


NewYork City Taxi Fare Prediction

Vipin Rai P, Harshabh Mahant, Saurabh Singh

November 2019

1 Data Overview

1. **Number of data points (TRAINING SET)** : 55 MILLION
2. **Number of features** : 8
3. **Features** :
 - (a) **key**: Unique identifier for a ride
 - (b) **amount**: Amount of fare charged for the ride
 - (c) **pickup datetime**: The date and time of the starting a ride.
 - (d) **pickup longitude**: The longitude coordinate of pickup location
 - (e) **pickup latitude**: The latitude coordinate of the pickup location
 - (f) **dropoff longitude**: The longitude coordinate of the dropoff location
 - (g) **dropoff latitude**: The latitude coordinate of the dropoff location
 - (h) **passenger count** : Total no of passengers in a particular ride



pickup_datetime	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
2010-10-20 23:26:26 UTC	-73.986910	40.739538	-73.991381	40.745614	2.0
2009-12-30 10:56:00 UTC	-73.961572	40.760283	-73.957438	40.769387	5.0
2012-07-20 11:24:00 UTC	-73.979437	40.746517	-73.984195	40.732117	1.0
2011-05-31 11:29:00 UTC	-73.964097	40.792508	-73.976422	40.785767	1.0
2010-05-25 17:57:00 UTC	-74.003943	40.725670	-73.988915	40.748370	1.0

Figure 1: Snapshot of dataset

2 Data Preprocessing

We have considered 20000000 rows for analysis and training the model. Preprocessing was done on the features to eliminate invalid data.

2.1 Latitude and Longitude values



Figure 2: Pickup and dropoff coordinates in the given data

1. It can be observed that some of the dropoff and pickups lie outside the boundary of NewYork city. Some coordinates are in ocean which is absurd and invalid.
2. NewYork is bounded by the coordinates (40.5774, -74.15) and (40.9176,-73.7004) so coordinates which are not within these range are not considered. Dropoff within NewYork are only considered.

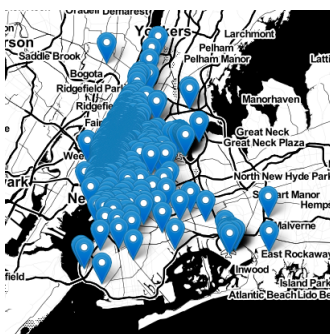


Figure 3: Pickup and dropoff coordinates after removing outliers

2.2 Fare amount

It was observed that fare amount ranges from 0 to 300. Most of the trips that are taken had fare amount less than 50. Distribution of fare amount were higher in the range 10 to 30.

2.3 Pickup Datetime

Pickup Date time is splitted into multiple columns:

1. Pickup Hour
2. Pickup Day of week
3. Pickup Year
4. Pickup Day of the month
5. Pickup Month

2.4 Passenger Count

1. Passenger count cannot be negative. So rows containing passenger count less than one was dropped.
2. It was observed that 69121 rows had values less than or equal to zero.

2.5 Trip Distance

1. This is the new feature created using pickup and drop latitude and longitude.
2. We have removed the outliers in latitude and longitude, which has led to removal outliers in distance feature.
3. Distance is calculated using python library "geopy.distance"

3 Exploratory Data Analysis

3.1 PDF

3.1.1 Distribution of fare amount

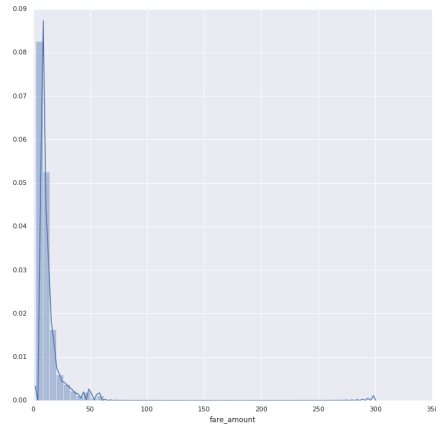


Figure 4: PDF of fare amount

It can be observed that fare amount ranges from 0 to 300. Most of the trips that are taken had fare amount less than 50. Distribution of fare amount were higher in the range 10 to 30.

3.1.2 Distribution of pickup hour

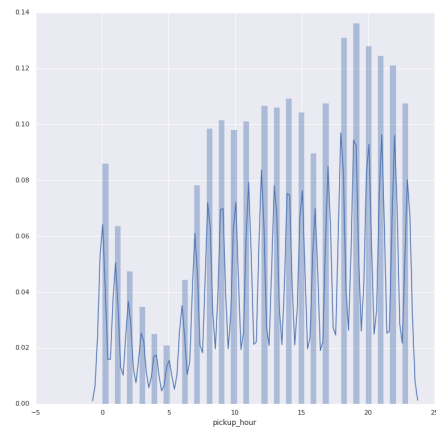


Figure 5: PDF of pickup hour

By observing the pickup hour of the ride it can be inferred that peak hours

were from 16:00 to 23:00. Very few rides were taken in the time slot from 02:00 to 06:00. So we can expect a high fare in this time slot.

3.1.3 Distribution of passenger count

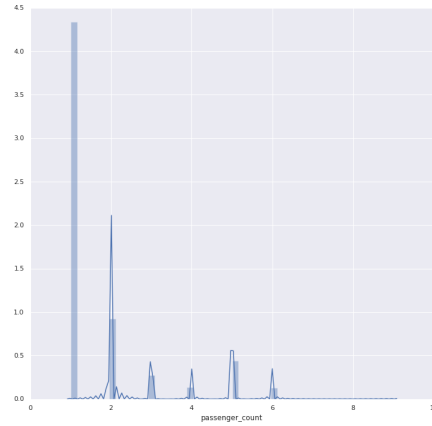


Figure 6: PDF of passenger count

Passenger count column having negative values were eliminated after pre-processing. It can be seen that most of the trips had a passenger count of 2.

3.1.4 Distribution of pickup day of week

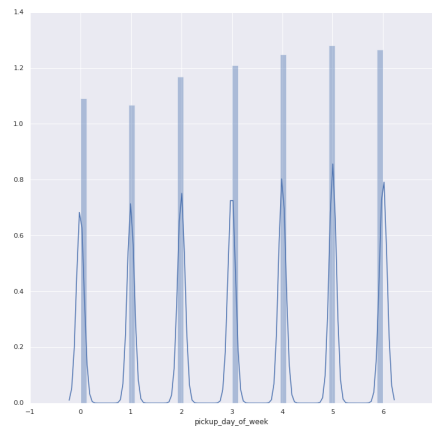


Figure 7: PDF of pickup day of week

Number of rides taken was almost same throughout the week, however there was only a slight increase in the rides for weekends (5: Friday and 6: Saturday).

3.2 Scatter plots

1. Scatter Plot shows positive co-relation between fare amount and trip distance. Increase in trip distance increases fare amount.

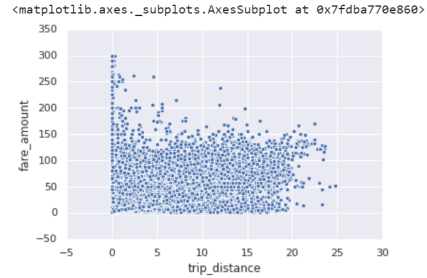


Figure 8: trip distance vs fare amount

2. Scatter Plot shows that fare amount is generally high from 12th to 15th hour of the day. It is low as compared to other hours from 5th to 8th hour of day.

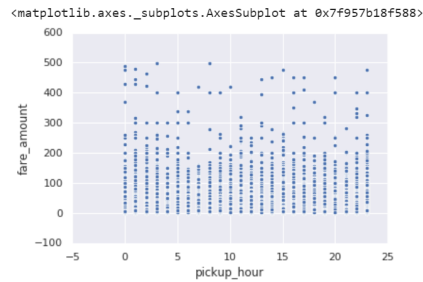


Figure 9: pickUp Hour vs fare amount

3. Scatter Plot shows negative co-relation between fare amount and passenger count. As no of passenger increases the fare amount drops.

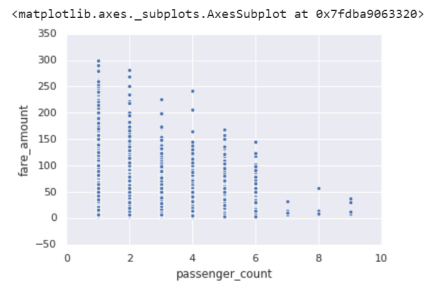


Figure 10: passenger count vs fare amount

3.3 Bar graphs

3.3.1 Pickups at different hours of Day:

It was observed that pickups are high at late hours generally from 18:00 Hrs to 23:00 Hrs. Pickups are minimal at early hours generally from 3:00 Hrs to 5:00 Hrs while fare amount is high during this time.

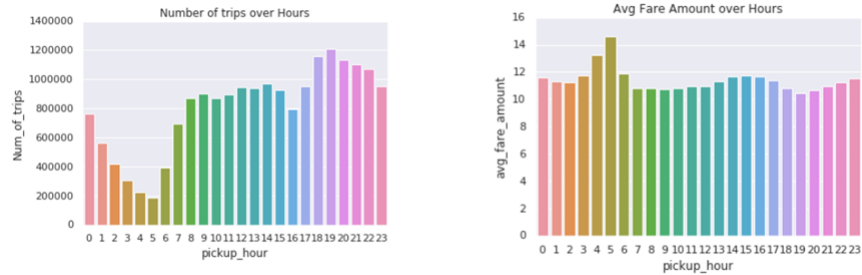


Figure 11: PickUp Hours

3.3.2 Pickups at different Days of week:

It was observed that pickups are generally high on 5th and 6th days of the week and slightly lower than normal pickups on 1st day of week.

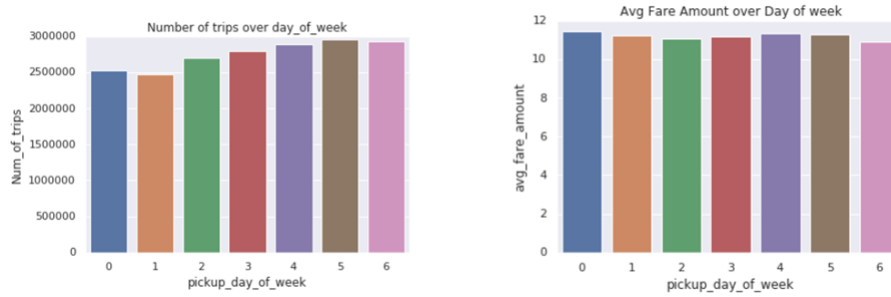


Figure 12: PickUp Day

3.3.3 Average fare amount over pickup Year



Figure 13: Distribution of fare amount over years

From the above figure we can see that average taxi fare has been increased over the years.

3.3.4 Pickups at different months:

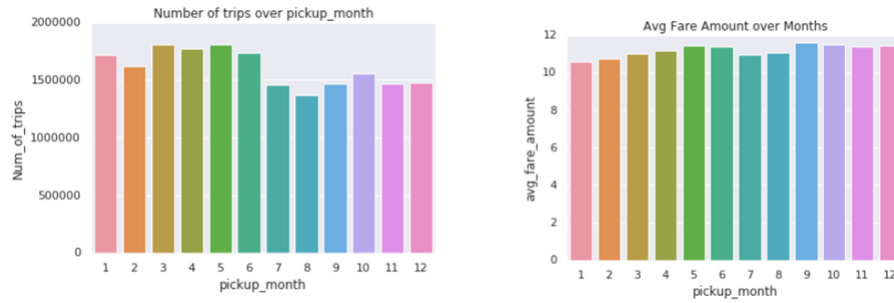


Figure 14: Distribution of fare amount over months

From the above figure we can see that there have been fewer pickups during the month of July to December but the average fare is almost constant

3.3.5 Average fare amount over passenger count

It was observed that Average Fare Amount was maximum 22 when passenger count was 8 and minimal 8 when passenger count was 0.

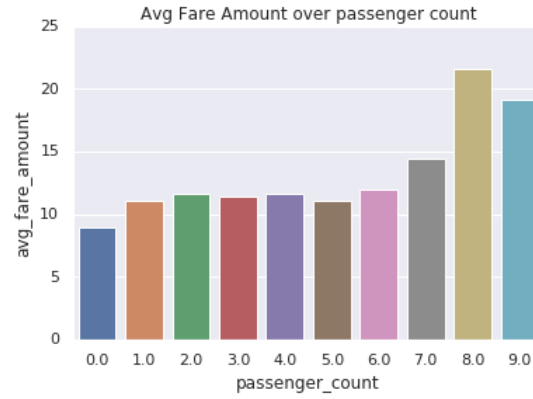


Figure 15: fare amount over Passenger Count

3.4 Fare amount distribution for airports



Figure 16: Pickup and dropoff fare - airport v/s non-airport

1. As we can see in above figure there is fare difference for airport pickup and drop versus non airport rides.
2. So we have added some columns to check if the ride is airport ride or non airport ride.
3. JFK and EWR are among the airports in NewYork and data for these airports is present in our dataset. So we have created new columns -
 - (a) is_pickup_EWR
 - (b) is_dropoff_EWR

- (c) is_pickup_JFK
- (d) is_dropoff_JFK

4 Process test data

Total number rows in test data are 11084772. As final submission is required to have same number of rows, we have not removed any rows but have below processing.

4.1 Splitting Pickup Date And Time

1. We have split pickup date time into below columns-
 - (a) pickup_year
 - (b) pickup_hour
 - (c) pickup_month
 - (d) pickup_day_of_week

4.2 Trip Distance

1. Trip Distance is calculated using pickup and drop latitude and longitude.
2. Exception handling is used to deal with outliers in latitude and longitudes.
3. For outliers distance is kept as zero.
1. As done in train data, we have added new columns to check if the ride is airport or non-airport.
2. Newly added columns are -
 - (a) is_pickup_EWR
 - (b) is_dropoff_EWR
 - (c) is_pickup_JFK
 - (d) is_dropoff_JFK

5 Model Building

Model used for prediction is XGboost. From the observations it can be inferred that features listed below are useful for prediction.

1. Trip distance
2. Pickup Longitude
3. Pickup Latitude

4. Dropoff Longitude
5. Dropoff Latitude
6. Pickup Year
7. Pickup Hour
8. Pickup Month
9. Passenger Count
10. Pickup Day of week
11. is_pickup_JFK
12. is_dropoff_JFK
13. is_pickup_EWR
14. is_dropoff_EWR

5.1 XGBoost

XGBoost model was trained with a sample size of 20 million rows. No of estimators used was 100 and trained with a learning rate of 0.15. Model was evaluated with root mean square metric. **RMSE value of 6.2974 was obtained on the kaggle test data.**

5.2 LightGBM

LightGBM model was used to predict the fare prices. Model was trained with 20 million train data with features listed above. Model was tuned to train with 5000 trees with 2000 number of iterations. Gradient boosting algorithm was used by the model to train. **RMSE value of 5.13099 was obtained on the Kaggle test data.**

6 Conclusion and Results

1. Data prepossessing was done on 20 million rows. Trips that are out of New York were eliminated and only within the city limits were considered.
2. New features like trip distance was derived from latitude and longitude values.
3. Exploratory data analysis gave insights about the data. Features containing invalid data were eliminated.
4. Model was trained on 20 million rows with features listed above. Root mean square of 5.13099 was obtained using LightGbm which was an improvement compared to XGBoost.