travel in either of the services in Manhattan on a weekday. It assumes the perspective of a person that has a choice between the two services and asks themselves "What would be the fastest mode of commute?". The report analyses data from July 2022 (from when the stay-at-home COVID order ended [5]) to April 2022 (latest available data) to model data that is potentially the best indicator of the near future and current trends. It also addresses data processing, preliminary analyses, and builds upon two models:

- **A classification model** - A Multi-Layer Perceptron Classifier based on which service had a shorter trip duration
- **A regression model** - A Gradient Boosted Tree Regression Model based on differences in trip duration between bikes and taxis

both generalized to areas in NYC (detailed in Section 2.1). The training and analysis will focus on the available data before 2022 and the test set will be comprised of the data available in 2022.

2 Data, Preprocessing, Analysis, and Geo-Spatial Visualisation

2.1 Dataset

The report uses yellow, green, and HVFHV datasets from the NYC Taxi & Limousine Commission [6], and CitiBike rider data [7] as the main points of comparison for trip durations from July 2022 to April 2022. Additionally, the report uses weather data that was web-scraped using a Selenium bot [8] for the same time period from the IBM Weather Company in NYC. Interesting trends were seen, as highlighted in the following sections, with regard to bike usage and weather throughout the year.

Also included was taxi zone information associated with the taxi data itself, which allowed the model to generalize the areas for both CitiBike and Taxi data

2.2 Preprocessing

To ensure consistency between both the main datasets, there were several steps required to process data. In addition to this, the weather data was also cleaned, imputed with weighted mean values, and uploaded as a link from where a script downloads the data.

2.2.1 Filters - Taxis

While the taxi data spans the whole of NYC, the CitiBikes are limited to only Manhattan and the edges of other boroughs because they need to be picked up and dropped off from designated stations only. Therefore, the taxi data was filtered by pick-up and drop-off location IDs that only contain bike stations in the area. And the same areas were used for all years even though bike stations were rapidly increasing over the years [9].

The taxi data was also filtered by passenger count (1-4 passengers only) to compare a bike that can fit one person against a taxi that is fit for one person. This eliminates possibly large vehicles that might be slow, and not fit to commute with just a single passenger. Values not in the data dictionary were removed, and so were airport trips, accessible trips, and shared trips to maintain the perspective of the person choosing the fastest mode of transport.

2.2.2 Filters - Bikes

Similar to the taxis, only the stations available in July 2022 were considered and the expansion into different stations and boroughs was not. And so all future data was filtered to keep the same number of stations. Also, in February 2021, CitiBike renamed all their station IDs, and so both new and old station IDs were matched to keep consistency.

2.2.3 Filters - Both

For both (Bikes and Taxis) to remove any outliers, a limit of 25 miles was applied to any trip (because each bicycle has can go up to 22-50 miles in one charge [10]), and also a trip duration limit of 3 hours was kept to keep it realistic for cyclists commuting on a daily basis. Only weekdays will also be considered to observe consistent patterns. All rows with null values were also dropped.

2.2.4 Feature Engineering

Features such as day, month, year, hour, and weekday were extracted from the pickup DateTime columns and trip durations were computed as the difference between pickup and dropoff times. The datasets were then merged on the date, pickup, and dropoff locations to compare and contrast times for the two modes of transport for similar rides. Also, merged into the dataset was the scraped weather data and a feature to indicate cold, warm, and hot parts of the year [11]. This can also be seen in Figure 1 as different parts of the graph have different bike trends. Another added feature was peak hours defined as 8-11 am and 4-7 pm.

2.2.5 Chosen Features

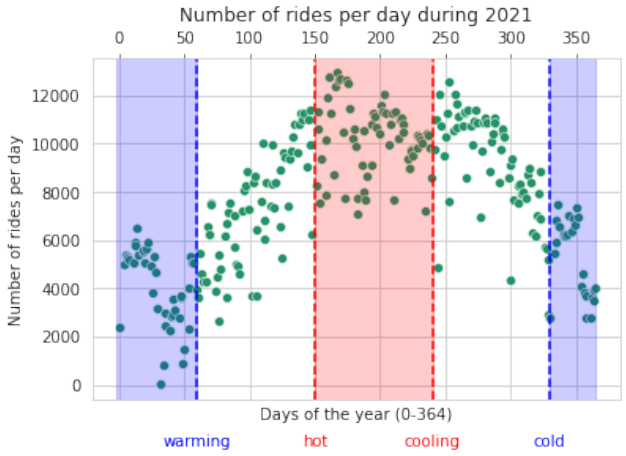

Figure 1: Seasonality of bike rides

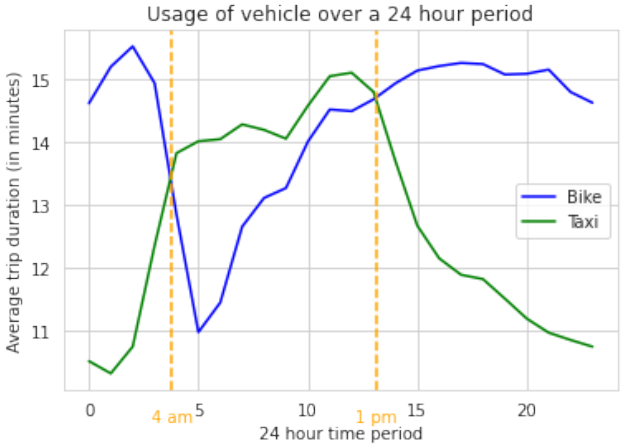

Figure 2: Average trip duration over a 24-hour period

Finally, 9,148,704 rows remained with 17 columns (including the target feature). The columns that remain in the final merged list are

- Month, Hour
- Day of week
- Seasonality pattern
- Pick Up Location ID
- Drop off Location ID
- Peak hour indicator
- Bike trip duration
- Taxi trip duration
- Trip difference

And from the weather data we have

- Temperature
- Humidity
- Wind Speed
- Wind Gust
- Precipitation
- Dew Point

2.3 Analysis

This section analyzes and understands the relationships between features, finds underlying trends and insights, and even disproves common misconceptions that may exist about the current dataset. For this section, additional features like number-of-vehicles-in-use count were engineered. Apart from this, some graphs were created with taxi or bike data before it had been merged and reduced. Due to most of the data being an aggregate function, no samples were required to be used.

2.3.1 Locations

The pick-up and drop-off locations can be thought of as a highly correlated features for the outcome as - the further the distance the longer it would take. Figure 3 shows a fascinating insight that further demonstrates the significance of Location IDs. Note, that red indicates that the taxis have shorter times and blue is in favour of the bikes having a shorter time. While it might be common for further away areas to be faster and more accessible by taxi, there are areas in the lower west of Central Park that seem to favour bike times a lot more than taxis (light blue shades for the morning peak times). And that is indeed the case! Those particular areas, Lennox Hill and Midtown East have been voted to be the most bike-friendly areas of NYC by the density of bike lanes, bike stations, and the number of bikes operating [12].

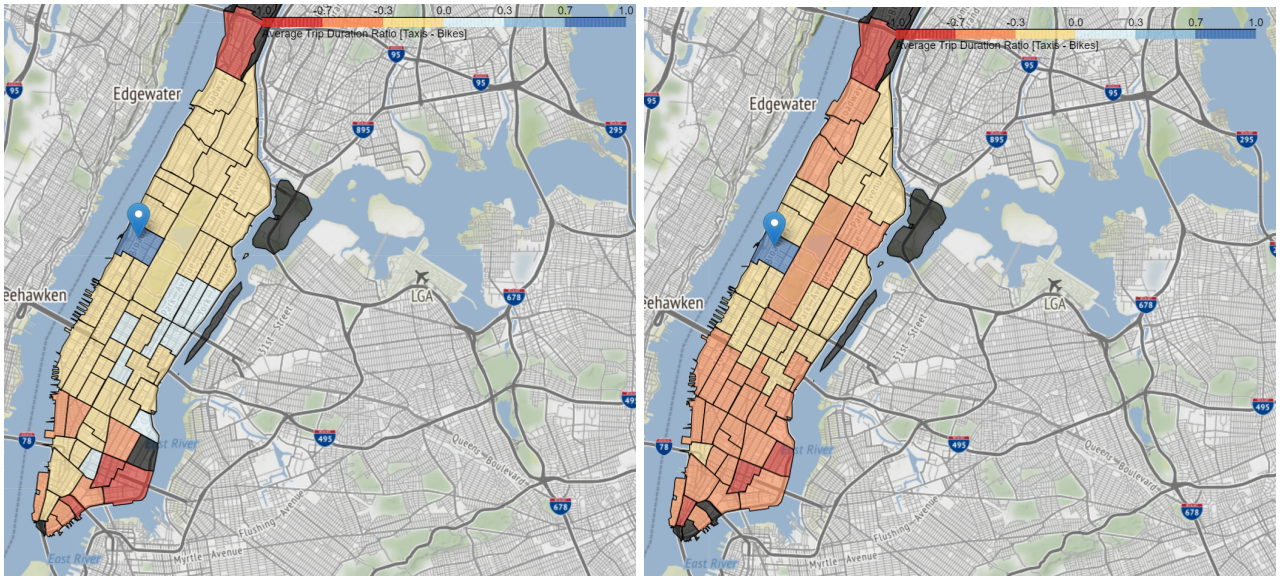

Figure 3: Time taken from marker to surrounding areas: from **7-10 am** (right), **4-7 pm** (left)

2.3.2 Peak Hours

It was first assumed that during peak hours, the traffic would slow down taxi trip durations and that would benefit those who use CitiBikes. This can be seen in both Figures 2 & 3. Figure 2 is based on the dataset which was not merged and contains the overall average trips for every hour. While the morning peak seems to be in favour of the bike, the evening peak favours taxis and exactly the same is reflected in Figure 3. An explanation for this could simply be that everyone starts their work between 7-9 am but ends at times they're comfortable with. More research into this would be helpful in building a more robust model.

2.3.3 Weather

As touched upon earlier, Figure 1 shows a seasonal trend in the graph that is rightly so as snowy and rainy conditions make it difficult to use bikes which, in fact, is also seen in the graph. On day 234 (22 August 2021), there is a sudden decrease in the number of riders (to around 5,000 from an average of 10,000) and that is because of a city-wide emergency declaration due to Hurricane Henry [13].

Further analysis was also done in ANOVA tables seen in Table 1 to check for feature relevance. And as the table suggests, there exists a strong relationship between weather information and trip durations. What is particularly surprising is that precipitation was found to be less relevant than any other factor, which goes against the idea that rain and weather would affect the ridership of bikes.

3 Model

This section goes into detail about the chosen models, predictors, and analysis of the performance of the said models. To compare which mode of transport is faster, we have one feature which could be made into two - trip difference. The trip difference on its own is linear and is simply the difference between times taken by taxis vs bikes. On the other hand, it can be converted to a categorical variable which simply shows which is faster but loses the quality of how fast it is. And so two models were used, a Multi-Layer Perceptron to classify the data and a Gradient Boosted Tree Regressor to quantify it.

	Sum Sq	**d.f.**	**F**	**P(>F)**
Temperature	4.748e+07	1	122	1.71e-28
Dew Point	2.293e+06	1	5	1.49e-02
Humidity	6.321e+06	1	16	5.34e-05
Wind Speed	7.405e+07	1	191	1.74e-43
Wind Gust	7.934e+07	1	204	1.83e-46
Pressure	2.572e+06	1	6	9.96e-03
Precipitation	1.824e+05	1	0.4	4.92e-01
Residual	1.772e+11	457617		

Table 1: ANOVA Table of weather variables against trip durations

For both models, categorical features were extracted, indexed, and encoded as SparseVectors while numerical features were standardized after being split into train, validation, and testing sets. The test set comprised of the first four months of 2022, the validation set from the last four months of 2021, and the rest was training. And while fitting the models, we assume independence between columns. But particularly for the gradient-boosted tree, with the increase in depth we also assume an increase in interaction between features.

3.1 Classification

A Multi-Layer Perceptron Classifier is a Neural Network consisting of a defined number of layers and nodes for each layer. Each node is a linear combination of the features from the previous layer after which the node is subjected to a sigmoid function being passed onto the next layer. For the problem, a shallow neural network with 100 nodes in 1 hidden layer was used. The model took 166 features as a SparseVector and metrics such as Precision, Recall, and Accuracy were used to evaluate the model. The model also trained over 50 epochs using a batch size of 128 with a defined seed.

The model performed fairly average with precision, recall, and accuracy of 58%, 49%, and 49% respectively. The model also overfits as training and validation accuracies were around 60% but test accuracies were less.

3.2 Regression

The Regression model used a Gradient Boosted Tree Regressor which is a flexible non-parametric statistical learning technique. In a simple context, it is used particularly in curve fitting and ensuring the curve is properly fit using Gradient Descent. Again, the model took 166 features as a SparseVector and fitted values with a max tree depth of 8 over 20 epochs. Metrics such as RMSE (Root Mean Squared Error) and R2 score were used to evaluate the model.

This model performed slightly better with regards to generalizing to the test data and not overfitting. The RMSE for the test data is 569, and the variance of the whole model is 644 which meant the R2 score was 0.016 which meant that there is still a lot more information to explain what the model is missing out on.

	0	1
0	674572	17970
1	723832	37975
	1398404	55945

Table 2: Confusion Matrix for classification

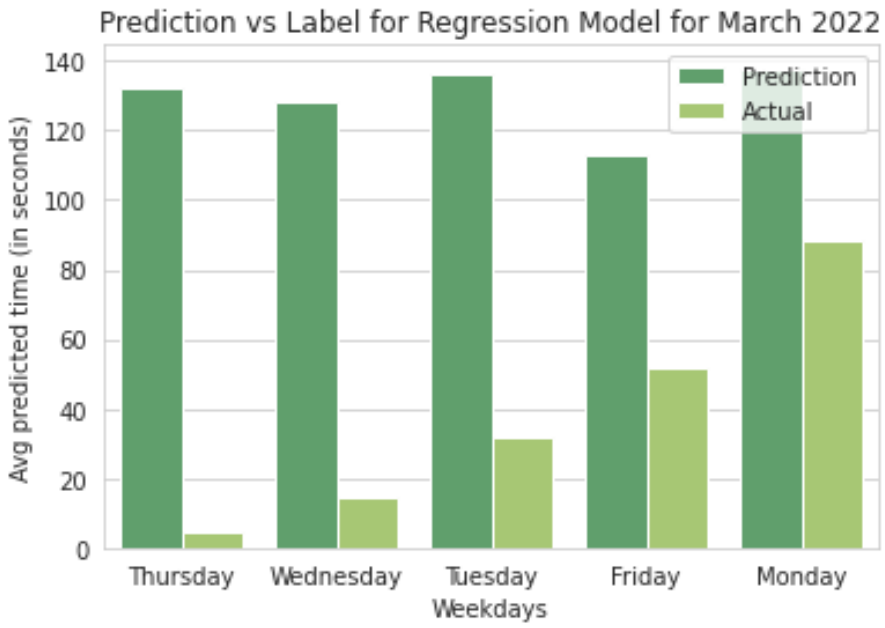

Figure 4: Average error (deviation from actual) for the weekdays in March

3.3 Error Analysis

From Figure 4, it can be seen the model is grossly underpredicting (since the values are actually negative) for all days in March showing large amounts of bias in the data. There is also little variance among the days of the weeks, but this can be expected from Linear Regression models.

On the other hand, in Table 2 we see that it got a lot of the true positives right but if we look closer, most of its predictions are 0 anyway. This is also an induced bias in the model and we can now see clearly that it is because of the large number of true values and very few false values. The skewed data has caused the model to not predict as accurately as it should, and should be something to look into further.

A trend that was missed earlier but highlighted in Figure 4 is that Mondays and Fridays tend to be the days with the longest usage. It could also have a higher count which would give more explanation and meaning to the dataset chosen.

4 Recommendations

As a person asking themselves on a weekday, which mode should I take to travel? Taxis seem to be faster on average. But if you further break it down, if one were to ask the question in the mornings (especially during peak hours in the morning) it is recommended to take the bike over the taxi for short distances. Otherwise, during nighttime and other periods, it is advisable to take a taxi for the quickest results.

Another recommendation is that if the person knows that they are passing through areas like Lennox Hill and Midtown East, it is preferable to ride a bike since these areas are the most bike-friendly which would allow one to have the quickest speeds when in the area.

5 Conclusion

This report focused on predicting trip durations and looked to model a difference in duration between taxis and bikes in an attempt to understand a post-COVID trend. It is without a doubt that the number of riders has been increasing and this could be an additional tool to help decide the best way to go forward every weekday. With over a thousand new stations and expansion plans to boroughs as well, the times become shorter and it only becomes easier to pick up and bike and travel wherever you want.

It is clear that both models are biased and need more fine-tuning and data cleansing before they should be used in a real-world scenario. And to add on, further study can also include bringing in costs and computing "the value" of the option rather than just the speed and time of transportation. Also, another point of interest would be to model data and trends on the weekend to give people a more meaningful travel experience when going out.

References

[1] Ilana Strauss. *Is the U.S. becoming more bike friendly?* https://www.nationalgeographic.com/environment/article/is-the-us-becoming-more-bike-friendly. Accessed: 2022-08-15.

[2] . *Citibike Information Page*. https://citibikenyc.com/. Accessed: 2022-08-09.

[3] U.S. Department of Transportation. *Bikeshare and E-scooter Systems in the U.S.* https://data.bts.gov/stories/s/Bikeshare-and-e-scooters-in-the-U-S-/fwcs-jprj/. Accessed: 2022-08-10.

[4] Sam Schwartz Data Analytics Team: Yana Chudnaya, PE, Matthew Dwyer, and Daniel Schack, AICP. *Shifting Gears: Citi Bike Recovers Faster Than Subway*. https://www.samschwartz.com/staff-reflections/2020/shifting-into-gear-citi-bike-overtakes-subway-in-ridership-recover. Accessed: 2022-08-10.

[5] Governor's Press Office. *Governor Cuomo Announces New York Ending COVID-19 State Disaster Emergency on June 24*. https://www.governor.ny.gov/news/governor-cuomo-announces-new-york-ending-covid-19-state-disaster-emergency-june-24#:~:text=Governor%20Cuomo%20Announces%20New%20York,June%2024%20%7C%20Governor%20Kathy%20Hochul. Accessed: 2022-08-12.

[6] NYC Taxi & Limousine Commission. *NYC Taxi Data*. https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page. Accessed: 2022-08-09.

[7] . *Citibike Data*. https://ride.citibikenyc.com/system-data. Accessed: 2022-08-09.

[8] Bojan Stavrikj. *Web Scraping Wunderground Weather History With Python - Fahrenheit*. https://bojanstavrikj.github.io/content/page1/wunderground_scraper. Accessed: 2022-08-14.

[9] . *CitiBike expansion plan*. https://ride.citibikenyc.com/blog/major-citi-bike-expansion-map-revealed. Accessed: 2022-08-13.

[10] Hugh Nash. *Citibike - Facts and Range*. https://fahrradshop24.org/bike/how-do-citi-bike-electric-bikes-charge.html. Accessed: 2022-08-12.

[11] . *2021 Weather History in New York City*. https://weatherspark.com/h/y/23912/2021/Historical-Weather-during-2021-in-New-York-City-New-York-United-States#Figures-Summary. Accessed: 2022-08-18.

[12] Mariela Quintana. *Top 5 NYC Neighborhoods for Bikers*. https://streeteasy.com/blog/top-nyc-neighborhoods-for-bikers/. Accessed: 2022-08-24.

[13] Mayor's Office. *Declaration of Local State of Emergency: Hurricane Emergency Declaration*. https://www1.nyc.gov/office-of-the-mayor/news/227-001/emergency-executive-order-227. Accessed: 2022-08-22.