INST 737- Introduction to Data Science

Predicting Taxi Ride Duration

COLLEGE OF
INFORMATION
STUDIES

Final Project Report

Reported By

Aakanksha Singh
Lindsay Huth
Mayanka Jha
Riya Chanduka

EXPECTATIVE SUMMARY: NEW YORK TAXI TRIP TIME PREDICTION

The dataset is obtained from the NYC Taxi and Limousine Commission (TLC). This report will be demonstrating on the taxi trip time prediction from the following predictors: pickup location, drop off location, and pickup date and time.

INTRODUCTION

There are roughly 200 million taxi rides in NYC each year. Analysis and understanding of taxi supply and demand could increase the efficiency of the city's taxi system. Predicting taxi ridership could present valuable insights to city planners and taxi dispatchers.

DATASET OVERVIEW

This dataset is collected by the NYC Taxi and Limousine Commission (TLC) and includes trip records from all trips completed in Yellow and Green taxis in NYC from 2009 to present. Records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts.

DATA STRUCTURE

- There are 1458644 rows with 11 variables in train.
- There are 625134 rows with 9 variables in test.

Train has two additional fields *trip_duration* and dropoff_datetime to the test set. The variable *trip_duration* is the independent, response variable we are trying to predict and is derived as the difference between dropoff_datetime and pickup_datetime. Each row in the datasets represent one taxi trip. All variable headings are populated.

train.csv

A	B	C	D	E	F	G	H	I	J	K
id	vendor_id	pickup_datetime	dropoff_datetim	passenger_c	pickup_longi	pickup_latitu	dropoff_long	dropoff_latit	store_and_fv	trip_duration
id2875421	2	3/14/16 17:24	3/14/16 17:32	1	-73.982155	40.7679367	-73.96463	40.7656021	N	455
id2377394	1	6/12/16 0:43	6/12/16 0:54	1	-73.980415	40.7385635	-73.999481	40.7311516	N	663
id3858529	2	1/19/16 11:35	1/19/16 12:10	1	-73.979027	40.7639389	-74.005333	40.7100868	N	2124
id3504673	2	4/6/16 19:32	4/6/16 19:39	1	-74.01004	40.7199707	-74.012268	40.7067184	N	429
id2181028	2	3/26/16 13:30	3/26/16 13:38	1	-73.973053	40.7932091	-73.972923	40.7825203	N	435
id0801584	2	1/30/16 22:01	1/30/16 22:09	6	-73.982857	40.7421951	-73.992081	40.7491837	N	443
id1813257	1	6/17/16 22:34	6/17/16 22:40	4	-73.969017	40.7578392	-73.957405	40.7658958	N	341
id1324603	2	5/21/16 7:54	5/21/16 8:20	1	-73.969276	40.7977791	-73.92247	40.7605591	N	1551
id1301050	1	5/27/16 23:12	5/27/16 23:16	1	-73.999481	40.7383995	-73.985786	40.7328148	N	255
id0012891	2	3/10/16 21:45	3/10/16 22:05	1	-73.981049	40.744339	-73.973	40.7899895	N	1225
id1436371	2	5/10/16 22:08	5/10/16 22:29	1	-73.982651	40.7638397	-74.002228	40.7329903	N	1274
id1299289	2	5/15/16 11:16	5/15/16 11:34	4	-73.991531	40.7494392	-73.956543	40.7706299	N	1128
id1187965	2	2/19/16 9:52	2/19/16 10:11	2	-73.962982	40.7566795	-73.984406	40.7607193	N	1114
id0799785	2	6/1/16 20:58	6/1/16 21:02	1	-73.956306	40.7679405	-73.96611	40.7630005	N	260

test.csv

id	vendor_id	pickup_datetime	passenger_c	pickup_longi	pickup_latitu	dropoff_long	dropoff_latit	store_and_fwd_flag
id3004672	1	6/30/16 23:59	1	-73.988129	40.732029	-73.990173	40.7566795	N
id3505355	1	6/30/16 23:59	1	-73.964203	40.6799927	-73.959808	40.6554031	N
id1217141	1	6/30/16 23:59	1	-73.997437	40.7375832	-73.98616	40.7295227	N
id2150126	2	6/30/16 23:59	1	-73.95607	40.7719002	-73.986427	40.7304688	N
id1598245	1	6/30/16 23:59	1	-73.970215	40.7614746	-73.96151	40.7558899	N
id0668992	1	6/30/16 23:59	1	-73.991302	40.7497978	-73.980515	40.7865486	N
id1765014	1	6/30/16 23:59	1	-73.97831	40.7415504	-73.952072	40.7170029	N
id0898117	1	6/30/16 23:59	2	-74.012711	40.7015266	-73.986481	40.7195091	N
id3905224	2	6/30/16 23:58	2	-73.992332	40.7305107	-73.875618	40.8752136	N
id1543102	2	6/30/16 23:58	1	-73.993179	40.7487602	-73.979309	40.7613106	N
id3024712	1	6/30/16 23:58	4	-73.968529	40.6784325	-73.966591	40.6357117	N
id3665810	2	6/30/16 23:58	1	-73.982773	40.7569084	-73.974693	40.7533302	N
id1836461	1	6/30/16 23:58	1	-73.921104	40.767292	-73.936859	40.774044	N
id3457080	2	6/30/16 23:57	1	-73.986801	40.7349167	-73.975899	40.7568932	N
id3376065	1	6/30/16 23:57	1	-73.996346	40.7481613	-73.950829	40.7828255	N
id3008739	1	6/30/16 23:57	1	-73.968025	40.7622833	-73.934792	40.7974358	N
id0902216	2	6/30/16 23:56	1	-74.007713	40.7406807	-73.968811	40.7538605	N

EXPLANATORY VARIABLES/ FEATURES

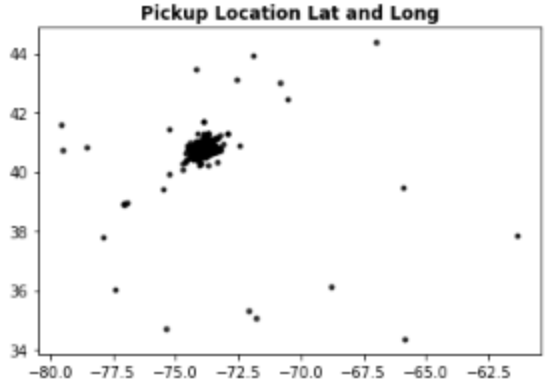
- **id** character. A unique identifier for each trip.
- **vendor_id** integer. A code indicating the provider associated with the trip record. There appears to be 2 taxi companies.
- **pickup_datetime** character. The date and time when the meter was engaged. This is currently a combination of date and time.
- **dropoff_datetime** character. The date and time when the meter was disengaged. As above this is a combination of date and time.
- **passenger_count** integer. The number of passengers in the vehicle (driver entered value). This is a count from up to 9.
- **pickup_longitude** numeric. The longitude where the meter was engaged. These are geographical coordinates and appear to be in the correct format.
- **pickup_latitude** numeric. The latitude where the meter was engaged.
- **dropoff_longitude** numeric. The longitude where the meter was disengaged.
- **dropoff_latitude** numeric. The latitude where the meter was disengaged.

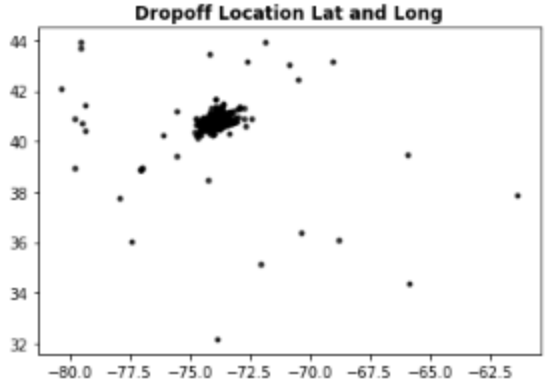
RESPONSE VARIABLE/ OUTCOME

- **trip_duration** integer. The duration of the trip in seconds.

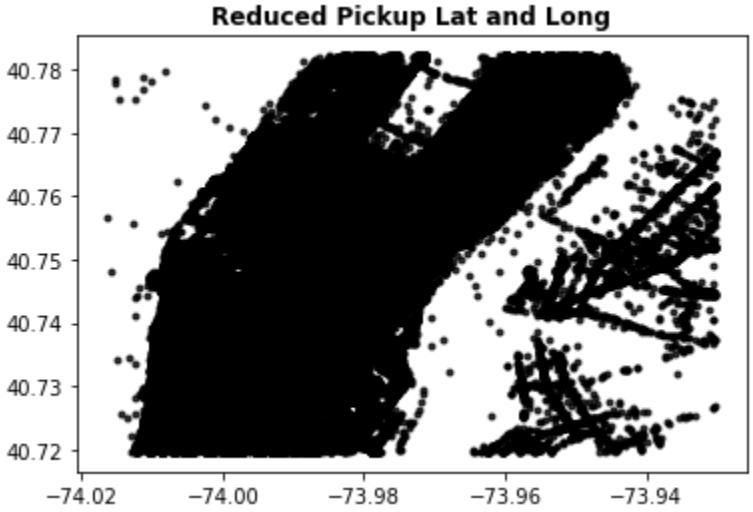
DATA CLEANING AND ORGANIZATION

Once we loaded our data, we checked for nulls -- luckily, we had none in our dataset. We also removed obvious outliers in the pickup and dropoff latitudes and longitudes.

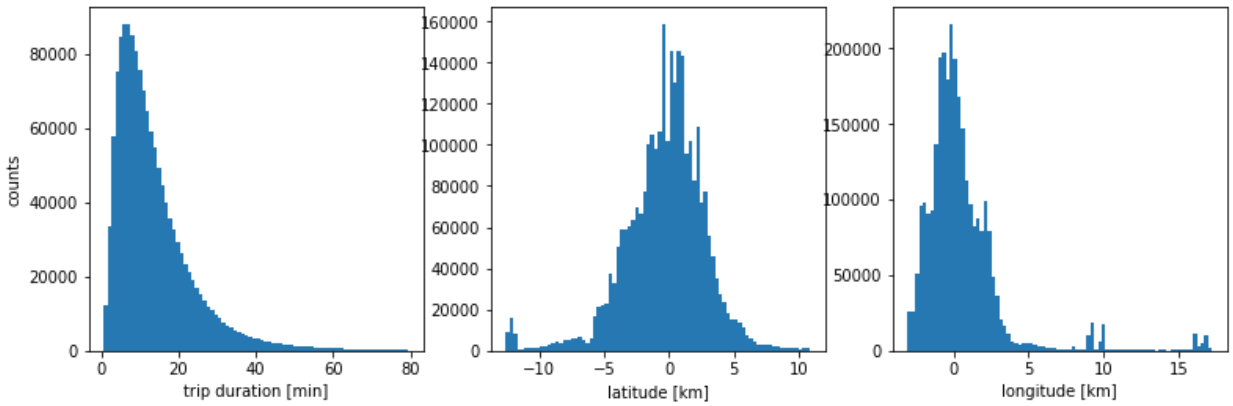




After removing the outliers from pickup and dropoff columns the following graph was obtained.

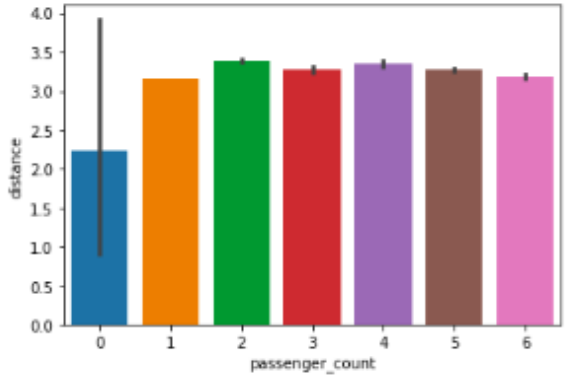


In order to make our data more usable, we converted the trip_duration from seconds to minutes and the pickup and dropoff locations from latitude/longitude to kilometers. Below are histograms of our values for trip_duration in minutes and the pickup_latitude and pickup_longitude in km:

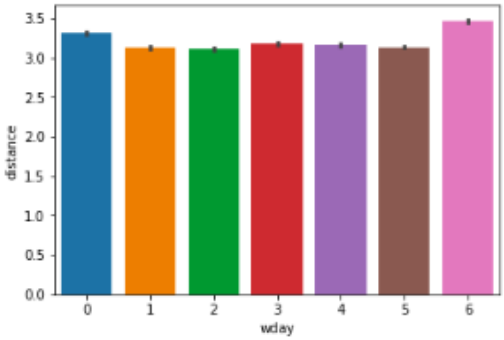


EXPLORATORY DATA ANALYSIS

To understand what factors might affect trip_duration -- which is related to the distance the car travelled -- we plotted the distance by the number of passengers. Trips with 0 passengers travelled shorter distances:



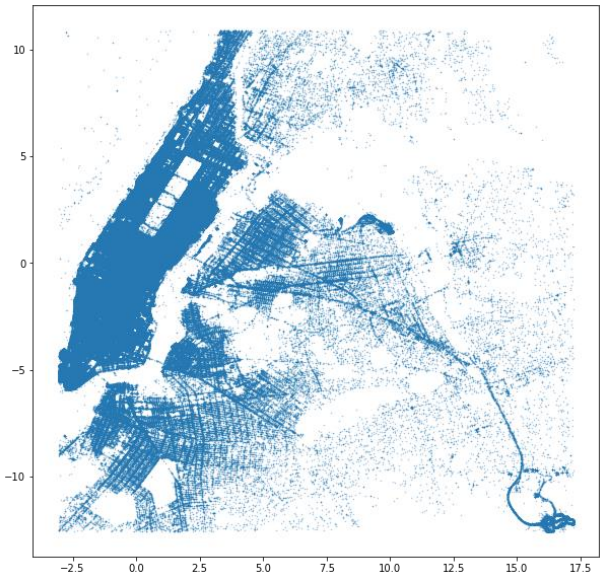
We also compared the distances travelled on different days of the week. Trips went the furthest on the weekends (days 0 and 6 on the graph):

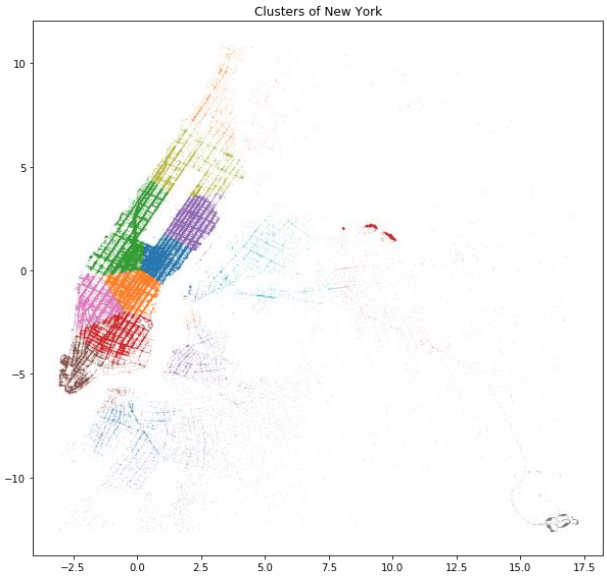


We also mapped the trips within New York City and clustered them based on the pickup and drop-off points for each ride. As we can see, the clustering results in a partition which is somewhat similar to the way NY is divided into different neighborhoods. We can see Upper East

and West side of Central park in light blue and green respectively. West midtown in pink, Chelsea and West Village in orange, downtown area in red, East Village and SoHo in purple.

The airports JFK and LaGuardia have there own cluster, and so do Queens and Harlem. Brooklyn is divided into 2 clusters, and the Bronx has too few rides to be separated from Harlem.



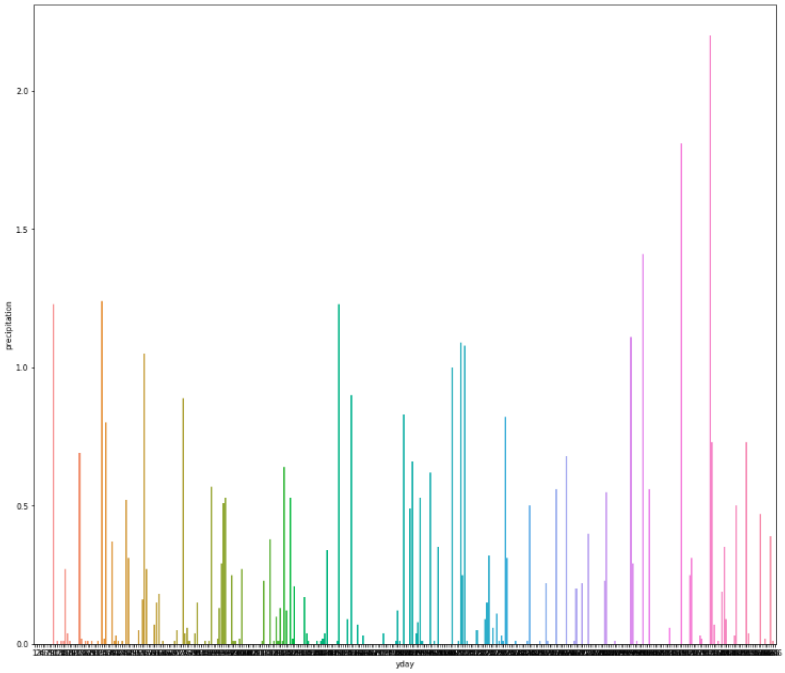


FEATURE ENGINEERING

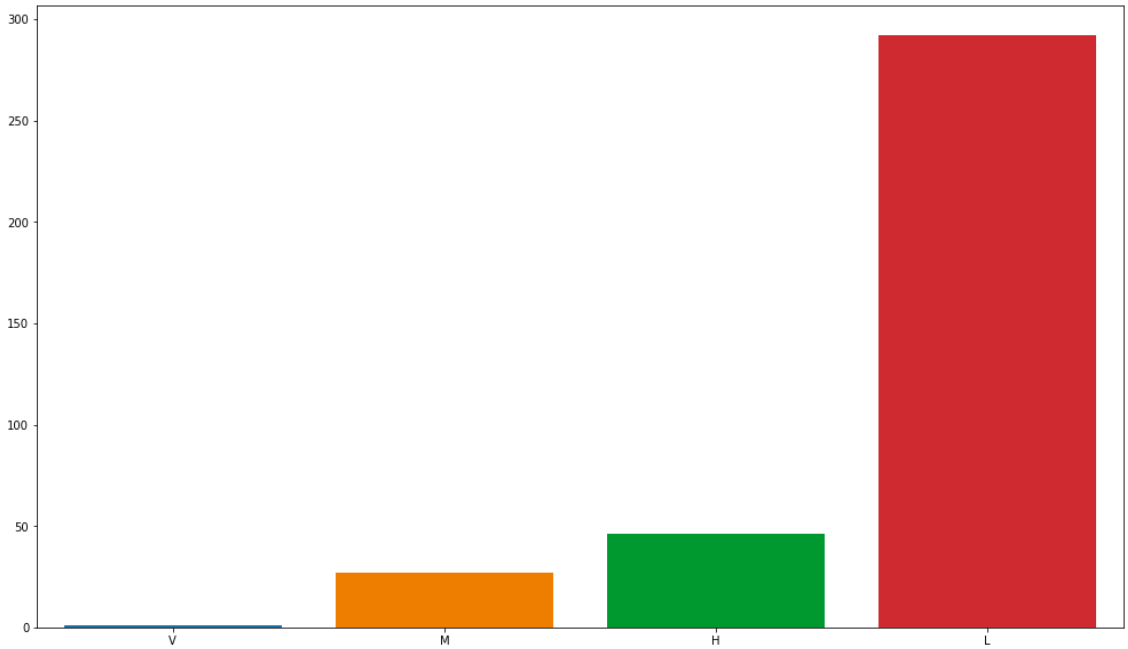
Weather can play an important role in influencing traffic. Both the increase of traffic, as well as the decrease of road conditions increases the travel time. Which in turn can have an effect on the trip duration of NYC taxi.

We added the New York Central Park 2016 weather data. It contains the first six months of 2016, for a weather station in central park. It contains for each day the minimum temperature, maximum temperature, average temperature, precipitation, new snowfall, and current snow depth. The temperature is measured in Fahrenheit and the depth is measured in inches. T means that there is a trace of precipitation.

We plotted the variation of precipitation over time and got a high level of variance in the graph.



So we extracted 4 new columns from the precipitation feature- High Precipitation, Medium Precipitation, Low Precipitation and Very High Precipitation.



Then we extracted the day of year from the date and merged the weather dataset with the NYC taxi duration dataset on the basis of this new column.

EVALUATION METRICS

To check the performance of our model on Kaggle we chose Root Mean Squared Logarithmic Error as our evaluation metric.

$$\epsilon = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_i + 1) - \log(a_i + 1))^2}$$

Where:

ϵ is the RMSLE value (score)

n is the total number of observations in the (public/private) data set,

Pi is your prediction of trip duration, and

ai is the actual trip duration for ii.

log(x) is the natural logarithm of x

BUILDING THE MODEL

Several models with different sets of features were built. The three Machine Learning models which we explored in this project were:

1.Linear Regression
2.Gradient Boosting
3.Lasso Regression

We observed that the Weather dataset didn't add any value to the model, so we built models without it.

The following features were selected to build our initial model :

- Vendor_id
- Passenger_count
- Pickup_longitude
- Pickup_wday
- Pickup_yday
- Pickup_hour

All the models were built using Cross Validation technique and RMSLE value were evaluated.

1) Linear Regression Model
   RMSLE of 0.731
2) Ridge Linear Regression Model
   RMSLE of 0.73

3) Gradient Boosting Regression Model

RMSLE of 0.732

Since, the performance didn't appear to be great, we tried to build models using the additional three features which were calculated using the latitude and longitude columns, **Haversine Distance, Manhattan Distance and Bearing Distance.** Also we split the store_and_fwd_flag into two new columns **N and Y**

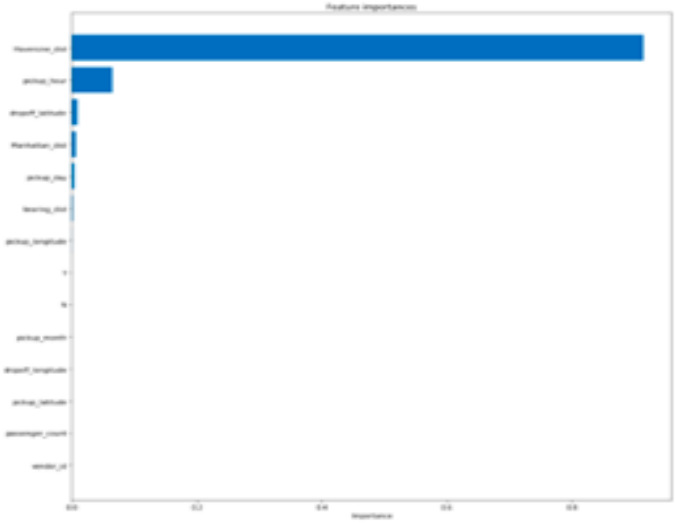
Features Selected:

- Vendor_id
- Passenger_count
- Pickup_longitude
- Pickup_latitude
- Dropoff_longitude
- Dropoff_latitude
- Trip_duration
- Haversine_dist
- Bearing_dist
- Manhattan_dist
- Pickup_month
- Pickup_day
- Pickup_hour
- N
- Y

Gradient Boosting (n_estimators=50, learning_rate=0.01, max_depth=5, random_state=0, loss='ls') - The model gave us a RMSLE of 0.68

FEATURE IMPORTANCE

We used Gradient Boosting feature importance parameters to check the important features:



According to the feature importance graph, we revised our features selection and chose below features for out third round of model building:

- Vendor_id
- Passenger_count
- Pickup_longitude
- Pickup_latitude
- Dropoff_longitude
- Dropoff_latitude
- Store_and_fwd_flag
- Month, Day
- weekday, hour, minute - extracted from pickup_datetime
- dist_long, dist_lat, dist - extracted from distance between pickup longitude
- dropoff longitude, pickup latitude and dropoff latitude
- Trip_duration

From round 2 model building, we observed that Gradient Boosting gave us the best result. Hence, in this section, we built the model only using Gradient Boosting: This gave us a RMSLE of 0.383.

COMPARISON OF MODELS

Feature Selection Stages	**Model**	**RMSLE**
Stage 1	Linear Regression	0.73
Stage 1	Ridge Linear Regression	0.73
Stage 1	Gradient Boosting	0.73
Stage 2	Gradient Boosting	0.68
Stage 3	Gradient Boosting	0.38

RESULTS

From above table, we chose the Stage 3 Gradient Boosting model as our final model. The 16 features which selected to build this model were:

- Vendor_id
- Passenger_count
- Pickup_longitude
- Pickup_latitude
- Dropoff_longitude
- Dropoff_latitude
- Store_and_fwd_flag
- Month, Day, weekday, hour, minute - extracted from pickup_datetime
- dist_long, dist_lat, dist - extracted from distance between pickup longitude, dropoff longitude, pickup latitude and dropoff latitude
- Trip_duration

CONCLUSION

Our model can:

- Help taxi dispatchers assign requested trips to available cars, making their services more efficient
- Help city planners understand duration of trips people are taking to better accommodate taxis

CONTRIBUTIONS FROM EACH TEAM MEMBER

Our team worked really well together, and everyone contributed equally towards the project.

Team Members	**Task Completed**
Aakanksha Singh	Documentation and Presentation Slides
Lindsay Huth	Documentation, Presentation Slides and Data Exploration
Mayanka Jha	Documentation, Presentation, Data Exploration, Predictive Modelling
Riya Chanduka	Documentation, Presentation, Data Exploration, Predictive Modelling