Historical Patterns are All You Need: Efficiently Forecasting Hourly Taxi Demand in New York City

Hanshi Tang
Student ID: 1266337
Github repo with commit

September 1, 2024

1 Introduction

Traditional taxi drivers in New York City (NYC) face unprecedented challenges since the rapid growth of High-Volume For-Hire Services (HVFHV) has drastically changed urban transportation patterns in recent years. According to the 2023 New York City Taxi and Limousine Commission (TLC) Annual Report, HVFHV has already occupied 70% of the total registered 115990 vehicle plates[1]. Yellow and green taxi service providers must raise their attention to optimise the dispatching of the taxi fleet to meet the demands of passengers. This will improve their operation efficiency and remain competitive in the industry. The beneficiary of the report can be extended to the NYC transportation regulator as a data-driven solution to improve passengers' satisfaction.

We have selected the **Yellow and Green Taxi Trip Record data published by NYC TLC from 2023-06-01 to 2023-12-31** to conduct our analysis. These datasets contain pickup locations, dates and times information segregated by taxi zones. The 6-month time-frame has also covered the 3 seasons from summer to winter, which counteracts any single seasonal effect on taxi demand. We could also observe how major holiday events such as Christmas impact taxi demand.

The external datasets used for supplementary analysis are **NYC Permitted Event Information – Historical from NYC Open Data**, and Past **weather dataset collected at Central Park, Manhattan by the National Centres for Environmental Information**[2]. The assumption made here is weather forecasts should be mostly consistent across the city, and especially the Manhattan borough occupies a central location. Note we have attempted to analyse the most up-to-date datasets in 2024, but the potential implication of sparsity of the events dataset on modelling outcomes forced us to research based on 2023 data.

Cautious pre-processing including handling missing values and outlier detections was completed, and features of interest were preserved. We have ensured no significant amount of data is discarded by checking the percentage of missing data at each stage. The 3 datasets were joined on Date, Hour, and Borough for the modelling stage. We have conducted a geospatial analysis to visualise different demand patterns for each taxi zone across NYC.

The trip record datasets were integrated with the respective weather and events datasets, and we conducted an in-depth analysis of features such as Temperature, Visibility, Number of Events, and temporal features including hours and days of the week. These features provide a multi-dimensional understanding of taxi demand here quantified as the Hourly Number of Trips.

Linear Regression (LR) as a statistical model and **Gradient Boost (GB)** as a machine-learning model were deployed to identify the major driving forces for traditional taxi demand.

2 Preprocessing

2.1 Outlier Detection

The Yellow Taxi Dataset contains 18816606 rows, and the Green Taxi dataset has 381880 rows. There are in total of **19184860** rows, however, there are several issues in the dataset that either do not have support from the business rules or are inconsistent with our target data frame, below are how they get addressed.

Records out of date range are filtered out as they are irrelevant to our analysis.

Records with passengers fewer than 0 or more than 6 are filtered out since there should be at least 1 passenger and according to the Official Website of the City of New York, the maximum number of passengers should be 5 [3]. However, given the exemption of having a baby sitting on an adult's lap, the threshold becomes 6.

The Fare amount and Total amount less than $3 [3] are filtered out. Since the starting fare for NYC Yellow and Green Taxi is $3. For other fares including Tips, Extra, MTA Tax, Improvement Surcharge, and Congestion Surcharge, anything below 0 should be anomalies and be filtered out.

The minimum Trip distance is set at 0.5 miles and the Minimum Trip Duration is set at 1 min since the assumption here is that 0.5 miles should be within walking distance, and people are extremely unlikely to take a taxi given the congestion in America's largest metropolitan.

The maximum Trip Distance is set at 50 miles, and the Maximum Trip Duration should be below 120 minutes. This assumption comes from the estimation of the potential most extreme case of taxi driving within NYC. The Northern End of the Bronx to the Southern End of Staten Island is below 50 miles and it takes less than 120 minutes.

Anomaly Detection based on Statistics was also performed on the Trip Record Dataset, so any entries outside **the 99% percentile** are filtered out, as they indicate outliers. We could therefore avoid their potential influence due to the skewness they bring to the dataset.

After pre-processing for TLC datasets, there is **86.78%** of data remaining. They should be able to represent the mainstream patterns.

2.2 Feature Engineering and aggregation

Temporal Features have been feature-engineered to separate dates and hours for pick-ups and drop-offs. This provides significant flexibility for data aggregation, whether interested in daily or hourly demand trends. Hence, we have conducted the same transformation for all 3 datasets.

Due to the limitation of the events dataset, we do not have access to the exact geographical coordinates or location ID for historical events. Hence, we mapped the Trip Record's Location ID with its corresponding borough for the prediction of hourly demand in each borough. We also aggregated the records to get the Hourly Taxi pick-ups during each date, hour, and borough. The hourly pick-up is the quantified feature representing taxi demand. For the event dataset, we aggregated the hourly number of events that occurred also based on date, hour, and borough.

We finally merged the 2 external datasets with the Trip Record datasets on date, hour, and borough through left joins.

2.3 Imputations

When merging the datasets, there is 4% missing data recorded as 999.9 or a similar value from weather features including Temperature, Wind Speed, etc. We first complete a forward filling since weather values are continuous over time. We also imputed the data with averages if forward filling could not solve the issue. When joining the events data, we observed some sparsity in the dataset. The assumption here is that for some regions with very small populations and during late nights of the day, there should be no events. Hence, we imputed these values with 0.

2.4 Feature Selection

To address the research aim of hourly demand, most features related to fares were dropped, the following 8 features in **Trip Record dataset** are retained for further analysis:

- Pick-Up Location ID
- Pick-Up Borough
- Pick-Up Hour
- Drop-off Hour
- Drop-Off Location ID
- Pick-Up Date
- Drop-off Date
- Hourly Trip Count

The following 6 features in **Weather Dataset** are retained for further analysis, since most columns are basically filled with Null Values, hence we only include the primary ones:

- Ceiling
- Visibility
- Dew Point
- Wind Speed
- Temperature
- Sea Level Pressure

The following feature from **Events Dataset** is retained for further analysis:

- Hourly Number of Events

For weather and event datasets, their date, hour and borough features have merged with the Trip Record dataset, and now also been represented with pick-up/ drop-off features.

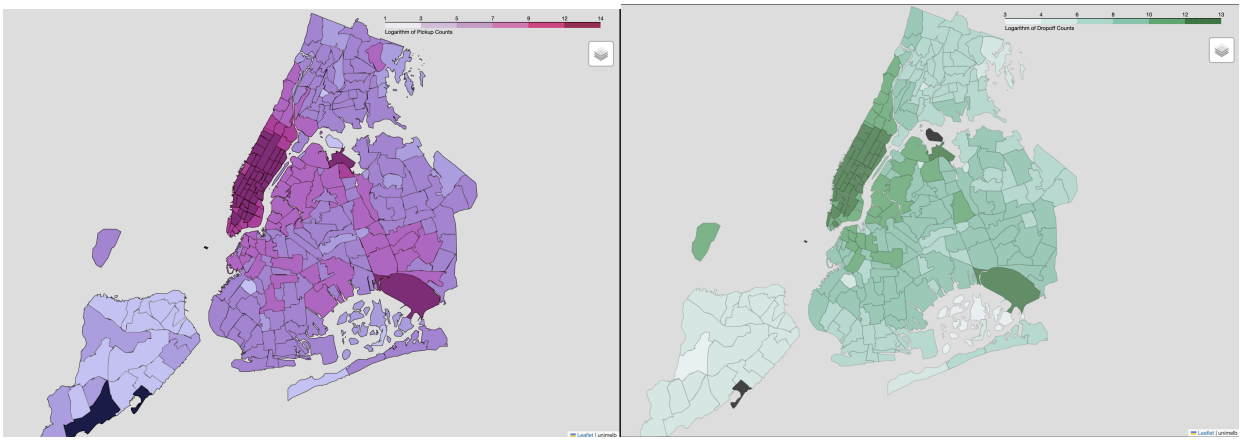
3 Analysis and Geospatial Visualisation

As a Preliminary Analysis, we first explored the demand trend and pattern solely based on the Trip Record Dataset and analysed the interaction between the demand and external factors including weather conditions and number of events.

3.1 Distribution of Pick-up Demand

Through geospatial analysis in Figure 1a,, we can obtain a visualised and intuitive overview of the hourly taxi demand across NYC. Being undisputed, Manhattan dominates most of the demand since it is the most commercial borough and attracts the greatest number of tourists compared to other areas as in Figure 2b. Manhattan has an average of 3918 taxi trips per hour under our analysis, which makes Staten Island's demand look negligible. The demand should be significantly affected by the borough, based on the map.

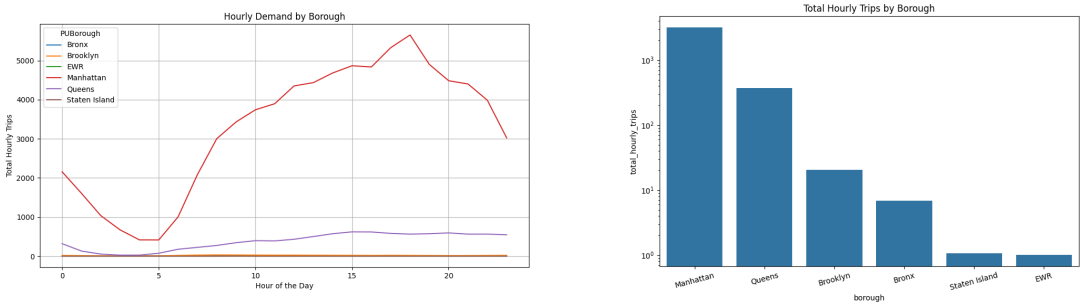
Outside Manhattan, LaGuardia Airport and JFK Airport also demonstrated much of the demand. On the geospatial map, both are marked in dark purple. This insight is consistent with the drop-off distribution as in in Figure 1b. We will observe the airport's demands further below.



(a) Pick up distribution (b) Drop off distribution

Figure 1: Pick up and Drop off distributions

There is also a clear hourly trend for Manhattan which differentiates it from the others as in Figure 2a. During evening rush hours at 6:00 pm, the pickup demand reaches its peak. This matches the working schedule of an office worker's life since they tend to finish their work from business districts and head back home. Taxi companies should allocate more of their fleets to Manhattan during peak hours to meet the demand.



(a) Hourly demand by borough (b) Total hourly trips by borough

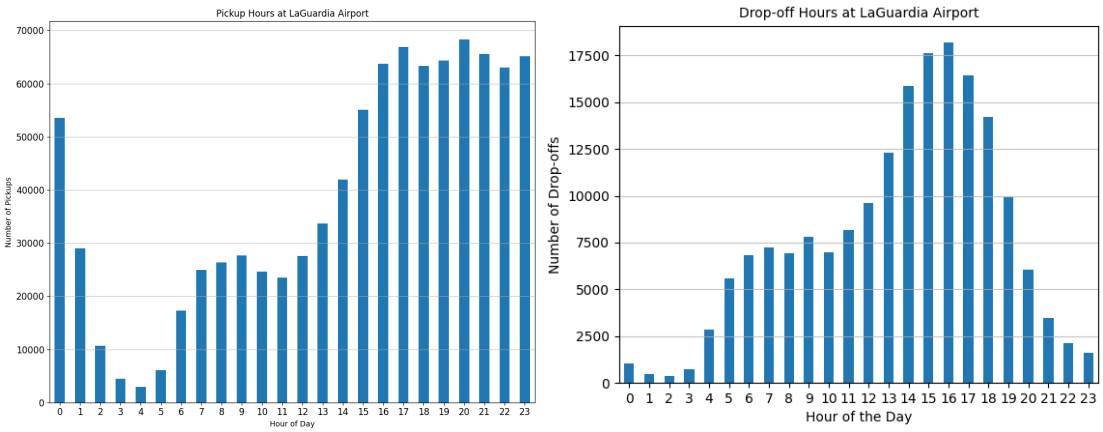
Figure 2: Hourly demand and total hourly trips by borough

3.1.1 Airport hourly demand

LaGuardia Airport is identified with the most amount of hourly trip demand outside Manhattan. Airports tend to have some distinctive differences between peak pick-up hours and drop-off hours. For pick-ups, the total demand per hour consistently remains high during afternoons and mid-nighti in Figure 3a. However, the drop-off hours only reach their peaks between 1:00 pm to 6:00 pm in Figure 3b. This phenomenon should be correlated with the schedule of landing and taking off at the airport. Taxi drivers and company management should take action and dispatch their fleet to meet the unique demand.

3.2 Weather conditions

Most weather features including Temperature, and Wind Speed demonstrated a very weak correlation with the hourly demand according to our analysis. However, the pattern becomes quite clear between



(a) LaGuardia pickup hours

(b) LaGuardia dropoff hours

Figure 3: LaGuardia pickup and dropoff hours

the demand and average visibility of the day, and there's a moderate positive correlation. Whenever the visibility is high, the demand also increases, and vice versa as in Figure 4. Hence, weather conditions should be included in the later modelling stage even though most of the time there are no direct linear patterns.

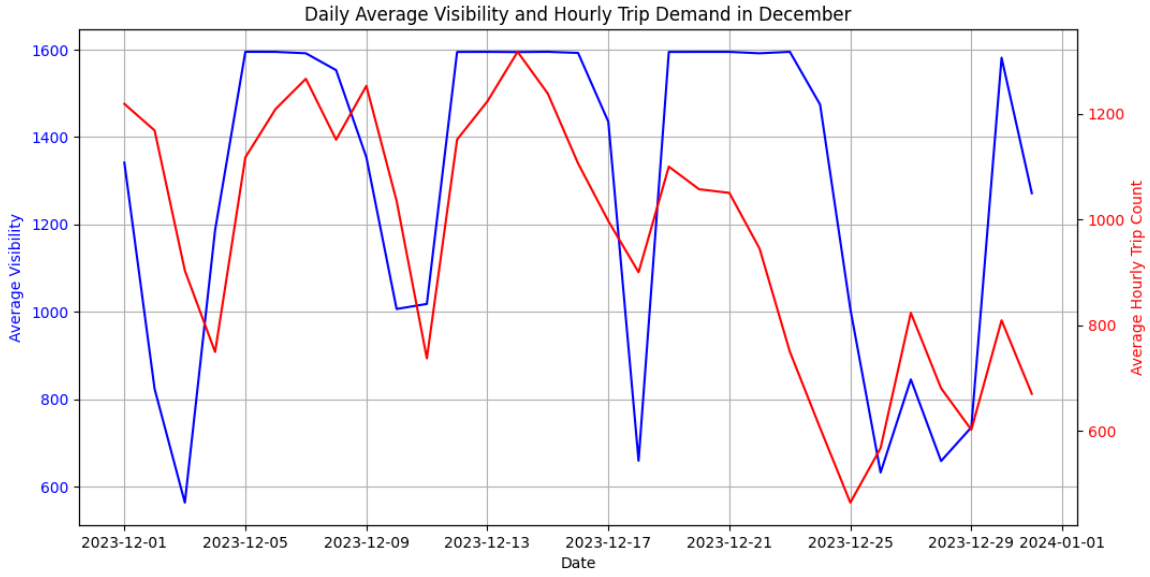


Figure 4: Daily average visibility and Hourly number of trips

3.3 Day of the week

In Figure 5, here's the general demand trend for a weekday, from Monday to Friday: The demand reaches its minimum at 5:00 am, which also becomes the starting time for a new bustling day. The demand then quickly increases to pick up the commuters from home to work. The trend steadily increases until reaches the evening rush hour at 6:00 pm.

However, people tend to stay up late on Friday and Saturday nights, and the remaining demand after 0:00 am is significantly higher compared to Monday to Friday. This matches the entertainment and relaxation demands after a week of hard work. Taxi drivers should focus on picking up late-night passengers from entertainment hotspot areas.

We can observe that Tuesday, Wednesday, Thursday, and Friday have the highest demand, so taxi drivers should avoid taking breaks during these days of the week when possible.

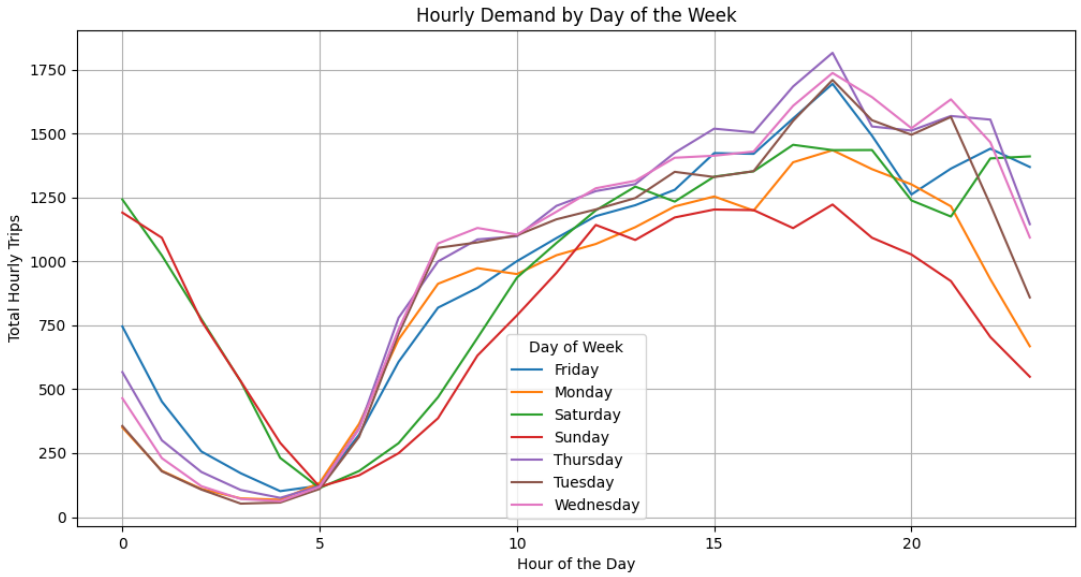


Figure 5: Hourly demand by day of week

‘

4 Modelling

LR and GB models are the 2 Machine Learning regression models implemented to forecast Yellow and Green Taxi hourly demand in NYC. We used the same attributes as input which includes the following:

- Pick-Up Borough encoded
- Pick-Up Hour
- Hourly Trip Count
- Hourly number of events
- Ceiling
- Wind Speed
- Visibility
- Temperature
- Sea Level Pressure

Note here we have used one-hot encoding on the Borough attribute since it is categorical and we must convert it to help the models understand.

The output feature is

- Hourly number of pick ups

We have properly used historical data as training and future datasets for validation and testing, so July to October data is used as the training set, November is used for the validation set, and December

is used for testing. This should prevent the potential implication of exposure to future data during training.

4.1 Linear Regression (LR)

Linear Regression is the fundamental statistical method to predict continuous variables. This model has its strength of perfect interpretability, while it assumes linear relations between its variables.

4.2 Gradient Boosting (GB)

Gradient Boosting is an ensemble learning method that leverages the strength of multiple decision trees to aim for more powerful prediction results. It improves accuracy significantly, at the cost of losing interpretability.

4.3 Error Analysis

We used Root Mean Square Deviation (RMSE) as the metric to evaluate the differences between the prediction value and the actual value for the 2 machine learning models. RMSE provides an error value in the same unit as the original value, which improves the interpretability of the outcomes.

The prediction outcome from the LR model offers a fundamental reference on which factors contribute the most to forecasting the hourly demand due to its perfect interpretability. However, since it heavily relies on the linear separability of the input features, the LR model should be deployed more to have a real-time analysis of demand change.

When comparing the RMSE scores for LR and GB models, very clearly linear regression has a higher error rate compared to GB. On the Testing Dataset, the GB model scores an RMSE of 736, and the LR scores 1027. Since multiple factors such as changes in weather, and events in the corresponding borough can all impact the taxi demand in a complicated manner, LR model's prediction capability might be quite constrained.

4.4 Feature Analysis

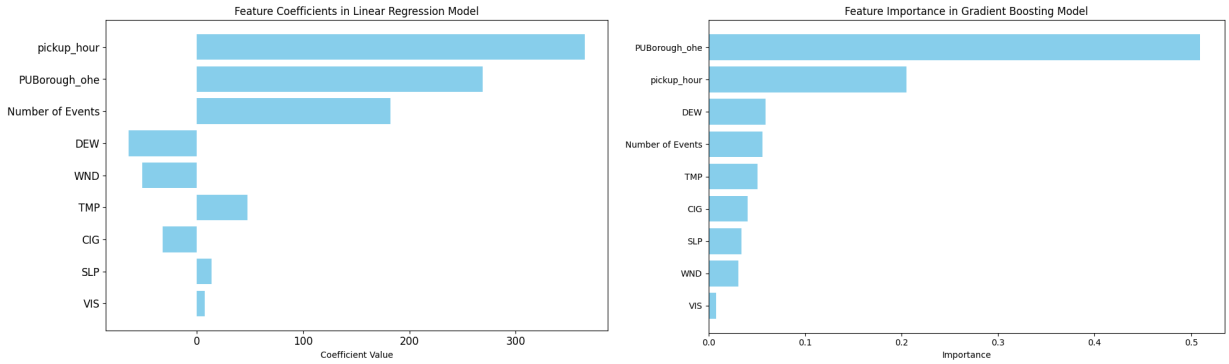
In LR model, the Hour has the highest coefficient among all features, which indicates that a certain hour in the day could significantly boost demand, which is widely anticipated. Both the Pick-Up Borough and Hourly number of events also contribute positively to the predictions. In the GB model, the Pick-Up Borough is the most important feature, so demand can be influenced by which region. Comparing feature analysis for the 2 models, we understand that both location and time matter very much as in Figure 6a and in Figure 6b. We should be able to observe how Dew Point Temperature is also a relatively important feature in GB, this further supports that complex models such as GB can better capture non-linear trends such as how both a decrease and an increase of temperature could result in a spike for taxi demand.

5 Recommendations

Here are some recommendations after completing all the analysis.

5.1 Optimize Fleet Allocation for Major Events and CBD Rush Hours

The geospatial visualizations and hourly demand trends by borough show specific times and locations with heightened demand, supporting the targeted fleet allocation during CBD rush hours and around



(a) LR feature coefficients

(b) GB Feature importance

Figure 6: LR feature coefficients and GB feature importance

major events.

Our GB model has demonstrated strong predictive capabilities in forecasting hourly taxi demand. By integrating this model into the operational workflow, taxi companies can anticipate demand spikes and strategically position their fleets in advance. There are still opportunities to tune the model and include more features such as day of the week which we have observed its correlation with demand in Section 3.3.

5.2 Implement Real-Time Demand Monitoring and Responsive Dispatching

Major airports, particularly LaGuardia, exhibit unique demand patterns with notable peaks for pickups and drop-offs. The data suggests that demand at airports remains elevated during afternoons and late-night hours. Implementing a system for real-time demand tracking would allow taxi companies to adjust fleet deployment dynamically, ensuring sufficient taxi availability when flights land, even during less conventional hours.

The detailed examination of airport demand patterns highlights the importance of a responsive dispatch system that can adapt to fluctuating taxi needs, particularly at airports.

6 Conclusion

In conclusion, we have developed strategies to forecast the hourly demand for Yellow and Green Taxis in NYC through 2 major machine-learning models including LR and GB. We have incorporated the relevant event information dataset since both are easily accessible for drivers and taxi company's management personnel. GB model has demonstrated outstanding performance and has captured complex patterns among the features. LR model might be limited by the linear separability of the dataset, and relatively less optimal and accurate predictions.

As for further research in the future, we can allocate computational resources to align the events' locations with the taxi zones in NYC. The granularity level of the models' predictions will be refined, and the insights will be even more practical since traditional taxi companies can better understand the hourly number of trips desired for each taxi zone. Additionally, we aim to implement more sophisticated feature engineering. These potential improvements will ideally match their market competitiveness with the HVFHVs.

References

[1] New York City Taxi and Limousine Commission. *2023 Annual Report*. Accessed: 2024-09-01. 2023. URL: https://www.nyc.gov/assets/tlc/downloads/pdf/annual_report_2023.pdf.

[2] National Centers for Environmental Information. *Past Weather Data for New York City*. https://www.ncei.noaa.gov/access/past-weather/NYC. Accessed: 2024-08-21. 2024.

[3] NYC 311. *Maximum Number of Passengers in a Taxi*. Accessed: 2024-09-01. 2024. URL: https://portal.311.nyc.gov/article/?kanumber=KA-01245.