Analysis and Forecasting of Trip Duration for High Volume For-hire Services In NYC

Luyao Chen
Student ID: 1266572
Github repo with commit

August 21, 2022

1 Introduction

The car sharing industry has grown rapidly in recent years, especially in large cities with traffic congestion and limited parking spaces. Study shows that in 2019, the "ride-sharing giant" Uber provided an average of 250,000 more rides per day than yellow taxis in New York City (NYC) [1]. However, with increasing demand of for-hire services, more and more customers start to complain about the inaccurate estimations of arrival time.

This report will focus on exploring the association between trip duration and other factors such as weather, trip distance, pickup location and destination. We aim to assist ride-sharing companies like Uber and Lyft to better predict the arriving time of a particular trip in NYC. Statistical model such as Generalised Linear Model and machine learning model like Random Forest are used in the modelling section.

2 Dataset

This report mainly utilises high volume for-hire vehicles trip data published by the New York City Taxi and Limousine Commission (TLC) [2]. This dataset contains trip records of TLC-licensed for-hire vehicles dispatched by NYC's High-Volume For-Hire Services (HVFHS)[1]. Before February 2019, TLC only received and published pickup and drop-off data/time and location for FHV trips. Given the limited information available up to 2019, we decide to uses 6 months of data from February 2019 to July 2019 inclusive, with over 129 million records and 24 columns.

The HVFHS dataset includes detailed information about license number, pick-up and drop-off dates/times, pick-up and drop-off locations, as well as trip fares. Among the 24 attributes, irrelevant features such as "driver_pay" and "shared_request_flag" are discarded as they are assumed to have no obvious relations with trip duration. Finally, the following attributes are chosen from HVFHS dataset as predictors:

- Pickup date/time
- Pickup location ID
- Drop-off location ID
- Trip distance (mile)
- Congestion surcharge (USD)

[1] HVFHS refers to FHV businesses that dispatch more than 10,000 FHV trips within NYC per day. The companies included in this dataset are Uber, Lyft, Via, and Juno.

As it is considered that extreme weather may result in longer trip duration, a daily NYC weather report recorded by the Weather Underground [3] is incorporated into the study. The weather data is collected from a weather station at Laguardia Airport and contains the following features:

- Average temperature (°F)
- Dew Point (°F)
- Humidity (%)
- Wind Speed (mph)
- Pressure (inch)
- Precipitation (inch)

3 Preliminary Preprocessing

Due to the high consistency and high completeness of the weather data, no preprocessing steps are required. However, a variety of data cleaning and feature engineering techniques are applied to the FHV trip data. It is expected that by removing outliers and dealing with missing values, the model produced can be more feasible to generalise on unseen data.

3.1 Data Cleaning

- **Trips with unrealistic trip duration.** Trips that take more than 5 hours or less than 2 minutes are removed.
- **Trips with zero trip distance.** A total of 60,178 records are discovered to have zero "trip miles" which indicates that the vehicles have not moved at all and thus, these trips are discarded from the data.
- **Trips with location ID 264 or 265.** According to the "Taxi Zone Lookup Table" published on the TLC website[2], location ID outside the range 1-263 suggests an "unknown" pickup or drop-off location. As this report only focuses on FHV trips within NYC, these records are considered as outliers and removed from the dataset.
- **Missing values in "congestion surcharge".** After removing outliers, there are 498,928 instances with Null values in "congestion surcharge" attribute. Considering that *Spark ML* cannot handle Null values and these records only form a small proportion of the whole dataset, it is decided to filter them out.

3.2 Feature Engineering and Selection

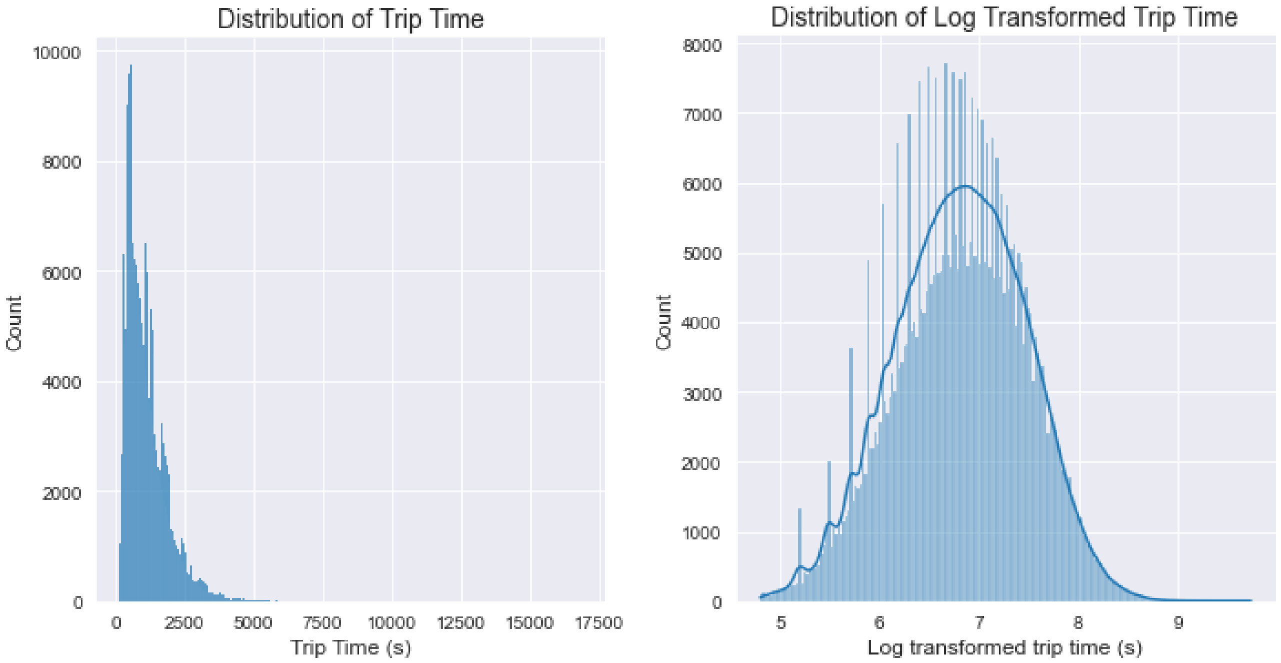
- **New attributes extracted from "pickup datetime".** Since travel time may be affected by different traffic conditions during different times, new features such as "day of week", "hour of day" are created based on pickup date.
- **New binary attribute "congestion zone".** From the website of Department of Taxation and Finance of New York State[4], the congestion surcharge applies to vehicles that begins in, ends in, or passes through the *congestion zone* [2]. The amount of surcharge depends only on the type of vehicle which is not a significant factor of predicting trip duration. It is therefore replaced by a binary attribute which indicates whether or not the trip passes through the congestion zone.
- **Merging FHV trip data with weather data based on pickup date.**

[2] The Congestion Zone is the borough of Manhattan Manhattan, south of and excluding 96th Street.

4 Analysis

To obtain a deeper understanding of the dataset and the correlations between attributes, we investigated the distribution of the target variable "trip time" and its relationships with individual features. The dataset is partitioned into training (Feb 2019 - May 2019 inclusive) and test (Jun 2019 - Jul 2019 inclusive) sets. To prevent data leakage, all feature analysis and selection processes in this section use only the training data. In addition, for visualisation purpose, some plots use 0.5% of data (422,831 instances) randomly sampled from the training set.

4.1 Trip Duration



(a) Before Log Transformation

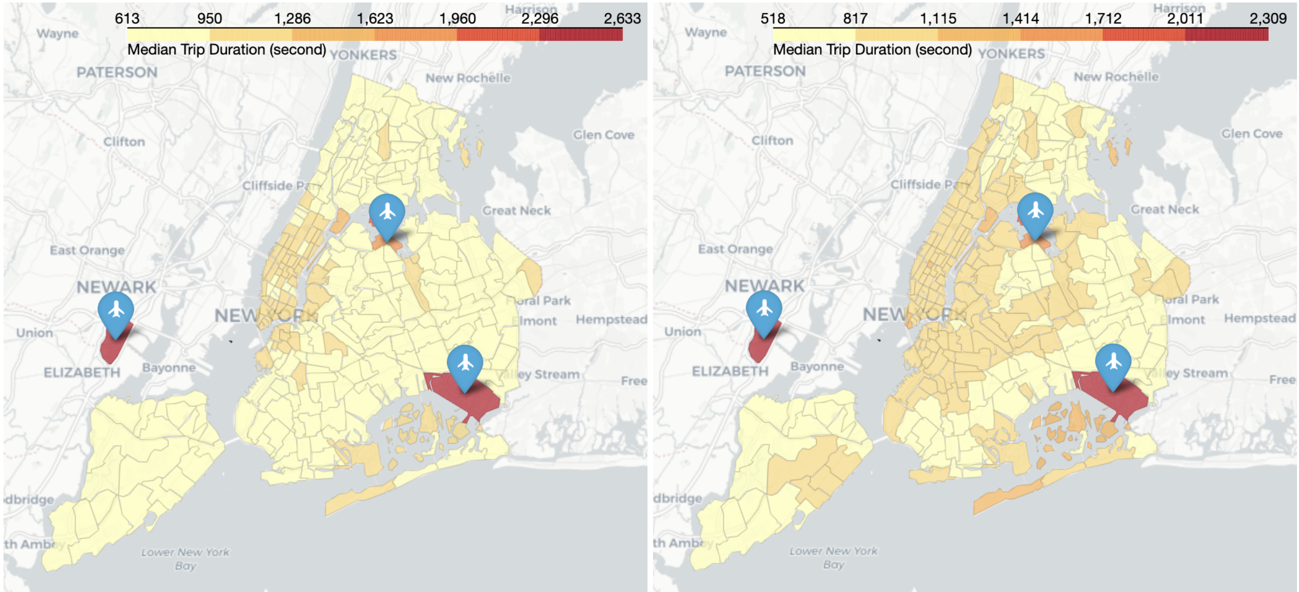
(b) After Log Transformation

Figure 1: Distribution of Trip Time Before and After Log Transformation

Figure 1a demonstrates that the response variable trip time is positively skewed, with majority of trips being less than 2500 seconds. To adhere the normality assumption of linear models, log transformation is applied. As shown in Figure 1b, the distribution of trip time after transformation is close to normal distribution.

4.2 Location ID

After examining the distribution of the target variable, we investigate the relationship between pickup/drop-off location and trip duration using geospatial visualisations. Median trip time in each taxi zone is plotted to avoid the effect of extreme values in trip duration. From Figure 2, it can be seen that trips start at or end in the three airports usually take longer time. In addition, trips in Manhattan and its surrounding area generally have longer duration. These findings are in line with our expectations. Since airports are generally far away from the city, and Manhattan is the economic centre of New York City and densely populated, it is reasonable to see longer journeys in these areas.



(a) Pickup

(b) Drop-off

Figure 2: Median Trip Duration for FHV in Different Pickup and Drop-off Locations

Due to limited computing power in the modelling section [3], categorical features "pickup location ID" and "drop-off location ID" are simplified and replaced with four binary attributes: JFK trip, Newark trip, LaGuardia trip and Manhattan trip.

4.3 Discrete Variables

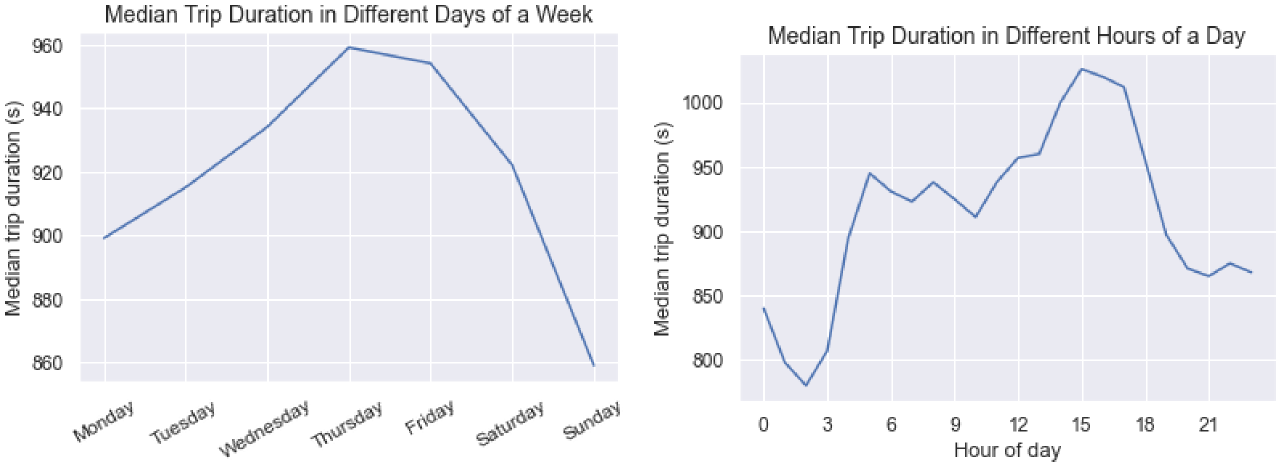
In Section 3.2, "day of week" is extracted from "pickup datatime". Also, a categorical attribute is created to indicate whether the vehicle has travelled through the congestion zone of NYC. In exploring the relationship between each discrete feature and the response variable, we use ANOVA to test the significance of them in predicting trip duration. Note that log transformation is applied to trip time in order to satisfy the normality assumption for ANOVA.

	Sum of Squares	DF	F-value	p-value
Day of week	4.97e+02	6	2.04e+02	2.34e-260
Congestion zone	9.23e+03	1	2.27e+04	0
Residuals	1.72e+05	4.23e+05		

Table 1: ANOVA for Categorical Attributes

As shown in Table 1, both features are identified to be highly correlated with trip time, having p-values extremely close to zero. Moreover, line plot for median trip duration in different weekdays is displayed in Figure 3a. It suggests that trips on Thursdays or Fridays usually take longer, while trips on Sunday have the lowest median duration. This is probably due to the fact that traffic is lighter on weekends, so vehicles are able to travel faster.

[3] When converting categorical features to numerical features through One-Hot Encoding, pickup/drop-off location will result in more than 500 features.



(a) Day of Week

(b) Hour of Day

Figure 3: Median Trip Duration for FHV by Weekday and Hour

4.4 Continuous Variables

To further investigate the correlation between each continuous predictor and trip time, Pearson correlation coefficients are computed. It can be seen from Figure 4 that trip miles has a strong positive relation (0.78) with trip time, while other features such as dew point, humidity, wind speed, pressure and precipitation are presented as trivial. We can also observe that temperature and dew point are highly correlated (0.87) which suggests that dew point might not make any additional contribution in interpreting the response variable. Upon examination of the correlation heatmap, we decide to drop the attributes that has little or no relation with trip time.

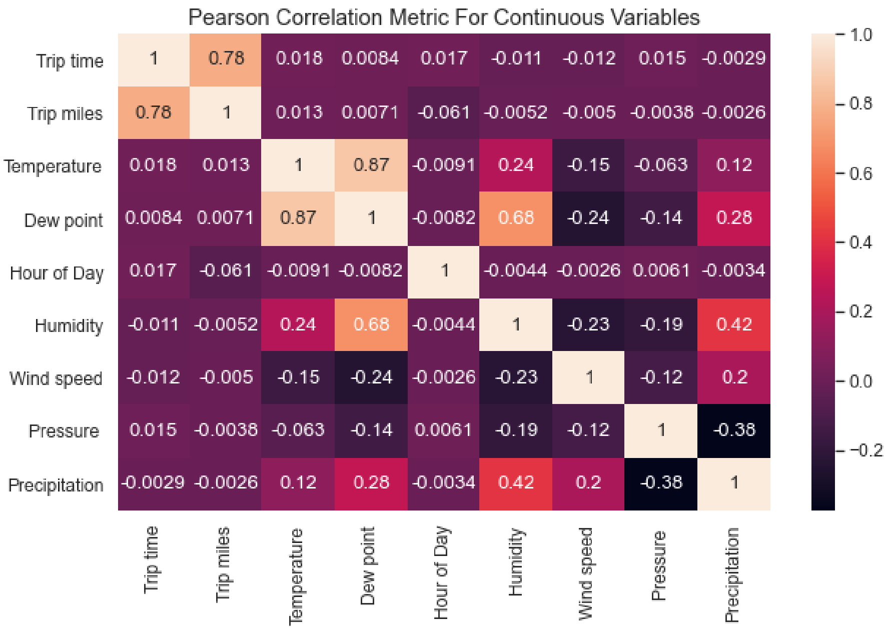


Figure 4: Pearson Correlation Matrix for Continuous Attributes

In common sense, travel times are usually longer during peak hours, and shorter in early morning.

However, trip time and “hour of day” exhibit an extremely low Pearson correlation of 0.017 (see Figure 4). This may be due to Pearson correlation’s inability to detect non-linear relationships. For this reason, we plot the median trip duration at different times of the day (Figure 3b), which illustrate a clear correlation between the two variables.

5 Modelling

After analysing the correlation between each feature and trip time, the following attributes are retained in the dataset:

- Day of week
- Hour of day
- JFK trip
- Manhattan trip
- Newark trip
- LaGuardia trip
- Congestion zone
- Temperature
- Trip miles

Before feeding the training data into the model, all discrete attributes are converted to numerical data via one-hot encoding, resulting in 14 individual features. In total, there are 84,626,689 instances in the training set and 39,881,431 instances in the test set.

5.1 Methods

5.1.1 Gamma Generalised Linear Model

Generalised Linear Model (GLM) is an extension of ordinary Linear Regression that it is able to predict response variables with non-normal distribution. In particular, Gamma Generalised Linear Model (Gamma GLM) is useful for predicting a non-negative, continuous and positive-skewed response [5]. Recall that the distribution of trip time (shown in Figure 1a) is very similar to a gamma distribution and hence, a Gamma GLM with log link is utilised in this report.

The hyperparameter to be tuned is the regularisation parameter λ which controls the amount of regularisation applied to the model and is used to reduce overfitting problems. Due to limited computing power and relatively large training set, we cannot perform cross validation in hyperparameter tuning process. Instead, several λ values are chosen and compared to obtain the optimal model.

5.1.2 Random Forest Regression

Random Forest Regression (RF) is an ensemble model that builds multiple decision trees and aggregates outputs of these base models. As base models are assumed to have uncorrelated errors, the aggregation of results is expected to effectively reduce model variance. Comparing to other ensemble models such as Adaboost, Random Forest can detect interactions and is more robust when encountering outliers and noise in the dataset [6]. Since decision trees usually work well with categorical attributes, we expect this model to have good performance on our dataset.

The hyperparameters of RF are number of trees and maximum depth of each random tree. Multiple RF models are trained with different number of trees and evaluated in Section 5.2.

5.1.3 Evaluation Metrics

The metrics chosen for model comparison and evaluation are R^2 and Root Mean Squared Logarithmic Error (RMSLE). R^2 score indicates the proportion of variation in the response variable that can be explained by predictors, while RMSLE measures the error in predictions. Thus, we prefer the model with a large R^2 score and small RMSLE value. RMSLE is used since it focuses on the relative error between predictions and actual values instead of absolute error. The wide range of target variable "trip time" (in seconds) and the fact that RMSLE penalises underestimation more than overestimation make it a good choice for assessing model performances.

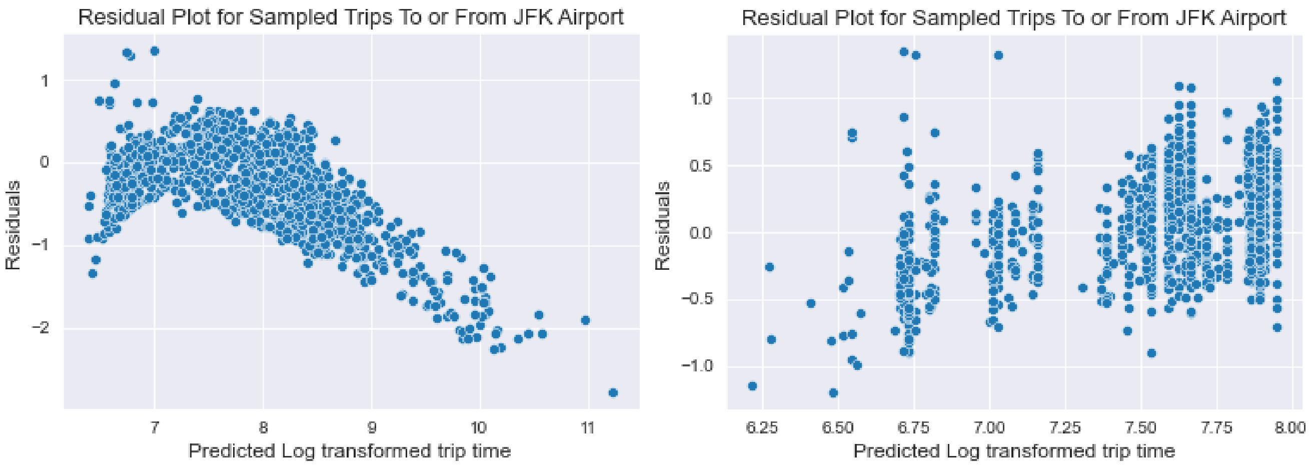
5.2 Results and Discussion

Table 2 presents the R^2 score and RMSLE for each model in hyperparameter optimisation. Through comparison of evaluation metrics computed, we can identify an increase in the GLM model performance as the regularisation parameter λ increases. For Random Forest, the model achieves the highest R^2 and lowest RMSLE when there are 5 random trees.

	GLM						RF			
	regularisation parameter						number of trees			
	0.01	0.1	0.3	0.5	0.7	0.9	3	5	10	12
R^2	0.4265	0.4534	0.4551	0.4562	0.4578	**0.4812**	0.7003	**0.7262**	0.7133	0.7208
RMSLE	0.4958	0.4841	0.4833	0.4828	0.4821	**0.4598**	0.3585	**0.3426**	0.3506	0.3460

Table 2: Hyperparameter Tuning Results

Comparing optimal results obtained by GLM and RF models, it is unsurprisingly that RF outperforms GLM in predicting trip duration with a R^2 score of 0.7262 and RMSLE of 0.3426. To further compare the investigate any error made by these two models, samples of test data are taken from trips involving JFK Airport (the busiest airport in NYC).



(a) Gamma GLM with Log Link

(b) Random Forest Regression

Figure 5: Residual Plots for GLM and RF Models with Sampled Trips To or From JFK Airport

From Figure 5a, we can clearly see a decreasing trend in residuals and notice that most of the data points are clustered in an inverted-U-shaped pattern. This indicates that trip time and chosen attributes may not be linearly associated and thus, a non-linear model might be more appropriate for this dataset. Furthermore, the large number of data with negative residual values suggests that GLM tends to overestimate trip duration in this particular case. In residual plot of the Random Forest model (Figure 5b), points are randomly scattered around zero and no clear pattern can be observed. This demonstrates that the RF model does a great job in capturing the deterministic component and dealing with non-linearity of the data. Therefore, it is proved to be a better method of modelling the relationship between features and trip duration. However, in real life applications, it should be taken into account that it takes approximately twice as long to train a Random Forest Regression model as a GLM model.

6 Conclusion

This report depicts that trip duration is highly correlated with pickup and drop-off location, day of week, trip distance and whether the vehicle has passed through the congestion zone. However, there exists interdependence between features and non-linear relationships with the target variable (see Figure 3b and 4). Comparing the two models fitted using the HVFHS data, it can be seen that Random Forest which has no distributional assumptions demonstrates a superior performance in predicting trip duration with a R^2 score of 0.7262 and RMSLE of 0.3426. As a result, we recommend companies in the ride-sharing industry to take into account the presence of non-linearity when building or refining their trip duration predictors. In addition, a Random Forest Regression model is suggested as it is robust to overfitting and performs well even on a small feature set.

It is also recommended that ride-sharing companies could use more data to produce an improved version of this model which allows users to choose a preferred departure time and utilise historical data to estimate travel time in the future. This model is able to help passengers to schedule their future trips more easily, improving service quality. Also, by adding historical data regarding demand and driver availability in a certain area at a particular time, the accuracy of predictions are expected to increase significantly.

Unfortunately, the weather dataset used in this report only contains daily weather conditions which is not specific to trips in a particular hour or minute. It is recommended that more detailed dataset, such as hourly weather reports, should be used in future studies. Additionally, more features could be explored, for instance, national holidays, real-time road and traffic conditions as well as large events. Another possible improvement would be to apply cross validation when adjusting the hyperparameters to obtain a more reliable result.

References

[1] Todd W. Schneider. *Taxi, Uber, and Lyft Usage in New York City*. `https://toddwschneider.com/posts/taxi-uber-lyft-usage-new-york-city/`. Accessed: 2022-07-26.

[2] New York City Taxi and Limousine Commission. *TLC Trip Record Data*. `https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page`. Accessed: 2022-07-28.

[3] Weather Underground. *New York City, NY Weather History*. `https://www.wunderground.com/history/monthly/us/ny/new-york-city`. Accessed: 2022-08-05.

[4] Department of Taxation and Finance. *Congestion surcharge*. `https://www.tax.ny.gov/bus/cs/csidx.htm`.

[5] Xin Chen, Aleksandr Y Aravkin, and R Douglas Martin. "Generalized linear model for gamma distributed variables via elastic net regularization". In: *arXiv preprint arXiv:1804.07780* (2018).

[6] Leo Breiman. "Random forests". In: *Machine learning* 45.1 (2001), pp. 5–32.