Determining Driver Tips for Rideshare Services in NYC

Manish Khilari
Student ID: 1173214
Github repo with commit

August 21, 2022

1 Introduction

With the explosion in demand for rideshare services over the past half decade, rideshare company revenue has exploded in parallel, and strategies to increase this revenue have become increasingly valuable. The rideshare market is projected to grow from USD 85.8 billion in 2021 to USD 185.1 billion by 2026. growing at a compounding annual rate of 16.6% within the 5 year forecast period. [1]

This rapid growth has resulted in fierce competition within the rideshare industry. As noted by The Economist [2], this growth has created an industry where driver tips are just as crucial, if not more crucial, than the base passenger fare.

In this report, we investigate the base passenger fare paid by rideshare customers in New York City (NYC), as well as driver tips as a proportion of this fare. We also build two statistical models, a Linear Model with Ordinary Least Squares and a Logistic Model, to predict both the base passenger fare and tip amounts using the selected input features.

We focus on only the For Hire Services (FHS) classified as High Volume (HV) by the NYC Taxi and Limousine Commission (TLC) [3], rather than a focus on all rideshare companies. This High Volume subset includes only FHV licenced companies that currently exceed, or plan to exceed, 10 000 trips per day in NYC. Within February to December 2019, the 4 HVFHV licenced companies (Juno, Uber, Via, and Lyft) achieved a collective total of over 23 million recorded trips (an average of over 63 000 per day).

2 Dataset

The first dataset used was FHS trip data published by the NYCTLC, recording features of each FHV trip taken within the city. This data included location information for both pickups and drop offs, enabling geospatial visualisation of fares and tip amounts.

We used only the High Volume subset of this data, rather than all FHV trips. In addition to the large volume of trips, this gave us access to more extensive records, which included detailed fare breakdowns rather than just a total trip cost. This further gives us the ability to provide targeted recommendations the top 4 rideshare companies by number of annual NYC trips.

We also used weather data published by the US National Center for Environmental Information (NCEI) in their openly available Integrated Surface Dataset [4]. This was because weather was expected to influence demand for rideshare services, with good weather expected to increase taxi demand as well as the length of each trip, and therefore, increase passenger fares.

We used only the subset of the weather data recorded at JFK Airport, rather than the weather specific to each trip pick up and drop off location. This was because of the extensive nature of airport weather data, providing hourly reports of temperature, precipitation, and visibility.

While weather at JFK Airport may not be representative of weather at all locations in NYC, since one location may have vastly different weather to another, this data gave a reasonable estimate of NYC wide conditions at the time.

2.1 Range Selection

Given the focus on only HVFHV data, the date range selected was February 2019 onwards. This was because HVFHV data was not present before and including January 2019.

As of 1 February 2019, FHV companies categorised as High Volume were required by NYC law to report more extensive records, ensuring a fair income for their drivers [5]. These more extensive records introduced tip amounts to each trip, a feature not present in earlier data. While earlier trips may not have necessarily had tips of size 0, the tip amount was unknown, making data from January 2019 or earlier irrelevant for tip prediction.

The date range selected for prediction was February to December 2021. This was because of the impact of the 2020 coronavirus outbreak in NYC, limiting the rideshare industry with lockdowns and the shift to working from home rather than commuting.

While the coronavirus outbreak may have had a permanent impact on rideshare demand and tip revenue, the more recent conditions seen in 2021 and 2022, as compared to conditions seen at the height of the outbreak in 2020, were more representative of the pre coronavirus conditions of 2019. This justified skipping 2020 data in favour of 2021 and 2022.

2.2 Preprocessing

While both the HVFHV and weather data published by the TLC and NCEI respectively were well structured, rather than unstructured, these datasets were not validated at the time of recording.

As a result, both datasets contained outliers, as well as instances outside the valid range. These invalid instances were removed to ensure a well founded analysis.

2.3 Feature Selection

We selected a subset of the provided features that had both a high proportion of valid values (with a low proportion of outliers or invalid instances), as well as a high relevance to the response. In this case, our responses were rideshare passenger fares, which meant selecting features relevant to rideshare service demand (time of day, temperature), rather than those that were irrelevant (airport runway status).

We also aimed to minimise redundant or highly correlated inputs, such as having both the total fare and the sum of the passenger payment breakdown as predictors in the same model, since these two values are expected to be equal.

There was a strong positive correlation ($\rho_{basePassengerFare,\ tipAmount} = 0.65$) between base passenger fare and tip amount. This was expected, as passengers are likely to pay a higher tip for a longer, and therefore, more costly trip. Passengers who can afford more costly trips are also more likely to be able to afford, and therefore, provide higher tip amounts.

To account for this correlation, we used tip amount as a proportion of base passenger fare, rather than tip amount as an absolute value, providing a better estimate of customer satisfaction.

While the weather data provided wind as a velocity vector, with both direction and speed, only wind speed was retained. Compared to speed, it was unreasonable to take wind direction as a significant predictor of rideshare fares.

The following features were selected as responses.

- Base Passenger Fare
- Tip Amount
- Tip Amount as a Proportion of Base Passenger Fare

The following features were selected as relevant predictors for the above responses.

- Date and Time
- Day of Week
- Pick Up Location ID
- Drop Off Location ID
- Request to Pick Up Wait Time
- Trip Time

Request to pickup wait time was considered to be a significant predictor of tip amounts, since an increased wait time is expected to reduce customer satisfaction, correlating with a reduced tip amount.

The rideshare company of a trip, which was represented by the TLC as an alphanumeric licence number (with HV0003 representing Uber), was cast to the actual company name for increased readability.

To ensure response validity, categorical predictors with k distinct values were represented by k distinct treatment contrasts, rather than the single parameter used for continuous predictors. In the case of the hour of day, each of the 24 hours were represented by categories rather than an ordinal value. This was because it was unreasonable to expect a (24 hour) time of day of 1200 to have double the weight of a time of 600 in predicting the fare.

The input features did not include year, allowing the trained models to generalise to future years.

2.4 Outlier Detection

Trip data timestamps were assumed to be valid and within the month and year specified in their filename.

Trips with a request time before the pick up time were considered invalid, with 367,578 of these trips removed. These removals also ensured a positive request to pick up wait time.

Trips with a trip time either negative or greater than 5 hours were considered invalid, with 4,567 of these trips removed. With the NYC CBD being less than 50km in diameter, it was unreasonable to expect a trip fully contained in the city to be greater than 5 hours.

Trips with a pick up time after the drop off time were also considered invalid. While HVFHV trips specified a trip time in addition to pick up and drop off times, a feature not present in non high volume FHV trips, a positive trip time was not always indicative of a pick up time before the given drop off time. Fortunately, after request time validation, there were 0 of these trips detected.

Trips with a pick up or drop off location ID outside of the valid range specified by the TLC (1 to 263 inclusive) were unable to be located, and therefore considered invalid, with 7,498,902 of these trips removed.

Trips with a negative base passenger fare were considered invalid, with 584,678 of these trips removed. While these negative fares may represent refunded trips, these were unreasonable to include as predictors of future fares, and were removed.

Trips with a negative tip amount were also considered invalid. Fortunately, after base passenger fare validation, there were 0 of these trips detected.

2.5 Imputation

After the removal of invalid trips, imputation was considered unnecessary. While there were trips present with missing values, these trips were removed completely, rather than kept with an imputed value. This was because the final number of valid trips was considered sufficiently large, without the need for the additional trips provided by imputation.

2.6 Weather Join

We joined the weather data with the trip data, giving the weather conditions for each trip. We performed an inner join on the combination of pickup hour, day of month, month, and year, providing each trip with JFK airport specific weather conditions at the relevant hour.

Weather conditions were determined only for trip pick up time, rather than both pick up and drop off times. While the weather may change between a pick up and drop off, with the chance of changed weather conditions increasing as trip time increases, changes are unlikely, and if a change occurs, is unlikely to be severe. Weather conditions at the time of pick up were selected, and gave a reasonable estimate for conditions over the full duration of the trip.

The following weather features were selected.

- Wind Speed
- Temperature
- Dew Point
- Visibility

There was no need for manual removal of weather instances with invalid timestamps. These were automatically removed following the inner join with the already validated trip data.

3 Analysis

Statistics beyond this point are generated using a random sample of the population. While a random sample is not a perfect representation of the population, analysing a sample rather than the entire population has significantly reduced runtime.

3.1 Location

The location of a trip, both at pick up and drop off, is expected to be a strong predictor of rideshare passenger fares. This is because location is a strong indicator of socioeconomic status, as well as the industry in which passengers work, in the case that the trip is a commute to work. Both these passenger attributes are expected to contribute heavily to the amount the passenger tips.

The average base passenger fare and tip amount by location is shown below.

Note: The heatmap scales have been excluded for greater map zoom, and can be found in the original map images.

Figure 1: Average Base Passenger Fare by Pick Up Location (February to December 2019)

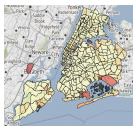

Figure 2: Average Base Passenger Fare by Drop Off Location (February to December 2019)

Figure 3: Average Tip Amount by Pick Up Location (February to December 2019)

Figure 4: Average Tip Amount by Drop Off Location (February to December 2019)

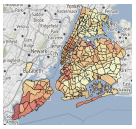

Figure 5: Average Tip Amount as a Proportion of Base Passenger Fare by Pick Up Location (February to December 2019)

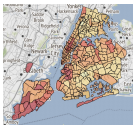

Figure 6: Average Tip Amount as a Proportion of Base Passenger Fare by Drop Off Location (February to December 2019)

As expected, pickup and drop off locations are strong predictors of rideshare passenger fares, with the 2019 data showing an average tip amount for pick ups from JFK Airport (coloured red) exceeding $50. This is a fare amount over 3 times greater than the nearby non airport locations shown.

Another observation is the strong correlation between base passenger fare and tip amount in several locations. To account for this correlation, the ratio of these two responses by location is also shown.

3.2 Day of Week

Day of week was also expected to be a strong predictor of passenger fares, with demand for ridehare services fluctuating across weekdays and weekends.

The average base passenger fare and tip amount by day of week is shown below.

3.3 Weather

Given the direct impact of weather on driving conditions, as well as the expected positive correlation between good weather and both trip length and passenger fares, we expected weather to be a good predictor of these fares.

The relevance of the selected weather features is shown below.

3.4 Pearson Correlation

The Pearson correlation between the two continuous predictors base passenger fare and tip amount was determined.

While a Pearson correlation value is present for each pair of relevant predictors, this matrix of values over the full population was not computed. This was because the time to compute the matrix over the 2019 specific data exceeded 1 day.

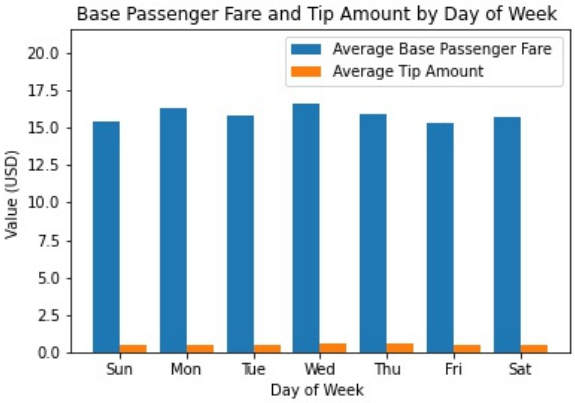

Figure 7: Average Base Passenger Fare in Comparison to Average Tip Amount by Day of Week (February to December 2019)

	SS	DF	F	P(F >f)
Wind Speed				
Temperature				
Dew Point				
Visibility				
Residual				

Table 1: ANOVA for Weather Predictors given the Reduced Model (with $P(F > f) < 0.05$ indicating predictor significance)
(This table was not generated due to time constraints)

4 Linear Model

The first model selected was the Linear Model with Ordinary Least Squares. While it was possible to use parameter penalisations such as Least Absolute Selection and Shrinkage Operator (LASSO) and Ridge, these penalisations were considered unnecessary for the relatively small number of parameters used. The link chosen was the identity, because it was the canonical link for the normal distribution.

Two distinct Linear Models were used to predict tip amount and base passenger fare.

A limitation of the Linear Model is the assumption of normally distributed errors with constant variance (homoscedasticity). This was a reasonable assumption, since the Central Limit Theorem ensures that the mean of an increasing number of independent and identically distributed random variables approaches a normal distribution. In this case, the response was average base passenger fare, which was the mean of a reasonably large number of fare amounts.

The Linear Model used is specified below.

$$(Y_{tipAmount}, Y_{basePassengerFare}) \sim N(\mu, \sigma^2) \tag{1}$$

$$g(\mu) = k^T \beta \tag{2}$$

$$\mu = k^T \begin{bmatrix} \beta_0 \alpha_{dayOfWeek} \\ \alpha_{dayOfMonth} \\ \alpha_{month} \\ \alpha_{pickUpLocationID} \\ \alpha_{dropOffLocationID} \\ \beta_{requestToPickupWaitTime} \cdot x_{requestToPickupWaitTime} \\ \beta_{tripTime} \cdot x_{tripTime} \\ \beta_{windSpeed} \cdot x_{windSpeed} \\ \beta_{temperature} \cdot x_{temperature} \\ \beta_{dewPoint} \cdot x_{dewPoint} \\ \beta_{visibility} \cdot x_{visibility} \end{bmatrix} \tag{3}$$

where α represents a categorical treatment contrast rather than an ordianl or continuous predictor.

The estimated average base passenger fare and tip amount in comparison to the true values are shown below.

5 Logistic Model

The second model selected was the Logistic Model. This provided a contrast to the Linear Model with Ordinary Least Squares. The link chosen was logit, because it was the canonical link for the more general Binomial Model.

The Logistic Model was used to predict tip amount as a proportion of base passenger fare.

A limitation of the Logistic Model is support for responses within the range [0, 1]. While it was possible for the tip amount to be greater than the base passenger fare, and therefore a tip amount as a proportion of base passenger fare to be greater than 1, these trips were considered outliers. A total of numOutliers of these outliers were detected.

Rather than removing these outliers, we chose to give trips with tip amounts greater than the base passenger a tip proportion value of 1. This ensured that the Logistic Model was given valid trips without the removal of these more generous tips.

The Logistic Model used is specified below.

$$g(p) = ln(\frac{p}{1-p}) = k^T\beta \tag{4}$$

where the parameter vector β is the same as the one found in the Linear Model.

The estimated average tip amount as a proportion of base passenger fare in comparison to the true value is shown below.

6 Discussion

Given the relevance of the selected trip and weather input features as predictors for rideshare passenger fares, both the Linear and Logistic Models can be valuable contributors to rideshare companies seeking to maximise their reputation and revenue, as they can be used to incentivise drivers to seek more well tipping and profitable trips.

Despite the significant impacts of the 2020 coronavirus outbreak on rideshare service demand, the trained models have managed to show reasonably accurate estimates of passenger fares in 2021, after these impacts. This survival of model relevance after 2020 gives strong confidence of model relevance well into the future, regardless of any unexpected changes conditions in the coming years.

While we were limited to only the rideshare companies considered High Volume by the TLC, these 4 companies comprised the vast majority of the NYC rideshare market. This tradeoff of analysing only a subset of the rideshare market, rather than the entire population including all competitors, came with the benefit of extensive and reliable input features. As a result, the predictions generated can be backed with even greater confidence, with this level of confidence being unachievable with minimally detailed or unreliable data.

We recommend that rideshare services further expand their already extensive trip data (with features such as in vehicle temperature, and average speed for each minute of the trip), while also maintaining, or even increasing, the reliability of their data through reducing the proportion of missing or invalid trip features. By increasing dataset size, as well as the number of relevant features, these rideshare companies can train more nuanced and accurate models for passenger fares in the future.

We also recommend that rideshare companies implement a validation step before reporting their data. Given that much of the outliers and invalid trip instances in the preprocessing stage were easily detected and removed, this step validation step would come at little cost. This low cost investment in data quality would benefit future analysis enormously. Given a commitment to keeping extensive and reliable trip data, we further recommend that rideshare companies invest in a Neural Network Model. A Neural Network with an itemised fare breakdown as an output layer may be able to identify complex relations between input features that are unidentifiable by Generalised Linear Models, such as the ones used here. These models for comapny income may prove valuable to rideshare comanies seeking to stay competitive in an increasingly competitive market.

7 References

[1] Markets and Markets, "The Ride Sharing Market - Forecast to 2026."

https://www.marketsandmarkets.com/Market-Reports/mobility-on-demand-market-198699113.html

2021. [Accessed 21-August-2022].

[2] The Economist, "Uber, Doordash and similar firms can't defy the laws of capitalism after all."

https://www.economist.com/business/uber-doordash-and-similar-firms-cant-defy-the-laws-of-capitalism-after-all/21806198

2021. [Accessed 21-August-2022].

[3] New York City Taxi and Limousine Commission, "TLC trip record data."

https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

2019 - 2021. [Accessed 21-August-2022].

[4] National Centers for Environmental Information, "Integrated surface dataset."

https://www.ncei.noaa.gov/access/search/data-search/global-hourly

2019 - 2021. [Accessed 21-August-2022].

[5] Daily News, "NYC to impose some of the world's toughest regulations on Uber and Lyft."

https://www.nydailynews.com/new-york/ny-uber-lyft-cap-fhv-regulations-tlc-20190612-pspo2afygje63mxm4lr55s57jq-story.html

2019. [Accessed 21-August-2022].