Analysis on Taxi Trip Duration in New York City

Shehan Ishanka, Dinesh Sandaruwan, Vibodha Vimarshana
Department of Computer Science and Engineering
University of Moratuwa
Sri Lanka
{shehan.20, dinesh.20, vibodha.20}@cse.mrt.ac.lk

***Abstract--* Taxi trip duration has become a vital figure for the taxi companies. All of them are recording the duration and are keen on knowing the facts underneath the recorded trip durations. In this analysis the factors that affecting the taxi trip duration has analyzed in three steps as Descriptive Analysis, Diagnostic Analysis and Predictive Analysis. Finally the result of the model has presented which trained with the factors identified during the analysis.**

***Keywords---* Trip Duration, Descriptive Analysis, Diagnostic Analysis, Predictive Analysis**

I. INTRODUCTION

Trip duration is an important metric with respect to a taxi company. Most of the operational decisions like taxi allocation for a given region, costing for taxi rides and strategic decisions like advertising would be based on this. In addition to that predicting the trip duration in advance of the trip would improve the standards of the taxi service.

Predicting the precision trip duration at the start of the taxi trip has become a challenge since the distance between the two locations is not the only factor affecting the duration. The time of the day, pick-up and drop-off location type (City or Suburb) and traffic conditions are some other factors that indirectly affect this duration. Therefore, by analyzing the pick-up and drop-off locations with the duration and the time of the day may provide more insight to the duration of a taxi ride.

Mainly two approaches can be used for predicting duration of the trip. Those two would be path-based method and origin-destination method. For the path-based method a rich data set describing the path of the ride would be required and this will cause a requirement of complex methodologies for handling those data. Therefore, in this paper origin-destination method will be used since that approach requires only the data related to the pick-up and drop-off points. This will reduce the calculation time and the complexity of the methodologies.

In this paper a descriptive analysis will be performed on the trip durations. Then a diagnostic analysis will be performed to identify the factors affecting the trip duration following a predictive analysis which will be based on the factors identified during the diagnostic analysis.

II. DATASET

A. Data description

Data set is collected from Kaggle competition of New York City taxi trip duration competition [1]. Data set contains 1458644 trip records which were collected over six months. Data fields are given below.

- id - a unique identifier for each trip
- vendor_id - a code indicating the provider associated with the trip record
- pickup_datetime - date and time when the meter was engaged
- dropoff_datetime - date and time when the meter was disengaged
- passenger_count - the number of passengers in the vehicle (driver entered value)
- pickup_longitude - the longitude where the meter was engaged
- pickup_latitude - the latitude where the meter was engaged
- dropoff_longitude - the longitude where the meter was disengaged
- dropoff_latitude - the latitude where the meter was disengaged
- trip_duration - duration of the trip in seconds

B. Data features

1) Column vendor_id

There are only two vendors in this data set. Vendor distribution over the average trip duration is as follows.

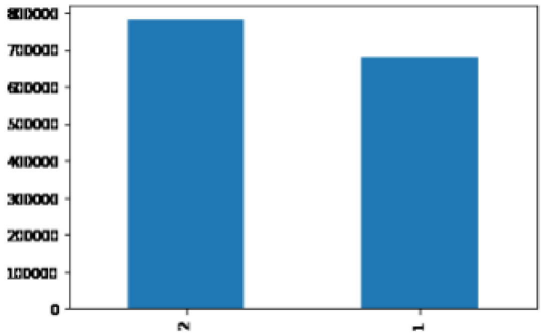

Fig. 1 vendor_id Distribution

2) Column passenger_count

Passenger count for each trip is mentioned in the data set and passenger count is ranging from 0 to 8.

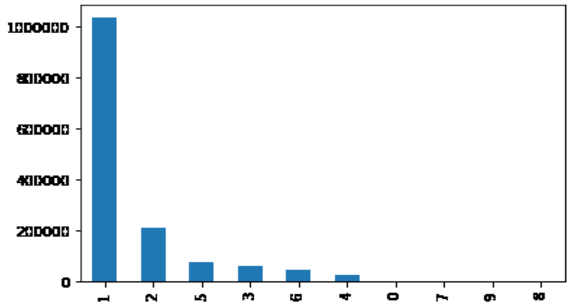

Fig. 2 passenger_count Distribution

3) Column pickup_datetime

Using the passenger picked up date time, year, month, day of the week, hour and minute data are extracted.

Day wise trip count for each hour is depicted below. As the figure shows in week days, the trip count pattern is in contrast with the weekend trip count pattern. On weekdays most of the trips are done in the evening while on weekends most of the trips happen in mid night.

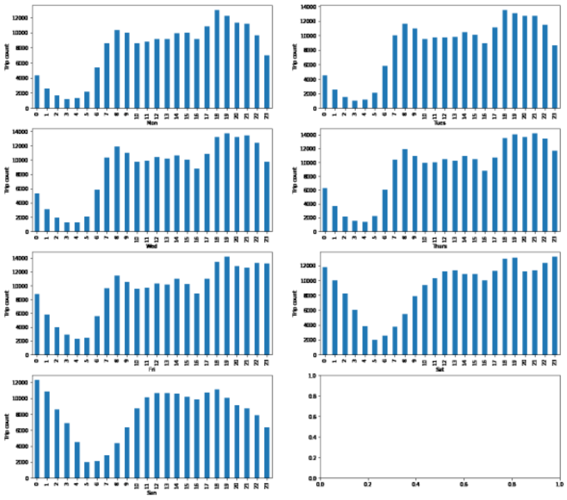

Fig. 3 pickup_hour distribution over days

Following graph is again broken down by month to identify patterns with respect to month.

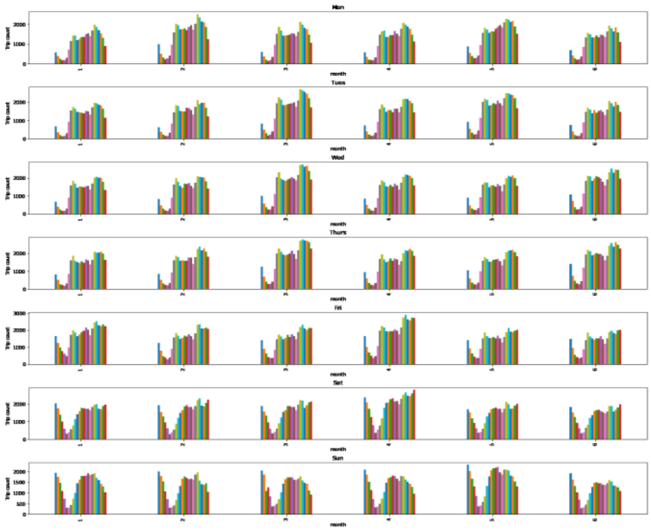

Fig. 4 pickup_month distribution over days

4) Pick-up and Drop-off points

Following are heat maps drawn using pick-up and drop-off points

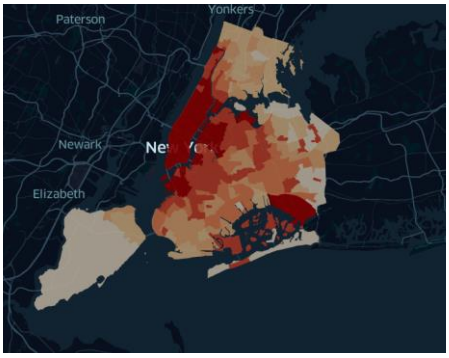

Fig 5 Pick-up locations

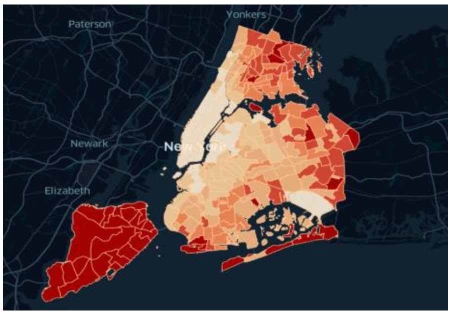

Fig 6 Drop-off locations

The data set is plotted Geo spatially using Hex bins[2]. Pick up Hex bins and Drop Hex bins are plotted separately.

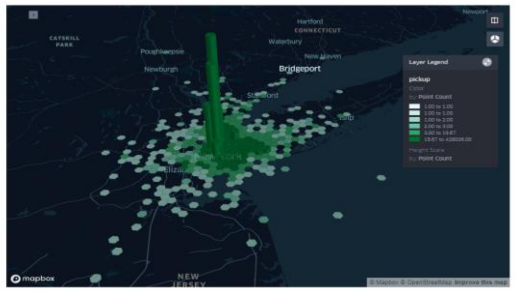

Fig. 7 Pick up Hex bins

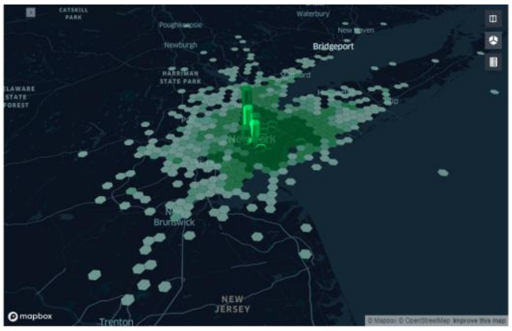

Fig. 8 Drop Hex bins

III. ANALYSIS

According to Gartners Maturity Model, there is a four step Analytical process which provides insights on a particular analysis.

- Descriptive analytic
- Diagnostic analytic
- Predictive analytic
- Prescriptive analytic

Therefore, regarding trip duration, each of those perspectives is analyzed according below.

A. *Preliminary analysis*

A basic analysis is gone through the trip duration feature. Using all the trip data, trip duration distribution is formed with different durations.

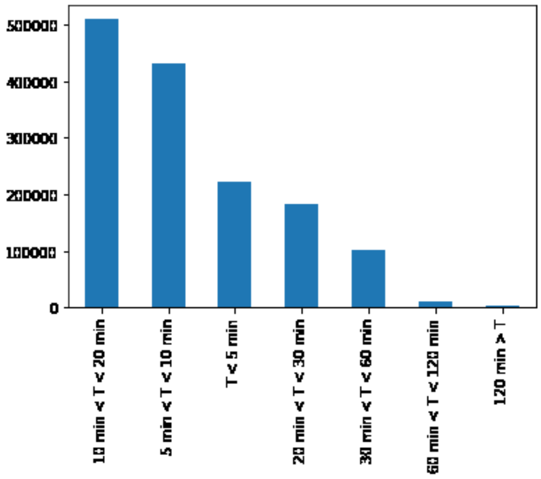

Fig. 9 Trip time duration distribution

It is apparent that most of the trips reside within 10 to 20 minutes time periods. Furthermore there are trips which have more than 2 hour duration. Geo spatial distribution of those trips with windows are as follows.

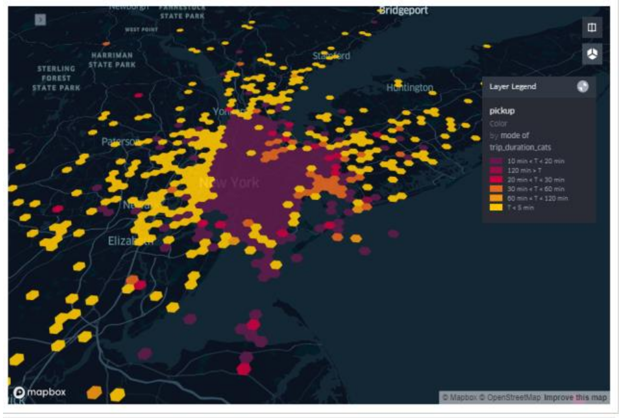

Fig. 10 Pick up Hex bin geo spatial view over trip duration bins

In addition to that, trip durations are analyzed throughout week days and months.

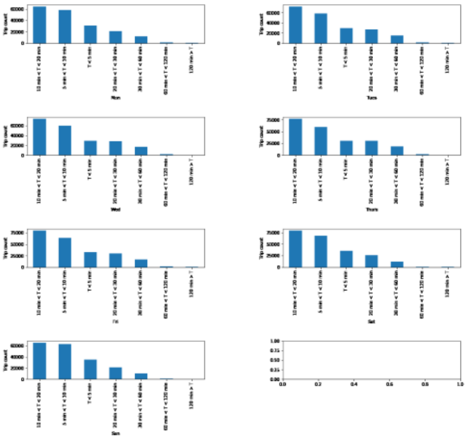

Fig. 11 Trip time duration distribution over days

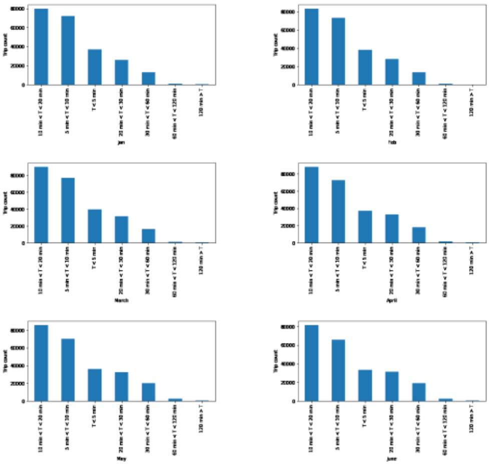

Fig. 12 Trip time duration distribution over months

Comparatively, towards mid of the week trips with duration less than 5 minutes are decreasing.

B. *Descriptive Analysis*

Descriptive analysis of trip duration is carried out analyzing “How much time has it been taken to go from some place to another place?” This is approached in two methods. In all these methods the training data set is used and it is again splitted into two subsets of train (0.8) and test (0.2) sets. Both these approaches are intended to find out duration estimation using historical data.

1. Duration estimate by location using six months data (A)

Trip duration is averaged from same pickup Hex bin to same drop Hex bin

2. Duration estimate by location using six months data with time factor (B)

Trip duration is averaged from same pickup Hex bin to same drop Hex bin in same week day and same hour.

The averaged values are compared with duration in the test and Root Mean Square Error (RMSE) in minutes is calculated in each approach.

TABLE I
RMSE OF APPOACH A AND APPROACH B

Approach A	Approach B
54.29 min	58.26 min

Further analysis proved that Approach A has lesser RMSE because in Approach B most of the trips could not be estimated due to lack of trips in train sets in certain hours in certain weekdays. Approximately 21% trips could not be estimated in Approach B while only 1% of trips are not estimated in Approach A.

C. *Diagnostic Analysis*

Diagnostic analysis of trip duration is carried out analyzing “What are the factors that affect to vary the time durations?” For this analysis, trips are filtered from pick Hex bin of ‘882a100d23fffff’ to drop Hex bin of ‘882a100d65fffff’.

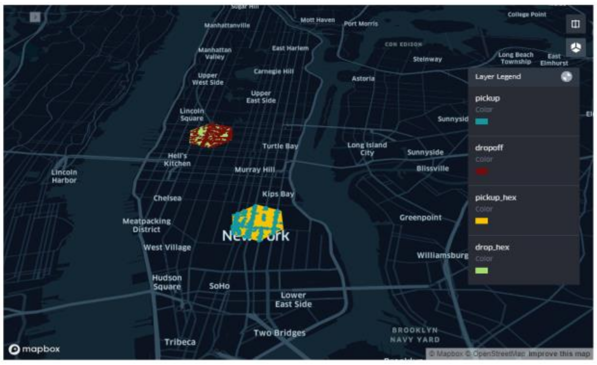

Fig. 13 Pick up and Drop Hex bin geo spatial view

649 trips are included in these locations. Trips between these two locations are distributed as follows.

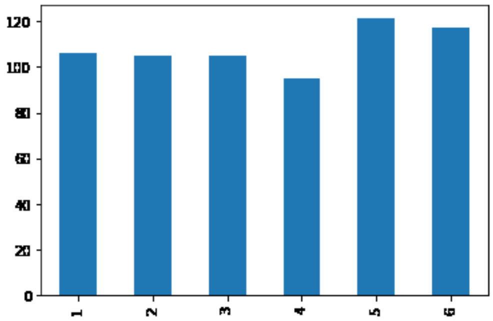

Fig. 14 Trip count distribution over months

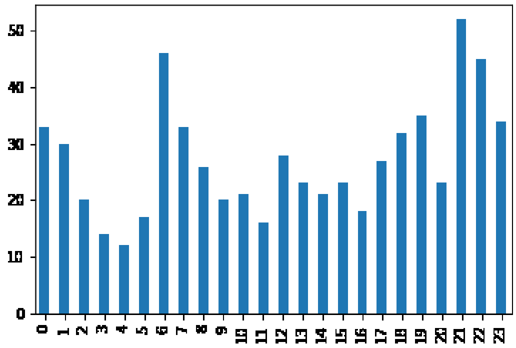

Fig. 15 Trip count distribution over hours

Trip duration distributions for each week day are as below.

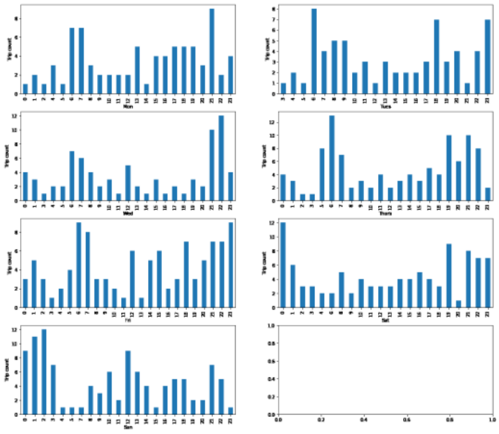

Fig. 16 Trip count distribution over days

Comparatively most of the trips are completed in 6th, 7th and 21st hours. Most of the trips are completed within 20 to 40 minutes between these 2 locations.

Using OSRM (Open Source Routing Machine) API[3] durations for each trip is estimated and trip paths are formed. OSRM duration is calculated without accounting traffic of the path and it assumes that the path generated between two locations is the shortest path.

An analysis is done regarding traffic in each weekday using the above data. The below assumptions are made for that assumption.

1. All the trips from pick up Hex to Drop Hex have used the route
2. Difference between actual trip duration OSRM duration resembles traffic.

Above diagram depicts the duration difference for each day. This kind of variation in traffic encompasses the trip duration variations in selected locations.

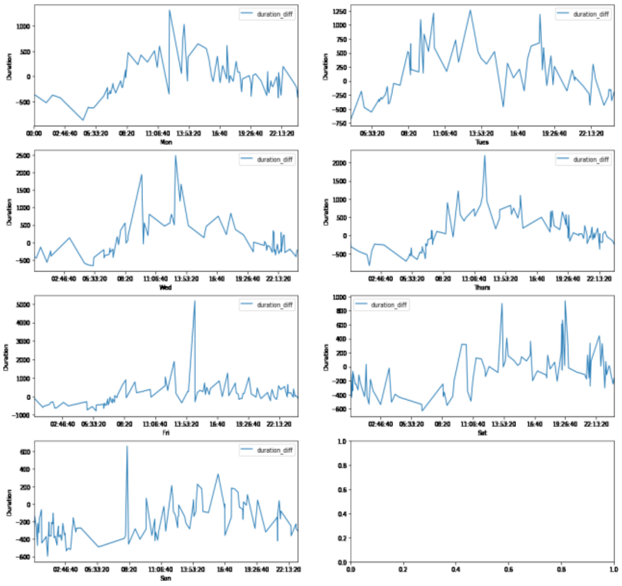

Fig. 17 Trip time duration box plot

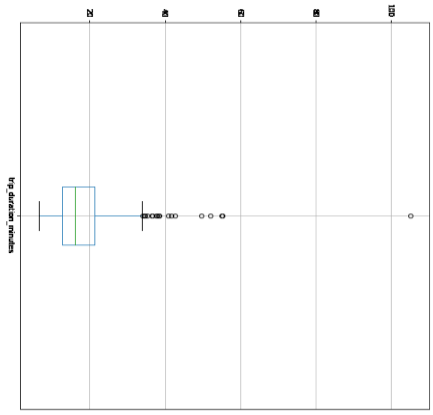

Fig. 18 Derived traffic approximation over days

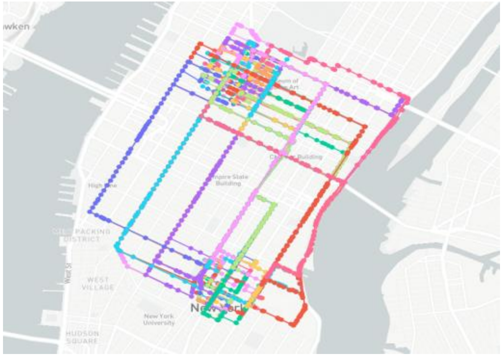

Fig. 19 Routes from Pick up to Drop

Further, OSRM routes are visualized to take an idea of the routes between these locations.

Following are the starting points of the trips having the same pick-up location and the same drop-off location in the space of 15 minutes in the month of April.

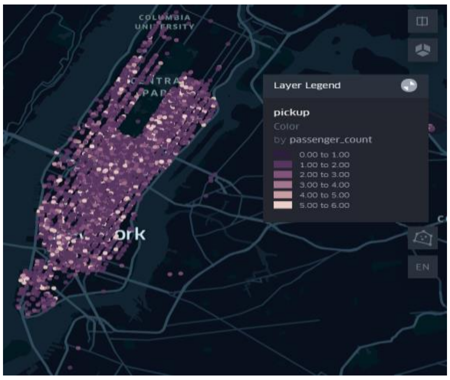

Fig. 20 Passenger counts of relevant trips

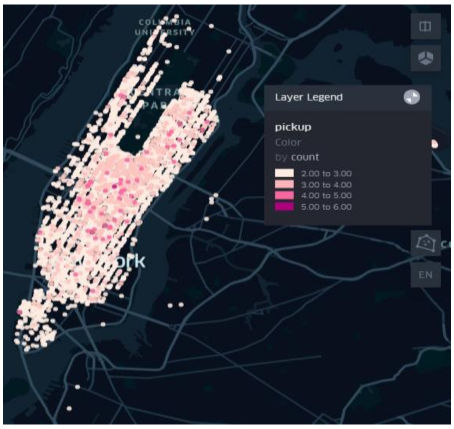

Fig. 21 Trip counts

As it can be observed the range of the count of the trips which share the same start location and end location in the space of 15 minutes in the month of April is from 2 to 6. The passenger count for these trips are mostly 1. From this what can be deduced is there are considerable rides that can be shared.

D. Predictive analytics

Predictive analytics is the most appropriate assessment which explains what will happen in future. Identification of the likelihood of future outcomes can be done by using statistical algorithms and machine learning techniques which are based on historical data.

1) Derived features

The features which are derived from raw data are mentioned below.

- Pickup_day_of_week - This feature is derived from pickup datetime
- Pickup_hour - This feature is derived from pickup datetime
- Pickup_cell_center_lat -This feature is derived from pickup latitude by using H3 library
- Pickup_cell_center_lon - This feature is derived from pickup longitude by using H3 library
- Dropoff_cell_center_lat - This feature is derived from dropoff latitude by using H3 library
- Dropoff_cell_center_lon - This feature is derived from dropoff longitude by using H3 library
- Osrm_distance - This feature is derived from pickup location and drop off location by using OSRM (Open Source Routing Machine) API
- Osrm_duration - This feature is derived from pickup location and drop off location by using OSRM (Open Source Routing Machine) API
- Average_velocity - This feature is derived from osrm_distance and trip_duration (actual) using formula as below.

$$\boldsymbol{Average\_velocity} = \frac{\boldsymbol{OSRM\ Distance}}{\boldsymbol{Trip\ Duration}}$$

2) *Data Preprocessing Steps*

1. Removed all trips which were not returned a route by OSRM API
2. Removed the all trips which the OSRM distance is more than 15 km (OSRM distance of 94% trips is between 0-15 km)
3. Removed the all trips which has the average velocity more than 50km/h and less than 5 km/h (average velocity of 96% trips is between 5-50 km/h)

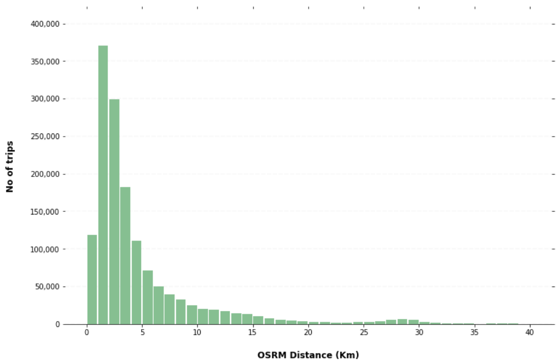

Fig 22 OSRM distance vs. No. of trips

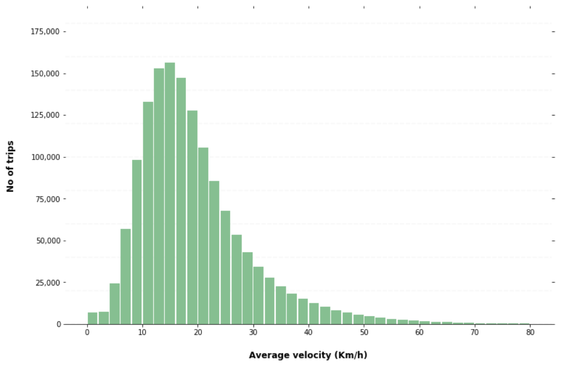

Fig. 23 Average velocity vs. No. of trips

3) *Extracted Features*

The features which are used to train the predictive model are mentioned below.

- Pickup_day_of_week
- Pickup_hour
- Pickup_cell_center_lat
- Pickup_cell_center_lon
- Dropoff_cell_center_lat
- Dropoff_cell_center_lon
- Osrm_distance
- Trip_duration (Label)

4) *Modeling*

As the modeling approach, we chose a simple linear regression algorithm and Gradient Boosting algorithm. Here we used Scikit-Learn, which is one of the most popular libraries in machine learning. Here we are going to predict the travel time of the trip based on its pickup day of week, pickup hour, pickup location, drop-off location and so on.

5) *Model Evaluation*

The various metrics used to evaluate the results of the predictions and here we used the following metrics.

Mean Squared Error (MSE)

MSE or Mean Squared Error is one of the most preferred metrics for regression tasks. It is simply the average of the squared difference between the target value and the value predicted by the regression model.

Root-Mean-Squared-Error (RMSE)

RMSE is the most widely used metric for regression tasks and is the square root of the averaged squared difference between the target value and the value predicted by the model

TABLE II
ACCURACY VALUES

Model	MAE (min)	RMSE (min)
Linear Regression Model	3.8462	5.4252
Gradient Boosting Model	2.9191	4.2967

6) *Benchmark Test*

There is a graph given below which represents MAE variation of OSRM estimation and Linear Regression model prediction along with distance. That provides better insight about trained Linear Regression model.

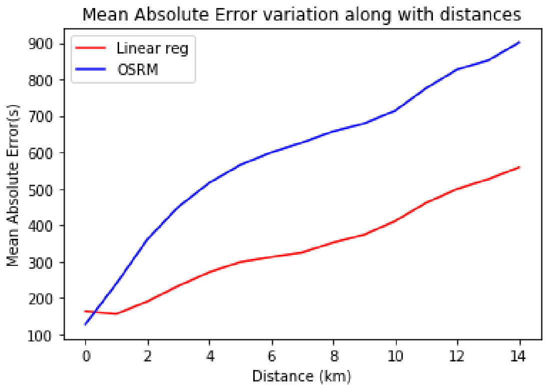

Fig. 24 MAE variation of OSRM estimation and Linear Regression model prediction along with distance

The following represents MAE variation of OSRM estimation and Gradient boosting model prediction along with distance. That provides better insight about trained Gradient Boosting model.

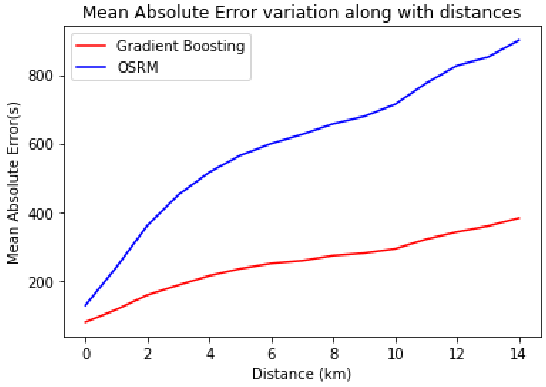

Fig. 25 MAE variation of OSRM estimation and Gradient Boosting model prediction along with distance

IV. CONCLUSION

The analysis is based on the New York City taxi trip data of first six months data of 2016. Preliminary analysis provides the insights on the data set on each column. The rest of the analysis is gone through according to Gartner’s Maturity Model. First the descriptive analysis is followed based on the question of “How much time has it been taken to go from some place to another place?” By splitting the data set, and averaging trip time duration on different conditions it was assessed. Then the diagnostic analysis is followed based on the question of “What are the factors that affect to vary the time durations?” Trips are identified from a particular Pick up area to drop and using them route analysis and traffic estimation is done. Then the predictive analysis is done using some models to predict a trip duration time.

Traffic estimation concept can be utilized to set markups on the trip when there is high demand. Furthermore, more markups can be set up from passenger pooling. The diagnostic analysis and the predictive analysis can be further improved by integrating weather data and route data.

REFERENCES

[1] Kaggle.com. 2020. New York City Taxi Trip Duration | Kaggle. [online] Available at: <https://www.kaggle.com/c/nyc-taxi-trip-duration/data> [Accessed 9 April 2020]

[2] H3geo.org. 2020. H3. [online] Available at: <https://h3geo.org/#/documentation/overview/introduction> [Accessed 10 April 2020]

[3] Project-osrm.org. 2020. OSRM API Documentation. [online] Available at: <http://project-osrm.org/docs/v5.5.1/api/#general-options> [Accessed 10 April 2020].