Forecasting NYC Taxi Revenues by Location and Hour

Michael Pollard
Student ID: 36667
Github repo with commit

August 20, 2023

1 Introduction: Business Case

New York City (NYC) is famous for many things - one of which are the ubiquitous Yellow (and Green) cabs which, whilst iconic, are being increasingly supplanted by rideshare providers such as Uber and Lyft. Despite most peoples' assumption that taxis and rideshare providers would be locked in a battle for dominance, in NYC we have actually seen cooperation between these cohorts to try and increase revenues. [1]

This report supposes that this consortium of transport providers (from here onwards referred to as taxis) are trying to determine how to increase their revenue further, by understanding where, and when, to direct drivers to capture the greatest total amount of fares. The target audience are the managers of these taxi companies, responsible for dispatching taxis around the city. As a simplifying assumption it is assumed that these companies are only interested in the total fare generated, not tips and other taxes payable, nor how many drivers are active without carrying passengers (this data was not available). This is probably a reasonable assumption given many taxi companies either have a business model of leasing cars to their drivers, or taking a cut of fares.

2 Datasets

2.1 NYC TLC Data

The primary dataset used in this report is from the NYC Taxi and Limousine Commission (TLC) [2]. This dataset includes information on pickup and dropoff location (out of 263 taxi zones across NYC) and time, details on fares and taxes collected, and other metadata. Three different sources are used: Yellow cabs, Green cabs, and rideshare providers (denoted For Hire Vehices High Volume / "FHVHV"). A fourth source was not used - providing data on limousines and similar vehicles, as this is a different kind of offering to that provided by the taxi consortium. This dataset contained some entries which were clearly suspect, and required some outlier detection.

2.2 NCEI Weather Data

The second dataset was provided by Iowa State University, and sourced from the National Centers for Environmental Infomration (NCEI) [3], detailing hourly weather recordings from JFK airport. The dataset contained a wide range of weather-related features from which a small number of features were chosen. Many of these features would be strongly correlated, so using a large number would be inadvisable. A simplifying assumption was made that the weather across NYC was identical to that at JFK, at any point in time.

2.3 NYSDOT Event Data

The final dataset utilised in this project consisted of events that affect traffic and transportation, sourced from data.ny.gov and originally produced by the New York State Department of Transportation (NYSDOT) [4]. This dataset includes a very wide range of events that might impact transport - from construction, to crashes, to sports and entertainment. Only the sports and entertainment events were used, as it was not clear how the other event types should impact taxi demand.

The dataset also included a record of longitude and latitude for each event, and temporal information. The dataset is of reasonably high quality, though there appears to be no constraints on how events are recorded (for example "boxing" and "boxing match" are listed as different event types), and some thought was required on how to decipher the termporal features (disucssed below). It would have been good to have been able to include information on public transport disruptions, but this data could not be found.

3 Preprocessing

3.1 Feature Selection

3.1.1 NYC TLC Data

Given the aim was to model where and when to dispatch taxis, in order to maximise fare collection, only pickup time (not dropoff), fare, and pickup location features were used for modelling. Revenue potential is likely to vary greatly between locations, so the 263 TLC zones were used as a categorical variable. In this way of all the features available in the datasets (Yellow = 19, Green = 18, FHVHV = 24), only 4 were finally used for analysis, though several others were used as indicators of suspect data, as described in the Outlier Detection section.

3.1.2 NCEI Weather Data

There were numerous features on offer in this dataset (32 in total) but temperature, wind speed, and precipitation were selected as the fields most likely to impact taxi demand.

3.1.3 NYSDOT Event Data

A lot of the features in this dataset were descriptive in nature, or metadata. The features that were used were the event type, those relating to the date and time that the event took place ("create_time" and "close_time"), and the longitude and latitude coordinates. Of 17 provided features, 5 were used.

3.2 Timespan

Data from May 2020 until May 2023 was used analysis. Whilst it would have been preferable to use data that was not heavily impacted by COVID, the pre-COVID data for the FHVHV dataset did not provide all the required features. It was also important to gain enough data points per location and per block of time, in order to generate a robust model, therefore three years of data was decided on. For purposes of modelling - May 2020 until May 2022 was used as train data, and May 2022 until May 2023 as test data. Given this is timeseries data it did not make sense to use a sampling process for train and test.

3.3 Filtering and Outlier Detection

3.3.1 NYC TLC Data

The following tests were used to filter and detect outliers in the TLC data:

- Fare: all negative fares were removed, and any fares greater than $500 were also removed. Conceivably someone might spend over half a day in a taxi, but these were only a very small amount of records and some were clearly wrong (fares with 5-6 digits)
- Distance: whilst distance wasn't used for the models, those trips with negative distance, or distance greater than 500 miles were removed, as being indicative of suspect records
- Time: once again, trip time was not used in the analysis, but records with negative trip time, or time greater than 5 hours were also removed as being suspect
- Pickup location: Those records with pickup location outside the 263 taxi zones were removed, presumably these zones were somehow entered in error
- Date: Those records outside 1st May 2020 - 31 May 2023 were removed

Table 1 provides data on how many raw records were present before the filtering process, what percentage were removed for each filter and for each data source, and how many records were left in a curated state at the end of the process. A reasonably modest number of records were removed.

Table 1: TLC data outlier detection

Service	Raw	Fare	Distance	Time	Pickup	Date	Cumulative	Curated
yellow	95,445,749	-0.65%	-1.35%	-0.23%	-1.13%	-0.07%	-3.38%	92,215,295
green	2,877,421	-0.46%	-4.54%	-0.36%	-0.19%	-0.05%	-5.54%	2,718,001
fhvhv	566,168,421	-0.15%	-0.03%	0.00%	-0.01%	-0.06%	-0.25%	564,757,294
Total	664,491,591	-0.23%	-0.24%	-0.04%	-0.17%	-0.06%	-0.72%	659,690,590

3.3.2 NCEI Weather Data

Checking the distribution (min, mean, max, and quartiles) of temperature, wind speed, and precipitation, all metrics looked as expected, so no filtering was required.

3.3.3 NYSDOT Event Data

Some records were filtered due to the provided longitude and latitude data being outside the TLC taxi zones, this is described in the next section. Otherwise the selection of sports and entertainment events, as opposed to all event types, and filtering by date range, reduced the total number of records from 2.93 million to 3,311 records.

4 Feature Engineering and Imputation

4.1 NYC TLC Data

It was decided that the time unit for analysis should be one hour, that is revenue should be analysed per location and per hourly block. Therefore pickup date and time was split into month and time of day as well as day of week. Further to this a weekend flag was generated, as revenue patterns are likely to differ greatly between weekdays and weekend, but perhaps less between individual weekdays.

Other than this - hours without any revenue data were not imputed, rather it was assumed no trips occurred during these times.

One further key feature that was engineered was a lag feature - which provided the revenue for each particular location, for the hour prior. The reason this was included was that there is clearly an overall upward trend in taxi revenues over this date range (see Initial Analysis), and clearly a tendency for autocorrelation in the timeseries. Including a lag feature is a somewhat crude way to incorporate this structure, further development might include looking at ARIMA models. There is also a fundamental justification for including this feature - if the various taxi providers are indeed pooling their resources, they may well have access to collective live revenue data, or failing that either Yellow can or FHVHV providers would be big enough that they could use only their own data.

4.2 NCEI Weather Data

The weather data had some dummy (non-numeric) entries where a valid recording was not available. As this data is timeseries data, with weather reasonably consistent from one hour to the next, it was decided to impute this data by taking the average of neighbouring values, or a single neighbouring value if only one was available.

4.3 NYSDOT Event Data

Some deciphering was required of the temporal features in this dataset. After investigation it appeared that the close_time represented the point in time when the event stopped impacting traffic, but only for past events. Whereas the create_time correlated to the start of the event for future events (that is - ahead of the Apr 2023 date when the database was last updated), but were not so for past events. This could be seen in the gap of multiple days between create_time and close_time for many past events. The approach was therefore to take the event impact window at +/- 1 hour around close_time for past events, and create_time+2 hours to create_time+4 hours for future events (predicated on an assumption that events are approximately 3 hours long). There are admittedly a number of approximations here.

Other than this the longitude and latitude coordinates also had to be mapped to the TLC taxi zones. This was achieved by examining the shape file to see which zone contained the provided event coordinates. At this stage a number of events were filtered out as they either occurred outside the greater NY area, or because the event coordinates were slightly off and sitting just outside a zone where it should have been located. Visual inspection highlighted some of these cases, though potentially a future development would be to allow same error in the coordinates and automatically pull them back into a zone. In this step the number of events records was reduced from 3,311 to 1,855.

5 Analysis and Geospatial Visualisation

5.1 Initial Analysis

Before modelling was conducted the TLC dataset was examined to provide further background to the topic, and a grounding for model development. It can be seen in figure 1 that the share of rideshare services (fhvhv) has indeed been growing since COVID, as passengers have embraced the ability to summon transport from their phones. Furthermore a clear upward trend is evident in taxi revenues, seen most clearly in the 30 day rolling average in figure 2. Clearly this has occurred as COVID-related shutdowns have largely become a thing of the past (for now). It was considered whether COVID cases should be included as a dataset, however the link between cases and mobility broke down as the pandemic progressed.

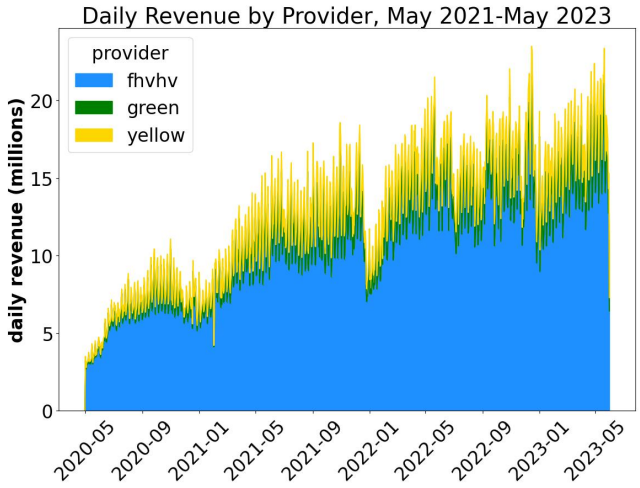

Figure 1: Shows revenue breakdown by provider

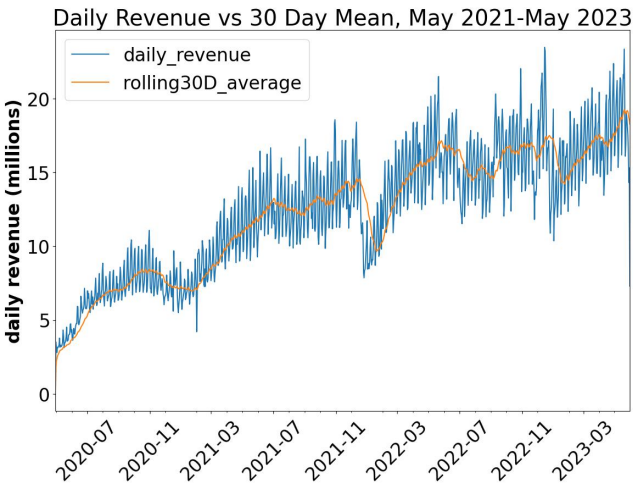

Figure 2: Shows upward trend in taxi revenues

In figure 3 it can be seen that the distribution of taxis fares between pickup locations displays a classic power law relationship, with many locations taking a small daily haul, and a small number of locations taking a lot (with airports generally the highest revenue generating venues, though presumably also having some of the highest numbers of idle drivers). Variability around daily revenue is high, as can be seen in figure 4. This boxplot of daily revenue distribution shows the top 10 revenue-generating locations. There is a wide spread, and a number of outliers identified (location 79, the East Village, having the greatest). The purpose of this project is to attempt to explain this variability.

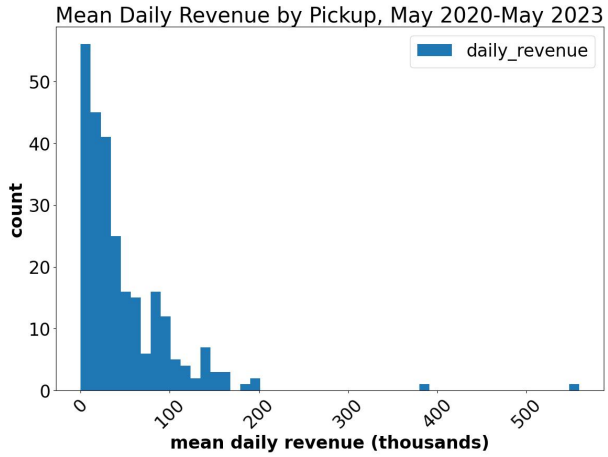

Figure 3: Revenue by location, a power law

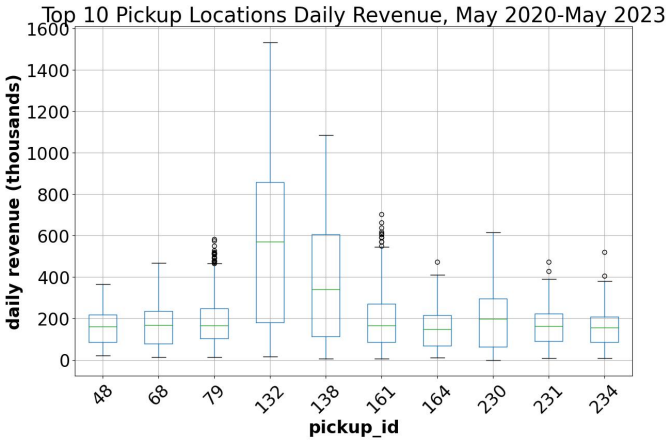

Figure 4: Revenue variation is high, with outliers

5.2 Geospatial Analysis

It is interesting to plot the number of events on a geospatial basis, as can be seen in figure 5. As might be expected there are a small number of zones where there are a large number of events over the three years of data, and much larger numbers of zones where there are a minor number of events, or none at all. Clearly the explanatory power of the events feature will be highly tied to the location in question, given the relatively small number of records.

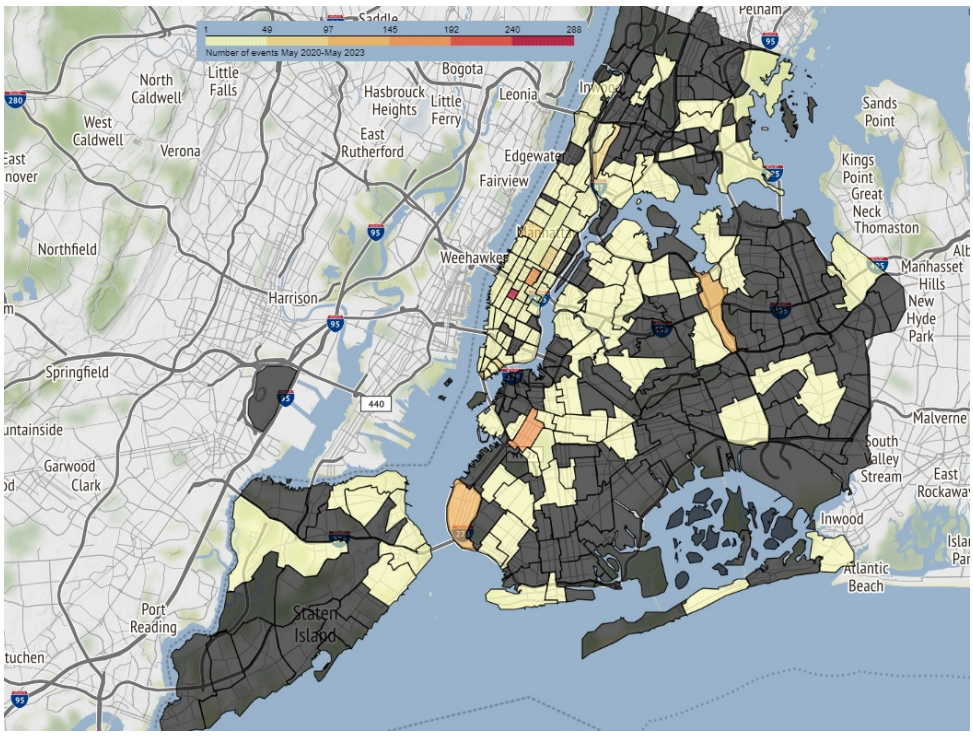

Figure 5: Events are sparsely distributed

6 Modelling

6.1 Linear Regression

The first approach taken was a linear regression, of the following form:

$$Y_{ijklm} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + \zeta_m + \sum_{n=1}^{4} a_n X_n + \epsilon_{ijklm} \tag{1}$$

Where Y is the hourly revenue; α=location effect, β = month effect, γ = hour effect, δ = weekend effect, ζ = event effect, all categorical variables; and the four continuous variables with X_1, X_2, X_3 representing the weather variables, and X_4 representing the lagged hourly revenue. As a simplifying assumption I have not modelled interaction terms, though there are likely to be some meaningful ones - for example between the weekend flag and hour. There is likely to be lots of demand for taxis in the morning on weekdays, and less on the weekend, for instance.

Before running the regression, an aggregation was run in order to bucket the data by pickup location, and hour. In addition one hot encoding was used to transform the categorical variables. The resulting large number of records and features suggested using the pySpark ml regression model. One downside of this approach was less availability of packages, for instance to run ANOVA across the variables to test for significance. Nonetheless a UDF was created to run this manually, by comparing the full model to a reduced model without each variable in turn, and computing the canonical F-statistic:

$$F = \frac{(SSR_{reduced} - SSR_{full})/(p_{full} - p_{reduced})}{SSR_{full}/(n_{obs} - p_{full})} \tag{2}$$

Where SSR = sum of squared residuals, and p_x is the number of parameters in model x.

Note in figure 6 that the residuals appear approximately normal (taking some liberty, and the y axis is log10), so the use of the F statistic is, perhaps, appropriate.

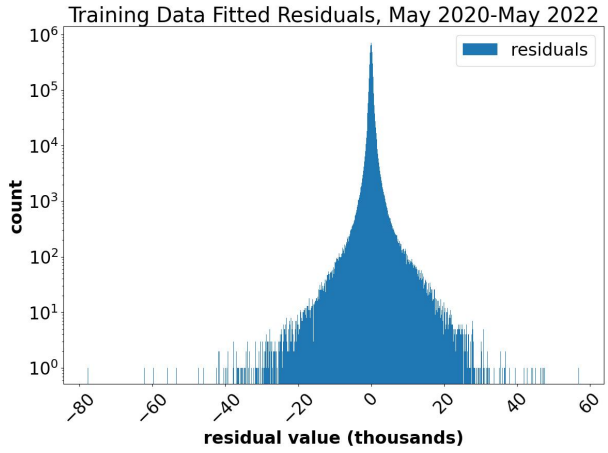

Figure 6: Residuals are approximately normal

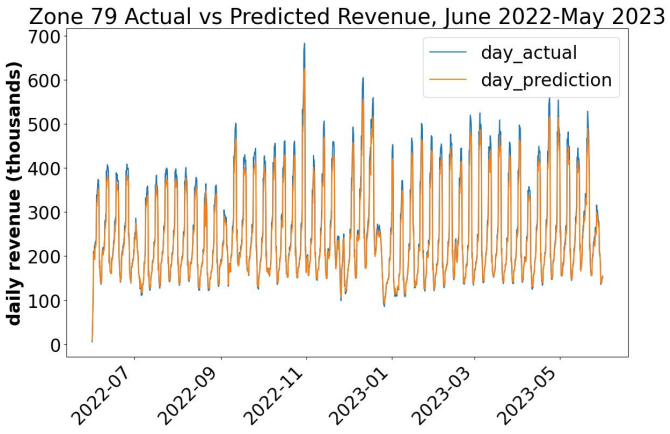

Figure 7: An example zone demonstrates a good fit

The ANOVA analysis in Table 2 at first glance would seem to indicate that all the chosen variables are indeed very significant. However it should be remembered that the training set of two years of hourly data, by location, generates a lot of records. With a large number of records there will always be a tendency for variables to be significant, though there is some question of whether they are indeed fitting some underlying feature of the dataset, or just fitting noise.

However, there is still some information that can be gleaned from this analysis, by looking at the relative value of the F statistics. Clearly the lag effect is very strong indeed, suggesting a powerful autocorrelation in the revenue data, which makes intuitive sense. The next most powerful variables appear to be weather and the weekend effect, with the month and event effect much less so. This also seems to be an intuitive result. Note that location and hour were not included in the analysis as the other variables essentially rely on the presence of these two features, in order to have meaning.

To complete the analysis the full model, and individually reduced models, were also run on the test data. Results are shown in the final two columns of Table 2. Unlike the ANOVA table it is clear from these results that the only meaningful variable is the lag variable, with perhaps weather making an extremely marginal contribution. The results for the full model are somewhat in line with the training data, where a RMSE of 882.6834 was achieved, and an R^2 of 0.8968. In Figure 7 we look back at taxi zone 79, which was highlighted as one with a lot of outliers, to see if the model does a good job of forecasting revenues. This was done by taking a rolling 24 hour number to produce a daily revenue. The fit is very close, with some misfitting around turning points in the data, which is to be expected when the lag variable is so dominant.

6.2 Decision Tree

A second model that was investigated was a decision tree. This was an interesting option as it does not require one hot encoding of the large number of categorical variables used in the regression model. To enable the process to run efficiently the categorical variables with more than 4 categories were treated as continuous variables. However, a max bin size of 300 was used such that these categories could still effectively be treated as distinct, if it was efficient for the decision tree to do so.

The decision tree produced a slightly worse result than the linear regression model with a RMSE of 1852.1500, and an R^2 of 0.8511. Potentially this may be due to lag variable being dominant, and the

Table 2: ANOVA analysis of linear regression model

Variable	RSS	df	F	Pr(> F)	RMSE	R^2
full	3.5577e+12				1505.1703	0.9016
event	3.5585e+12	1	1087.8352	0	1505.4947	0.9016
weather	3.5696e+12	3	5124.6387	0	1506.4133	0.9015
lag	1.7275e+13	1	17604880	0	3613.9666	0.4330
month	3.5612e+12	12	377.6042	0	1505.1886	0.9016
weekend	3.5619e+12	1	5403.1669	0	1505.4115	0.9016

max bin size of 300 being applied to the spread of values on this. With normalisation of the data potentially this source of variance might be reduced.

It should also be noted that the random sampling involved in the Decision Tree model produced a different output each time - for instance the final run saved in the notebook produced a RMSE of 1974.89, and R^2 of 0.8307. Potentially a number of runs should be executed with an averaging process.

7 Discussion and Recommendations

Despite incorporating a set of features that held some promise for predicting revenues, it seems that perhaps the simplest, and most intuitive predictor of revenues was indeed the prior hour's revenue for that particular location. Incorporating weather forecasts may also add some minor value. However, perhaps a more fruitful target might be to work further on ARIMA models that are designed to deal with such timeseries data and autocorrelation.

Recommendations for the taxi companies would be to structure their processes for dispatching of taxis around the city, in order to maximise revenue by location and time of day. It has been demonstrated in this report that there is great variability across these metrics. The most informative feature used to predict this revenue turned out to be the lagged revenue series. Most likely using such a feature effectively incorporated the impact of other features that were being modelled. On the upside - such a system would be easy to implement, as presumably the companies would already have this data to hand. Perhaps they are already doing this, but it certainly would make sense to systematise the process to automatically dispatch cars accordingly. As mentioned previously it may certainly be the case that taxi companies are not willing to share their live data, but both Uber and Yellow cabs would still be large enough that their own datasets would be indicative of the overall revenue pool available.

References

[1] New York Times. *Uber Partners With Yellow Taxi Companies in N.Y.C.* https://www.nytimes.com/2022/03/24/business/uber-new-york-taxis.html. Accessed: 2022-08-10.

[2] New York City Taxi Limousine Commission. *TLC trip record data and geospatial data.* https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page. Accessed: 2022-08-10.

[3] Iowa State University, NCEI ISD. *Hourly weather data from JFK airport.* https://mesonet.agron.iastate.edu/request/download.phtml?network=NY_ASOS. Accessed: 2022-08-10.

[4] data.ny.gov, New York State Department of Transportation (NYSDOT). *511NY Events Data, all recorded events impacting traffic and transport.* https://data.ny.gov/Transportation/511-NY-Events-Beginning-2010/ah74-pg4w. Accessed: 2022-08-10.