Predicting The Number Of High Value Trips For Yellow Taxis At Different Times In NYC

Un Leng Kam
Student ID: 1178863
Github repo with commit

August 25, 2022

1 Introduction

New York City is an area with more than 65 million visitors per year, and most of those visitor tends to stay in Manhattan. For such reasons, the other four boroughs of NYC often get neglected and taxi services become less available. Green Taxis has therefore been introduce to resolve such issue where they are restricted to only pickup in the other four boroughs[1]. However, ride-share services has proved to be a strong competitor where Uber alone handled more than 84,000 trips in a single neighbourhood, contrasting to 25,693 trips citywide for Green Taxis[2]. It is crucial to provide taxi services in other boroughs to ensure accessibility for other NYC's residents, however, the New York State Government should also take into account the taxi drivers' income when improving the city's infrastructure.

Therefore, this reports aims to **predict the number of high value trips for Yellow taxis at a given time in day to encourage them to provide taxi services in other boroughs but also receive a better income.**

To conduct such predictions, the datasets are pre-processed to remove invalid trip entries, features are then visualised and engineered to produce each trip's value through a self defined formula. Quantiles of trip value is computed to create a metric to determine high value trips, and the data is aggregated by pickup location and time. The curated dataset is used to train two different regression models to provide predictions and evaluated to give recommendations.

1.1 Data selection

The New York Taxi and Limousine Commission (NYCTLC) provided the data used in this research[3]. Yellow taxi trip data is selected for analysis due to the purpose of this research, where each data entry is an representation of a trip. Yellow taxis are densely located in Manhattan, through predicting high value trips for Yellow taxis, we can encourage certain drivers to divert towards other boroughs while guaranteeing an adequate income. Green taxis and FHV trips are excluded from this research as they already carried the purpose in promoting transportation outside of Manhattan to a certain extent.

There are 19 attributes in the Yellow taxi data, the attributes of interest are the pickup date-time, location of Pickup, fare amounts, tips, and extras.

New York City's daily weather data provided by Visual Crossing[4] is also incorporated into this research as the weather conditions do have an impact on the decision of taking a taxi.

There are 28 attributes in the NYC weather data, the attributes of interest are maximum temperature, minimum temperature and precipitation.

1.2 Data range selection

The datasets selected ranges between January 2016 to May 2017, resulting in 180,457,103 yellow taxi trip entries. 2016 and 2017 is chosen as the years to research on our purpose to reduce the uncertainties inflicted by the COVID pandemic, to allows the general behaviour of the distributions of high value trips to be first understood, in-order to better model trip values under a more unstable setting. Also, it is also the years before 2018 where Green taxis experience a 55 percent drops in trips[2], suggesting that For-Hire services has already affected the taxi demand but not drastically. Such conditions allows a meaningful result to be derived under a hypothetical situation of where For-Hire services are not putting taxis out of business, allowing this research to study the distribution of high-value trips.

1.3 Assumptions

This study is based on such assumptions as there exists real life factors that are too complex to capture:

1. All passengers' payment are through credit cards.
2. NYC's daily weather condition can be represented by only temperature and precipitation.
3. The value (defined in detail later) of a trip is greater when the duration of the trip is short but the amount received is high.
4. Only temperature and precipitation affects the fuel economy (the distance a car can travel given an amount of fuel).
5. The value of a trip degrades by the same percentage proportion when the fuel economy degrades.
6. Airport fees, taxes, surcharges, tolls amount are not received by the Yellow taxi driver.

2 Pre-processing

2.1 Removal of invalid data

The following section removes invalid data entries to create better visualisation and a modelling result free from entries that do not have a meaningful impact.

- 61,037,121 entries that are not payment type 1 (credit card payments)
- 10,083 entries with total amount and fare amount less than $2.5 as $2.5 is the minimum charge.
- 5,807 entries with Vendor ID that does not belong to the vendors mentioned in the data dictionary provided NYCTLC .
- 49 entries with negative fare amounts.
- 2,084,487 entries with pickup and drop-off location IDs out of NYC (1-263).
- 507,084 entries with negative trip duration and longer than 10 hours is removed, as NYCTLC stated that the maximum aggregated drive period per 24 hour is 10 hours.
- 220,947 entries with trip dates outside of the defined dataset range.

In total 63,865,578 invalid data entries, approximately 35 percent of the raw data has been removed.

2.2 Feature Engineering

The following section creates new predictive attributes from the previous cleaned data.

Average temperature The weather dataset only provided maximum and minimum temperature of each day in NYC, hence, an average is taken to be the presume temperature at each hour of the day.

Normalize time Each trip is group into a 24 hour bin according to their pickup time, 11:59:59 or 11:00:00 is group into the 11th bin, and converted into a float between 0 and 1 to reduce the effect of magnitude on the regression model, this is further discussed in *Aggregation*. It is intuitive that under different times the amount of trips will change accordingly, thus, it is important to capture such information for the model.

Normalize day-time From figure 1 it is evident that across different times and weekdays, the amount of pickups varies. Such information is captured by representing weekday from 0 (Monday) to 6 (Sunday) and add in the time float and divide by 7 days a week. Giving 168 representations of values between 0 and 1 of the combination between hours and weekday.

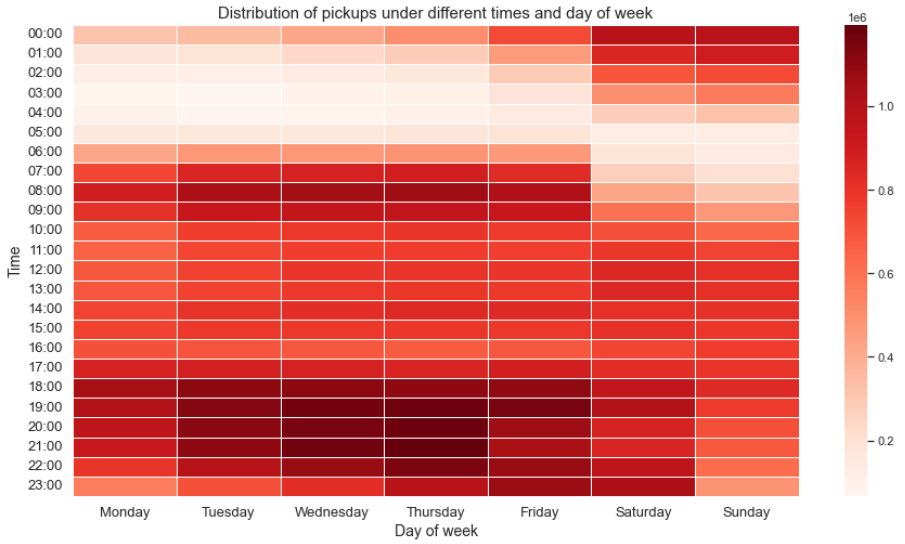


Figure 1: Heatmap of amount of pickups under different times and day of week

Normalize month From figure 2 it shows the number of trips being significantly different across different months. Such information is captured by representing month from 0 (January) to 11 (December) and divide by 12 months a year.

Holidays A Boolean column denoting whether the taxi trip was taken during a United States Federal holiday.

Cosine and Sine date-time floats To capture the cyclical patterns behind date-time features cosine and sine is used to transform the features according to a $0 - 2\pi$ cycle. Through the normal ordering of 23:58 and 00:02, it results in a 1336 gap where they are only 4 minutes apart. [5]

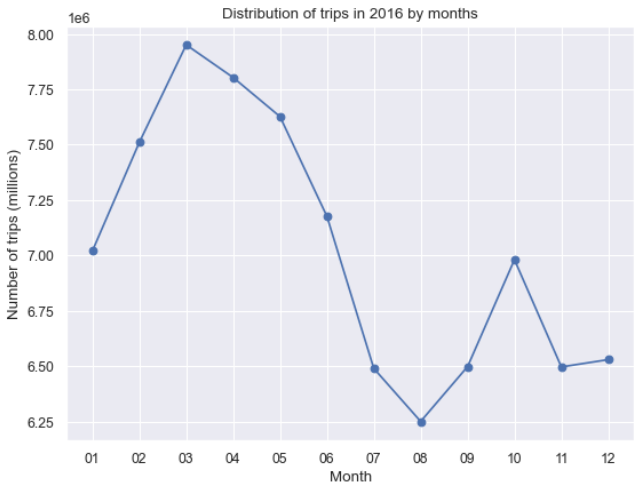


Figure 2: Monthly pickups in 2016

Trip duration Time taken between pickup and drop-off in minutes.

Trip Value A trip's value is defined as follow, it is the sum of the amount received by driver divided by the trip duration.

$$trip\ value = \frac{fare + tips + extra}{duration(minutes)} \tag{1}$$

The fuel economy suffers such percentage losses which also degrades the trip value when operating in such weather condition [6] [7] [8] , shown by table 1 and table 2.

	Temperature < 20F	20F < Temperature < 75F
Fuel economy loss	15%	10%

Table 1: The percentage loss of fuel economy under different temperature conditions

	Light Rain (<2.5 mm/hr)	Moderate rain (2.6 to 7.5 mm/hr)	Heavy rain (>7.6 mm/hr)
Fuel economy loss	1%	2%	4%

Table 2: The percentage loss of fuel economy under different precipitation conditions

2.2.1 Trip Value Analysis

To define high value trips, quantile analysis was performed to study the value of trips.

Minimum	First Quantile (Q1)	Median	Third Quantile (Q3)	Maximum
0.0036487211	0.7377777	0.9007679	1.1019744	1579.8486

Table 3: Quantiles of trip values

From table 3 it can be seen the values seems to be located around 0.7 to 1.1, extreme values like 1579 is possible under situations where the customer gave a huge tip to the driver. However, such situations are too scarce and could cause certain locations to be predicted to have a high number of high value trips. Intequartile Range (IQR) outlier detection is performed to allow the model to provide a more generalise result.

The standard method uses 1.5 times IQR as the bound for outliers. However, trip values that are higher than the third quantile or lower than the first quantile may have hidden information that provides ideas for the model to capture, such as the socioeconomic status of passengers at different pickup location which may lead to better tips, raising the value of the trip. Hence, the bound is widen to capture such information as seen below.

$$Q1 - 5 \times IQR < \text{trip value} < Q3 + 5 \times IQR \tag{2}$$

After the removal of outliers, table 4 shows the quantiles and it is decided that trips with value greater than the third quantile is determined as high value trips.

Minimum	First Quantile (Q1)	Median	Third Quantile (Q3)	Maximum
0.0036487211	0.7377777	0.9007679	1.1019744	2.8788321

Table 4: Quantiles of trip values

2.3 Aggregation

The purpose of this research is to predict the number of high value trips in NYC given a time period. It is decided that the data will be aggregated base on pickup location, day, and time. The granularity of time was chosen to be a per hour block, this allows drivers to maneuver around NYC with enough time to locations with abundant high value trips, such as heading from central Manhattan at 2:00pm to Staten Island at 2:45pm, indicating that at 3:00pm there are a large amount of high value trips.

After aggregation there are 1,291,280 entries. It is expected that there are locations at certain dates and time with no taxi trips taken, imputation was not performed for these structurally missing data to prevent bias to predict 0 high value trips on such locations at the time and date, as the entries are missing due to nonexistence.

3 Preliminary Analysis

3.1 Pickup location and the amount of high value trips

It is clearly evident of the relationship between pickup location and high value trips 3, the behaviour of Yellow taxi being centered around central Manhattan is justified as seen by large amounts of high value trips. High value trips also tends to be abundant around JFK and LaGuardia airport too, such as Woodhaven and Foresthills.

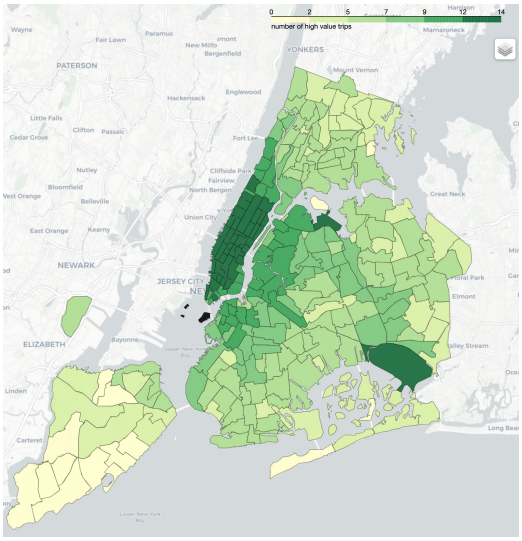


Figure 3: Total high value trips by location in 2016 under log scale

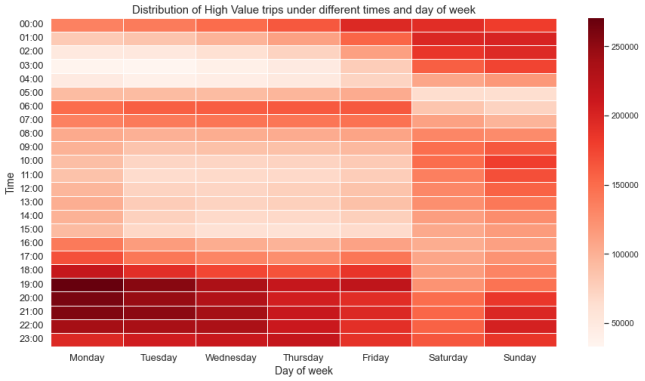


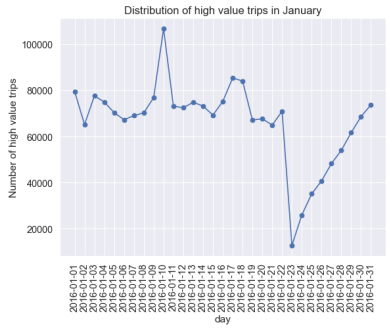
Figure 4: Total high value trips by week day and time in 2016

3.2 Datetime and the amount of high value trips

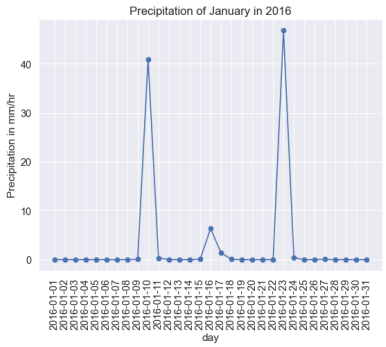
Figure 1 shows an abundance of trips at 7:00 to 16:00 across the weekdays, however, figure 4 suggests that at such times, most of the trips are not high value. This provides an intuition that although drivers working between 7:00 to 16:00 may have more passengers to pickup, the value of their trips may be lower than drivers working between 19:00 to 03:00. As passengers may be generous under the influence of alcohol or their eager to get home after work can cause them to take taxi trips back home that are a long distance from work.

3.3 Weather conditions and the amount of high value trips

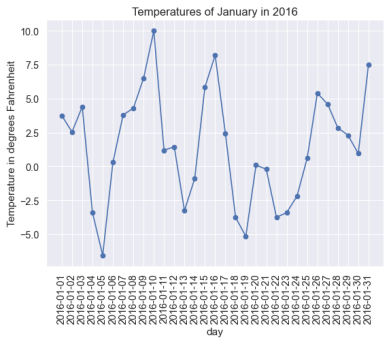
January and July was chosen as they are in months in the Winter and Summer seasons respectively. As the temperatures and precipitation are drastically different, providing a more intuitive understanding of the correlation between weather conditions and amounts of high value trips. Overall, in both seasons the number of high value trips peaks when precipitation is high, and drops when the temperature is too high or low. January 23 is a special case where precipitation is the highest, around 45 mm/hr(figure 5b), and temperature drops below 0 Fahrenheit(figure 5c), but a huge drop in amount of high value trips, it was caused by an extreme snow storm in NYC which paralysed the city's road. It is noted that extreme weather conditions may divert the outcome against the general trend of weather conditions and amount of high value trips.



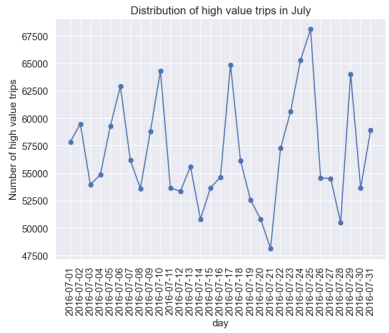
(a) Daily high value trips in January



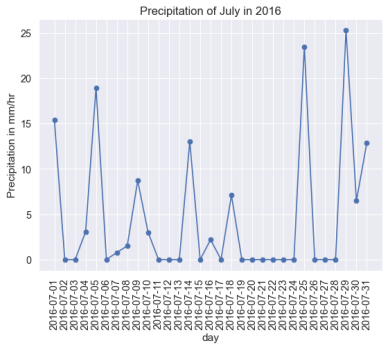
(b) Precipitation in January



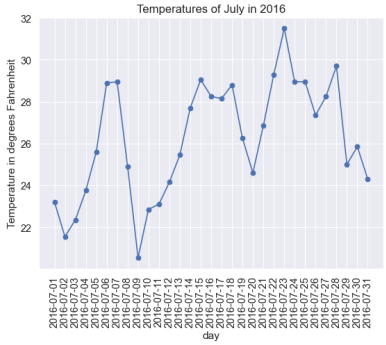
(c) Temperature in January



(d) Daily high value trips in July



(e) Precipitation in July



(f) Temperature in July

Figure 5: Daily weather conditions and amount of high value trips in January and July 2016

4 Modelling

4.1 Data preparation

The dataset was split into 2016 January to December for training, 2017 January to February for validation, and 2017 March to May for testing.

Pickup location ID was one-hot encoded due to being a categorical variable. The prediction target which is the aggregated amount of high value trips goes under a natural logarithmic transformation so that predictions are correct to an order of magnitude. As predicting 10 trips for an area with only 1 high value trips and 110 trips for an area with 100 high value trips, are treated as the same with out the logarithmic transformation.

4.2 Model training and results

4.2.1 Random Forest Regressor

Without tunning, the Random Forest Regressor achieved a R^2 training score of 0.99 with a training $RMSE$ of 1.16, however, its validation score was 0.904 with a validation $RMSE$ of 1.61. The training score exceed but approximately 0.9 suggest a moderate overfit by the model. With a training time of more than a hour, it is decided to reduce the number of trees and the depth to explore. It is expected that the model's performance may drop, but such decision is taken due to limited computational and time resources. After tuning the hyperparameter under restricted conditions, table 5 presents the results.

Table 5: Random Forest model results

	Train	**Valid**	**Test**
R^2 score	0.97967	0.98342	0.98221
$RMSE$	1.248	1.220	1.226

4.2.2 XGboost Regressor

XGboost Regressor is a tree based supervised learning algorithm which utilise boosting and gradient descent when predicting a target variable. Creating regression trees for predictions, the model minimizes a regularized objective function, and the process continues iteratively by adding new tress that predicts errors from prior tree. The trees are then combined for the final prediction[9]. After tuning the hyperparameter, table 6 presents the results.

Table 6: XGboost model results

	Train	**Valid**	**Test**
R^2 score	0.97718	0.98267	0.98236
$RMSE$	1.265	1.226	1.225

4.3 Model evaluation

Both models produced a convincing result of being approximately 98% accurate, there is a slight trace of underfitting for the models, however, it is welcomed in this research. It shows that the model did not overfit the training data which could damage the generalisation performance of the model, the purpose of this research is to improve the city's transportation infrastructure but guarantee a moderate source of income for taxi drivers. If the predictive performance of the models are degraded, drivers who are directed to wrongly predicted areas with high amounts of high value trips, they would have a high chance of picking up trips that may not cover the cost of them heading to such areas.

From the results presented in section 4.2, it seems that the Random Forest model performed better compared to the XGboost model. However, the boosting technique employed by XGboost may be more suitable in this scenario, boosting is an iterative learning process where more weightage is given to data points where it is wrongly predicted. Thus, when a prediction is given by the XGboost model it is assured that the prediction did not occur randomly, but with a strong understanding of the underlying information in the dataset. However, with Random Forest, due to the random element given to subset the dataset in creating decision trees, there are still random chances when making predictions. Such

randomness should be minimised, as there is a chance in harming the drivers' income, by 'randomly' directing the drivers to locations that are not beneficial.

Figure 6 presents predictions where they are round down to the nearest integer as we do not want to over estimate for the purpose of this research. It is seen that Random Forest Regressor tend to predict a higher amount of high value trips.

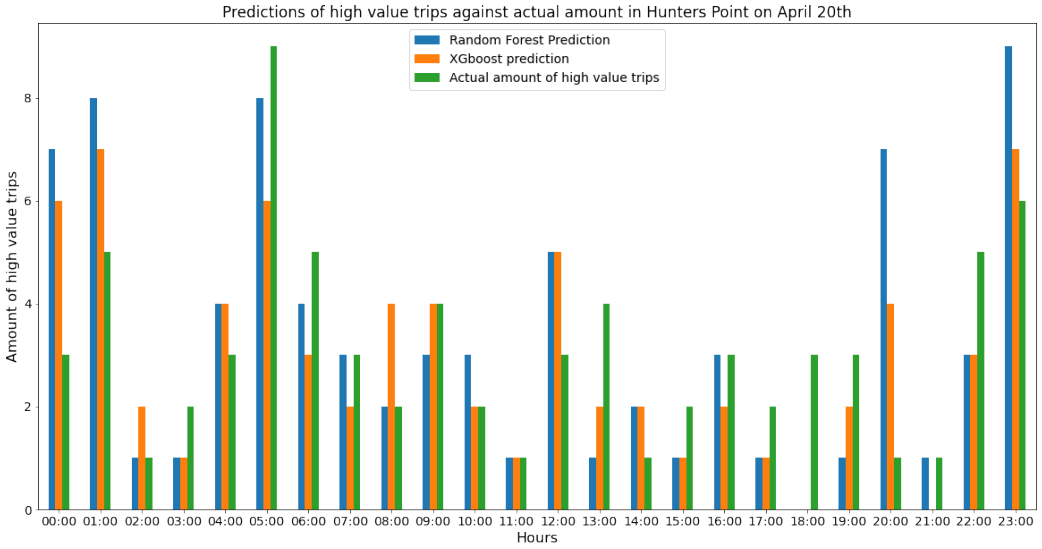


Figure 6: Model predictions against the actual amount of high value trips

5 Recommendations

It is recommended for the NYC government to look into producing such a model with a similar purpose of this research, to ensure policies that are designed to improve the city's infrastructure have a guarantee source of income for the stakeholders that are relied on.

For taxi companies and ride share services, it is recommended for them to deploy such models for their taxi drivers. The company could charge a usage fee on drivers for using such model to gain a higher revenue, but also to better allocate resources around NYC. The companies can employ a XGboost model using cloud computing, to reduce the cost of installation of external devices on taxis. Drivers can read-off predictions just from their phones, and arrive at high value locations more efficiently.

The definition of high value trips in this research is rather naive, real life factors are fully considered, cost of fuel prices, effect of added weight of customers and luggages on usage of fuels, and distance between current location and targeted location, these are some examples that could affect the value of trips. Further investigation on a trip's value should be proceeded to provide a more significant result, so that in the future models, it may be used for decision making for companies and governments as it could provide an understanding of NYC's transport structure. Allowing the development of different pricing or taxing strategy which they could used to provide a better infrastructure or more jobs.

In short, this report presented convincing results that aims to benefit Yellow taxi drivers, thus, in the future such finding could be manipulated to benefit NYC's stakeholders.

References

[1] newyorksimply. *What's With the Green Taxis in NYC?* https://newyorksimply.com/green-taxis-nyc-cab/. Accessed: 2022-08-11.

[2] The New York Times. *Where Yellow Cabs Didn't Go, Green Cabs Were Supposed to Thrive. Then Came Uber.* https://www.nytimes.com/2018/09/03/nyregion/green-cabs-yellow-uber.html. Accessed: 2022-08-12.

[3] NYCTLC. *TLC Trip Record Data.* https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page. Accessed: 2022-07-22.

[4] Visual Crossing. *Weather Data Services.* https://www.visualcrossing.com/weather/weather-data-services. Accessed: 2022-07-26.

[5] Pierre-Louis Bescond. *Cyclical features encoding, it's about time!* https://towardsdatascience.com/cyclical-features-encoding-its-about-time-ce23581845ca. Accessed: 2022-08-16.

[6] U.S. Department of Energy. *Fuel Economy in Cold Weather.* https://www.fueleconomy.gov/feg/coldweather.shtm. Accessed: 2022-08-16.

[7] FreightLiner. *How Adverse Weather Affects Fuel Economy.* https://freightliner.com/blog-and-newsletters/how-adverse-weather-affects-fuel-economy/. Accessed: 2022-08-16.

[8] Jan Barani. *Rain rate intensity classification.* https://www.baranidesign.com/faq-articles/2020/1/19/rain-rate-intensity-classification. Accessed: 2022-08-16.

[9] AWS. *How XGboost Works.* https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost-HowItWorks.html. Accessed: 2022-08-17.