Predicting Hourly Fare and Tip for Yellow Taxi in NYC

Phuong Trang Tran
Student ID: 1409465
Github repo with commit

August 31, 2025

1 Introduction

In the era when booking a taxi is deemed old-fashioned, the drivers of New York City taxis are getting older too, with the average age being 50 [8]. Balancing between health and money has become increasingly difficult in this economy. Unfortunately, with the weather getting more extreme and a shift in passenger preference, these drivers will need to plan their day with precision.

In this report, we step into the shoes of yellow taxi drivers by using data from September 2024 to April 2025. Key features like pickup date and time, fare amount, and tip amount will help us estimate the average fare and tip per trip. We aim to help driver utilize their time to make time for breaks and maximize their daily income (fare and tip). To do this, we employ two distinct regression models, Random Forest Regressor and Histogram-based Gradient Boosting Regression Tree, as they are flexible in handling both categorical and numerical features.

2 Data

This report will use two data sets. The primary one is the TLC Taxi Trip Record Data published by the NYC Taxi & Limousine Commission [4]. This data set captures rich information on the fare amount, tip amount, pickup date and time, etc. Yellow Taxi was chosen because it is the most iconic and widely studied taxi type in New York, offering consistent data quality and reliable coverage across the city..

Additionally, we also incorporated weather data from the Central Park of NY City, NY US, data sets recorded by the National Centers for Environmental Information [7]. The provided features, such as an hourly observation of temperature, dew point, air pressure, and precipitation, are critical to understanding how environmental conditions might influence the fares and tipping behavior. To capture the newest change in consumer behavior and weather, we decided to use the trip information from September 2024 to April 2025, as it captures the whole Autumn and Winter as well as gives us a peek at Spring, ensuring a well-rounded analysis.

Datasets	**Instances**	**No. features**
TLC Yellow Taxi Trip Record Data	34,541,965	19
NCEI NY City Centre Park	7,484	93

Table 1: Datasets shape

3 Preprocessing

Due to the nature of our study, several high-level preprocessing steps have been done on the two datasets. This section will give us the first look into the data, point out the inconsistencies, and explain how we handled those inconsistencies.

3.1 TLC Taxi Trip

3.1.1 Removing invalid target variable

To save energy on the feature selection since there were 34,541,965 rows of data, we had removed the invalid target variable first. Fare amounts that total under 3 USD were removed from the raw dataset, as this does not match the updated initial charge from the Taxi Fare from TLC [3].

- The filtering had removed 1,454,608 rows or $\tilde{4}$.21% of data.

3.1.2 Feature Selection

There were only a few features that we are interested in, as we are focusing mostly on time and using data from another data set instead. We will not include any fee and only take fare and tip because the driver only takes home this money. Other unwanted features include Vendor ID, number of passengers, the final rate code, store and forward, miscellaneous extra and surcharges, MTA tax, total amount of toll paid in trip, total amount charged to passengers, not including cash tips, trip distance, and improvement surcharge.

As such, these features remain as our final data for exploration:

- Pickup date time
- Dropoff date time
- Fare amount
- Tip amount
- Payment type
- Dropoff Location ID
- Pickup Location ID

During this selection, we have also done a double check on the dataset to ensure the imported data falls within the research range from September 1, 2024, to April 30, 2025. This filtering and selecting left behind 28,864,486 entries with 7 features to continue further processing.

3.1.3 Fixing invalid value

Unlike normal, we will not focus on cleaning invalid data from unwanted features, but on the features we are interested in. These are what we did for the next cleaning process:

- **Swapping the pickup and drop-off date time columns** in cases where the pickup time is recorded as later than the drop-off time. This approach minimizes the removal of 1129 potentially valid trips, retaining as much useful information as possible.
- **Remove data extreme duration**: Trip lasting more than 10 hours violated the business rule from the TLC rule for preventing fatigued driving [2] while trips with duration under 1 minute were excluded as such trips are unlikely. With this removal, the potentially useful data comes to 28,508,957 or 82.534% of the original TLC Yellow Taxi records.

3.2 Holiday

The holidays in this analysis were manually selected based on their significance and expected impact on the taxi demand and tipping behavior. We had combined federal holidays (e.g., Christmas, New Year) with globally popular events (e.g., Valentine's, Halloween) as they are culturally significant and are suspected to have an impact on travel patterns. Manual selection allowed us to focus on holidays with a measurable and meaningful effect on the data, rather than including every holiday listed. This approach allows the analysis to concentrate on a smaller, high-impact set of events.

3.3 NECI Intergrate Surface Dataset(Global)

This dataset initially contained 18,865 entries, including those outside our target time range. After sorting and standardizing the data and time from this set has approximately 93 features, though we only investigate some main ones. Using the cleaned timestamp, we were able to filter the desired date and time of September 2024 to April 2025, resulting in 7,484 rows.

3.3.1 Feature filtering and Aggregation

Just like the TLC Taxi data, we use a small selection of attributes. New data like Precipitation Occurrence

- Timestamp
- Date
- Time
- Wind speed
- Air temperature
- Dew point
- Precipitation
- Precipitation occurrence

These attributes were also converted to standard measurement units by multiplying the values by 10, following the processing guideline in the paper NOAA Integrated Surface Database Historical Weather Data by visual crossing [5]. Moreover, placeholder values (999 or 9,999) used to represent the missing data were identified and removed from the dataset. After this cleaning process, the final weather dataset was reduced to 7,191 rows with 8 features.

3.4 Feature Engineering and data aggregation

During the preprocessing, date and time were separated into two columns. After that, we aggregated the weather data and taxi metrics (fare and tips) by date and hour to capture the daily pattern as well as the studied timeline. To account for the holiday effect, we link the holiday with the hour feature of the average fare and tip, and observe the effect of the holiday compared to a normal day. Additionally, a weekday feature was included later on to capture the weekly pattern, which can improve the result of the Machine Learning Model

4 Analysis and Geospatial Visualisation

This section will examine the processed fare and tip and their relationship with the weather and holidays. During the analysis, aggregation was used, reducing the number of rows while still capturing the overall pattern, eliminating the need for sub-sampling.

4.0.1 Average tip and fare trend

While the season or month only has a limited impact, the time of the day can be seen to significantly influence the income of our taxi driver.

We can clearly see from Figure 2, for example, that at 5 A.M, the demand for expensive trips soars along with the higher likelihood of receiving a substantial tip. Therefore, the early-morning hours can contribute substantially to drivers' daily income. This pattern is likely driven by the night surcharge and early travellers, airport passengers, or early commuters, who will take longer and more expensive rides. In contrast, the latter morning and times like 1 A.M. seem to receive a lower income, suggesting a low demand for expensive trips during these hours.

Looking into the yearly pattern from Figure 1, fare and tip remain fairly consistent within each season. However, there was a slight dip during the winter, followed by a high-income period observed at the end of December and the middle of January, possibly due to holiday travel or other special events like the New Year, Christmas. The income gradually recovered, returning to the levels similar to before the winter season, indicating a seasonal cycle rather than a permanent decline.

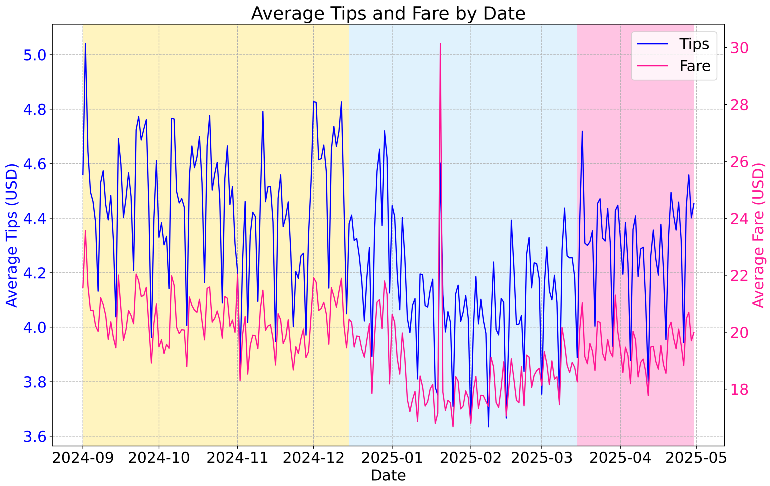

Figure 1: Average Tip and Fare by Date of the Year

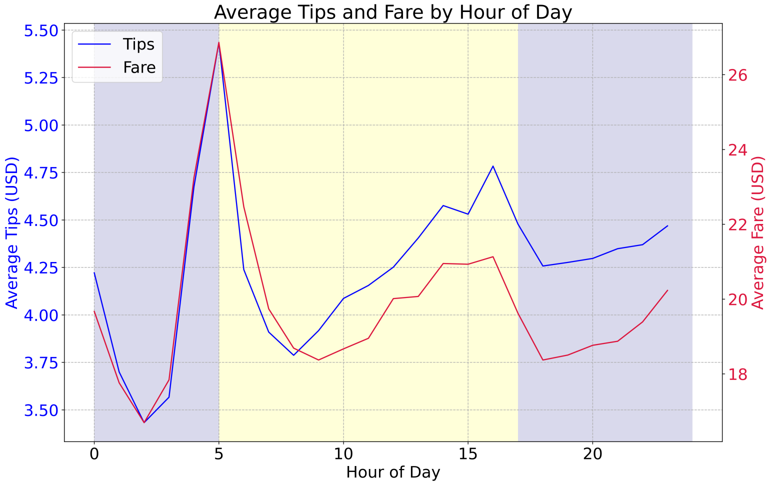

Figure 2: Average Tip and Fare by Hour of the Day

4.1 Holiday effect

On top of the temporal analysis of taxi demand, we examined the relationship between income levels and selected holidays. As shown in Table 2, fares on holiday such as Valentine's Day and Halloween did not differ much from regular days. This finding suggested that taking trips on these days is comparable to non-holiday periods. Similarly, official holidays such as Thanksgiving and Patriot Day had high Pearson correlation coefficients with the non-holiday pattern at 0.934 and 0.896, respectively. While the fare trends on these days are closely aligned with regular days, tipping behavior showed more variation. Nevertheless, tips on these days still exhibited a strong resemblance to the typical daily pattern and were excluded from further analysis and modeling. This outcome may be explained by the nature of the holiday. Thanksgiving and Patriot Day are primarily occasions for remembrance rather than large-scale celebration, while Valentine's Day and Halloween are celebrated by a smaller population, often within local neighborhoods. As a result, these holidays are less likely to generate a substantial increase in high-priced trips.

Holiday	**Pearson Corr on Fare**	**Pearson Corr on Tips**
April Fool's Day	-0.055	-0.494
Black Friday	0.876	0.716
Christmas Day	0.865	0.649
Christmas Eve	0.657	0.504
Earth Day	0.472	0.186
Easter Sunday	0.692	0.653
Halloween	**0.914**	**0.824**
Labor Day	0.71	0.58
New Year's Day	0.253	0.213
New Year's Eve	0.82	0.623
Patriot Day	**0.896**	**0.886**
St. Patrick's Day	-0.047	-0.371
Thanksgiving	**0.924**	**0.77**
Valentine's Day	**0.909**	**0.746**
Veterans Day	0.69	0.671

Table 2: Correlation between holiday income compared to a normal day income

4.2 Weather effect

Regarding the effect of weather, only precipitation and dew point appear to affect cab fee (Figure 3) and therefore are retained for the final model. On the other hand, Figure 4 shows that temperature has a gradual impact on tips. As a result, the average temperature was used in the tipping prediction model. This effect may reflect the influence of weather on mood, with more pleasant temperatures potentially encouraging higher tipping. Additionally, temperature is a seasonal indicator: as shown in Figure 1, winter brings less income to the driver.

An intriguing observation arises from the precipitation data. While precipitation does not exhibit a linear relationship with either fare or tip, it produced one of the most pronounced effects on both. The variability in precipitation suggests that consumer behavior is influenced by rain intensity: light rain will reduce trip passage and tips, a heavier rain attracts costlier rides and higher tips as travellers show consideration for the driver. Surprisingly, this trend collapses in extreme rainfall, discouraging most travel. However, at the heaviest rain point, we have the most valuable trips along with the largest tips. Maybe only critical trips remain, only premium ones paying the driver the most tips and fare,

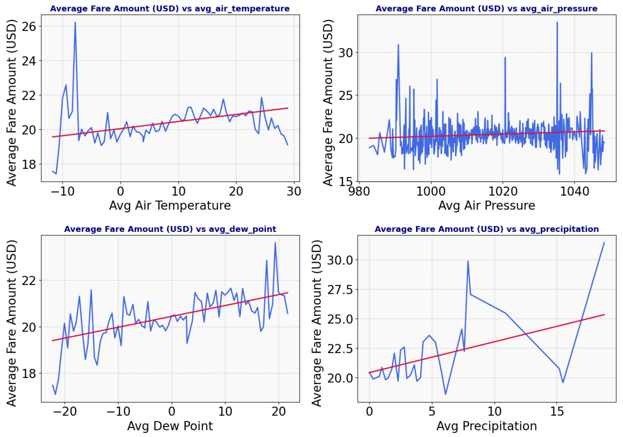

Figure 3: Average fare vs weather

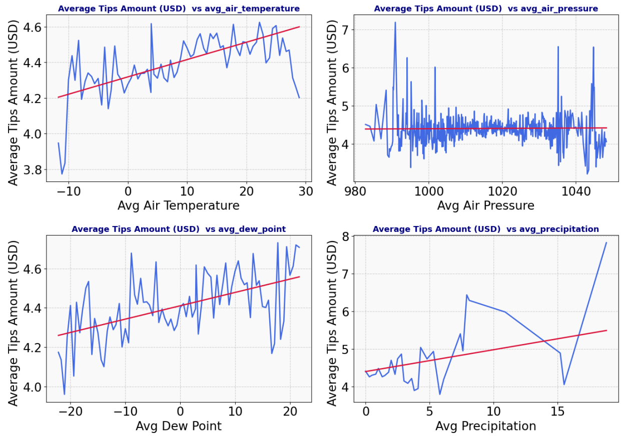

Figure 4: Average tip vs weather

5 Modelling

In the modeling stage, two algorithms were implemented to predict hourly fare and tip: Random Forest Regressor (or RFR for short) and Histogram-Based Gradient Boosting Regressor (or HGBR). The dataset spanning six months, from September 2024 to the first of March 2025, was used for model training and validation, while the last 2 months, March and April 2025, were reserved for model testing.

The same predictors were used for the two models, with a minor adjustment, as the tips predictor also includes the average temperature (numerical) due to the significant relationship with tipping behavior. The other predictors include hour (ordinal), day of week (categorical), holiday flag (categorical), season (categorical), and weather attributes (all continuous): average dew point, average precipitation. Attributes with categ Weather data is assumed to be provided in advance through weather forecasting. Models' hyperparameters were optimized through a grid search combined with cross-validation using Time Series Split, ensuring that the temporal order was preserved and no future data was used to predict the past.

5.1 Random Forest Regressor

To handle the mix of categorical and numerical input, a random forest regressor was used to predict the numerical output [1]. This algorithm is an ensemble method based on bagging of decision trees. This helps reduce the high variance usually associated with individual decision trees and can capture the non-linear relationships between variables, like between precipitation and the interested features 4 and 3. Using grid search, we tuned our parameters for the fare and tip model. For fare prediction, RF achieved its best performance with unrestricted tree depth, at least 1 sample on one branch, and 400 trees in the forest. Tips also need 400 trees and 1 sample, but perform best with a shallower max depth of 20.

5.2 Histogram-Based Gradient Boosting Regressor

The Histogram-based gradient boosting is an ensemble machine learning algorithm that builds models sequentially, where each tree learns from the residuals of the previous one [6][9]. Since it also bins continuous features into histograms, it is more efficient in terms of speed and memory on large datasets like ours. Similar to RFR, HGBR was refined using grid search on max depth, the shrinkage rate, the number of boosting rounds, and minimum samples in a leaf node. A shrinkage of 0.1, no max depth, 500 iterations, and at least 5 samples per leaf are the best estimators for Fare. It is the same for tips, except for 3 samples on min_samples_leaf. We used the default metric, the square error, for our loss metric.

6 Discussion

To assess the performance of the predictive model, this report considers both training and validation results. As shown in Table 3 and Table 4, the two algorithms exhibit comparable accuracy in general. On the test set, HGBR achieved a lower RMSE for tip amount prediction, while the Random Forest Regressor had a slightly lower RMSE for fare prediction. This trend was reversed in the validation set. Both machine learning models display a tendency to overestimate fares and tips during peak hours and underestimate during off-peak periods. Nevertheless, they could be seen to follow the general daily pattern and were able to predict accurately around 4 a.m. Even though neither mode demonstrates perfect generalization, the errors remain relatively small compared to the typical values. While the highest RMSE was 2 USD in testing 3, the typical values can go up to 28 USD. Similarly, the error for the tip prediction is only 0.465 dollars against the actual value reaching 5.75 dollars or roughly 8% 3.

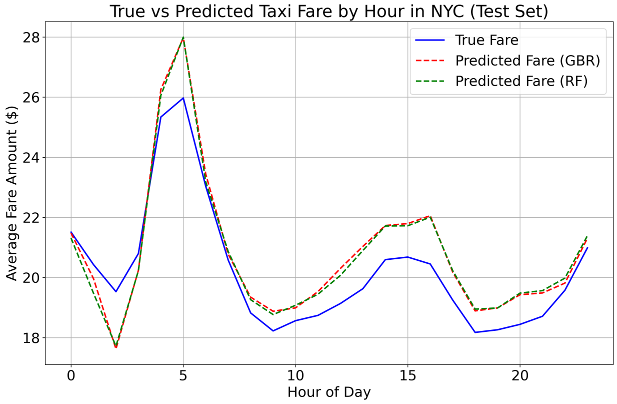

Figure 5: True vs predicted fare

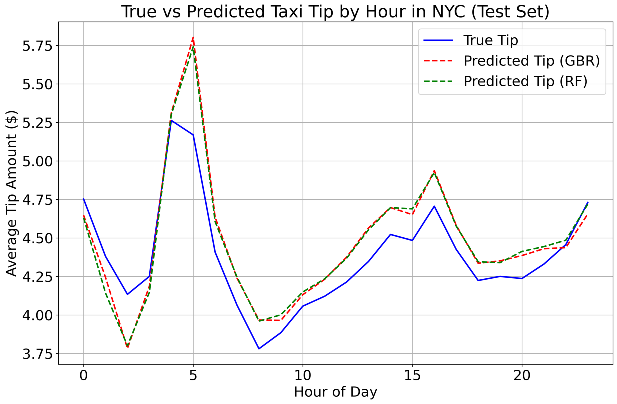

Figure 6: True vs predicted tip

	RMSE	
Income	**RFR**	**HGBR**
Fare	**1.976**	2.000
Tip	0.465	**0.462**

Table 3: Performance of the model (RMSE) on testing

	RMSE	
Income	**RFR**	**HGBR**
Fare	3.14	**2.836**
Tip	**0.582**	0.638

Table 4: Performance of the model (RMSE) on validation

Both models indicate that drivers can expect predictable income trends throughout the day. With earnings peaking early in the morning hours, slowing when it is midday, and spiking up again during the golden hour, declining during the transition to the next day. This information can help older drivers plan their shift more effectively to maximize their income while still allowing for adequate rest.

By providing a reliable forecast of the hourly fare and tip levels, the two regressors can support taxi drivers in balancing work with personal well-being, rather than working excessively long and irregular hours while bringing in less money.

7 Recommendations

Taxi drivers should optimize their work schedule around the daily income peaks depicted in Figures 2,5,6. Starting before 5 a.m. to capture the morning peaks in both fares and tips, then have a break during midday off-peak hours, and resume work during the afternoon rush hour. This strategy helps align the working hours with the periods of highest income potential, maximizing drivers' earnings, and reducing wasted time. Taxi drivers should also plan their work during holidays such as Christmas, April Fool's Day, or St. Patrick's Day, as these days have a different pattern from a normal day, as we can see from Table 2.

An annual plan should also be in place. From Figure 3 and Figure 4, we recommend drivers to prioritize working during the warmer conditions (temperatures above $-9°C$) and in moderate rainfall conditions (up to 7 mm), like during Fall or Spring, as these are associated with higher fares and tips. In contrast, extreme cold (under $-9°C$) and heavy rainfall (over 7 mm) introduce significant safety risks while the demand is reduced.

8 Conclusion

This analysis has explored how some factors impact NYC yellow taxi drivers' income and used them to predict future trends. External factors like hourly weather conditions, temporal features were incorporated into the two models to capture their influences on drivers' earnings. Random Forest Regressor has demonstrated better performance for fare prediction; at the same time, Histogram-Based Gradient Boosting Regressor achieved lower RMSE for tip prediction. Both models successfully express the overall daily income patterns, with some overestimation during high-income hours and underestimation during slow hours.

While the results are proficient, further refinement of features is needed. Incorporating more detailed weather patterns, focusing on location and pick-up zones, to improve performance and better reflect real-world travel demand. Overall, the regressors give actionable insights for the drives, helping to plan more efficient scheduling, balancing work opportunities, and personal well-being.

References

[1] Basim Azam. *Classifier Combination.* Semester 1 2024 COMP30027 Machine Learning Lecture Notes. 2024.

[2] New York City Taxi & Limousine Commission. *Fatigued Driving Prevention - Frequently Asked Questions.* Accessed: 2025-08-29. 2025. URL: https://www.nyc.gov/site/tlc/about/fatigued-driving-prevention-frequently-asked-questions.page.

[3] New York City Taxi & Limousine Commission. *Taxi Fare.* Accessed: 2025-08-29. 2025. URL: https://www.nyc.gov/site/tlc/passengers/taxi-fare.page#:~:text=Trips%20between%20Manhattan%20and%20John,50%20cents%20MTA%20State%20Surcharge.

[4] New York City Taxi & Limousine Commission. *TLC Trip Record Data.* Accessed: 2025-08-13. 2025. URL: https://www.nyc.gov/site/tlc/passengers/taxi-fare.page#:~:text=Trips%20between%20Manhattan%20and%20John,50%20cents%20MTA%20State%20Surcharge.

[5] Visual Crossing. *How We Process the NOAA Integrated Surface Database Historical Weather Data.* Accessed: 2025-08-16. Feb. 2025. URL: https://www.visualcrossing.com/resources/documentation/weather-data/how-we-process-integrated-surface-database-historical-weather-data/.

[6] GeeksforGeeks. *HistGradientBoostingClassifier in Sklearn.* Accessed: 2025-08-29. 2025. URL: https://www.geeksforgeeks.org/machine-learning/histgradientboostingclassifier-in-sklearn/.

[7] National Centers for Environmental Information (NCEI). *Global Hourly Surface Data.* Accessed: 2025-08-16. 2025. URL: https://www.ncei.noaa.gov/access/search/data-search/global-hourly?bbox=40.959,-74.251,40.469,-73.761%5C&pageNum=1%5C&stations=72505394728%5C&startDate=2024-10-01T00:00:00%5C&endDate=2025-03-31T23:59:59.

[8] NYC Taxi & Limousine Commission. *TLC Factbook 2020.* Accessed: 2025-08-16. 2020. URL: https://www.nyc.gov/assets/tlc/downloads/pdf/2020-tlc-factbook.pdf.

[9] Fabian Pedregosa et al. "Scikit-learn: HistGradientBoostingRegressor in Python". In: (2012). URL: %5Curl%7Bhttps://scikit-learn.org/stable/modules/generated/sklearn.ensemble.HistGradientBoostingRegressor.html%7D.