Using Linear Statistical Modelling to Predict New York Taxi Prices

Harry Amad
Student ID: 1082047

August 16, 2021

1 Introduction

In recent years the ride-sharing business, once dominated by taxis, has become a particularly competitive environment. In New York City in particular, trips with ride-sharing services such as Uber and Lyft have increase by 46% from 2014 to 2017.[1] It seems that the use of smart phone apps to identify the cost of services is a reason for this shift towards ride-sharing services, as this experience is 'more convenient and reliable than some other modes'. [2] If this service could be adopted by the taxi industry, expected costs could be compared with the prepaid competition.

The purpose of this report is to investigate whether a linear regression model can be developed to predict the cost of a planned taxi trip by a customer, given a desired pick up and drop off location. I have used data for both yellow and green taxi trips from the first six months of 2018 and 2019 from the NYC TLC datasets. 2018 and 2019 data was used to avoid the effects of COVID-19 introducing any confounding factors to degrade the model. The TLC data used holds over 58 million instances, each with 21 features.

As well as the TLC data, external data sources were used to investigate whether factors which may affect the traffic in NYC at a given time—namely the weather and car crashes—would influence the price of taxi trips. The weather dataset from the National Centers for Environmental Information [3] has hourly measurements of 124 variables for the 362 days investigated from LaGuardia airport, while the car crash dataset from NYC OpenData [4] has 200,000 instances of crash reports with 32 features.

2 Preprocessing

2.1 NYC TLC Dataset

For the taxi trip dataset, cleaning was conducted to remove instances with:

- `total_amount` less than \$2.50 or over \$50 (the 95th percentile of the data)
- `passenger_count` less than 0
- `trip_distance` less than 0 miles or over 20 miles (the 99th percentile of the data)

[1] https://journals.sagepub.com/doi/10.1177/0361198119835809
[2] https://journals.sagepub.com/doi/10.1177/0361198119835809
[3] https://www.ncdc.noaa.gov/cdo-web/datatools/lcd
[4] https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95

- A non-standard rate
- Payment not with cash or card
- Unknown pick up or drop off locations

This cleaning removed approximately 4 million instances, or 7% of the data. This is an acceptable amount, because not only extreme outliers were removed but trips above the 95th percentile of `total_amount`, because this model is primarily for standard trips within New York.

Then, numerous features were engineered for visualisation or modelling purposes, including:

- `trip_cost` which is the total amount without the tip amount. The tip amount was removed so the model did not attempt to predict the tip a customer would give.
- `year`
- `month`
- `dow`—the day of the week of the trip
- `tod`—time of day. Either 'Morning' (2am - 5am), 'Day' (6am - 5pm), 'Evening' (6pm - 9pm) or 'Night' (10pm - 1am)
- `PUBorough` and `DOBorough`—the borough where pick up and drop off occurred
- `approx_dist`—the approximate distance of a taxi trip, based on the straight line distance between the centres of the pick up and drop off boroughs. Instances which occur within one zone will be given an `approx_dist` of 0.5 miles, as this is the case for about 5% of instances and 0.5 is the 5th percentile of `trip_distance`
- `PU_rides_in_zone` and `DO_rides_in_zone`—the number of taxi trips occurring in the pick up or drop off zone in the same hour

2.2 Weather Dataset

For the weather dataset, each record for the hour 23:00 - 24:00 included the daily totals for that day. All rows that did not have these daily totals were removed, as only the daily weather data was used, since weather such as snow or rain may affect traffic conditions many hours after occurring. Some imputation had to be done, as in some instances instead of a number recording rainfall, snowfall or snow depth, the letter 'T' was the only record — likely indicating that some rain or snow did fall, but the measurement was not taken. These values were imputed with the mean values of the relevant attribute.

2.3 Car Crash Dataset

For the car crash dataset, instances which did not have a location specified with a latitude and longitude were dropped, because only crashes which could be mapped to specific taxi zones were useful. For the remaining instances, the feature `zone` was engineered, which, using the taxi zones shapefile from TLC, maps the specific location of the crash to the general taxi zone in which it occurred. Then, for each instance in the taxi data, `PU_crashes` and `DO_crashes` was created, which is the number of crashes in that zone in that day.

3 Preliminary Analysis

The main attribute of interest for this investigation is of course the trip cost, so before attempting any modelling its distribution must be investigated. From (Figure 1) it is clear that the `cost` variable is right-skewed, with most trips having a cost of less than \$20. The mean trip cost is around \$14.

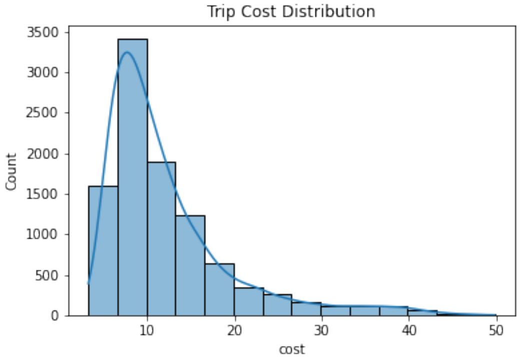

Figure 1: Histogram using Sturges' binning method with n = 10000

Before a linear model can be fit to predict `trip_cost`, it should be transformed so it is approximately normally distributed. Taking the log-transform achieves this, as (Figure 4) shows.
Given a standard rate taxi fare increases by 50 cents for every 1/5th of a mile travelled over 12mph[5], the `approx_dist` attribute will likely have high predictive power. Indeed, as (Figure 2) shows, there is a strong correlation between `cost` and `approx_dist`.

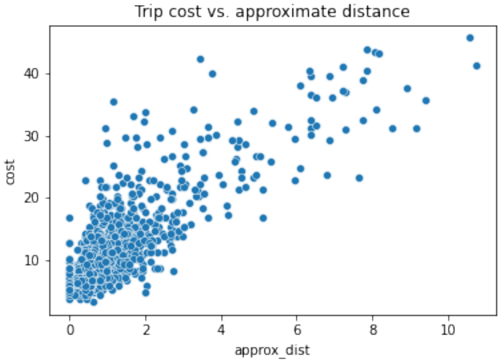

Figure 2: A clear correlation between trip cost and approximate distance

Of course, there are limitations to the `approx_dist` attribute as it is quite a crude estimate for the trip distance, especially in areas where the taxi zones are large, as their centroids may be quite a distance from the actual location of the pick up or drop off location. As (Figure 3) shows, the `approx_dist` measure consistently underestimates the actual trip distance, which is expected given it is based off straight line distance.

[5] Taxi Fare - TLC, https://www1.nyc.gov/site/tlc/passengers/taxi-fare.page

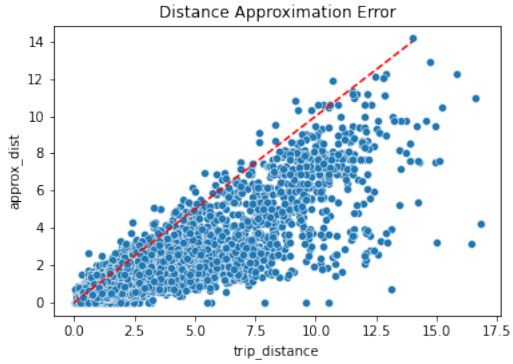

Figure 3: Distance estimation vs. actual values (the red line is y = x)

Also of interest is how prevalent taxis are in the different boroughs of New York. Manhattan dominates the taxi trip landscape, encapsulating 89.9% and 87.3% of pick up and drop off trips respectively. The remaining boroughs each take up small portions of the remaining trips, with Staten Island and Newark Airport (EWR) having by far the fewest with less that 0.1% of all pick ups and drop offs.

3.1 Attribute Relationships

Investigating the relationships between key attributes and the trip cost is very important before fitting a linear regression model, to ensure the the predictor variables are linearly related to the response variable. As (Figure 6) shows, after log-transforming `cost` and `approx_distance`, there appear to be some linear relationships of varying strengths between the relevant attributes and `cost`.

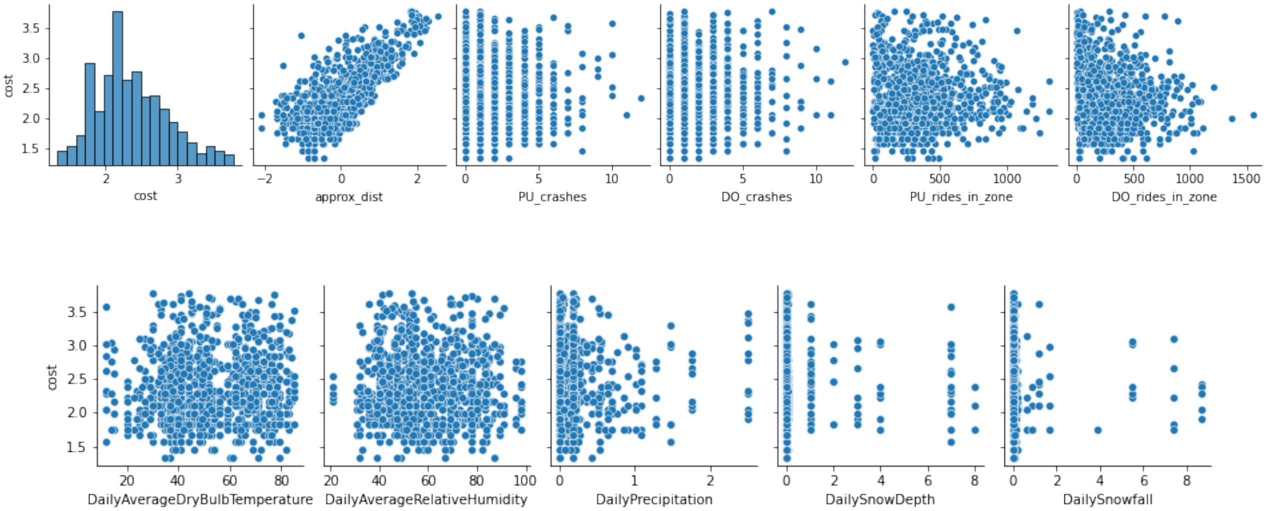

Figure 4: Attributes plotted against trip cost

3.2 Geospatial Visualisation

Using the TLC taxi zones, attributes can be explored geospatially. (Figure 5) shows how trips originating in the outer areas of New York, such as East Queens, are more expensive than those beginning in Manhattan by upwards of $20 on average. This is likely due to trips within Manhattan being very short—only 2 miles on average—while trips in the outer boroughs are usually longer because of the decreased density.

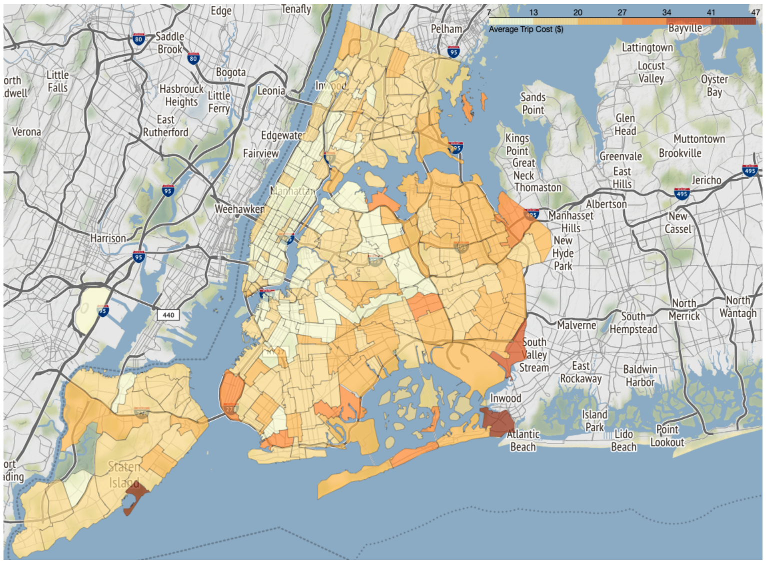

Figure 5: Average trip cost for pick ups in each taxi zone

The car crash data is also very interesting to explore geospatially. As can be seen in (Figure 6), particular spots in Manhattan have very many car crashes. These areas are at the end of bridges coming in and out of Manhattan, which would likely become bad choke points for traffic in the event of a crash. Taxi customers should be aware of this, and potentially avoid trips going through these areas as they may result in higher than expected fees due to traffic.

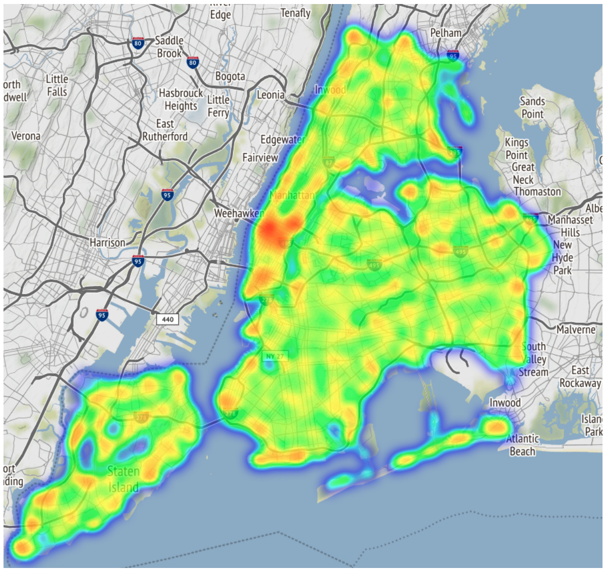

Figure 6: Density of car crashes

4 Statistical Modelling

A linear regression model was built to model the cost of taxi trips, using a sample of 10 million instances from the 2018 data as a training set and testing on samples of the 2019 data. The continuous attributes considered for use in the model were all the variables identified in (Figure 4) as having a potential linear relationship with `log(cost)`.The categorical attributes of `PUBorough`, `DOBorough`, `tod`, and `month` were also all considered.

The continuous attributes were standardised, so that any differences in scale would not affect the relative size of their parameters and potentially their significance in the model.

4.1 Model Development

To begin developing the linear regression model, a simple additive model was fit with all the attributes. All the attributes were statistically significant in this model according to the p-value of their respective t-tests except for `DailyAverageWindSpeed`, which had a p-value of 0.506.

Interaction between some attributes was then considered. Interaction between the attribute `approx_dist` and the categorical predictors `PUBorough` and `DOBorough` intuitively seems likely, since the different boroughs have different traffic conditions, and therefore the effect of distance in each would not be the same. Indeed (Figure 7) and (Figure 8) do suggest this is the case, as the fitted lines relating `cost` and `approx_dist` in each borough do have different gradients.

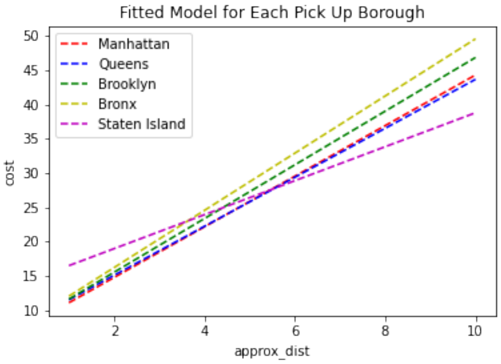

Figure 7: The effect of distance in pick up boroughs (EWR removed because of so little data that trendline was clearly erroneous)

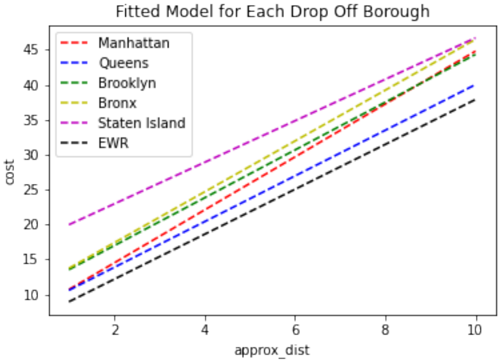

Figure 8: The effect of distance in drop off boroughs

Along with this, interaction was also considered between the time of day and day of week attributes, as well as between those two attributes and the pick up and drop off boroughs. This was just motivated by intuition that the effect that the time of day had on taxi prices would differ on different days and in different boroughs—for example on weekends in Manhattan late night taxis may be more expensive than late night taxis during the week because there may be more people out. As (Table 1) shows, these new parameters decreased the AIC and BIC of the model.

Following adding these interaction terms, the `DailyAverageRelativeHumidity` (p-value 0.855) was removed from the model, and this was the final change that was made to the model.

The final model satisfies the assumptions of a linear regression model, as (Figure 9) shows the residuals appear to have a mean of zero and a relatively constant variance given the fitted value. (Figure 10) also shows that the residuals closely follow a normal distribution.

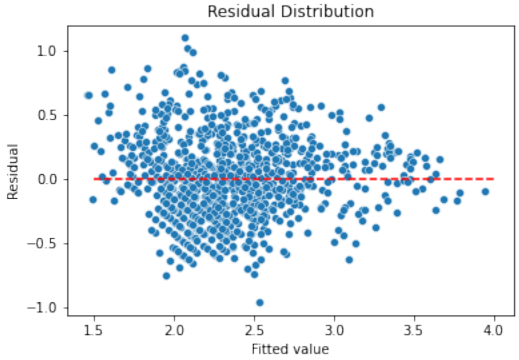

Figure 9: Residual distribution against fitted values (n = 1000)

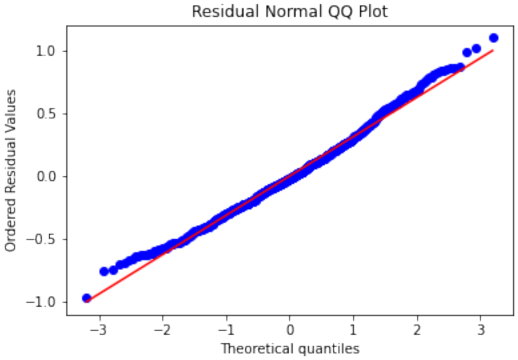

Figure 10: Residual values compared to theoretical normal distribution values

4.2 Prediction Results

For the final fitted model, predictions were made for 100 random samples of size 10000 from the 2019 dataset, with an average root mean squared error (RMSE) of $5.06 and mean absolute error (MAE) of $3.60. A baseline model which simply predicts the mean of `cost` for every instance produces an RMSE of $7.98 and MAE of $5.10. Clearly, the final fitted model performs much better than this, improving on these baseline metrics by 36% and 30% respectively.

Model	AIC (x10^6)	BIC (x10^6)	R^2	RMSE ($)	MAE ($)
Additive	6.037	6.038	0.569	5.12	3.61
With interaction	5.842	5.844	0.577	5.12	3.58
Final model	5.842	5.844	0.577	5.06	3.60

Table 1: Evaluation metrics for each model developed

4.3 Discussion

(Figure 11) shows that the model's predicted costs tend to be lower than the actual cost, which is not a significant surprise. Indeed, comparing (Figure 11) to (Figure 3), they are quite similar and it seems that a large reason for the underestimation of trip costs is because of the underestimate of trip distance.

As (Figure 12) shows, the cost predictions were worst in the taxi zones in outer New York, such as areas of Staten Island and areas near Jamaica Bay, while predictions within Manhattan were much more accurate. This behaviour is expected, as the vast bulk of taxi trips occur within Manhattan and therefore the model will tend to be slightly 'overfit' to the average Manhattan trip profile.

In the final model, both the `PU_crashes` and `DO_crashes` parameters were negative, suggesting that trips in areas with many crashes were likely cheaper. This was unlikely a causal relationship, however,

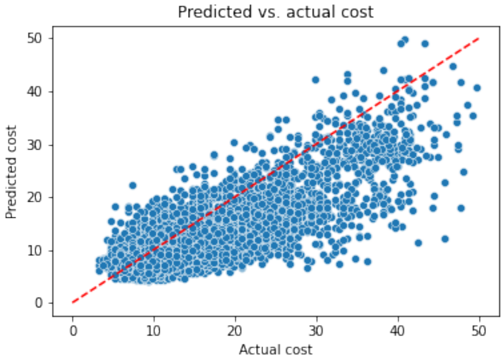

Figure 11: Final model predictions vs. actual values

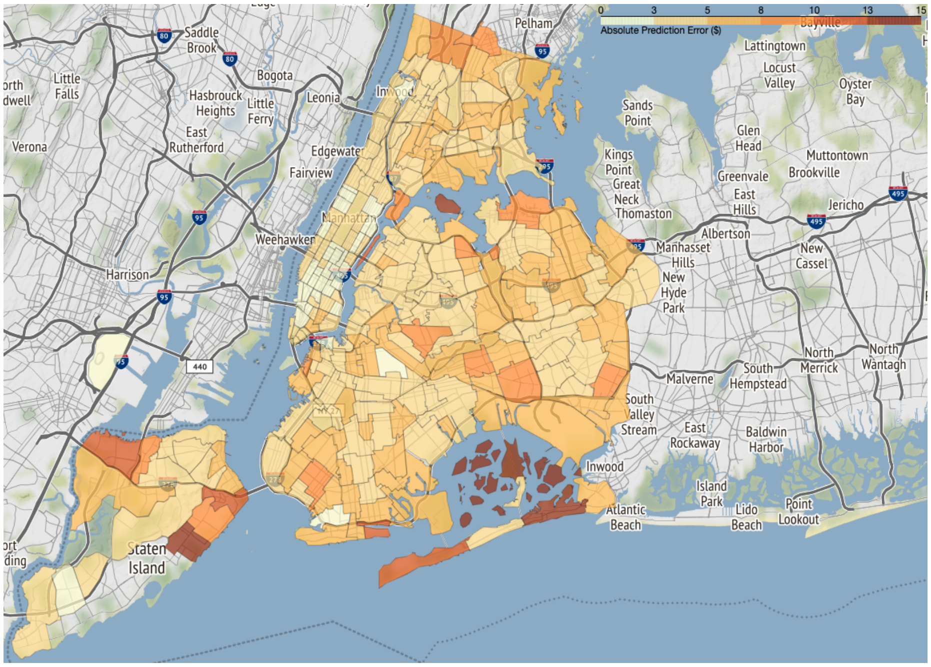

Figure 12: Absolute prediction error by taxi zone

and may just be because most car crashes occur in Manhattan, where taxi trips are usually cheap.

All of the parameters relating to the weather data were very small, suggesting these attributes have little impact on the cost of a taxi trip.

5 Recommendations

While the results from the linear regression model produced in this investigation were an improvement on a baseline model, the error in predictions is still too large to justify developing a product for commercial use to predict taxi prices for customers. There is promise in the idea, however, as these relatively good results were achieved with quite crude estimates of important factors which affect trip price. The approximate distance of the trip can easily be improved by not using naïve straight line distance, but instead using routing software to find the most likely route a taxi will take.

Should real-time crash data be available, and be able to be mapped onto the route a taxi may take, this would prove much more useful than the approach taken here, as such crashes would be very likely to impact the traffic conditions at the time of the taxi trip.

The weather data did not have very much predictive power, despite the expectation that heat and precipitation do affect traffic flow.[6]. Weather data should not be used directly in any improvements to this model.

Ultimately, if a service is to be developed to predict taxi costs, it should seek to use traffic data directly to assess how long a taxi trip will take, and therefore how expensive it will be. Using peripheral data such as car crashes and weather to simulate traffic data to a limited degree does not provide enough predictive power to develop a highly accurate model.

6 Conclusion

The idea to give customers an approximate price for a taxi ride is worth continued exploration. Indeed, some apps—such as Weave—have been created in recent years to offer prepaid services for taxi rides in New York, lending credibility to the idea. This exploration serves as an interesting investigation into the predictability of taxi prices and how it can be broken into two components, the 'easy'—approximating the distance of the trip—and the 'hard'—approximating the traffic conditions at the time of the trip.

[6] Cools, Mario, Elke Moons, and Geert Wets. Assessing the Impact of Weather on Traffic Intensity. Weather, Climate, and Society 2, 1 (2010): 60-68 https://doi.org/10.1175/2009WCAS1014.1

References

[1] Taxi Fare - TLC. Accessed August 9, 2021. https://www1.nyc.gov/site/tlc/passengers/taxi-fare.page.

[2] Atkinson-Palombo C, Varone L, Garrick NW. Understanding the Surprising and Oversized Use of Ridesourcing Services in Poor Neighborhoods in New York City. Transportation Research Record (2019);2673(11):185-194. doi:10.1177/0361198119835809

[3] Cools, Mario, Elke Moons, and Geert Wets. Assessing the Impact of Weather on Traffic Intensity. Weather, Climate, and Society 2, 1 (2010): 60-68 https://doi.org/10.1175/2009WCAS1014.1

[4] New York City Taxi and Limousine Commission. 2018 Annual Report. https://www1.nyc.gov/assets/tlc/downloads/pdf/annual_report_2018.pdf