Modelling Tips - How an average driver can get ahead

A Qualitative Study into NYC Taxi Data

Arjun Rajkumar
Student ID: 910 941
https://github.com/arjunrajkumar-ds/ADS_Project_1/commit/8e57fba5a212c255ed9e0a3cc793b35

August 28, 2022

1

1 Introduction

In the recent deluge of American financial news, it is constantly observed that the mainstream media announce proudly how Fortune 500 businesses are seeing ever-higher profits, all while worker equity is being ground out. As an Australian observer, my sympathies go out to the middle and lower classes of America.

The world has gone through a tremendous amount of stress and uncertainty in the last few years. COVID-19 rocked the roots of trade and commerce to their core, and glaring issues with supply chain and infrastructure were harshly brought to light. As the number of active cases increased, the workforce shrank, and roles critical to a functioning society were highlighted. The term 'essential worker' was coined, and these jobs were highlighted. However, these same record profit-breaking companies offered cosmetic rewards only, in the form of public thank-yous and pittances. Instead, the younger generations received a shaky economy, featuring ever-rising inflation (Figure 1), increasing debt to GDP and other worrying markers.

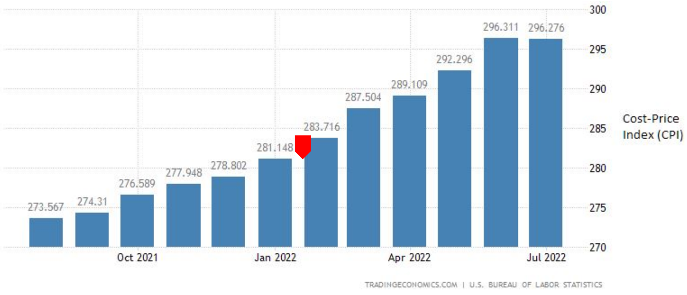

Figure 1: Increase in CPI over the last year. Source: FRED

This study aims to empower those individuals who field these roles and increase their quality of life, by maximising profits. This analysis goes over NYC Taxi drivers specifically, utilising trip data provided by the New York Taxi and Limousine Commission (TLC) and property data scraped from ://www1.nyc.gov/assets/finance/downloads/pdf/rollingsales

1.1 Datasets

1. TLC Yellow Taxi Data:

   This dataset is from the NYC TLC, and is (12,671,164 rows x 19 columns) raw. After pre-processing, 8,933,245 rows remain - 70.88% of the original size. It's comprised of the following features:

   - Pickup / Dropoff Timestamps
   - Trip Distance (KM)
   - Pickup / Dropoff Zone IDs
   - Fare Amount (AUD)
   - Tip Amount (AUD)
   - Total Amount (AUD)

2. NYC Gov. Rolling Sales Data:

   This dataset is from the NYC Government, and originally had (98525 rows x 21 columns). Due to suitable available imputations, no rows had to be dropped. Its features are:

   - Borough
   - Neighbourhood / Zone
   - Building (Tax) Class
   - Year Built
   - Sale Price (AUD)
   - Square Area [2]

1.2 Assumptions

1. Property valuations consider the area of the land, not of the building built upon that land
2. That the area of a property followed a random distribution. However looking back, this was an incorrect assumption and the randomness was caused by the presence of various different (Tax) Classes of Buildings present in the data. These ranged from farms to commercial buildings to residential units. I should've filtered out non-residential codes.
3. The volume of Taxi trip data is large enough that the Central Limit Theorem applies, as approximately 280,000 trips in Yellow taxis take place each day [3].

2 Preprocessing

- The Taxi data was read into a Spark DataFrame, whereas Pandas was used for Property
- The raw data came with several columns that were deemed irrelevant to this particular study. Some examples are:
  - Passenger Count: It was reasoned that even if multiple people were to travel in the same cab, the total fare would still be the same - taking into account distance travelled, time taken and any applicable tarrifs

  - Several other attributes were dropped intuitively, and can be seen in the `preprocessing` notebook

- The raw data was quite inconsistently named. Columns were renamed in line with the `snakecase` format

- Null / N/A data had to be dealt with for the Property dataset only. 2 features (of residential units of commerical units on the property) were dropped, as there were a moderately high percentage (25-44%) of values missing. It also didn't make sense to impute the missing values, as the values were not normally distributed.

  The fields that did warrant imputing were Square Feet Year Built. Square Feet had 47% of rows missing, however outlier removal at the time was ineffective, due to my incorrect assumption (See 1.2 Assumptions).

- Simple outlier removal was attempted, by using quantiles, the Inter-Quartile Range and creating upper & lower bounds for each feature. However an anomaly was observed - for each numerical feature in the Taxi dataset, attempting to impose an upper bound resulted in loss of data. One explanation could be that the data's true distribution is quite left-skewed, although further discussion is welcomed.

- Columns were appropriately cast to their respective type.

- As both datasets are from American sources, the units used were not metric. Conversions were made for ease of interpretability to an Australia-based audience

- Finally, filtering based on feature values was carried out. According to the Data Dictionary provided by the TLC, only specific data was relevant to this study: trips where the RateCodeID $\bar{1}$ or 6. These values capture trips that are Standard Fare and Group Rides. Group Rides were retained due to an assumption listed above Further, tip amounts were only captured when the method of payment was credit card. Cash or non-credit transactions would've recorded a tip amount of $0, and so were filtered out.

3 Exploratory Data Analysis

The distributions of the data were examined first. It was observed that log transformations on all numerical features in the Taxi dataset resulted in approximate Normal distributions of those variables (See Figures 2 - 5 below)

Similar transformations were made to the numerical features in the Property dataframe. One interesting observation was seen when analysing the distribution of the Year that properties were built. In Figure 6, you can see 2 major dips, signifying a massive halt to property development in NYC. The first dip starts around 1938 and ends 1945. This lines up perfectly with WWII, which would explain a slowdown in development. The second dip is a bit more ambiguous. My current theory is that the aftereffects of the Cold War (1947 - 1991) had an effect on development in NYC.

When looking at the distribution of building class codes present in the data (Figure 6), 4 groups make up the majority of properties present. They are:

- 01: One Family Dwellings
- 02: Two Family Dwellings
- 10: Co-ops - Elevator Apartments

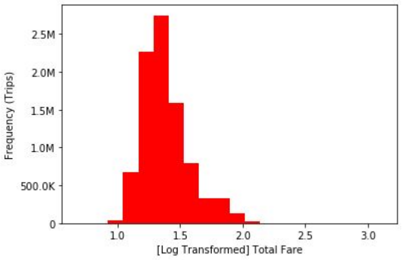

Figure 2: Histogram of transformed Total Fare

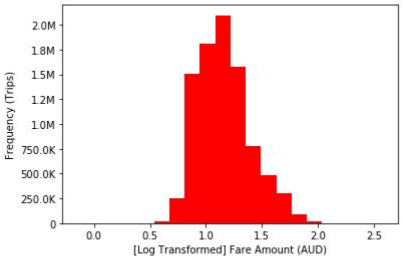

Figure 3: Histogram of transformed Fare Amount

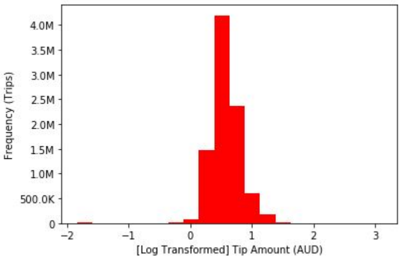

Figure 4: Histogram of transformed Tip Amount

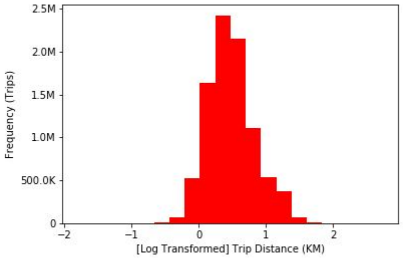

Figure 5: Histogram of transformed Trip Distance

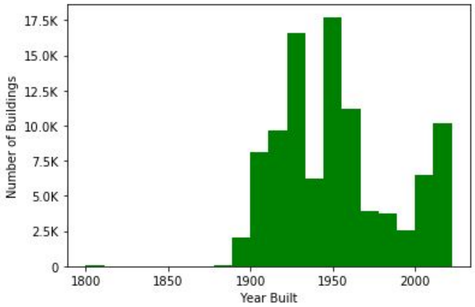

Figure 6: Distribution of when properties were built in NYC

- 13: Condos - Elevator Apartments

Once again, looking back it would've been more prudent to filter out irrelevant building codes for a less noisy dataset, however our data is representative enough of each borough (see Figure 7), so optimistically, noise from each borough's various types of buildings would contribute approximately equally, and not skew the data in any certain way.

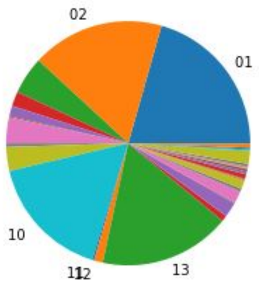

Figure 7: Building Class Code Representation in Property Data

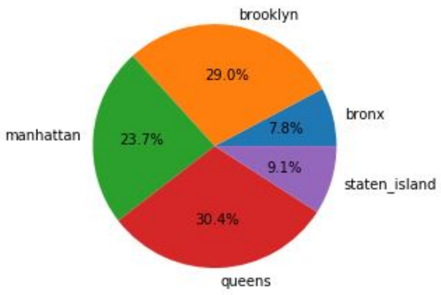

Figure 8: Borough Representation in Taxi Data

4 Bivariate Analysis

The numerical features from both datasets were examined next. By plotting a scatter matrix (Figure 9), it can be observed that each numerical feature in the Taxi dataset is approximately normally distributed, and have moderate to strong positive relationships with each other. Figure 10 displays a heatmap visualisation of the Pearson correlation coefficient between the Property dataset's numerical features. One can see a very strong positive relationship between the Sale Price of a property and its (transformed) Area. This was to be expected, as even in current market conditions, the area of a property plays a big part in pricing. Although this may seem elementary, we can rest assured that the relationship has now been mathematically shown. On the other hand, almost no relationship exists between the year a property was built and its sale price. This is quite interesting, and serves to prove that assumptions from day-to-day life don't always hold when statistically evaluating some data.

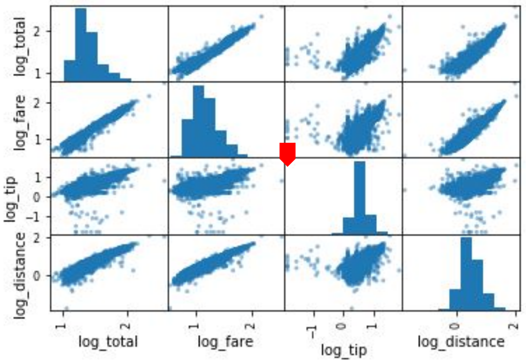

Figure 9: Scatter Matrix of Pairwise Features

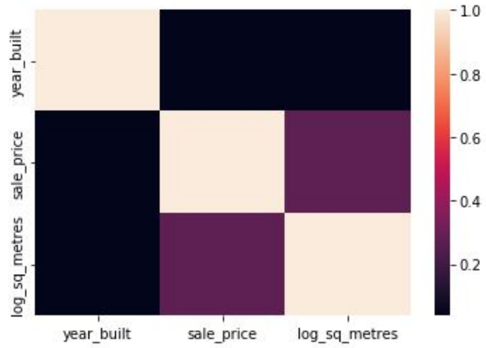

Figure 10: Correlation Heatmap of Numerical Property Features

5 Feature Generation

To get more of an insight into trip activity, features were generated from the timestamps provided. The day and hour that each trip occurred on was extracted and can be seen in Figures 11 & 12. It can be seen that Thursday, Friday and Saturday are a taxi driver's busiest days, with the hours of 2PM - 7PM experiencing heavier traffic than otherwise, with the rides peaking around 5-7PM. This is most likely attributable to office workers heading home for the day, taking a taxi. This spike in traffic isn't observed around 7-9AM, which indicates that most individuals commute to work another way in the morning.

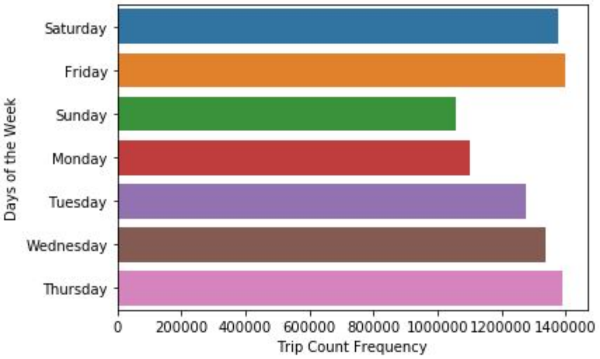

Figure 11: Distribution of Trips grouped by Day

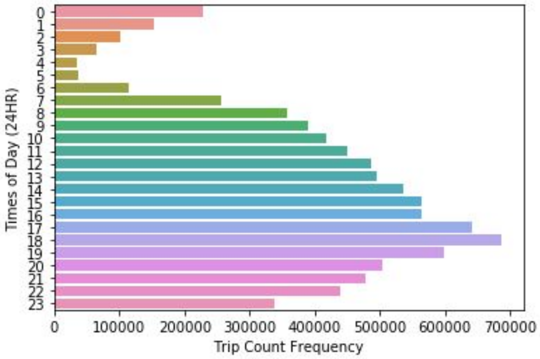

Figure 12: Distribution of Trips across the day (24HR)

6 Modelling

6.1 Linear Regression

A linear model was fit to the numerical data present in the Taxi dataset. Its objective was to predict the tip amount based off the other numerical variables present: the fare, the total and the distance travelled. In retrospect, it was quite naive in its approach, arguably, two of the variables (fare and total) could be said to be linear combinations of each other, however with more and more various charges and fees being added to the total amount payable for a NYC taxi[2], it would be fair to say that while tip is very closely related to fare, there's enough variation and randomness introduced, when human factors (that would influence tipping behaviour) are considered.

The fitted model is:

$$log[TipAmount] = \beta_0 + \beta_1 \cdot log[FareAmount] + \beta_2 \cdot log[TotalAmount] + \beta_3 \cdot log[DistanceTravelled] + \epsilon \tag{1}$$

The parameters are:

- β_0 = -1.7842
- β_1 = 3.0781
- β_2 = -1.6397
- β_3 = -0.0141

To evaluate this model, the R^2 and the Root Mean Squared Error (RMSE) were examined. R^2 is also known as the coefficient of determination, and is used to quantify the proportion of variance in the target variable that can be explained by the design variables. It's commonly used to assess Goodness of Fit. The value of R^2 always falls between [0, 1], and the R^2 of our model is 0.6837. Generally speaking, the larger the R^2 the better, however that rule cannot be followed blindly. In the context of this study, we have a moderately high R^2, which indicates that a good proportion of the variability in amount tipped is explained by the numerical features I chose.

The RMSE is an aggregation of the distance between the observed and predicted data points when fitting and testing the model’s predictive power on unseen variables. The RMSE of this model is 0.1425, and (generally speaking), the lower the better. This RMSE indicates that our model has good predictive power, even when generalising to unseen data. The distance between our predicted tip amount and the actual is quite small.

No feature selection was performed due to the few numerical features retained. Initially, I attempted performing stepwise selection over all the numerical variables present in the raw (cleaned) dataset. This proved to be too computationally demanding, and it was then that I trimmed features on intuition, leading to a parsimonious model to begin with.

7 Reflection and Recommendations

In most failed ventures, hindsight is 20/20. Through this entire process, one of the most crucial aspects of a study was overlooked - there was very little focus on an objective question when initially examining the data. I got too excited during the Exploratory Analysis phase, wherein I observed relationships between variables, and created visualisations for several different observed phenomena. Some unused visualisations can be found in my notebooks, and they were purposely omitted in this report to avoid creating even more ambiguity.

I set out to form a basis upon which drivers could be empowered with the knowledge about what key drivers facilitate higher payments. I was somewhat successful in creating that foundation - the linear model performed adequately, for a naive approach. Ultimately, the open-endedness of this study worked against me. The area in which I delivered the least was drawing insights from external data - when initially looking for external datasets, I failed to consider how the data would be joined. There was no common column shared between my two datasets, and that left simple uni bivariate analyses performed entirely within those respective datasets. However, the knowledge gained about manipulating data in both Pandas and PySpark dataframes was tangible, and is demonstrated in my notebooks.

In future, seeing as the Taxi trip data comes with timestamped data, further analysis into related timeseries features would yield very promising results. For example, if a weather dataset was used, depending on the granularity of timed weather data provided, one could obtain weather conditions for every ride, in each ’section’ of the day - morning, noon and night. This area could lend itself to exploring interesting questions.

8 References

- 1 - https://techxplore.com/news/2022-03-uber-space-nyc-taxi-cabs.html
- 2 - https://www.thecity.nyc/2022/8/22/23313635/struggling-cabbies-congestion-pricing-toll
- 3 - https://ny.curbed.com/2017/10/13/16468716/uber-yellow-cab-nyc-surpass-ridership