Taxi fare forecasting using Gradient Boosted Trees

Noah Sebastian
Student ID: 911150
Github Repository (Remote Branch) [6738f5c]

August 22, 2022

1 Introduction

On demand vehicle services play an essential role in providing convenient and affordable travel for individuals. Taking advantage of consumer demand for convenience and flexibility, technologically progressive companies such as Uber and Lyft have exploded onto the scene and now capture a significant proportion of the ride hail service, leaving the traditional taxi behind in the process.

An aspect of the New York City (NYC) taxi service, and one that differs from the more modern on-demand-vehicles, is their inability to present end-to-end fare prices and inform individuals pre-travel about their future incurred costs. This is an area where for-hire-vehicles (FHV's) have a significant competitive advantage over the taxi industry.

As such, this report will assume the perspective of a research and development company working on behalf of the New York City Taxi & Limousine Commission (NYCTLC) to provide insights and prospectus integration's for the current NYCTLC system and, more specifically, will attempt to offer a method for forecasting trip fare amounts for varying circumstances around New York City.

For this task, a Gradient-Boosted-Tree Regressor is suggested as such a model since it is capable of handling imperfect data sets, can generalize well and offers high flexibility. It is also currently one of the more mainstream algorithms adopted by the data science community. It will be contrast against a Random Forest Decision Tree Regressor, with the intention of exploring their differences, and effectiveness'.

2 Dateset

The NYC Taxi & Limousine Commission record and report monthly statistics [1] on the trips undertaken by NYC's taxi's (Yellow, Green), for-hire-vehicles and high-volume-for-hire-vehicles (Uber, Lyft, Juno, Via). This dataset provides rich information into the process' that dictate taxi travel and fares. The dataset contained a variety of features. These were reduced for forecasting and analysis according to the case objective, as detailed in Section 2.2.

In addition to this, hourly historical NYC Meteorological Aerodrome Report (METAR) [2] data was also acquired and integrated into the model as it is hypothesized that its features could have a statistically relevant relationship with the predictor variable. The features selected include temperature, relative humidity and dew point, as from these, most common weather types can be inferred.

2.1 Range Selection

As this report assumes the perspective of an innovative analysis, it is necessary to provide a representation of what could be capable and achievable by the TLC using existing technologies, and limited resources. As such, a subset of the Yellow Taxi dataset was selected. All trips that occurred during 2019, ($\approx$ 84.6m) records were used for pre-processing, model training and analysis. Further, all trips recorded in 2021 ($\approx$ 30.9m) records were used to validate and evaluate model performance.

A more recent feature set would have been preferred for training, but due to concerns over the extreme effects COVID-19 & Lock downs may have had over the taxi industry, they had to be ommitted.

2.2 Feature Selection

Due to the specific intentions of the report and subsequently the modelling, significant restrictions had to be imposed on the feature set. That is, any features that would not be available before a taxi ride commences, were removed from forecasting. Further, to comply with this scenario, and still propose insightful findings, some realistic assumptions were made about the features retained for forecasting:

- **Trip distance/duration** could be estimated by use of a modern GPS system e.g. (Google Maps, TomTom) and as such can be retained as a feature.
- **Congestion and their consequent surcharges** are considered to be similarly estimable.
- **Instantaneous weather data** is available and accessible.
- **Location specific variables** such as Ratecode etc. could be generated from destination information.

This gave the following feature set:

Retained Features:		**Weather data Features:**
Date & Time	Ratecode ID's	Temperature
Passenger Counts	Location ID's	Dew Point
Trip Distance	Payment Types	Relative Humidity
Fare amounts	Congestion surcharge	

In order to increase information from the data, the following features were engineered using the original set:

- Hour of day (PU/DO)
- Day of week (PU/DO)
- Day of month (PU/DO)
- Month of year
- Trip Time (minutes)
- Trip speed (mph)
- Fare per minute

2.3 Collation

Aggregation was required when joining the weather and taxi data sets. The METAR data came with sub-hourly samples. Since the lowest time unit that had been engineered was hourly, the weather data was mean aggregated across these intervals before merging.

Since the weather data was relatively comprehensive and due to the abundance of trip records, any inconsistencies in matches resulted in the records being dropped from analysis.

2.4 Outlier Analysis

Although the data sets were generally well formed, the taxi dataset contained a significant amount of outliers and discrepancies that would hinder the performance of the model if left unaddressed.

In accordance with sound data literacy principles, the taxi dataset underwent significant processing and outlier handling:

Trips with negative and over 312 (miles) were removed. Any trips longer than 312 (miles) were considered to be beyond the scope of analysis.

Trips with Location ID's outside the range 1-263 were removed. This is in line with the specifications provided in the data dictionary [3].

Trips with Fare Amounts greater than $500. These trips were investigated and concluded to be predominantly noise/input error.

Trips with passenger counts greater than six people. According to the NYCTLC guidelines [4] regarding 'Overloading Vehicles'. A driver must not permit more than five passengers, but is allowed an additional passenger riding on the lap of their guardian, giving at most 6.

Trips with negative tip amounts were removed. These were consistent with similar negative values found in other features. No pattern was identifiable to these outliers and as such they had to be dropped.

Trips with Vendor ID's outside the specified range. Guidelines state that only vendor codes 1 and 2 are registered as valid TPEP providers[3].

Trips filtered to only contain credit card or cash Payment Types.

Trips with Ratecode ID's outside valid bounds were removed. Specifications report Ratecodes 1 to 6 as the only valid rate codes being in effect subject to the end of the trip.

Trips with average speed (mph) greater than 65 mph and less than 1 mph were removed. NYC state guidelines [4] report the inner-city speed limit as 50 mph, and the interstate speed limit as 65 mph and it is assumed that taxi drivers mostly adhere to these restrictions.

Trips with travel speeds under 1 mph were considered noise/outliers.

Trips with Standard Ratecodes and Fare amounts under $2.50 were removed. In line with NYCTLC guidelines trips with standard ratecodes have an initial fee of $2.50, so any fare's fitting this description, under this price, were removed.

After processing the data sets contained $\approx$ 71.7m and $\approx$ 26.7m records for 2019 and 2021, respectively.

2.5 Imputation

After strict outlier removal process, there were no longer any missing values and no imputation was required.

3 Preliminary Analysis

This section will serve to identify relationships between the feature set and the variable of interest (Fare Amount) and also the available investigate categorical descriptors.

In order to feasibly provide visualization, sub-sampling without replacement was performed to down-sample the set to a more manageable size. The sampling was performed at random from the entire collated data set, with the intention of capturing enough distributional information to present graphically.

The primary question this section intends to answer is: **what constitutes a taxi's fare amount** and of the features we have available: **how can we forecast this feature?**

To help aid the direction of analysis, a correlation matrix was plot for the relevant features (Figure 1).

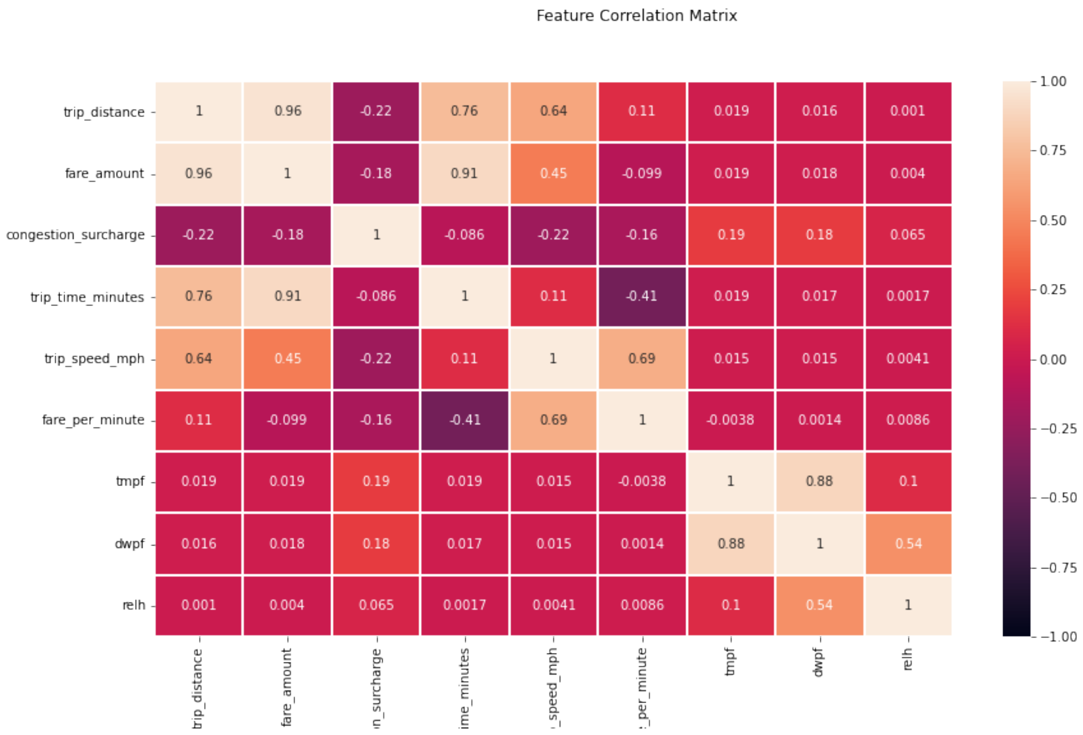


Figure 1: Correlation matrix of the feature set

3.1 Trip Duration

It was assumed that trip duration would have significant correlation with cost of the trip. This is confirmed when observing the correlation matrix. We can see that the trip duration and fare amount features are highly correlated (0.96). This is unsurprising due to distance travelled being a main constituent in how taxi fares are formulated; Specifically, the charge is increased by 50 cents for every 0.2 miles when travelling above 12 mph[5].

3.2 Trip Time

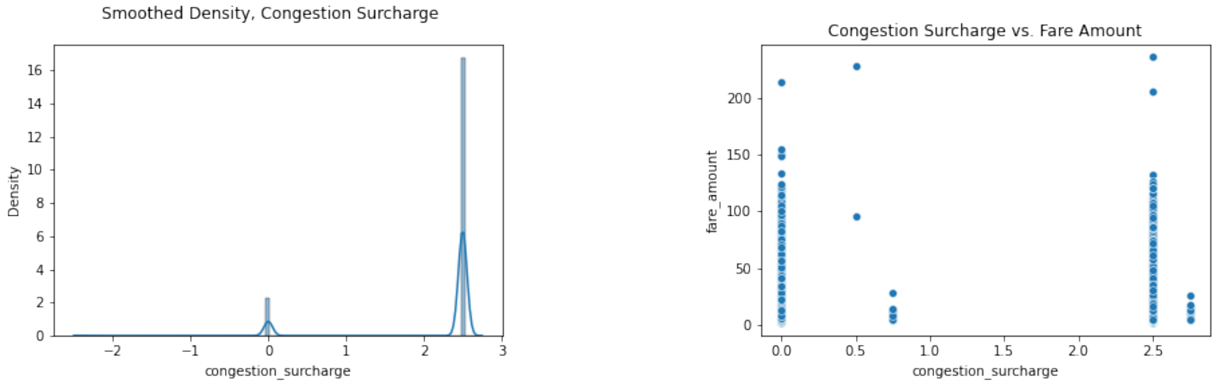
Again, it would be assumed that trip time and fare amount should be highly correlated; and they are (0.91). Trip time is also a constituent in fare calculation, but is only used when travelling below 12 mph. More specifically, for every 60 seconds that is spent under 12 mph, an additional 50 cents is charged[5].

This is interesting, as one would imagine that a significant amount of trips would average speeds much faster than this, and therefore trip time should not be such a strong predictor.

An explanation for this, is trip times extremely high positive correlation with distance (0.96); and we know that trip duration is an excellent predictor of fare amount, hence so must be trip time.

3.3 Congestion Surcharge

Congestion surcharge had a reasonably significant negative correlation with fare amount (-0.18). This garnered uni/bivariate investigation of the feature (Figure 2). It can be seen that there are two prominent bands in the distribution of the feature. This banding was found to be explained by the way congestion surcharge is calculated. Fare guidelines state that any trips that at any point pass south of 96th Street, Manhattan, incur a $2.50 charge[5]. It could be considered to convert this into a Boolean or binned categorical variable.



(a) Congestion Surcharge Distribution

(b) Congestion Surcharge vs. Fare Amount

Figure 2: Congestion surcharge distribution and relationship with Fare amount

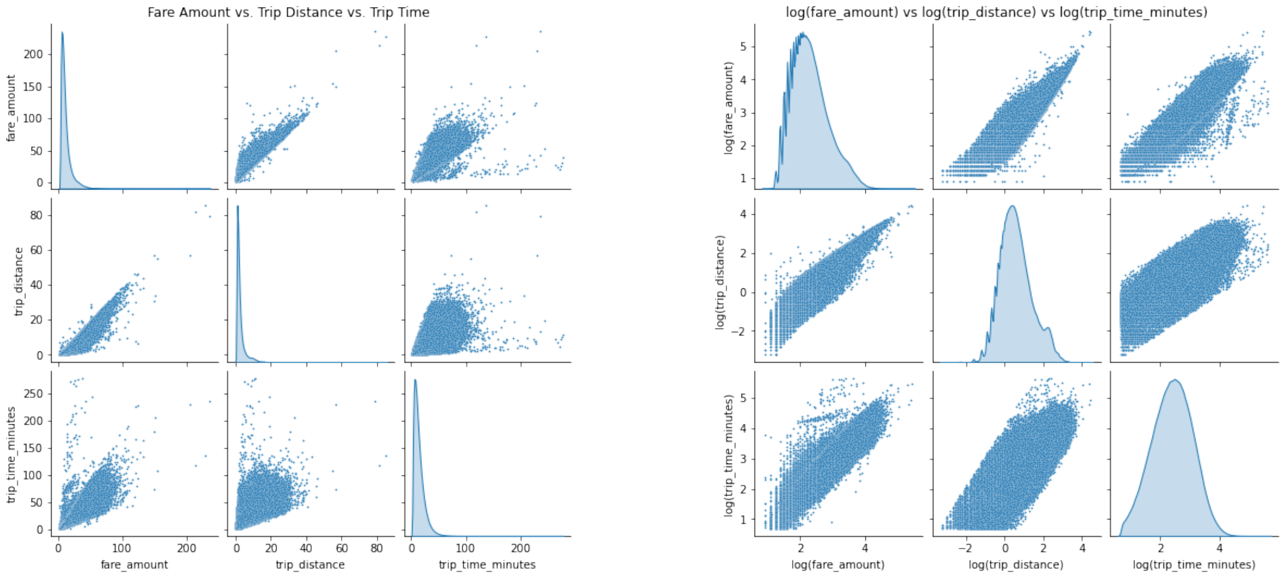
3.4 Distributional Analysis

Trip distance and duration were selected for a closer look as they showed compelling correlations with fare amount.

The pair plot (Figure 3) shows their relationships. It can be noticed in (Figure 3)(a) that their distributions all present long right-tails, and are positively skewed.

It was hypothesised that a natural log transformation could present a more normal depiction of these features, and observing (Figure 3)(b) it can be seen that the extreme tail behaviour is much more Gaussian.

Further, the pairs show strong linear relations, suggesting they will be effective in forecasting their counterparts.



(a) No transformations

(b) Natural Log transformation applied

Figure 3: Pairwise plot's of the Fare Amounts, Trip Times, Trip Durations and their respective Log transformations.

	Trip Distance	Tip Amount	Passenger Count	Fare Amount	Congestion Surcharge
Mean	2.5890	2.0484	1.5950	11.9264	2.2457
St. dev	2.8275	2.1992	1.2037	8.3485	0.7611
Max	0.04	0.0	1.0	2.5	0.0
Min	149.5	500.0	6.0	394.5	2.75

Table 1: Post processing descriptive statistics of the features

4 Gradient Boosted Tree (GBT) & Random Forest (DT) Regression

In Gradient Boosting Decision Trees multiple weak form models (Decision Trees) are combined via ensemble methods to produce a singular greater model. The individual decision trees are connected in series, where each new tree's goal is to minimise the error of its predecessor.

The sequential nature of this 'boosting' process means that the algorithms are typically slower to learn, but can produce higher accuracies. This will be analysed when comparing the bagging methods of Random Forests with the GBT methods.

Due to decision trees being flexible with their inputs, no encoding or scaling was performed, but the features were vectorised in order to implement the machine learning libraries of PySpark[6].

The models were trained on the entire 2019 data set. This may not typically be the case, but to be in line with subject specification, additional data was required for testing the models.

Predictions were then made on the set and evaluated. The metrics used to evaluate the models were: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Squared Error (MSE), Explained Variance and Coefficient of Determination (R^2).

	RMSE	MAE	MSE	Exp.Variance	R2
GBT	1.7055	0.7775	2.9086	63.7937	0.9563
RF	1.9801	0.9406	3.9207	52.6606	0.9411

Table 2: Prediction Results for the Gradient Boosted Tree model and the Random Forest model,

The results (Table 2) followed expectations regarding GBT and RF trees. The Gradient Boosted Tree outperformed the Random Forest across the metrics, although it cannot be said that the Random Forest Model performed poorly. In fact, both models performed remarkably well; that is, they were able to on average predict true fare amounts to within (0.7775) and (0.9406) (USD) respectively over the test set. The feature importance's were then investigated to understand which attributes are contributing to the successes of the models (Table 3).

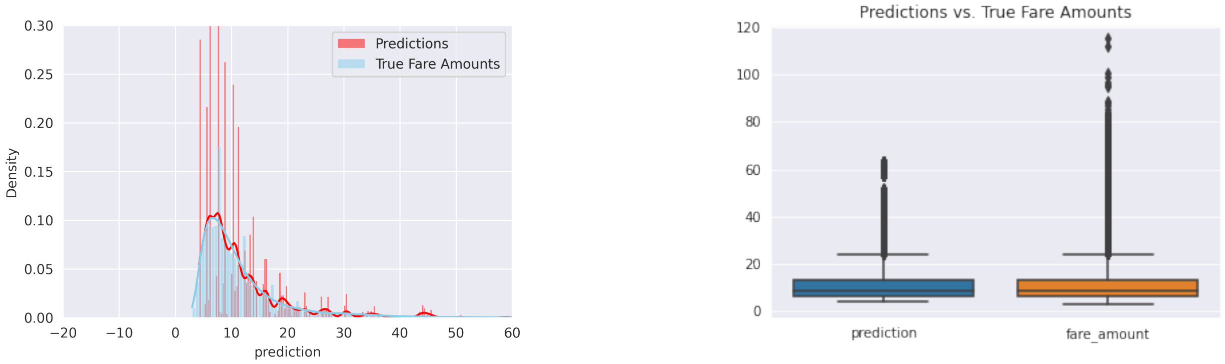
Random Forest		**Gradient Boosted**	
Feature Name	Feature Importance (%)	Feature Name	Feature Importance (%)
Trip distance (miles)	0.5394	Trip distance (miles)	0.7513
Trip time (minutes)	0.3640	Trip time (minutes)	0.2260
Trip speed (mph)	0.0610	Trip speed (mph)	0.0213
Pick-up LocationID	0.0252	Congestion Surcharge	0.0009

Table 3: Top 4 highest scoring features (importance) for the Gradient Boosted and Random Forest Models.

Both models used similar features, the only difference being RF favoring Pick-up Location ID's over GBT's Congestion Surcharge as it's 4th highest ranking predictor. What is more interesting is the distribution of importances'. The GBT model put heavier emphasis on Trip Distance (0.7513%) against the RF model's (0.5394%) importance. The correlation between trip distance and fare amount, combined with GBT's higher weighting could help to explain its out performance of RF.

4.1 Prediction Error Analysis

In order to understand the errors in prediction and the model, it's predictions were visualised against their true values (Figure 5).



(a) insert caption

(b) Predictions vs. True Fare Amount's Boxplot

Figure 4: (Left) Histogram of Gradient Boosted Tree model's predictions and the true Fare Amounts for 2021 and respective box plots (Right)

When visualising the GBT predictions versus the true fare amounts, it is worth noting the differences in concavity of the distributions. Unlike the true fare value distribution, the predictions do not follow a smooth slope, and instead display significant fluctuations, above and below, over and under estimating around the true values.

(Figure 4)(b) gives further insight to the limitations of the model. The whiskers of the box plots show that their is a deficiency in predicting large fare amounts. That is, at the tails of the data the model underestimates the true price. It is suggested that this could be addressed by collecting more data about these regions in turn allowing the model to make better informed predictions at the extremes. Further, transformations could be applied to the model inputs to regularize their behaviour.

5 Discussion

To reiterate, the aim of this report is to propose a method for forecasting taxi fare amounts in order to help the NYC Yellow Taxi overcome a significant disadvantage it has against HVFHV's. Considering the predictive power the models were able to achieve, and their general simplicity, this should be encouraging for NYCTLC or any offline for-hire-vehicle-service to pursue implementing or investigating this technology.

If the NYCTLC were able to incorporate this technology and perhaps extend into the digital booking fields, they could potentially create a new style of taxi driver. One who harmoniously moves between 'on-the-fly' pickup's and digitally requested rides.

If this field of interest was to be extended, it is highly recommended to broaden the range of data. Space constraints limited the abilities for significant processing and training. Further outlier analysis, noise reduction and aggregation would be necessary, especially regarding older data sets.

Further improvements would have to be made in understanding the data sets distribution. Perhaps a GBT model, although generally applicable, would be outperformed by a model that more accurately fits the distribution and as a consequence can correctly approximate the tail behaviour. The tail behaviour is a significant pitfall in the models predictive capabilities.

Also, the GBT and RF Models would both benefit from hyper-parameter tuning. Due to the complexity of the data set, the computational costs of performing a Grid Search type operation could not be completed, but it would no doubt help to consolidate or improve results.

6 Conclusion

This report aimed to offer an insight into the feasibility of implementing an estimative service for fare amounts that could be appreciated by customers. Using NYCTLC trove of data regarding the trips by Yellow Taxi's, an in-depth analysis and learning process was able to showcase the power of modern decision tree algorithms, where they may under perform and explore their potential use cases.

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: 2022-08-01.

[2] Iowa State University. *New York ASOS METAR reports (New York City).* https://www.mesonet.agron.iastate.edu/request/download.phtml?network=NY_ASOS. Accessed: 2022-08-01.

[3] New York City Taxi and Limousine Commission. *"TLC Yellow Taxi Data Dictionary"*. https://www1.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf. Accessed: 2022-08-01.

[4] New York City Taxi and Limousine Commission. *"NYCTLC Rule Book, Chapter 54, "* https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page. Accessed: 2022-08-01.

[5] New York City Taxi and Limousine Commission. *NYCTLC Fare Page.* https://www1.nyc.gov/site/tlc/passengers/taxi-fare.page. Accessed: 2022-08-01.

[6] Apache. *PySpark Documentation.* https://spark.apache.org/docs/latest/api/python/. Accessed: 2022-08-01.