Report

- Dropped the rows that have null values
- Removed outliers in fare amount. Fare amount < 0 & > 500 are removed.
- Removed outliers in passenger count. A taxi at maximum can take 6 people at a time.
- Pick up and drop off latitudes which are not in range (-90,90) are removed.
- Pick up and drop off longitudes which are not in range (-180,180) are removed.
- Found Distance using pickup and drop off latitudes and longitudes (Haversine Distance)

Now, for **EDA**. The following are my considerations -

- Does the number of passengers affect the fare?
- Does the date and time of pickup affect the fare?
- Does the day of the week affect the fare?
- Does the distance travelled affect the fare?

Single passengers are the most frequent travelers, and the highest fare also seems to come from cabs which carry just 1 passenger.

Split the datetime field 'pickup_datetime' to the following -

- year
- month
- date
- hour
- day of week

There are values which are greater than 100 kms! In NYC I am not sure why people would take cabs to travel more than a 100 kms. Since the number of bins for 100-200 kms is quite high, I will keep these. These outliers could be because of typos or missing values in the latitude or longitude. Remove fields of the following -

1. Pickup latitude and pickup longitude are 0 but dropoff latitude and longitude are not 0, but the fare is 0
2. vice versa of point 1.
3. Pickup latitude and pickup longitude are 0 but dropoff latitude and longitude are not 0, but the fare is NOT 0. Here I will have to impute the distance values in both the train and test data.

Checked the H_Distance fields which are greater than 200 kms cause there is no way that people would travel more than 200 kms at the most in NYC in a CAB.

A quick Google search gave me the following prices -

- $$2.5 base-price + $1.56/km --> 6AM to 8PM Mon-Fri
- $$3.0 base-price + $1.56/km --> 8PM to 6AM Mon-Fri and Sat&Sun

Replaced them with distance = (fare_amount - 2.5)/1.56

When distance is 0

- Fare and Distance are both 0. According to the table above, we shall delete them as they do not provide us with any info with regards to the data. (22 rows)
- Fare is not 0 and is less than the base amount, but Distance is 0. Deleted these rows as the minimum is $2.50, and these fares are incorrect values. (10 rows)
- I understood that the distance is 0, but the fare is all the minimum fare of $2.5. This could be because the passenger booked the cab but ended up cancelling to pay the base fare (not sure how this works in NYC, but I'm assuming that's how it is)
- Fare is 0, but Distance is not 0. These values need to be imputed.I can calculate the fare as I have the distance. I shall use the following formula *fare = 2.5 + 1.56(H_Distance).*
- Fare is not 0, but Distance is 0. These values need to be imputed. *distance = (fare_amount - 2.5)/1.56.*

Modelling:

Boosting with lgbm gave me 3.50603

Random forest regressor gave me 3.53203