Estimating Uber surge pricing from a particular location at an hour of a day in New York City

Ngoc Minh Vu
Student ID: 1375708
Github repo with commit

August 25, 2024

1 Introduction

A few years after the launch date of FHVs[1] in NYC[2], in 2017, this service surpassed yellow taxis - one of the city's most recognizable symbols - to become the most popular mode of transportation. Based on the 2023 annual report of the NYC TLC[3], there were about 13,500 drivers with yellow cab licenses, while the number of drivers with HVFHVs[4] licenses was roughly 6 times that amount[1], which indicates that HVFHVs are capable of transporting an enormous number of passengers. However, at certain hours, the demand can exceed the supply in some locations. As a result, fares for trips from these locations will be higher, encouraging more drivers to come to these places. This process is known as **surge pricing**, which helps balance supply and demand.

In this paper, two Machine Learning models will be discussed and evaluated based on their performances in predicting dynamic pricing imposed by **Uber** - the leading HVFHV service - at a specific location and time. The result of this paper can be used by Uber to accurately estimate surge pricing, helping them achieve the equilibrium point of supply and demand[5]. Besides the company-level benefits, the paper also allows drivers to determine when and where they should operate trips to maximize their profit.

2 Preprocessing

2.1 Dataset

The main dataset that will be used throughout this paper is the HVFHV Trip Records (including trip length, trip duration, pick-up location ID, etc.), which is provided by NYC TLC[2]. According to the FHV license report published in early 2024, after the restrictions imposed during the pandemic period were removed, the annual growth in HVFHV trips slightly increased, which implies the stability of the demand for ride-sharing services[3]. Due to this reason, it can be assumed that any intervals after the COVID-19 era with sufficient length can be chosen as the representative period for analyzing factors related to HVFHVs. Therefore, this paper will use the trip records from the most recent 6 months as the primary dataset **(December 2023 to May 2024)**.

[1] For-Hire Vehicles

[2] New York City

[3] NYC Taxi & Limousine Commission

[4] High Volume FHVs

[5] The point where supply and demand intersect at

Aside from the trip records, weather conditions are also a related factor. Based on earlier research[4], when it rains, there is a substantial rise in both supply and demand for rides, which is directly associated with surge pricing. Hence, hourly weather data collected from NCEI[6] (including wind speed, air temperature, etc.) will be included in this paper[5]. This data is recorded at the NY Central Park station, which is assumed to be the most reliable one since it is the first station in NYC, and it also has long, largely continuous measurement intervals over several decades.[6]

2.2 Data Cleaning

2.2.1 HVFHV Trip Records

This dataset is significantly enormous with over 120 million records, and they were also sent directly from ride-sharing services to NYC TLC, so the errors are inevitable. Therefore, the following steps are applied to the raw dataset to address these issues:

- **Missing Values**: In this paper, it is assumed that any null value is missing completely at random. Therefore, they will be discarded.
- **License Number**: Since only trips operated by Uber are considered, all the trips with license number HV0003 will be kept.
- **Date Order**: Based on the data dictionary provided by NYC TLC[7] and the range of the research, only records that satisfied the chronological sequence below will be kept:
  2023-12-01 $\leq$ request_datetime $\leq$ on_scene_datetime $\leq$ pickup_datetime $<$ dropoff_datetime $<$ 2024-06-01 (a $\leq$ b means that a is before b)
- **Location ID**: Since the location ID is only defined between 1 to 263, any record outside the range will be discarded.
- **Non-Negative Values**: Records with negative tolls, bcf, sales_tax, congestion_surcharge, tips, or driver_pay will be discarded.
- **Airport Fee**: The price is $2.50 for either drop-off or pick-up at any airport. Therefore, trips with an airport fee that is a multiple of 2.5 are retained.
- **Attributes with Y/N values**: shared_request_flag, shared_match_flag, access_a_ride_flag, wav_request_flag, or wav_match_flag not receive a Y/N value, records will be discarded.
- **Base Passenger Fare**: It is assumed in this paper that this fare is the price appearing on the customer's app after he/she books a ride. Uber has set the minimum fare of $7.00 for a trip[8]. Therefore, any trip with the base_passenger_fare ≥ 7 will be kept.
- **Distance and Duration**: The dataset contains some trips with remarkably short distances; therefore in this paper, it is assumed that they were all system errors, and a person only considers booking a ride if the distance is ≥ 0.1 miles. Additionally, the duration for one Uber trip is limited to 8 hours (28800 seconds)[9]. This limitation implies that if the driver can ride at 45mph (maximum speed within the city) for 8 hours, the maximum distance they can ride is 360 miles. Therefore, any trip with $0.1 \leq$ trip_miles ≤ 360 and $0 <$ trip_time ≤ 28800 will be kept.

After removing values outside defined ranges, the dataset is still considerably large. To reduce the impact of extreme outliers and keep the generalization of the model, values outside approximate 0.1st to 99.9th percentiles will be excluded. In the end, there are **68.1134%** of the original data kept.

[6] National Centers for Environmental Information

2.2.2 Weather Data

This dataset contains a huge number of attributes, but the majority is not available. Since building models requires as many non-null values as possible, only features belonging to the Mandatory Data Section, defined in the NCEI data documentation[10], will be kept. Therefore, only basic element factors will be retained, including wind speed, air temperature, dew point temperature, and sea level pressure. After that, data will be inverse transformed to original values, with scaling factors shown in the data dictionary. The dataset is relatively small so all the data should be kept. Therefore, any missing value will be imputed from the data in the previous hour, because it can be assumed that the weather does not remarkably change after one hour.

2.3 Feature Extraction

Besides available features, the hour when the rider requests to be picked up and the type of day (whether weekday or weekend) are assumed to be effective predictor variables. This is because surge pricing occurs during peak hours, but peak hours differ between weekdays and weekends. Therefore, request_datetime will be split into hour and date, and the type of day will be stored as is_weekend, where weekday is 0, and weekend is 1.

3 Preliminary Analysis

3.1 Train Test Split

Surge pricing at a location is solely based on the number of taxi drivers available at that place (supply) and the number of requests for rides (demand). While the supply is not provided, the demand can be estimated by the total of rides that occur from that location.

Month	Total Demand	Month	Total Demand
12/2023	13,187,764	03/2024	14,521,744
01/2024	13,470,015	04/2024	13,584,403
02/2024	13,422,763	05/2024	14,389,672

Table 1: Monthly Total Demand

From **Table 1**, it can be seen that the total demand did not notably fluctuate month to month. Hence, we assume that the demand remains consistent each month. Therefore, the records of the first 5 months in the dataset will be used to train the models, while the data from May 2024 will be used to evaluate their performances.

3.2 How to calculate Surge Pricing

As discussed in **3.1**, demand has a strong influence on surge pricing, in which increasing in demand would lead to a rise in surge pricing to balance the supply and demand. According to **Figure 1**, different areas will experience different ride request volumes, with places marked on the map having higher demand than others.

Surge pricing at location L is derived by using a surge multiplier, which can be calculated as:

$$\text{SurgeMultiplier}(L) = \text{SurgeAverageFPM}(L)/\text{StandardAverageFPM}(L) \tag{1}$$

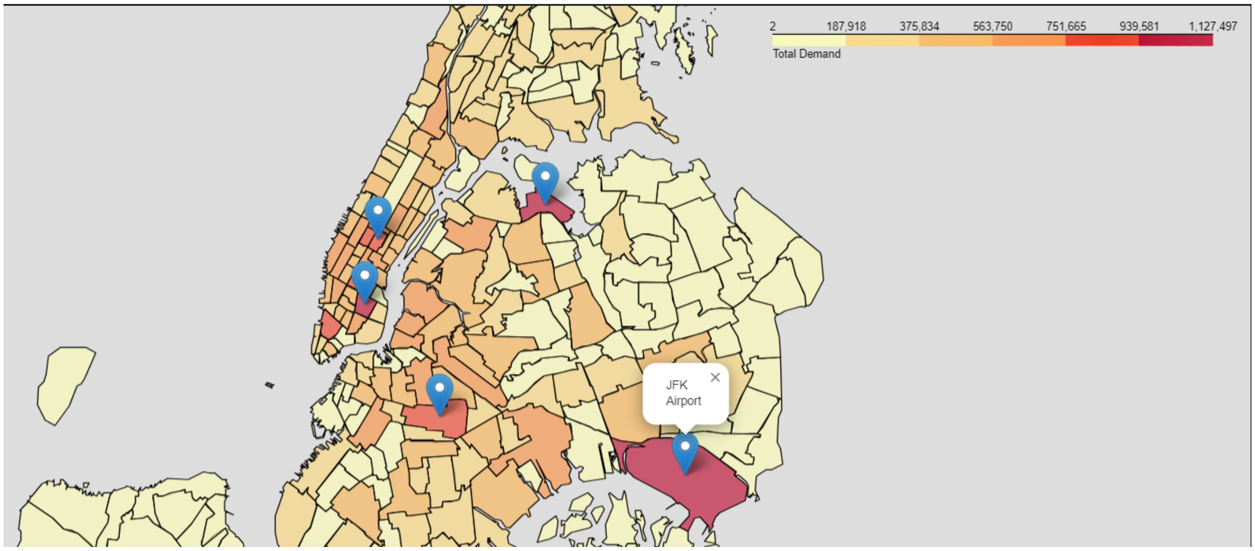

Figure 1: Total demand at each location from Dec 2023 to Apr 2024

StandardAverageFPM[7] represents the average FPM without the surge pricing strategy, while SurgeAverageFPM shows the average FPM with it. However, there is no explicit formula for StandardAverageFPM. In this paper, it will be calculated by the average FPM at the lowest-demand hour at location L. The average FPM and hour are chosen as key factors in determining surge multiplier because based on **Figure 2**, it can be seen that the average FPM is highly correlated with the total demand, and it varies between different hours.

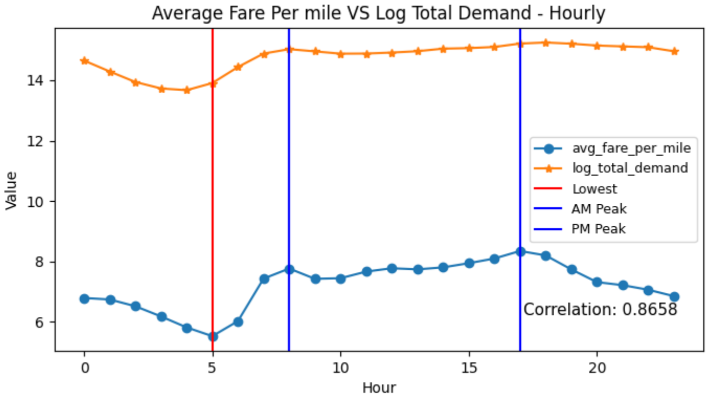

Figure 2: Average FPM VS Log Total Demand - Dec 2023 to Apr 2024

Since various locations and times affect dynamic pricing differently, surge price in location L at time t will be computed as:

$$\text{SurgeMultiplier}(L,t) = \text{SurgeAverageFPM}(L,t)/\text{StandardAverageFPM}(L) \tag{2}$$

With the surge multiplier, the surge price at location L and time t can be calculated as:

$$\text{SurgePrice}(L,t) = \text{SurgeMultiplier}(L,t) \times \text{StandardFare}(L) \tag{3}$$

[7] Average fare per mile

3.3 Features Selection

Numeric: To estimate surge pricing, base_passenger_fare will be chosen to be the **target variable**, which is reasonable since this shows the dynamic price when the surge pricing strategy is applied. Because the dataset is substantially huge, fitting a complex model with all the features will lead to significant time consumption, and possibly overfitting. Considering features with the highest correlations with base_passenger_fare shown in **Figure 3**, bcf, sales_tax, and driver_pay will be excluded. This is because these values can be derived directly from the target variable[11, 12]; using these features to predict the target will lead to data leakage, resulting in inflated performance and a non-generalization model. Other features will be chosen as they are assumed to be significant.

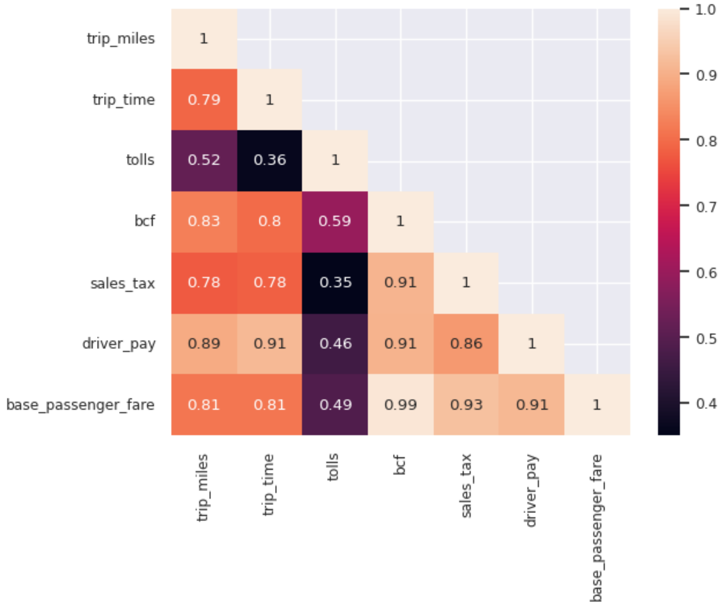

Figure 3: Correlation Heatmap (Top Highest Correlations) - Train Data

Categorical: hour, pulocationid, dolocationid, and is_weekend are considered as nominal variables. However, OneHotEncoded will not be applied since the feature dimension will rise explosively, but with a slight performance improvement. The possibility of creating artificial orders among the values will be accepted in this paper.

Weather Information: All 4 features will be kept.

4 Modelling

4.1 Feature Transformation

Figure 4 illustrates the distribution of base_passenger_fare from Dec 2023 to Apr 2024. The distribution is extremely skewed and should be transformed to stabilize variance. Therefore, a log transformation will be applied to resolve the issue.

4.2 Linear Regression with LASSO Regularization

There are a few regression models that can deal with a mixture of numerical and categorical attributes, and Linear Regression is the simplest model among them. This can be used as the benchmark for

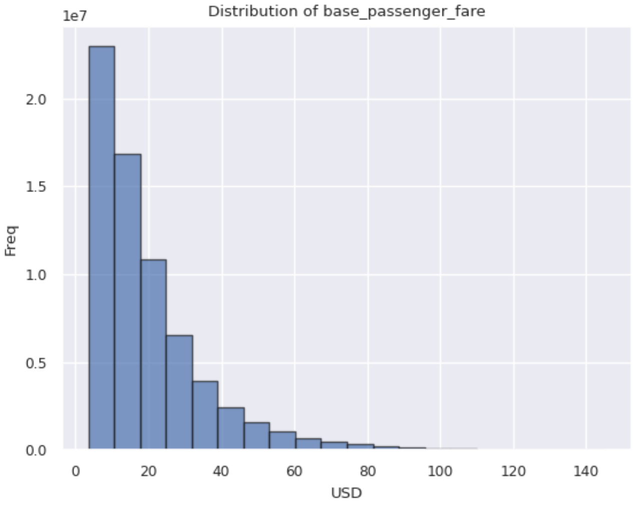

Figure 4: Distribution of base_passenger_fare - Train

evaluating more sophisticated models. According to **Figure 3**, some independent variables are highly correlated with others, which may cause multicollinearity. As a result, LASSO regularization will be used to address this problem.

4.3 Gradient-Boosted Tree (GBT)

GBT Regression is one of the most powerful models that can outweigh simple Linear Regression by capturing complex, non-linear relationships between features and the target variable. It is a technique based on boosting strategy and predicts the target by combining predictions of weak models. Since it is a complicated model and time-consuming, running different configurations to tune hyperparameters takes a huge amount of time. Therefore, in this paper, it is assumed that the default GBT is relatively good for this dataset. Finally, although GBT is powerful, it sacrifices intelligibility and interpretability because following the path of several trees is non-trivial[13]. However, this will not be a major problem since only Uber data scientists will need to understand the underlying model details, while others will only need to use the results discussed in the next section.

5 Discussion

RMSE and R^2 will be used as 2 metrics to evaluate the performances of the models. RMSE is a loss function with high interpretability, where lower RMSE implies better performance. Besides RMSE, R^2 will also be used to give a more comprehensive view of performance, as it shows the amount of variance explained by the model, with higher R^2 normally indicating a more efficient model. It can be seen from **Table 2** that although the simple linear model has a relatively good performance, GBT outweighs it in all metrics, which is logical since GBT has a more complex structure than Linear Regression.

Note: After **Table 2**, test data (May 2024) will be used for analyzing and visualizing.

The metrics shown in **Table 2** are from models predicting log(base_passenger_fare), so they are not straightforward for non-expert audiences to derive meaningful statistics. Hence, exponential transformation is used to convert values to their original scales. The RMSE of the GBT model

Metric	**Data**	**Linear Regression**	**GBT**
RMSE	Train	0.2541	0.2369
RMSE	Test	0.2697	0.2528
R^2	Train	0.7785	0.8163
R^2	Test	0.7791	0.8060

Table 2: Metrics - log(base_passenger_fare)

after transformation is 8.7101, while the RMSE of the LR model is 9.3942. This shows that on average, the predicted passenger fares of the **GBT** model deviate from the actual fares by **$8.7101**. Since the average passenger fare is $24.7612, the RMSE is considerably large. This is reasonable because the models were built based on the whole dataset of all locations in NYC. However, as we discussed in section **3.2**, each area has a different number of ride requests, and higher demand usually leads to higher average passenger fares. Based on **Table 3**, it can be observed that locations with notably higher average fares than the mean fare of the whole dataset will have higher prediction errors. Thus, the original dataset should be subsampled based on the demand, so the average fare at each location in the dataset will have approximately the same scale, which can potentially improve the model performance.

pulocationid	**avg_actual_fare**	**avg_pred_fare**	**demand**
138	49.6206	42.4051	275729
132	58.6269	53.2716	235872
61	20.3495	21.2741	189048
110	22.8968	23.7093	40
2	29.5594	27.3718	36

Table 3: Location ID - Avg Actual Fare - Avg Predicted Fare - Demand

5.1 Estimating Surge Pricing

In order to estimate surge pricing, we need to estimate the surge multiplier. The following steps will show how to determine it with an example:

1. Choose a location L: From **Table 3**, it can be seen that pulocationid = 61 (Crown Heights North) is considered a high-demand area but has a low prediction error. Therefore, it will be chosen to analyze.

2. Determine the lowest-demand hour at L: Assuming that this hour does not change day by day. Based on **Figure 5**, the lowest-demand hour is 3 am.

3. Based on section **3.2**, StandardAverageFPM(L) will be calculated as the average FPM at 3 am. **SurgeMultiplier(L,t)** will be derived as shown above.

The surge multiplier difference on 2024-05-15 is selected for visualization in **Figure 6** under the assumption that the pattern is consistent daily. It can be seen that although 3 am is assumed to be the lowest-demand hour, there are some durations where the surge multiplier is smaller than 1, which indicates that the surge average FPM is lower than the standard average FPM. This means that our formula shown in **3.2** is somewhat imprecise. Not only that, while the figure shows that the surge multiplier is estimated fairly well (RMSE = 0.0985), the predictions are highly underestimated when there is a significant gap between the actual average FPM and the standard average FPM. Considering both **Figure 5** and **Figure 6**, this happens at AM peak hours, but not at PM peak hours, implying that the average FPM in the morning peak is higher than in the evening peak, despite higher demand in the evening. This can happen because surge pricing is affected by both supply and demand, but only demand information is available. It possibly suggests that the supply at PM peak hours is higher than at AM peak hours, which does not lead to significantly high surge pricing. Besides, as explained earlier, high passenger fares are not accurately predicted under this model because of using the full dataset.

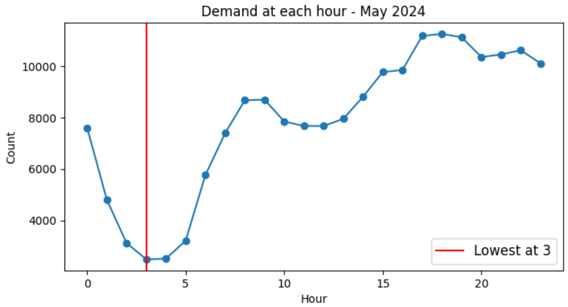

Figure 5: Demand at each hour - May 2024

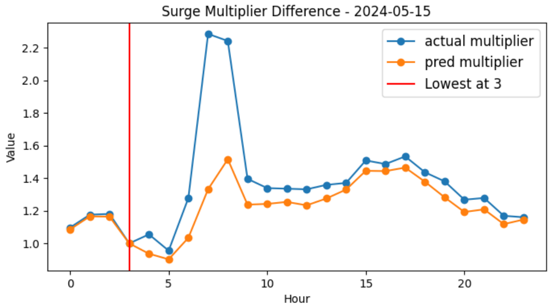

Figure 6: Surge multiplier difference - 2024-05-15

6 Recommendations

Throughout this paper, the benefits of surge pricing are explored. Although the 2 models discussed are not fully optimized, they can still be used to provide meaningful recommendations:

- As a ride-sharing company, based on our model, Uber can develop a more reliable model to predict surge pricing. After that, they can include the surge multiplier next to the fare displayed on the customer's app to inform them that there is a dynamic pricing strategy applied. Uber can also add the next duration when the surge multiplier is lower. This improvement will likely increase the trust and loyalty of users toward the firm, as it will reduce the possibility of riders facing unexpectedly expensive rides, which will enhance the customer experience with the service.
- As a driver, they can find appropriate durations to gain maximum profit based on what we discussed earlier. For example, in **Figure 6**, it seems that the demand exceeds the supply in the morning peak at Crown Heights North - a high-demand location. If the driver operates trips within this interval, they would potentially earn remarkable revenue.

7 Conclusion

In this paper, 2 machine learning models were discussed and evaluated to estimate Uber surge pricing at a specific location and hour in NYC. These models used attributes of trip records collected from NYC TLC and NCEI weather data to predict the passenger fare. Although the metrics show that these 2 models are reasonably accurate, they can perform better by subsampling the original dataset based on the demand of locations. With a more powerful device's configuration, categorical variables can be OneHotEncoded to ensure there is no order between the values, and the hyperparameters of the GBT model can be tuned by using the cross-validation method to find the most optimal model.

References

[1] New York City Taxi and Limousine Commission. *2023 Annual Report*. Accessed: 2024-08-25. URL: https://www.nyc.gov/site/tlc/about/industry-reports.page.

[2] New York City Taxi and Limousine Commission. *TLC Trip Record Data*. Accessed: 2024-08-25. URL: https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page.

[3] New York City Taxi and Limousine Commission. *February 2024 FHV License Review*. Accessed: 2024-08-25. URL: https://www.nyc.gov/site/tlc/about/industry-reports.page.

[4] Abel Brodeur and Kerry Nield. "An empirical analysis of taxi, Lyft and Uber rides: Evidence from weather shocks in NYC". In: *Journal of Economic Behavior Organization* 152 (2018), pp. 1–16. ISSN: 0167-2681. DOI: https://doi.org/10.1016/j.jebo.2018.06.004. URL: https://www.sciencedirect.com/science/article/pii/S0167268118301598.

[5] National Centers for Environmental Information (NCEI). *Global Hourly Data Search*. Accessed: 2024-08-25. URL: https://www.ncei.noaa.gov/access/search/data-search/global-hourly?pageNum=1&bbox=40.959,-74.251,40.469,-73.761&startDate=2023-12-01T00:00:00&endDate=2024-05-31T23:59:59.

[6] Luis E. Ortiz et al. "High-resolution projections of extreme heat in New York City". In: *International Journal of Climatology* 39.12 (2019), pp. 4721–4735. DOI: https://doi.org/10.1002/joc.6102. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.6102.

[7] New York City Taxi and Limousine Commission. *High Volume FHV Trips Data Dictionary*. Accessed: 2024-08-25. URL: https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page.

[8] Christina Majaski. *Uber vs. Yellow Cabs in New York City: What's the Difference?* Accessed: 2024-08-25. 2024. URL: https://www.investopedia.com/articles/personal-finance/021015/uber-versus-yellow-cabs-new-york-city.asp.

[9] Steven John. *How Far Can Uber Take You?* Accessed: 2024-08-25. URL: https://www.businessinsider.com/guides/tech/how-far-can-uber-take-you#:~:text=There%20isn%27t%20a%20limit,the%20time%20limit%20kicked%20in.

[10] National Centers for Environmental Information (NCEI). *ISD Format Document*. Accessed: 2024-08-25. URL: https://www.ncei.noaa.gov/data/global-hourly/doc/isd-format-document.pdf.

[11] Automarketplace. *Have You Seen The Black Car Fund (or BCF) Fee on Your Uber or Lyft Pay Stub?* Accessed: 2024-08-25. URL: https://automarketplace.substack.com/p/have-you-seen-the-black-car-fund.

[12] Noam Scheiber. *How Uber's Tax Calculation May Have Cost Drivers Hundreds of Millions*. Accessed: 2024-08-25. URL: https://www.nytimes.com/2017/07/05/business/how-uber-may-have-improperly-taxed-its-drivers.html.

[13] Wikipedia. *Gradient Boosting*. Accessed: 2024-08-25. URL: https://en.wikipedia.org/wiki/Gradient_boosting#Penalize_complexity_of_tree.