Demand Estimation for Trips from JFK Airports

New York Yellow Taxi

Patrick Lourenz
Student ID: 1470744
GitHub repository

August 21, 2022

1 Introduction

Yellow taxis have been very closely acquainted with the city of New York. However, New York City Taxi and Limousine Commission (TLC) has been struggling since ride-sharing like Uber and Lyft disrupted the marketplace. It is evident with the decreasing taxi cab medallion licenses. There were 51,398 active taxi licenses as of March 2014. As of July 2016, this number has dropped significantly to 13,587 [1].

Substantively, taxis still do have an advantage over ride-share services. When demand is higher than supply for ride-share services, this can benefit taxis as there is a high chance that the price will be more affordable to take a taxi. Uber and Lyft, utilise demand spikes to formulate dynamic surge prices whereas NYC taxis do not have a surge pricing mechanism [2]. Therefore, TLC needs to be able to meet demand at certain time and strategic location, such as JFK Airport.

This report will analyse and predict from the viewpoint of a taxi company. As the data reflect users' behaviour, the aim is to assist this company to understand the marketplace behaviour in JFK Airport and be able to fulfil supply in specific hours utlising the estimated demand. The Prediction used a Multiple Linear Regression model, as a base model [3], and a Random Forest Regression model [4]. Random Forest Regression was used as it is expected to be able to capture complex decision boundaries and relationships between features through multiple decision trees majority voting.

2 Dataset

2.1 Yellow Taxi Trip Record

This report fundamentally used the monthly **Yellow Taxi Trip Record** dataset released by New York TLC from **2018-2019**[5]. In March 2020, New York City went into lockdown. This certainly impacted taxi demand negatively. Therefore, this report did not include trip data from 2020 and beyond. Green Taxi Trip Record was not included in this report as they are not allowed to pick up passengers from airports.

Initially, this dataset contains 187,469,831 rows and 19 columns. However, after cleaning and selecting relevant features dataset was reduced to 10,308,204 rows and for the purpose of this report, five columns were selected; **Pick-up Date and Time, Drop-off Date and Time, Pick-up Location ID, and Drop-off Location ID**.

2.2 JFK Airport's hourly observed weather data

Supporting the trip dataset, external datasets were also included. The first dataset was JFK Airport's hourly observed weather data from 2018-2019 retrieved from United States National Centers for Environmental Information (NCEI)[6]. This data has 27,170 rows and for the purpose of this report seven columns were selected; **Date and Time, Temperature, Dew Point, Precipitation, Wind Direction, Pressure, and Visibility**.

2.3 New York Public School Holiday

Another external dataset included was the New York Public School Holiday from 2018-2019 retrieved from the New York Department of Education website [7]. Hypothetically, the school holidays may increase airport pickup demand as more people travel to New York for vacation purposes, assuming most cities have the same holiday dates. Moreover, approaching the end of a long school holiday, *New Yorkers* will return to their city. This, expectedly, will also increase airport pickup demand.

3 Preprocessing

3.1 Yellow Taxi Data (Outlier Detection)

Out of range pick-up dates
There were trips with pick-up date outside of the selected year 2018 and 2019. These data were filtered out.

Negative trip duration
There were trips where pick-up date timed after the drop-off date, intuitively this is not possible. These data were removed.

Trips with duration more than 5 hours
A trip from JFK, LaGuardia or Newark Airport to the furthest point destination in New York during peak afternoon hours is estimated by Google Maps to take 1.5 - 2 hours. According to the New York TLC, drivers are not permitted to refuse passengers with more than one stop (taximeter will run as one continuous trip). Therefore, trips that have more than 5 hours (3 hours added to assume multiple stops trip) in trip duration were eliminated.

Only trips from airports
As the report aim is to estimate demand from airports, only trips from Newark, JFK, or LaGuardia Airport were selected.

Check for null values
After previous filter steps taken, there were no more null values present in the taxi data.

3.2 Weather Data

Imputations for missing data
According to the weather dataset documentation, there are consecutive 9s (i.e. 999.9, 999, 999.0) in the dataset and these were missing values. Therefore, these data were converted to null values.

Imputations for null values
There are null values across all weather features with precipitation being the most (6,214 rows). Null values were replaced using the median value within a day's 6 hours time frame; 12 AM - 6 AM, 6 AM - 12 PM, 12 PM - 6 PM, 6 PM - 12 AM.

4 Analysis

4.1 Airport is an important location for TLC

2018-2019	Proportion of Trips	Proportion of Total Amount
All Airports Pick-up	5.51%	15.58%
JFK Airport Pick-up	**2.81%**	**9.00%**
Other Airports Pick-up	2.70%	6.58%

Table 1: Proportion of Airport Pick-up Trips from all yellow taxi trips in New York 2018-2019

JFK Airport is a strategic location for TLC. This is due to trips from and to the airport being highly valuable. Table 1 describes the proportion of trips from JFK Airport proportion in terms of trip count and total amount. JFK Airport pick-up trip contributed to 2.81% of the total trips but contributed 9.00% to the total generated total amount. This explains that JFK airport trips, generally, are high average selling price trips (even higher than other airports).

4.2 Drop-off Locations

Figure 1 displays the drop-off locations of yellow taxi trips that started from JFK Airport. It appears that most trips end around the Manhattan area. This is as expected as Manhattan is a tourist destination and Lower Manhattan is the central borough for business. Moreover, TLC charges a flat rate of $52 for trips between Manhattan and JFK Airport. This could be the reason why people commuting to Manhattan prefer taxis to ride-share. Hence, the demand for a trip to Manhattan from JFK Airport is high.
Interestingly, LaGuardia and JFK Airport itself are also high drop-off locations from JFK Airport. This could be due to transit flights to LaGuardia Airport or between terminal commuting in JFK Airport.

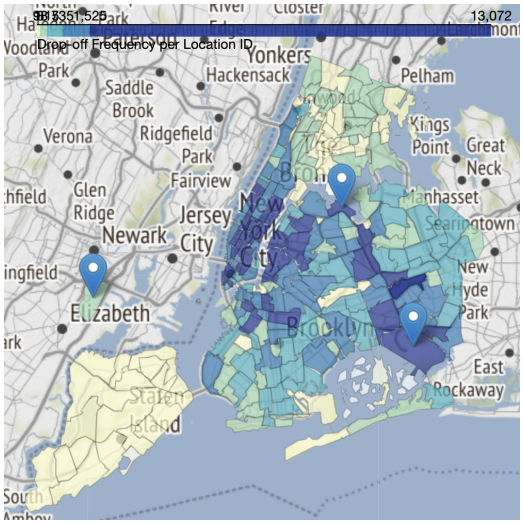

Figure 1: Drop-off Frequency per Location ID. Markers denote airport locations.

4.3 Correlation between features

There are clear evidence where there exists strong evidence that features are correlating with each other. For instance, in Figure 2, it shows a strong correlation between `temperature` and `dew_point`.

The p-value between these two features was also lower than 0.05, shown in Figure 3. Therefore, there is enough evidence that there is a correlation between these two features. These strongly correlated features will be assessed later before modelling.

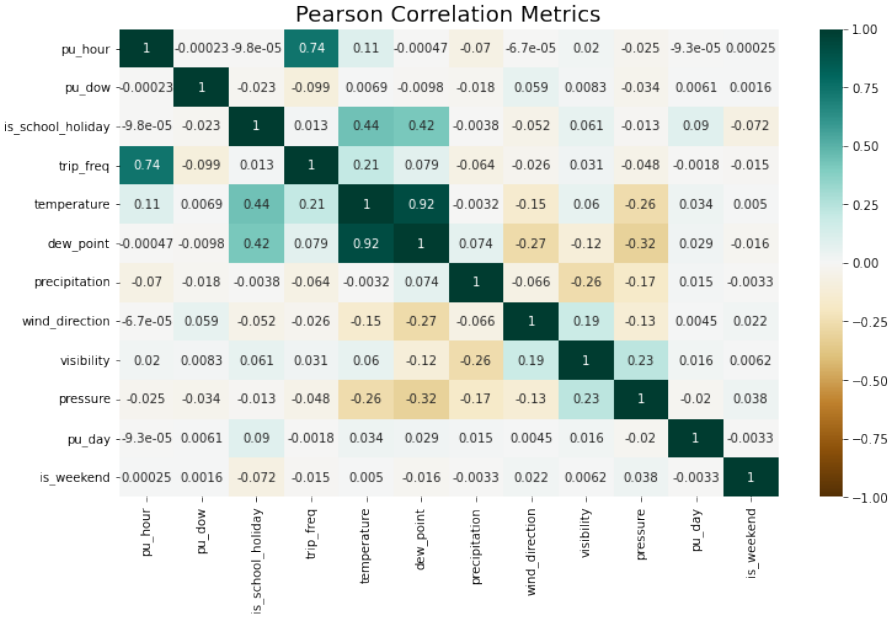

Figure 2: Pearson Correlation Heat-map between features

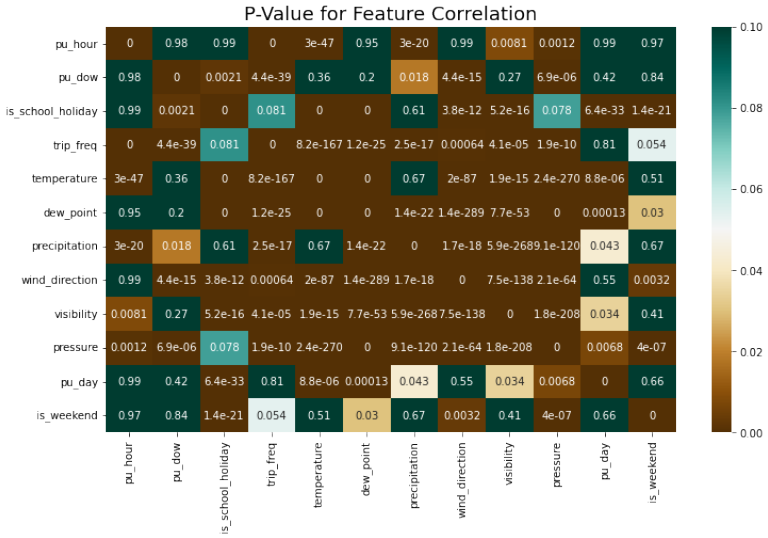

Figure 3: P-value for Correlation between features

4.4 Insight 1: Most airport pick-up ride passengers are business travellers

The hourly taxi pick-up rate was grouped by day of the week to further analyse the behaviour of users. Figure 4 shows that the median trip frequency shows that Saturdays, in general, have the lowest daily trips, this insight may show that most airport pickup trips are business travellers. Sundays and Mondays are exceptionally high as this day, most business travellers come to start their working week.

Tuesday and Wednesday, demand drop is as expected as not many business travellers travel during in the middle of a working week.

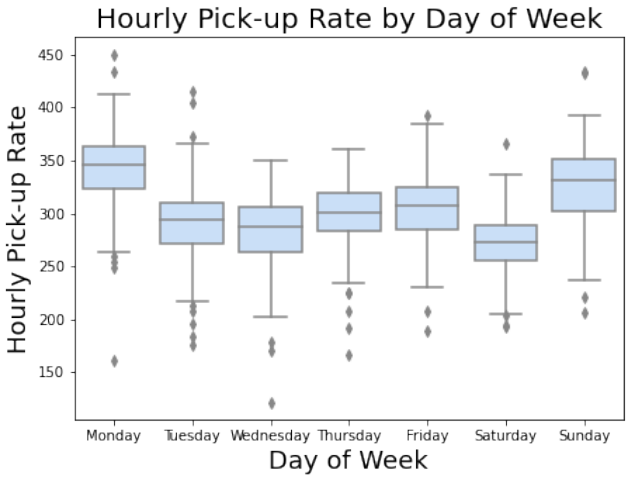

Figure 4: JFK Airport Pick-up Hourly Rate by Day of Week

4.5 Insight 2: Airport pick-up demand peak times in a day

Figure 5 captured the demand drops and peaks within the hours in a day. Lowest JFK Airport pick-up demand is during graveyard hours, 1 am - 5 am, which is as expected. In contrast, demand spikes multiple times; in the morning (6-7 am), afternoon (2 pm - 5 pm), and night (8 pm - 9 pm).

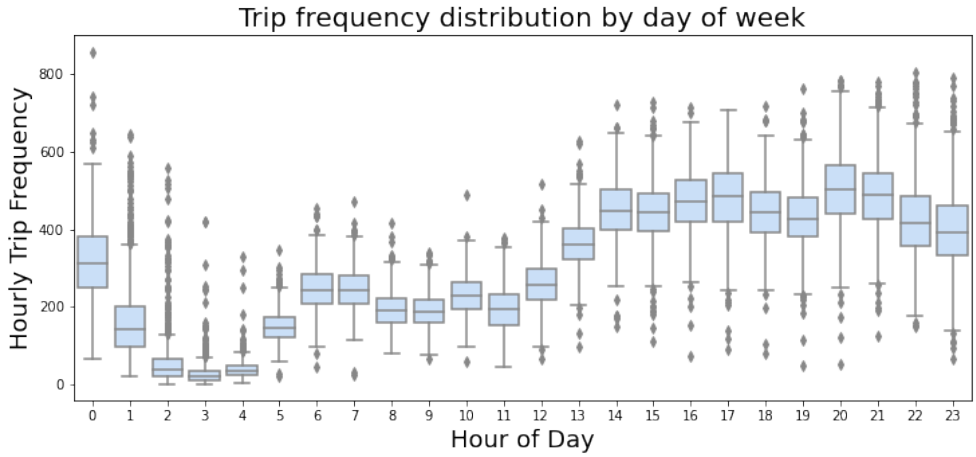

Figure 5: JFK Airport Pick-up Hourly Rate

4.6 Insight 3: School holidays do impact airport pick-up demand

School holidays may impact positively or negatively airport pick-up demand. In Figure 6, demand was compared between December 25, 2018, and the previous 2 weeks. The result showed that overall demand decreased throughout the day, but spiked at 11 pm.

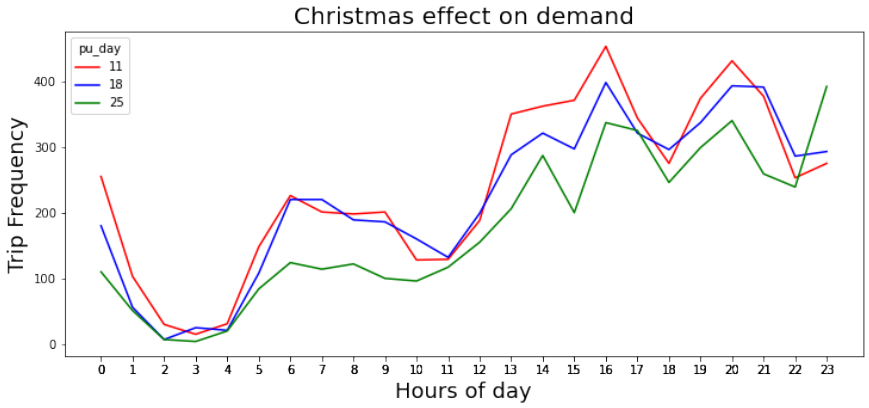

Figure 6: Christmas Day impact on JFK Airport taxi pick-up demand

Figure 7 visualises when a longer holiday period, such as summer break, was compared to *non*-summer break days. The result was summer break increased airport pick-up demand overall.

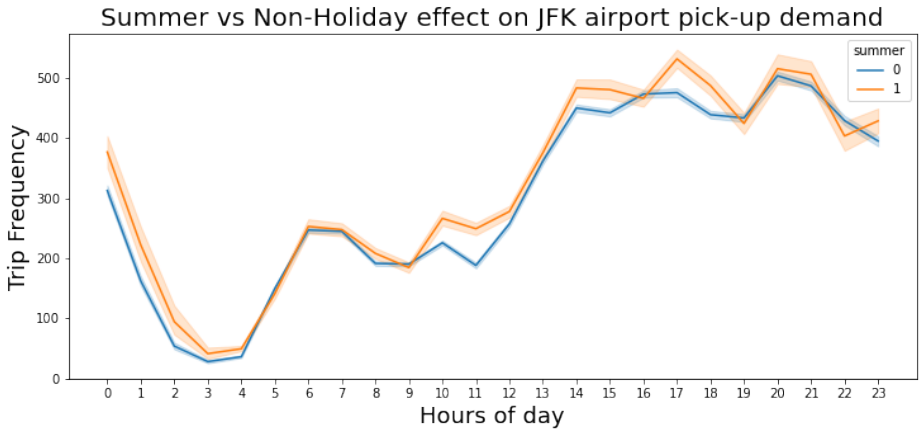

Figure 7: Summer break impact on JFK Airport taxi pick-up demand

5 Data Aggregation, Feature Selection and Engineering

Prior to statistical modelling, taxi dataset are grouped by the date and hour and a new feature, `trip_frequency`, aggregates the record count. Next, weather data was merged using inner join to eliminate missing hours. `is_school_holiday` boolean feature was also added to identify if the pick-up date was a school holiday. Following the correlating features; `temperature` and `dew_point` from the Analysis section, two ordinary least squares (OLS) regressions were fitted to compare which model has the lowest Akaike Information Criterion (AIC) [8]. The result was the model without `dew_point` has the lowest AIC. Therefore, `dew_point` was eliminated.

Ordinal features were one-hot encoded and numerical features were standardised to avoid bias introduced to the model.

6 Statistical Modelling

Test data used was a yellow taxi, weather and school holiday data from January 2020. It was pre-processed following the same steps as the train data.
Metrics used to assess the model performance are root mean squared error (RMSE), mean absolute percentage error (MAPE), and R^2 score. RMSE explains how far off, in terms of the number of trips, the model predictions to the actual target value, while the MAPE described the error in terms of percentage from the actual value. R^2 score showed the amount of variability in the data that can be explained by the model.

6.1 Multiple Linear Regression (baseline model)

Multiple linear regression enables the model to assess multiple features by conducting linear regression with each feature and the target feature which follows this equation:

$$Y_i = f(X_i, \beta) + e_i$$

The model resulted in an RMSE of 71.48 trips, a MAPE of 30.68%, and R^2 score of 0.8061. These metrics will be used as a baseline for the Random Forest Regression. Considering this linear regression may not capture the complex relationship between features, a R^2 score of 0.8061 was a promising sign.

6.2 Random Forest Regression

Random forest regression is a supervised machine learning method that ensembles $\boldsymbol{n}$ decision trees on various sub-sample of the dataset. A majority vote was then taken from all trees to generate a predicted result.

There are two hyper-parameters; `max_features`, `n_estimators`. These were tuned by selecting the lowest RMSE and highest R^2 score. `max_features` and `n_estimators` were set to 18 and 300 respectively.

Random forest regression resulted in an MAE of 66.13 trips, MAPE of 26.17% and R^2 score of 0.8344. These metrics were understandably much better than the linear regression results.

Figure 8 shows that on most dates, the difference between predicted and actual demand was not far off. The demand spike that the model could not predict was for the 2nd, 5th and 20th of January. It appears that there was an NHL sporting event on the 5th and 20th date, while on the 2nd there was an NBA game. This was further supported with Figure 10 showing the model underestimated 3 consecutive hours, 6 pm - 8 pm. This was when the NHL game began.

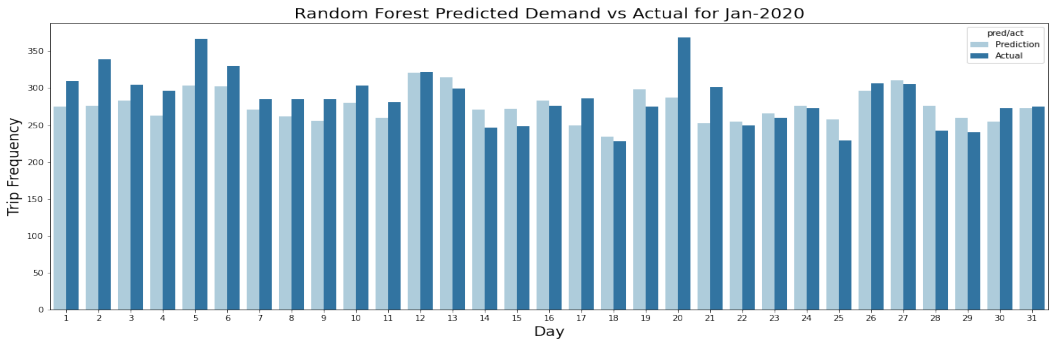

Figure 8: Random forest regression predicted demand for January 2020

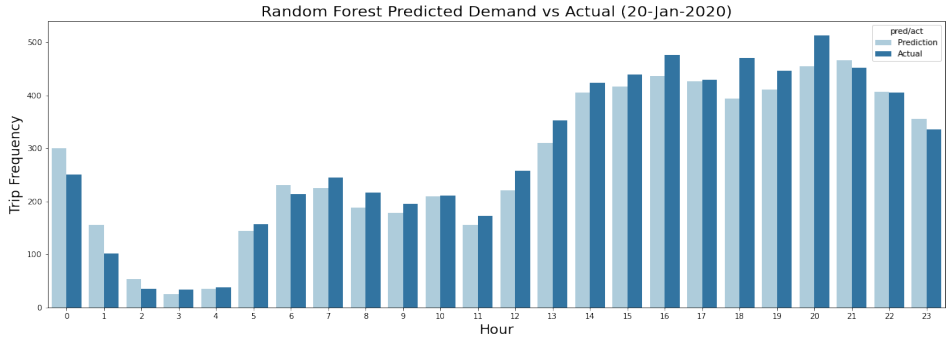

Figure 9: Random forest regression predicted demand compared to the actual trips at January 20, 2020

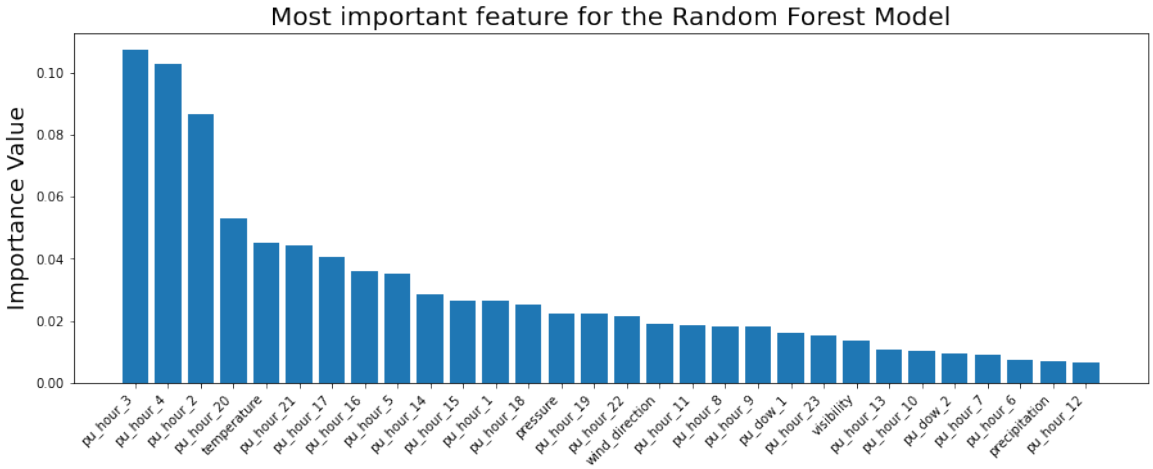

Figure 10: Important Features according to the Random Forest model

7 Recommendations and Conclusion

This report has presented how JFK Airport pick-up demand can be estimated using Linear Regression and Random Forest Regression method utilising New York Yellow Taxi dataset, JFK Airport weather dataset, and New York School Holiday. The model error may be slightly too high for TLC to implement this for their decision-making as it has an MAE of 66.13 trips and a MAPE of 26.17%. In future implementations, more effort should be made in cleaning data outliers for both weather and taxi. Furthermore, random forest regression performance may need to be compared with other machine learning models that are more sophisticated in terms of understanding the data, such as deep neural networks.

There were demand spikes that the model could not explain, as seen in Figure 8. Most of these dates were where major sporting events were played. Future implementations could add New York major sporting events dates data, such as NHL and NBA.

Moreover, it can be interesting to investigate how yellow taxi demand influences ride-share demand and vice versa. Moreover, looking deeper into how dynamic price surge ride-share services in conjunction with demand estimation can be beneficial for the New York TLC. This will help them to know where to direct supplies to a specific location at a specific time where taxi price can win over ride-share.

References

[1] Doren, Peter Van. *Should Taxi Medallion Owners Be Compensated?* https://www.cato.org/blog/should-taxi-medallion-owners-be-compensated. Accessed: 2022-08-21. 2014.

[2] Akhtar, Allana. *As Uber and Lyft fares surge, NYC taxis are becoming popular again.* https://www.businessinsider.com/e-hailing-taxis-curb-increasing-faster-than-uber-lyft-nyc-2021-8. Accessed: 2022-08-21. 2021.

[3] T. L. Lai, Herbert Robbins and C. Z. Wei. *Strong Consistency of Least Squares Estimates in Multiple Regression.* Proceedings of the National Academy of Sciences of the United States of America (Vol. 75, No. 7, pp. 3034-3036). 1978.

[4] Ho, T. K. *Random decision forests.* In Proceedings of 3rd international conference on document analysis and recognition (Vol. 1, pp. 278–282). 1995.

[5] New York Taxi and Limousine Commission. *TLC Trip Record Data.* https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page. Accessed: 2022-08-21. 2018-2019.

[6] National Centers for Environmental Information. *Integrated surface dataset.* https://www.ncei.noaa.gov/access/search/data-search/global-hourly. 2018-2019.

[7] New York Department of Education. *2018-19 School Year Calendar.* https://infohub.nyced.org/docs/default-source/default-document-library/school-calendar-2018-2019. 2018-2019.

[8] Akaike, H. *Canonical Correlation Analysis of Time Series and the Use of an Information Criterion.* https://doi.org/10.1016/S0076-5392(08)60869-3. 1976.