Recommendations for Rideshare Drivers in NYC

Lachlan Macartney, repo (embargoed)

September 25, 2022

1 Introduction

The intended audience of this report are those considering (or already) working for a ride sharing service such as Uber or Lyft to supplement their income. This type of driver has flexibility in their schedule and aims to maximise their income per hour of work. The ridesharing industry is notoriously opaque, by analysing this data we aim to provide transparency to the benefit of drivers. The intention is to provide actionable advice. For this reason, model explainability is to be prioritised. The advice should be relevant today, consequently the most recent 6 months (11/2021-4/2022) of New York City Taxi and Limousine Commission (TLC) data [1] was chosen for analysis. The New York Department of Transport (DOT) provides granular traffic data for over 140 links in the city [2], this data was also considered for the same period.

The attributes studied included:

- The times a trip was requested, commenced and finished
- The area a trip commenced and finished in
- The trip distance
- The time taken to cross traffic links
- The ridesharing company (ultimately Uber or Lyft)

Approximately 110,000,000 trips occurred during this period. The data was filtered for outliers and sources of potential bias. Imputation was used on the traffic dataset. The profitability of a trip was estimated by performing feature engineering and making some assumptions. Analysis was conducted that looked at the impact of different features on this profitability metric. The distribution of features was examined. The data was trained on a Generalised Linear Model (GLM) and a regression tree - interpretable models. The performance of these models is briefly analysed and recommendations are made for rideshare drivers.

2 Preprocessing, Analysis and Geospatial Visualisation

2.1 Preprocessing of TLC Data

- 11,000,000 (11.35%) Trips with extreme distances, time taken or driver pay were removed
- 16,000 (0.02%) Shared trips were removed, this included Uber Pool and Lyft Shared

2.2 Analysis of TLC Data

Exploratory feature extraction was done to get a feel for the data. Table 1 makes note of an interesting observation.

Service	Average Commision	Average Pickup Time
Lyft	\$5.05	5m 46s
Uber	\$4.28	4m 54s

Table 1: exploratory analysis

In Table 1 we observe that, on average Lyft takes longer to connect drivers with riders and takes more commission than Uber. Commission was calculated as the difference between the base fare and the driver's pay.

The goal of this report is to provide recommendations to rideshare drivers wanting to maximise their income per unit of hour work. Thus, it was necessary to engineer a feature that captured this.

The following assumptions were made:

1. Vehicle depreciation, maintenance and operation costs are 58.5 cents per mile[1]
2. Vehicle travels at the same average speed during pickup and trip
3. Drivers have already sunk the fixed costs of vehicle ownership (garaging, insurance, registration etc)[2]
4. Drivers immediately accept a pickup request upon completing a trip[3]

The hourly profitability for a trip was estimated with:

$$\text{hourly profit}(\$/\text{h}) = \frac{\text{driver pay}(\$) - 0.585(\$/\text{mile}) * \text{estimated total distance(miles)}}{(\text{trip time} + \text{pickup time})(\text{h})} \tag{1}$$

$$\text{where, estimated total distance} = (\text{trip time} + \text{pickup time}) * \frac{\text{trip distance}}{\text{trip time}}$$

It was decided not to include tips in the driver's pay. The reason being is that tips are often paid with cash and thus missing from the data. Consequently we felt it would be misleading to include in-app tips as part of the driver's pay, but omit cash tips.

As a sanity check for our profitability metric, we confirm in Figure 1 that Lyft is on average, less profitable for drivers than Uber.

In Figure 2 we note that drivers tend to earn more early in the morning, particularly on the weekends. However, they earn similar amounts on weekends and weekdays overall. The early morning and weekend variables interact here.

In Figure 3 we see that trips going to or from an airport are more profitable.

[1] Based on the tax deductible mileage rate published by the Internal Revenue Service for 2022[3]

[2] The target audience would keep their vehicle, whether they work for a rideshare service or not

[3] Problematic, discussed in conclusion

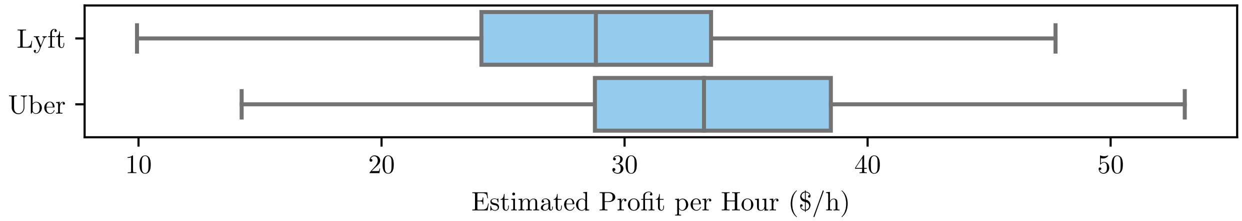


Figure 1: Profitability by Service

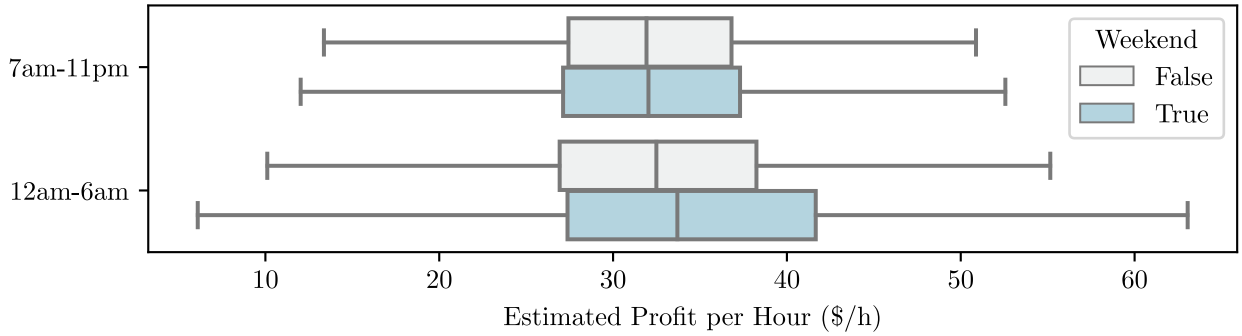


Figure 2: Profitability by Weekend, Early Mornings

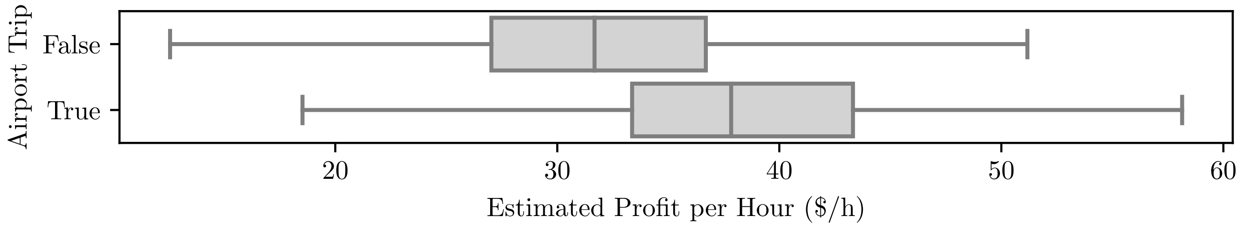
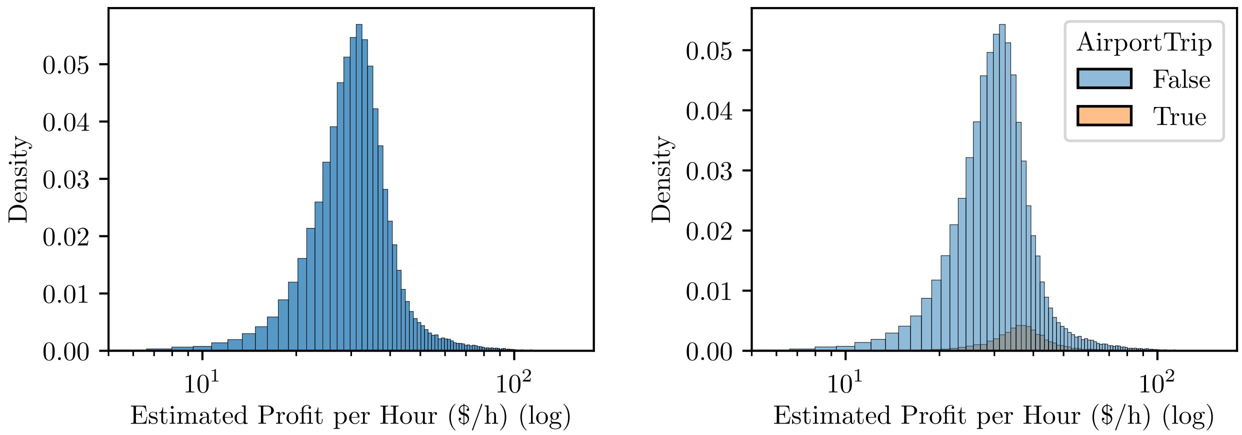


Figure 3: Profitability by Airport Trip



(a) overall distribution

(b) airport trips highlighted

Figure 4: log-normal distribution of profitability

In Figure 4(a) we note that our profitability appears to be normally distributed upon taking a logarithmic transformation. It should be noted that this distribution was skewed before the proper removal outliers. Figure 4(b) shows that the distribution of profitability after filtering for airport trips likewise appears log-normal. The same was confirmed for the features examined in Figures 1, 2. Plots in Figures 1, 2, 3 & 4 were 0.1% samples but confirmed representative by repeated sampling.

Figure 5 shows the New York Metropolitan area grouped into four classes of profitability. These classes of interest contain the following areas. Newark International Aiport (New Jersey), is the yellow island in Figure 5(a). The two other isolated yellow regions are the airports shown on 5(b). Inner Manhattan and the airports are clearly where drivers should aim to be. Figure 5(a) was based on trips departing from the area, but it was confirmed to be a similar map when trips arriving in the area was plotted instead.

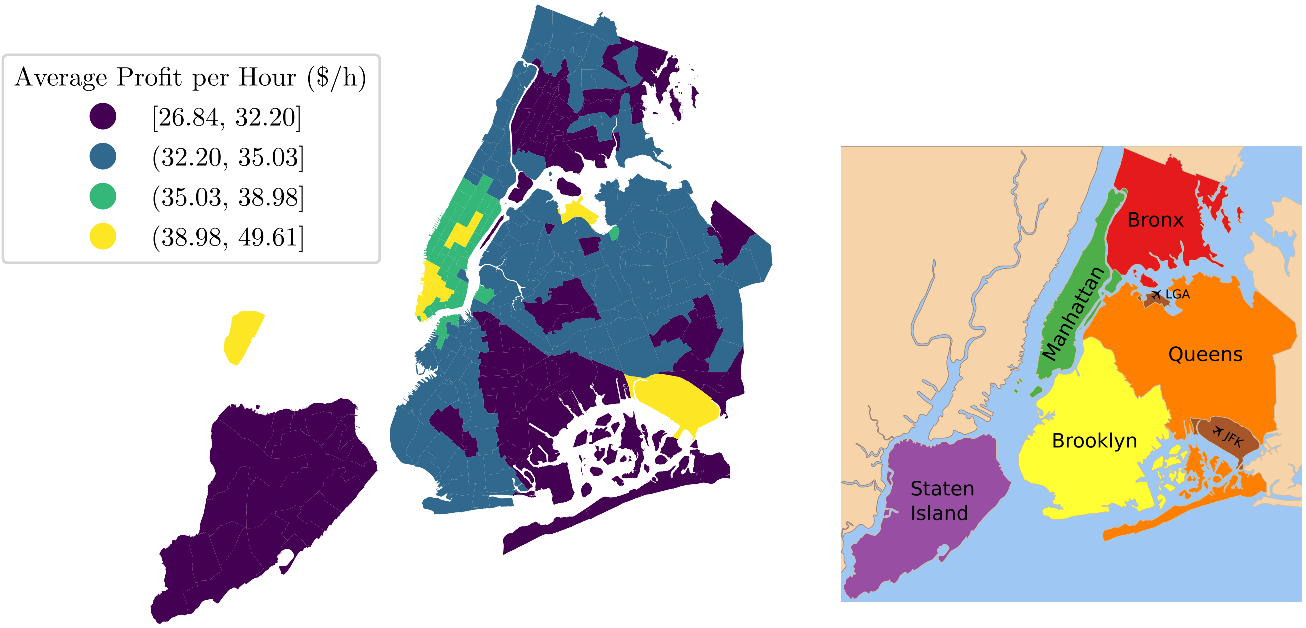
2.3 Preprocessing of DOT Data

The DOT Data was a series of observations on the time taken for a vehicle to traverse a traffic link (point to point) at a given time. Each observation was given a link ID which could be mapped to geographic coordinates. The data appeared to be of poor quality, with many links not reporting consistently.

- 26/143 links either didn't report any data or reported data intermittently, these were removed
- Of the remaining observations, 5.65% were missing data, the majority of these were imputed using linear interpolation
- 0.007% of observations were dropped because imputation failed

This congestion data was aggregated hourly, by the borough and summated. The physical meaning of this statistic is the time it would take to transverse every link in the borough each time it was reported in the hour. This statistic appeared to be log-normal. After taking logarithims and standardising we obtained the histograms in Figure 7.

The decision to aggregate by the borough (rather than be more precise) was for simplicity and ultimate model explainability.



(a) profitability map

(b) borough reference[4]

Figure 5: Geospatial Profitability

2.4 Analysis of DOT Data

The motivation for choosing traffic data is that it would capture information missing from other attributes, for example a protest or roadworks in the area.

Figure 6 is a reassuring sanity check for our congestion rating, congestion is the least during early hours of the morning and rises to a peak during the afternoon rush.

Figure 7 shows that the congestion rating is reasonably normally distributed. Before adequate preprocessing, these histograms were skewed.

3 Modelling

3.1 Generalised Linear Model

Our analysis so far had shown the hourly profit to be log-normally distributed (Figure 5). This suggests the use of a GLM, with a Gaussian Family and log link function. In mathematics this is:

$$\begin{aligned} \boldsymbol{y} = e^{X\boldsymbol{\beta}+\epsilon}, \text{ where } X &= \text{feature matrix} \\ \boldsymbol{y} &= \text{hourly profit} \\ \boldsymbol{\beta} &= \text{parameters} \\ \epsilon &\sim N(0, \sigma^2) \end{aligned}$$

The X matrix contained the following features:

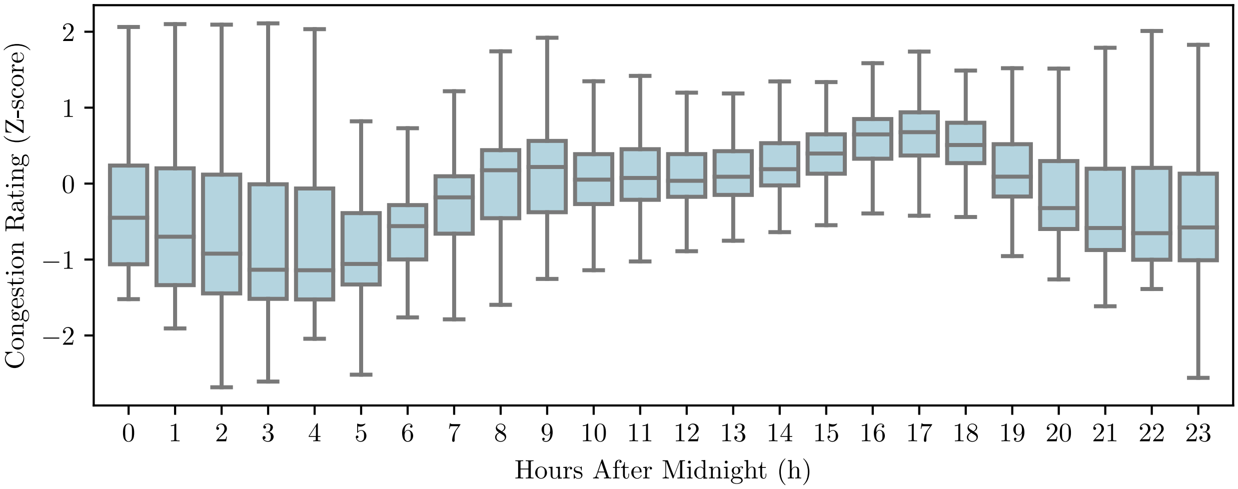
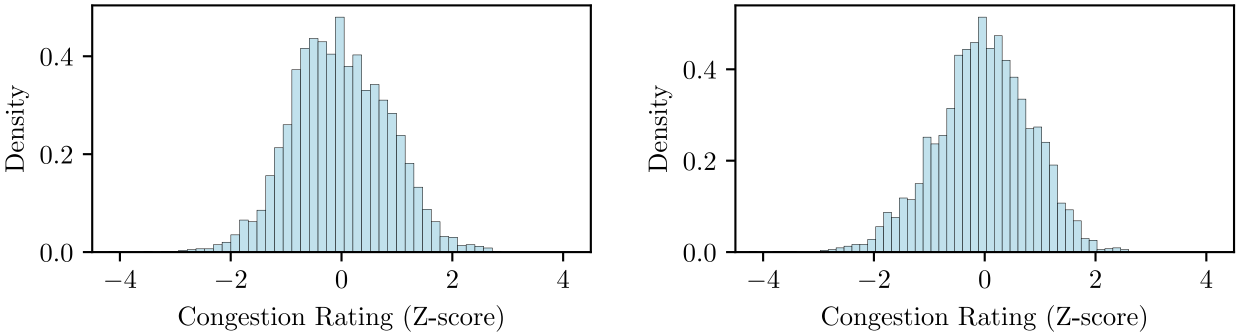


Figure 6: Congestion Rating by Hour in Brooklyn



(a) Queens

(b) Manhattan

Figure 7: Distribution of Congestion Rating

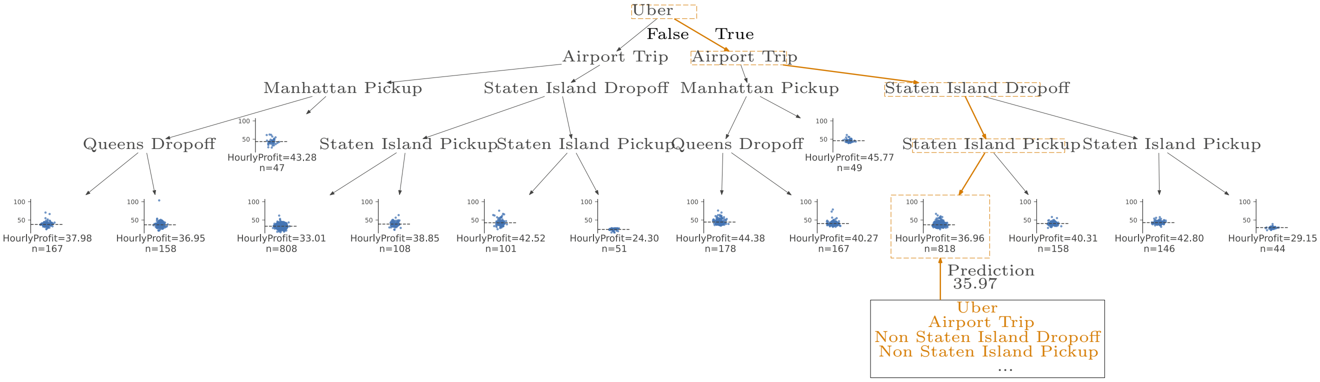


Figure 8: Regression Tree Diagram

- Day (category)
- Hour (category)
- Rideshare type (boolean)
- Airport trip (boolean)
- Pickup borough (category)
- Dropoff borough (category)
- Pickup congestion rating (continuous)
- Dropoff congestion rating (continuous)

This model was trained on data aggregated by features of the X matrix.

3.2 Regression Tree

Care was taken to encode the binary variables (rideshare service, airport trip) properly. The X matrix and response y were unchanged. Figure 8 Shows the regression tree run on a sub sample of the test data.

3.3 Comparison

A benefit of the regression tree is it is highly interpretable, even for someone with a limited knowledge of statistics - in contrast to the GLM. An average ride share driver could reasonably print off Figure 8 and refer to it while driving to increase their earnings.

Drawing our attention to Table 2, we see some of the most significant parameters are expected from our prior analysis, such as the whether it was an Uber, Airport Trip or on the weekend. However, the value of this model is that it reveals hidden insight which is highly implementable. For example, suppose we have a trip from Manhattan to Staten Island, summing up the corresponding parameter coefficients $0.1535 + 0.4156 - 0.1542 \approx 0.42$ - this suggests, this trip is likely to be more profitable than average.[4]

Notably, the congestion scores had a slight positive effect on earnings ($p < 10^{-3}$) - this can can be interpreted as it it being more profitable to drive when it's more congested than usual in the area.

A *important caveat* of the linear model is that these parameter estimates apply only in the *presence of all parameters*. So technically, every parameter of the trip must be taken into accounted for - it is however useful in a loose sense.

[4] See caveat ahead

parameter	p-value	coeficient
Intercept	$< 10^{-3}$	3.3148
Uber	$< 10^{-3}$	0.1065
Airport Trip	$< 10^{-3}$	0.1077
Monday	0.002	0.0030
Saturday	$< 10^{-3}$	0.0188
Sunday	$< 10^{-3}$	0.0170
Thursday	$< 10^{-3}$	0.0112
Tuesday	$< 10^{-3}$	0.0080
Wednesday	$< 10^{-3}$	0.0085
Brooklyn Pickup	$< 10^{-3}$	0.3369
Manhattan Pickup	$< 10^{-3}$	0.1535
Queens Pickup	$< 10^{-3}$	0.2408
Staten Island Pickup	$< 10^{-3}$	0.3883
Brooklyn Dropoff	$< 10^{-3}$	0.3229
Manhattan Dropoff	$< 10^{-3}$	0.1142
Queens Dropoff	$< 10^{-3}$	0.2259
Staten Island Dropoff	$< 10^{-3}$	0.4156
Brooklyn Pickup and Brooklyn Dropoff	$< 10^{-3}$	-0.6188
Manhattan Pickup and Brooklyn Dropoff	$< 10^{-3}$	-0.2407
Queens Pickup and Brooklyn Dropoff	$< 10^{-3}$	-0.3917
Staten Island Pickup and Brooklyn Dropoff	$< 10^{-3}$	-0.5201
Brooklyn Pickup and Manhattan Dropoff	$< 10^{-3}$	-0.2491
Manhattan Pickup and Manhattan Dropoff	$< 10^{-3}$	-0.0770
Queens Pickup and Manhattan Dropoff	$< 10^{-3}$	-0.1261
Staten Island Pickup and Manhattan Dropoff	$< 10^{-3}$	-0.1818
Brooklyn Pickup and Queens Dropoff	$< 10^{-3}$	-0.3819
Manhattan Pickup and Queens Dropoff	$< 10^{-3}$	-0.1087
Queens Pickup and Queens Dropoff	$< 10^{-3}$	-0.4152
Staten Island Pickup and Queens Dropoff	$< 10^{-3}$	-0.2961
Brooklyn Pickup and Staten Island Dropoff	$< 10^{-3}$	-0.4752
Manhattan Pickup and Staten Island Dropoff	$< 10^{-3}$	-0.1542
Queens Pickup and Staten Island Dropoff	$< 10^{-3}$	-0.2916
Staten Island Pickup and Staten Island Dropoff	$< 10^{-3}$	-0.8953
Pickup Congestion (z-score)	$< 10^{-3}$	0.0023
Dropoff Congestion (z-score)	$< 10^{-3}$	0.0042

Table 2: GLM Parameter Estimates

In Table 3 we can compare the mean absolute and squared error on unseen future test data. November through to March was used for training, April was used for evaluating these statistics. The GLM performed better overall than the regression tree. The choice of features and hyperparameters was done heuristically to meet the goals of explainability and to uncover new, actionable insight. For example we have already shown the interaction of weekend and early morning in Figure 2, so it was not necessary to complicate the presented models with these additional parameters. However, it was verified that interaction does occur. If we were concerned about performance, aggregating congestion data by the more precise zones in the TLC dataset would likely result in far greater performance. The takeaway from this section is the additional insight regarding impact of congestion and pickup and drop off boroughs that was not obvious in the analysis section.

4 Recommendations

The ridesharing industry is opaque for drivers, and intentionally so - as it stands to benefit the oligopolies which control it. However, given this data we are given a rare glimpse into the ridesharing universe and can make evidenced based recommendations for drivers. Based on the assumptions, analysis and modelling in this report we would recommend drivers in New York:

- Work for Uber, not Lyft
- Prioritise trips to airports
- Trips within Manhattan are to be prioritised
- Drive between 12-6 am, as it has the highest earning potential, particularly on weekends

It should be emphasised that these recommendations only apply under assumptions 1,2,3 & 4. The assumption that drivers immediately accept a job upon completing a previous one is problematic and difficult to address given the dataset. To get a more realistic picture of driver income we would need to factor in the time spent waiting for a trip and driving home at the end of the day.

Model	Mean Absolute Error ($/h)	Mean Square Error ($/h)^2	R^2
GLM	2.86	21.60	0.749
RT	3.85	30.87	0.321

Table 3: Model Evaluation

References

[1] New York City Taxi and Limousine Commission. *TLC trip record data*. `https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page`. Accessed: 2022-08-20.

[2] City of New York Department of Transportation. *Archived Real Time Traffic Speed Data*. `https://data.beta.nyc/dataset/nyc-real-time-traffic-speed-data-feed-archived`. Accessed: 2022-08-20.

[3] Internal Revenue Service. *IRS issues standard mileage rates for 2022*. `https://www.irs.gov/newsroom/irs-issues-standard-mileage-rates-for-2022`. Accessed: 2022-08-20.

[4] Julius Schorzman. *5 Boroughs of New York City*. `https://commons.wikimedia.org/wiki/File:5_Boroughs_Labels_New_York_City_Map.svg`. Accessed: 2022-08-20.