Capstone Project - Regression

NYC Taxi Trip Time Prediction

Submitted by - Deepak Solanki

Problem Statement

Objective - To build a model that predicts the total ride duration of taxi trips in New York City.

Data Description - The dataset is based on the 2016 NYC Yellow Cab trip record data made available in Big Query on Google Cloud Platform. The data was originally published by the NYC Taxi and Limousine Commission (TLC). The data was sampled and cleaned for the purposes of this project. Based on individual trip attributes, we need predict the duration of each trip in the test set.

NYC Taxi Data.csv - the dataset contains 1458644 trip records

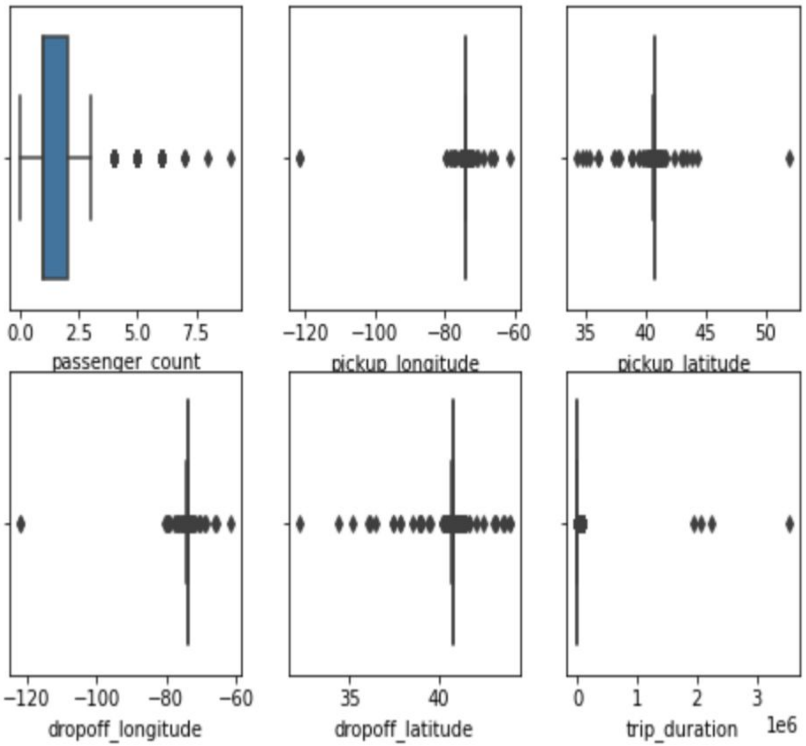
.

Data Fields

- **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
- **store_and_fwd_flag** - This flag indicates whether the trip record was held in vehicle memory before sending to the vendor because the vehicle did not have a connection to the server - Y=store and forward; N=not a store and forward trip
- **trip_duration** - duration of the trip in seconds

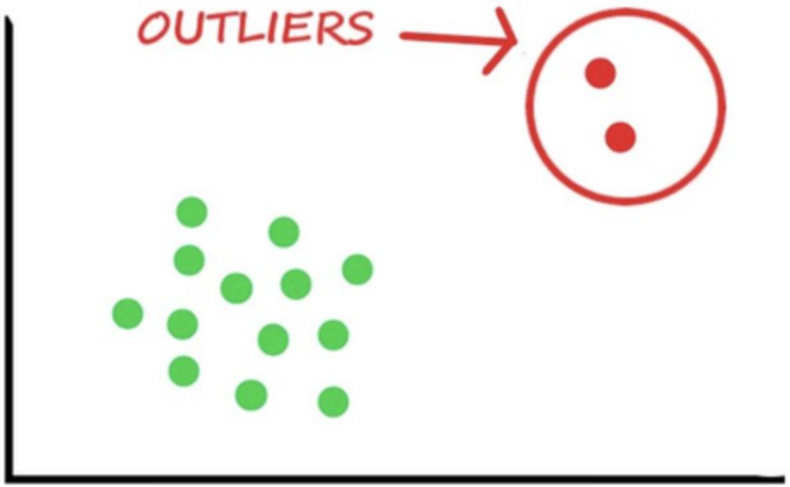
Outliers detection (Box Plots)

- Box plots shows the presence of outliers in a couple of features



Missing values checking & Outliers removal

- We checked and found that there were no missing values in the whole dataset.
- We removed outliers from following features using various methods like predefined threshold and interquartile range.
  - passenger_count
  - pickup_longitude
  - pickup_latitude
  - dropoff_longitude
  - dropoff_latitude
  - trip_duration
  - distance_km

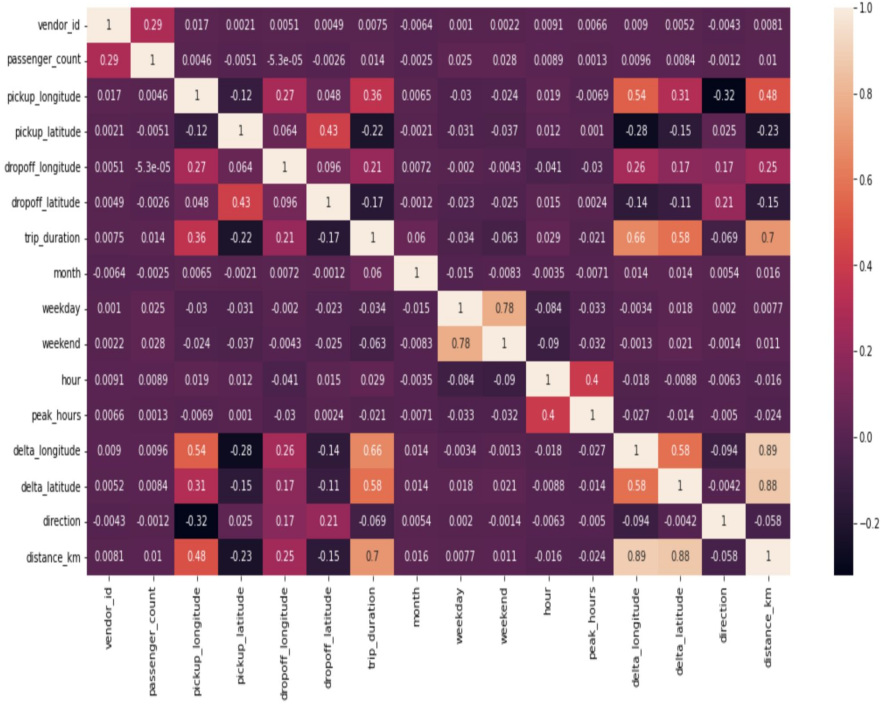


Feature Engineering

- Features extractions from “Pickup Datetime”
  - Month
  - Weekday
  - Weekend
  - Hour
  - Peak_hours
- New Derived features from Latitudes and Longitudes
  - Delta longitude / latitude
  - Direction (in degrees)
  - Distance_km (haversine distance)

EDA (Feature correlation matrix)

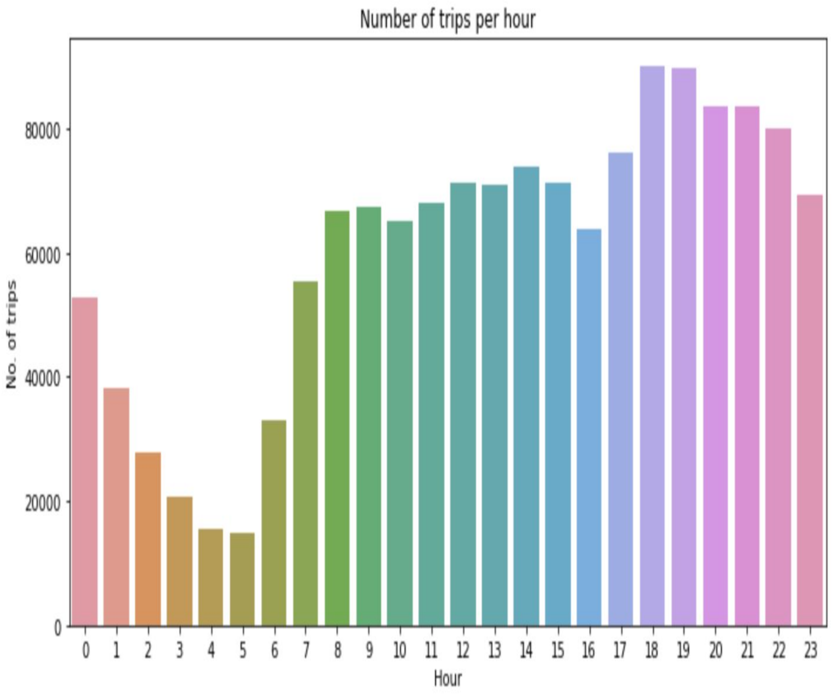
- There are few couplets of variables that are significantly correlated.
- For ex.
  - ○ Corr(distance_km, trip_duration) = 0.7
  - ○ Corr(delta_longitude, trip_duration) = 0.66
  - ○ Corr(delta_longitude, pickup_longitude) = 0.54



	vendor_id	passenger_count	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	trip_duration	month	weekday	weekend	hour	peak_hours	delta_longitude	delta_latitude	direction	distance_km
vendor_id	1	0.29	0.017	0.0021	0.0051	0.0049	0.0075	-0.0064	0.001	0.0022	0.0091	0.0066	0.009	0.0052	-0.0043	0.0081
passenger_count	0.29	1	0.0046	-0.0051	-5.3e-05	-0.0026	0.014	-0.0025	0.025	0.028	0.0089	0.0013	0.0096	0.0084	-0.0012	0.01
pickup_longitude	0.017	0.0046	1	-0.12	0.27	0.048	0.36	0.0065	-0.03	-0.024	0.019	-0.0069	0.54	0.31	-0.32	0.48
pickup_latitude	0.0021	-0.0051	-0.12	1	0.064	0.43	-0.22	-0.0021	-0.031	-0.037	0.012	0.001	-0.28	-0.15	0.025	-0.23
dropoff_longitude	0.0051	-5.3e-05	0.27	0.064	1	0.096	0.21	0.0072	-0.002	-0.0043	-0.041	-0.03	0.26	0.17	0.17	0.25
dropoff_latitude	0.0049	-0.0026	0.048	0.43	0.096	1	-0.17	-0.0012	-0.023	-0.025	0.015	0.0024	-0.14	-0.11	0.21	-0.15
trip_duration	0.0075	0.014	0.36	-0.22	0.21	-0.17	1	0.06	-0.034	-0.063	0.029	-0.021	0.66	0.58	-0.069	0.7
month	-0.0064	-0.0025	0.0065	-0.0021	0.0072	-0.0012	0.06	1	-0.015	-0.0083	-0.0035	-0.0071	0.014	0.014	0.0054	0.016
weekday	0.001	0.025	-0.03	-0.031	-0.002	-0.023	-0.034	-0.015	1	0.78	-0.084	-0.033	-0.0034	0.018	0.002	0.0077
weekend	0.0022	0.028	-0.024	-0.037	-0.0043	-0.025	-0.063	-0.0083	0.78	1	-0.09	-0.032	-0.0013	0.021	-0.0014	0.011
hour	0.0091	0.0089	0.019	0.012	-0.041	0.015	0.029	-0.0035	-0.084	-0.09	1	0.4	-0.018	-0.0088	-0.0063	-0.016
peak_hours	0.0066	0.0013	-0.0069	0.001	-0.03	0.0024	-0.021	-0.0071	-0.033	-0.032	0.4	1	-0.027	-0.014	-0.005	-0.024
delta_longitude	0.009	0.0096	0.54	-0.28	0.26	-0.14	0.66	0.014	-0.0034	-0.0013	-0.018	-0.027	1	0.58	-0.094	0.89
delta_latitude	0.0052	0.0084	0.31	-0.15	0.17	-0.11	0.58	0.014	0.018	0.021	-0.0088	-0.014	0.58	1	-0.0042	0.88
direction	-0.0043	-0.0012	-0.32	0.025	0.17	0.21	-0.069	0.0054	0.002	-0.0014	-0.0063	-0.005	-0.094	-0.0042	1	-0.058
distance_km	0.0081	0.01	0.48	-0.23	0.25	-0.15	0.7	0.016	0.0077	0.011	-0.016	-0.024	0.89	0.88	-0.058	1

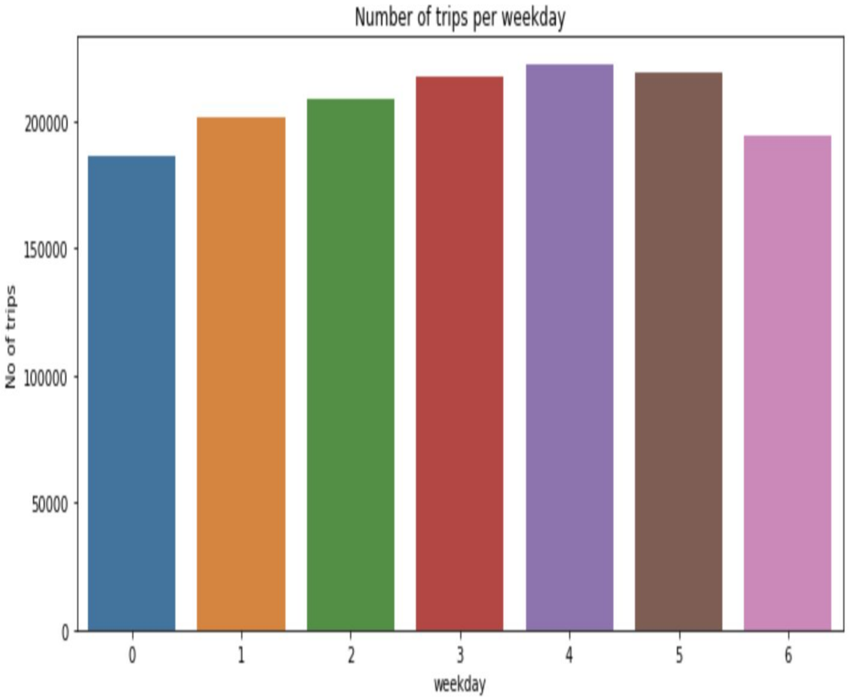
EDA (Continue)

- Maximum no. of trips occurs in the evening time slot **(18:00-20:00 hours)**. Then it has an decreasing trend till 5:00. After that, It start increasing till evening.



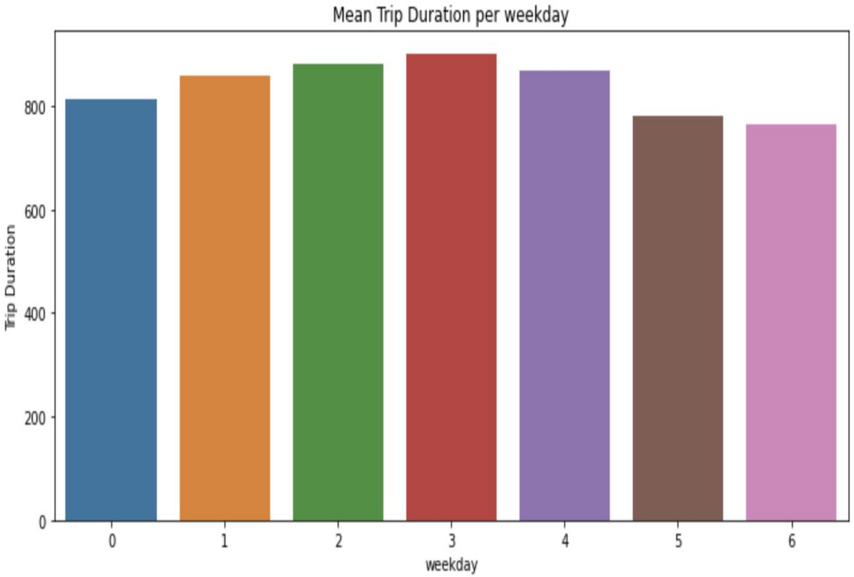
EDA (Continue)

- Number of trips has an increasing trend from Monday to Friday. And it’s relatively low on Sunday as compared to other days.



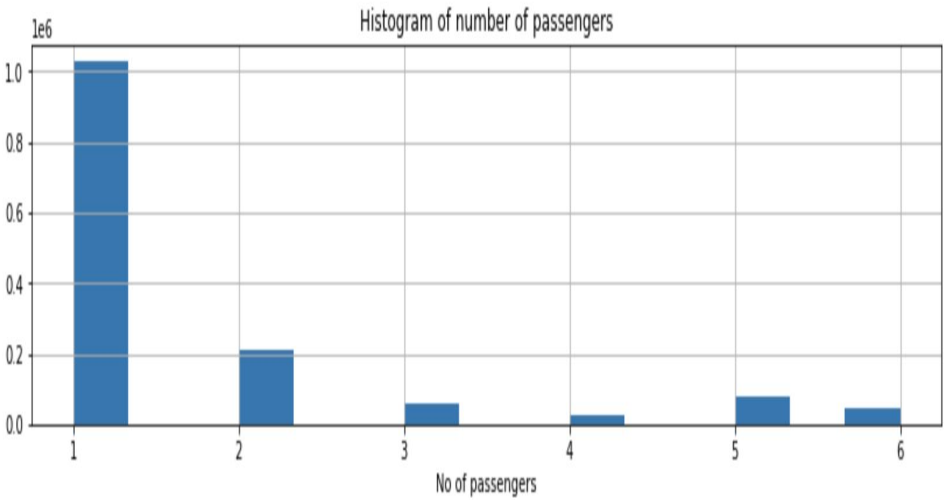
EDA (Continue)

- Mean trip duration on weekdays is higher than that of on weekends. It is highest on Thursday (weekday = 3).



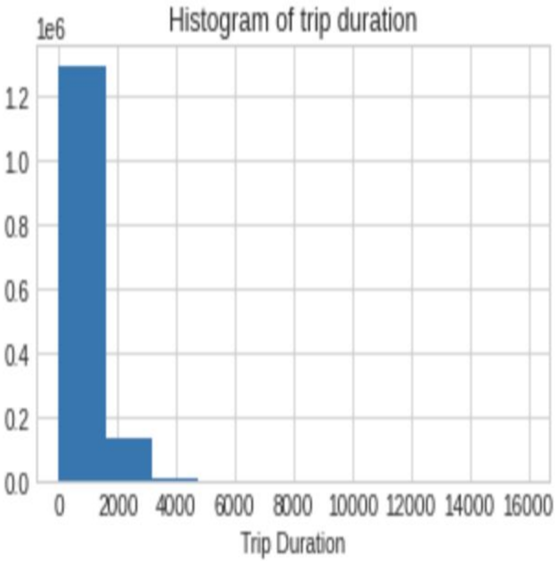
EDA (Continue)

- No. of passengers ranges between 1 and 6.
- **71%** of the trips are having only 1 passenger.



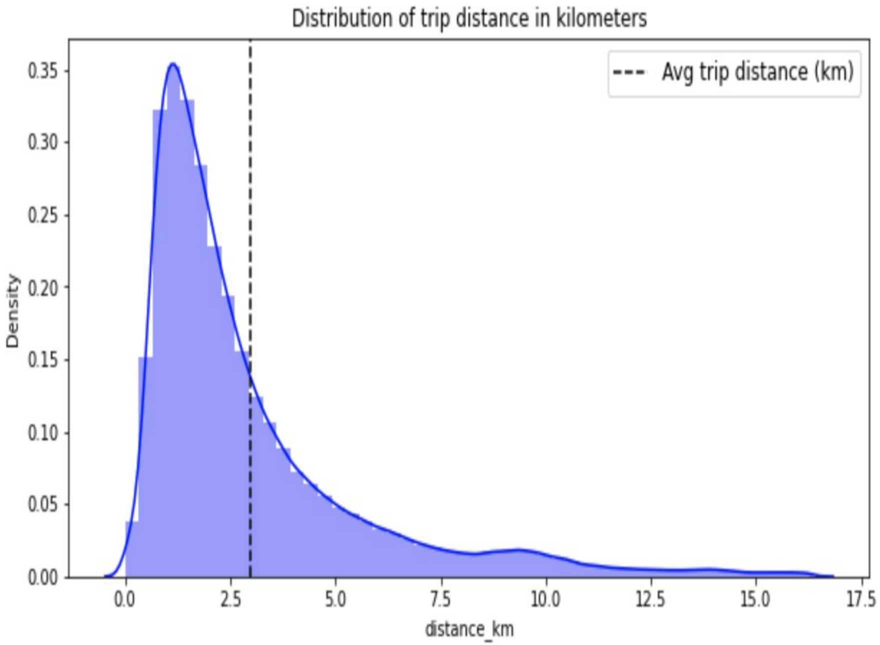
EDA (Continue)

- Majority of the trip durations are below 1600 seconds. Also, a few of trips are above 3000 seconds.

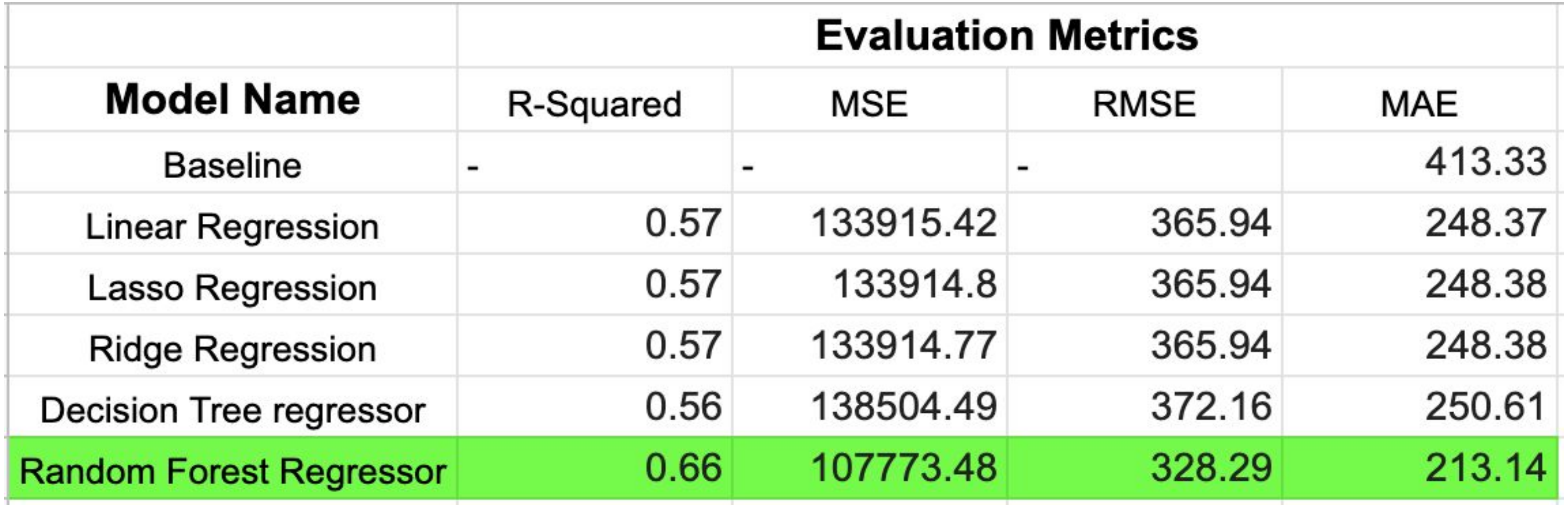


EDA (Continue)

- The distribution of **distance_km** is right skewed. And avg distance is ~3 km.



Modeling



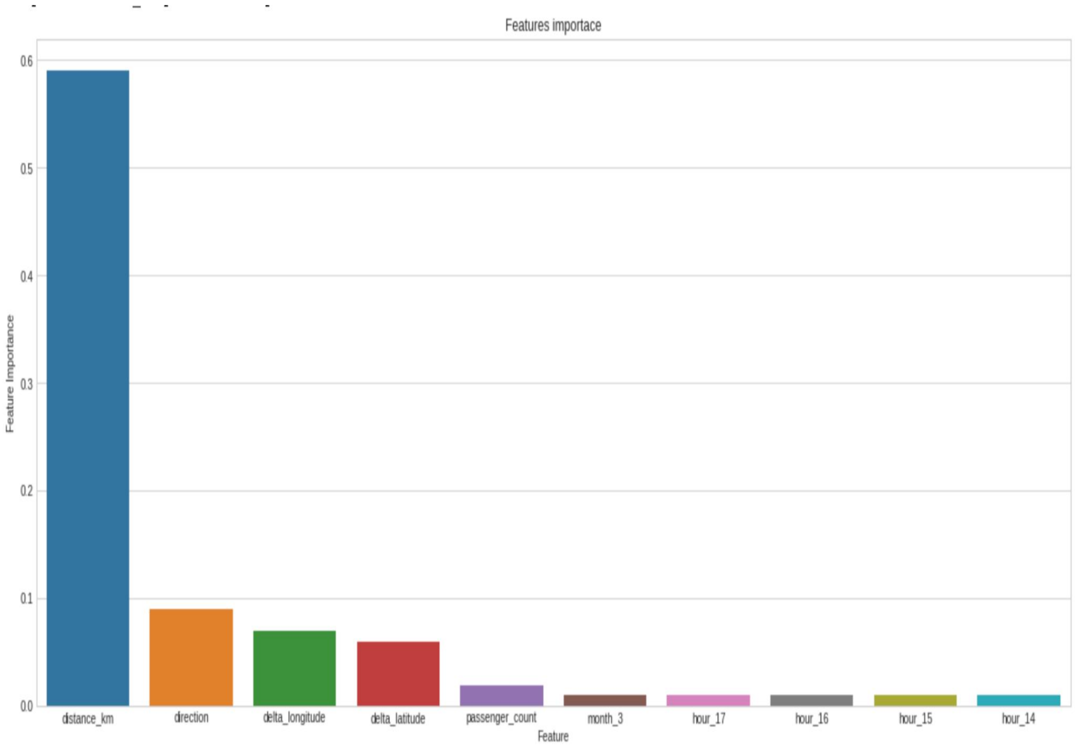
	Evaluation Metrics			
Model Name	R-Squared	MSE	RMSE	MAE
Baseline	-	-	-	413.33
Linear Regression	0.57	133915.42	365.94	248.37
Lasso Regression	0.57	133914.8	365.94	248.38
Ridge Regression	0.57	133914.77	365.94	248.38
Decision Tree regressor	0.56	138504.49	372.16	250.61
Random Forest Regressor	0.66	107773.48	328.29	213.14

Random Forest Regressor (Best Model)

- We used **GridSearchCV** cross validation technique for Hyperparameter tuning.
- param_grid
  {**'n_estimators':** [10, 25],
  'max_features': [5, 10],
  'max_depth': [10, 50, None],
  'bootstrap': [True, False]}
- Random Forest **best parameters** are
  {**'bootstrap'**: True,
  'max_depth': 50,
  'max_features': 10,
  'n_estimators': 25}

Random Forest Top 10 important features

- **distance_km** is the most important feature contributing around 58%.
- Other important features are direction (9%), delta_longitude (7%) and delta_latitude (6%).



Conclusions

1. Linear, Lasso and Ridge regressions are giving the similar results. And They are performing better than the baseline model.
2. In Decision Tree regressor, the results are slightly poorer than that of linear models.
3. Out of all tried models, Random Forest is giving the best result.

Challenges Faced

1. Deriving new features from geospatial data (using given Latitude and Longitude coordinates).
2. Extracting useful features from pickup datetime column.
3. High computation time in hyperparameter tuning.

Further scope for improvement

- **Feature Engineering** - More features can be derived from the Geospatial data which can improve the models performance.
- Neural Network based regression model can be built to improve the results further.

Thank You