Final Report on Machine Learning Capstone Project

NYC Taxi Fare Prediction with Deep Learning

Definition

Project Overview:

Majority of residents of New York city rely on public transportation for commute. Only 22% of residents own vehicle as compared to national average of 91% Americans who at least own one vehicle. Taxi companies in New York city are fourth largest provider of public transportation providing over 1.2 million rides to city dwellers and visitors. Therefore, taxis are lifeline of transportation in New York city area than the rest of America.

Recently, it has been noticed that traditional taxi companies are struggling to stay in the business due to stiff competition from ride hailing apps. Accurate fare pricing could benefit the profitability of the taxi owners. Moreover, new ride hailing apps also need to calculate the price of the ride accurately in order to maintain the profitability for struggling taxi business in New York city. Ride hailing apps have 65% more trips than taxi trips. Ride hailing through apps (Lyft, Uber etc.) has soared to 15 million rides and taxi rides are fewer than 10 million. Ride hailing app popularity increased taxi usage mainly due to cheaper pricing and easy availability.

Motivation of this project is to familiarize with current mapping and pricing techniques such as openstreetmap for plotting data. Another part of the motivation includes use the machine learning techniques such as deep learning, for drawing meaningful results and make informed decisions from city data. This has a huge implications in near future as more and more cities are opening their datasets.

This is a very recent competition on Kaggle ended just 3-4 weeks ago. It offers very interesting information in terms geospatial information and price prediction challenges.

Problem Statement

Objective of the project is to predict the fare accurately, given taxi pick up and drop off locations, on a given date with given number of passengers. Root mean square error is calculated and serve as a metric to check the accuracy of the model.

Evaluation Metric:

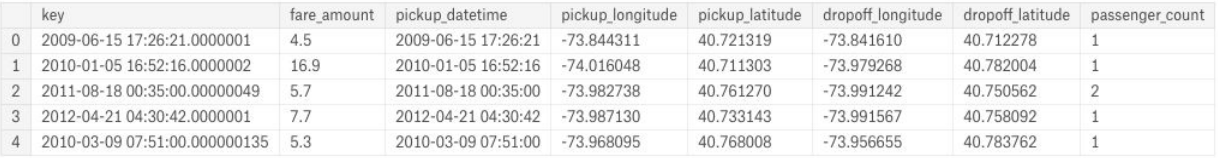
Root mean squared (RMSE) error is used as the evaluation of the model. This is typically done for regression model. RMSE measures difference between prediction by the machine learning model and actual value.

Analysis

In this section, data is explored through various ways and plotted through the graphical as well visual maps. Dataset comes from the Kaggle competition, New York City Taxi Fare Prediction, used to predict the fare amount (including tolls) for the ride, given the pickup and drop off location. This dataset most likely originated from the NYC Taxi & Limousine Commision (TLC) dataset for yellow and green taxis.

Data Exploration

To understand the data better first we will simply load the data to check the columns and what it means.



	key	fare_amount	pickup_datetime	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
0	2009-06-15 17:26:21.0000001	4.5	2009-06-15 17:26:21	-73.844311	40.721319	-73.841610	40.712278	1
1	2010-01-05 16:52:16.0000002	16.9	2010-01-05 16:52:16	-74.016048	40.711303	-73.979268	40.782004	1
2	2011-08-18 00:35:00.00000049	5.7	2011-08-18 00:35:00	-73.982738	40.761270	-73.991242	40.750562	2
3	2012-04-21 04:30:42.0000001	7.7	2012-04-21 04:30:42	-73.987130	40.733143	-73.991567	40.758092	1
4	2010-03-09 07:51:00.000000135	5.3	2010-03-09 07:51:00	-73.968095	40.768008	-73.956655	40.783762	1

ID in dataset

- key - Unique string identifying each row in both the training and test sets. Comprised of pickup_datetime plus a unique integer.

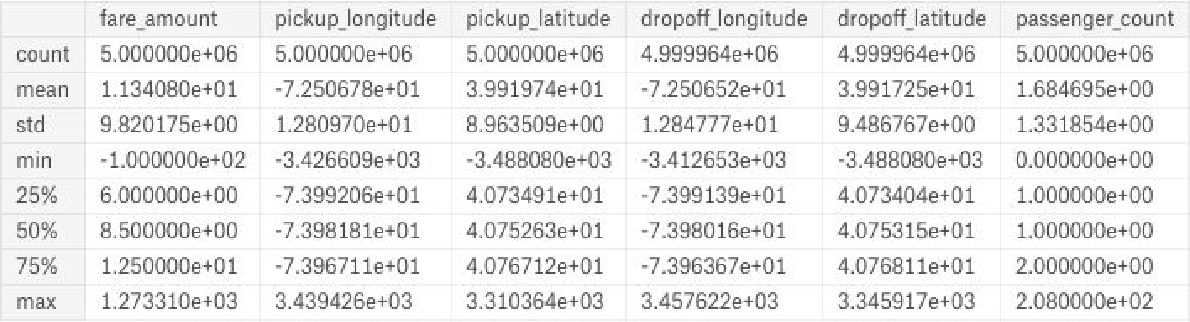
Features in dataset

- pickup_datetime - timestamp value indicating when the taxi ride started.
- pickup_longitude - float for longitude coordinate of where the taxi ride started.
- pickup_latitude - float for latitude coordinate of where the taxi ride started.
- dropoff_longitude - float for longitude coordinate of where the taxi ride ended.
- dropoff_latitude - float for latitude coordinate of where the taxi ride ended.
- passenger_count - integer indicating the number of passengers in the taxi ride.

Target feature

- fare_amount - `float` dollar amount of the cost of the taxi ride. This value is present for the training dataset and predicted by the machine learning model using test dataset.

We now describe the data statistically to note the mean, standard deviation and minimum values.



	fare_amount	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
count	5.000000e+06	5.000000e+06	5.000000e+06	4.999964e+06	4.999964e+06	5.000000e+06
mean	1.134080e+01	-7.250678e+01	3.991974e+01	-7.250652e+01	3.991725e+01	1.684695e+00
std	9.820175e+00	1.280970e+01	8.963509e+00	1.284777e+01	9.486767e+00	1.331854e+00
min	-1.000000e+02	-3.426609e+03	-3.488080e+03	-3.412653e+03	-3.488080e+03	0.000000e+00
25%	6.000000e+00	-7.399206e+01	4.073491e+01	-7.399139e+01	4.073404e+01	1.000000e+00
50%	8.500000e+00	-7.398181e+01	4.075263e+01	-7.398016e+01	4.075315e+01	1.000000e+00
75%	1.250000e+01	-7.396711e+01	4.076712e+01	-7.396367e+01	4.076811e+01	2.000000e+00
max	1.273310e+03	3.439426e+03	3.310364e+03	3.457622e+03	3.345917e+03	2.080000e+02

Fig: Statistical description of the data provide after loading first 5 million lines

We can notice that mean fare is $11.30. However, there are some outliers such as negative fares and maximum fares of $1273. Data with number of fares larger than $100 is really small (0.034%) when we load 5 million lines of data. Hence, we cannot see data with large fares in frequency chart.

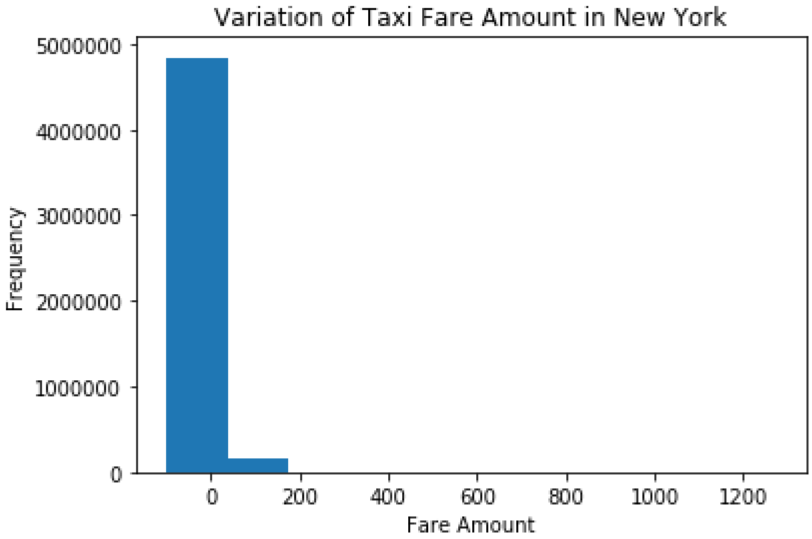
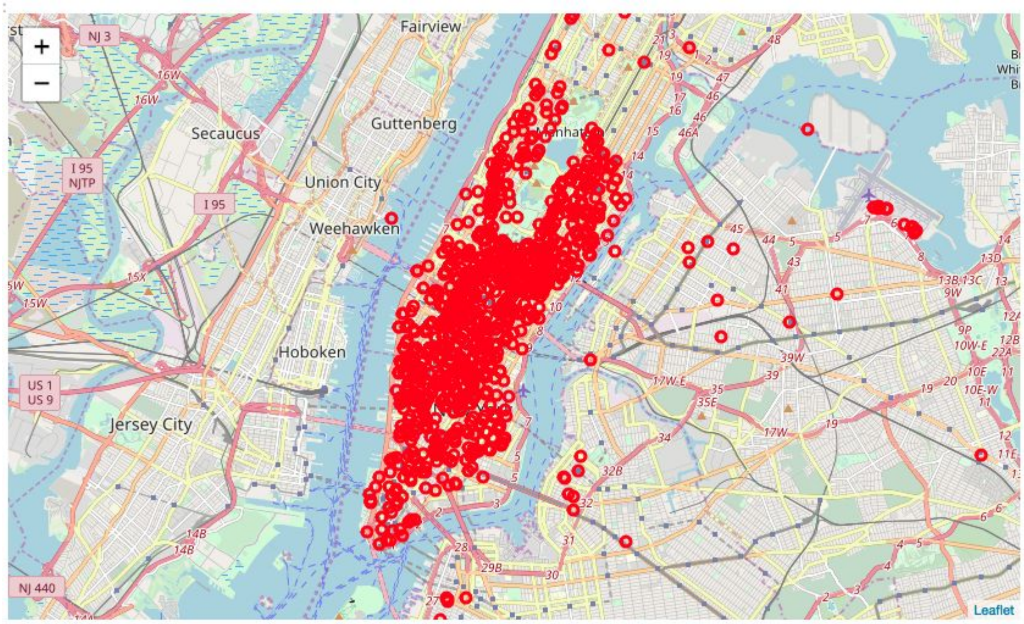


Fig.: Variation Taxi Fare Amount in New York

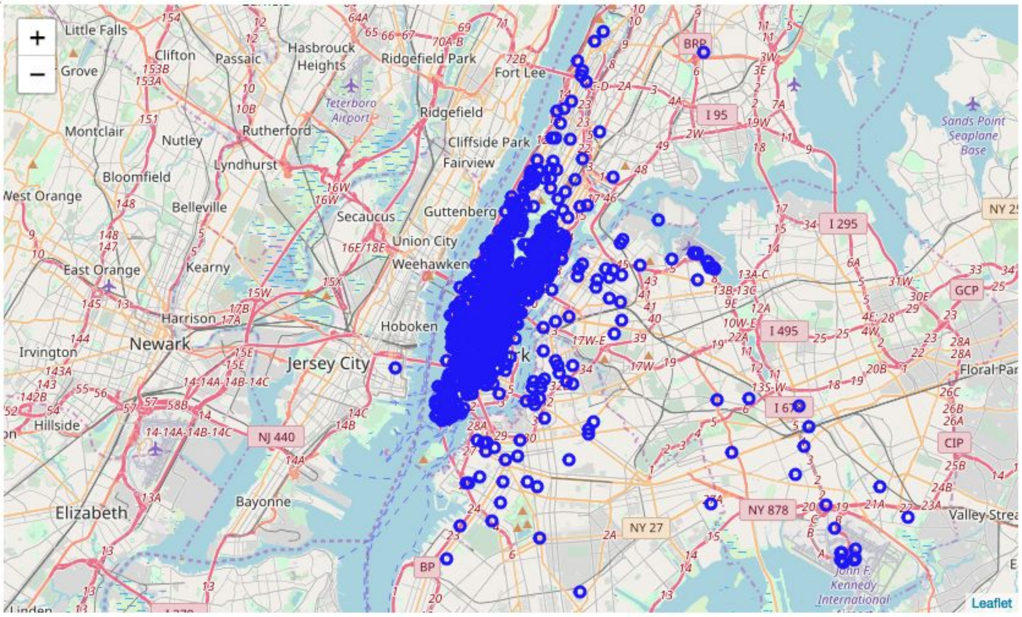
Also note that there are incorrect (large negative and positive) values for Latitude and Longitude. New York city coordinates are 40.7128° N, 74 ° W, so we expect pickup and dropoff data around the same values. Any data beyond these will be of incorrect values and will be discarded. Similarly, passenger count for some cases is zero which is puzzling and fare could be for hauling the goods.

In the two image below, pickup and dropoff locations of first 1000 taxi rides are plotted using Folium library. They look highly concentrated in Manhattan area and sprodiac outside.

Pickup location of first 1000 taxi rides in New York city



Drop Off location first 1000 of taxi rides in New York city



Time history contains important information of pickup date, time, day, year, but it is in a string format and cannot be plotted to understand the variation of fares with time of the year, day of the week, time of the day.

New features of distance and bearing will also be added later during the data preprocessing stage. Time history, distance, bearing features will be then explored further after we preprocess the data and create new features.

Algorithms and Techniques

Solution will include the building deep neural network model using Keras and Tensorflow for this multidimensional data. Model takes cleaned and scaled data free from any outliers and new engineered features. Model then predicts the fare for these inputs with highest possible accuracy. Details of the model features such as number of layers, activation function, regularization, dropout layers and optimization setup are discussed in the Implementation section.

Benchmark Model

Multivariable linear regression model will serve as a benchmark model. It will provide a solution with simple approach towards fare prediction problem. Coefficients of the model can provide the indication of feature importances.

Evaluation Metric

Root mean squared (RMSE) error is used as the evaluation of the model. This is typically done for regression model. RMSE measures difference between prediction by the machine learning model and actual value.

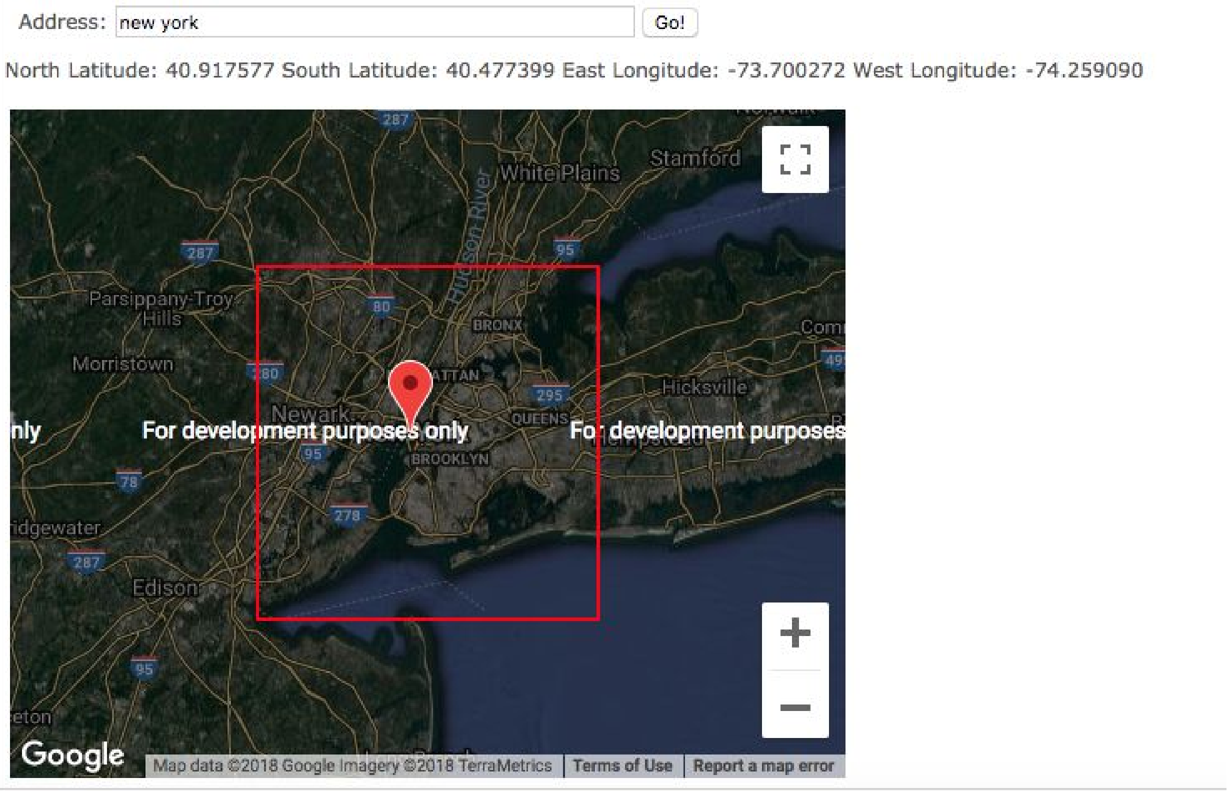
Methodology

Data Preprocessing

During this step, the data is cleaned by removing the outliers such as,

- Data with unusually large fares and negative fares:
    - Number of rides with fares more than $100 are negligible, 0.04%, compared to the rest of the data
    - Minimum taxi fare is $2.50 for the taxi operating in New York city as per NY TLC commission fare information page.
    - Hence all the fares above $100 and below $2.5 are ignored for the modeling
- Data with same starting to ending locations
    - These are rides with incorrect information or cancelled rides so they are removed from dataset
- Any null/na fields - incomplete information data is removed from the training dataset.
- Data with pickup/drop off beyond NY city limits are taken from this reference. Any ride with pickup and drop off is dropped if it is beyond New York city limits.

- Limits of bounding box of New York city are,
    - North Latitude: 40.917577
    - South Latitude: 40.477399
    - East Longitude: -73.700272
    - West Longitude: -74.259090

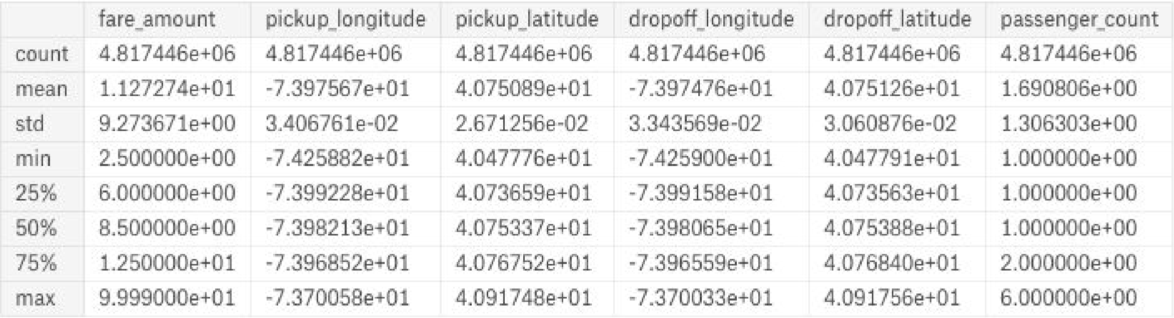


Address: new york Go!

North Latitude: 40.917577 South Latitude: 40.477399 East Longitude: -73.700272 West Longitude: -74.259090

- Data with negative person count:
    - Any rides with passenger count less than zero and more than 6 are considered as outliers and data is ignored.

After these steps, 3.65% of the data is removed and, summary of the data looks as below. You can note that outliers of negative fares, large LAT/LONG are dropped out of the data.



	fare_amount	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
count	4.817446e+06	4.817446e+06	4.817446e+06	4.817446e+06	4.817446e+06	4.817446e+06
mean	1.127274e+01	-7.397567e+01	4.075089e+01	-7.397476e+01	4.075126e+01	1.690806e+00
std	9.273671e+00	3.406761e-02	2.671256e-02	3.343569e-02	3.060876e-02	1.306303e+00
min	2.500000e+00	-7.425882e+01	4.047776e+01	-7.425900e+01	4.047791e+01	1.000000e+00
25%	6.000000e+00	-7.399228e+01	4.073659e+01	-7.399158e+01	4.073563e+01	1.000000e+00
50%	8.500000e+00	-7.398213e+01	4.075337e+01	-7.398065e+01	4.075388e+01	1.000000e+00
75%	1.250000e+01	-7.396852e+01	4.076752e+01	-7.396559e+01	4.076840e+01	2.000000e+00
max	9.999000e+01	-7.370058e+01	4.091748e+01	-7.370033e+01	4.091756e+01	6.000000e+00

Feature Engineering:

Taxi fare of course is determined by the distance, peak hour and special airport rides. In this section, new features are created from locations of pickup/dropoff and time data. They allow us to look more closely into effect of these new variables on fare.

It is also necessary to consider the fare methods used by New York Taxi and Limousine Commission (NY TLC).

- As per which, there is a surcharge of \$1 for taxi pickup between peak hours, Monday - Friday after 4:00 PM & before 8:00 PM.
- In addition, there is a night surcharge of \$0.50 after 8:00 PM & before 6:00 AM. This means we need to create a feature for this.
- There is also a flat fare of $52 plus toll charges plus surcharges for pickups to and from between JFK and Manhattan.
- There is a also a fixed surcharge of $17.5 for drop off to Newark airport.

New Features:

Distance:

Ideally we would like calculate the driving distance between two locations. However this requires the support of paid services such as Google or Bing maps.

Instead using the pickup and dropoff locations, distance can be Haversine distance which is calculated as straight line distance between two locations on earth.

Bearing:

Similarly we will calculate the bearing which is indicator of heading. Formula comes from another reference.

Time history features:

Similarly date-time time is split into the various parts for time history analysis.

- Year - to capture year over year increase in the taxi fare,
- Day of the year - to capture the effect of seasonal weather changes on taxi fares,
- Day of the week - to capture effect of weekly patterns on taxi fare, such as work week Vs weekend
- Time of the day - to capture effect of daily work schedules on taxi fares

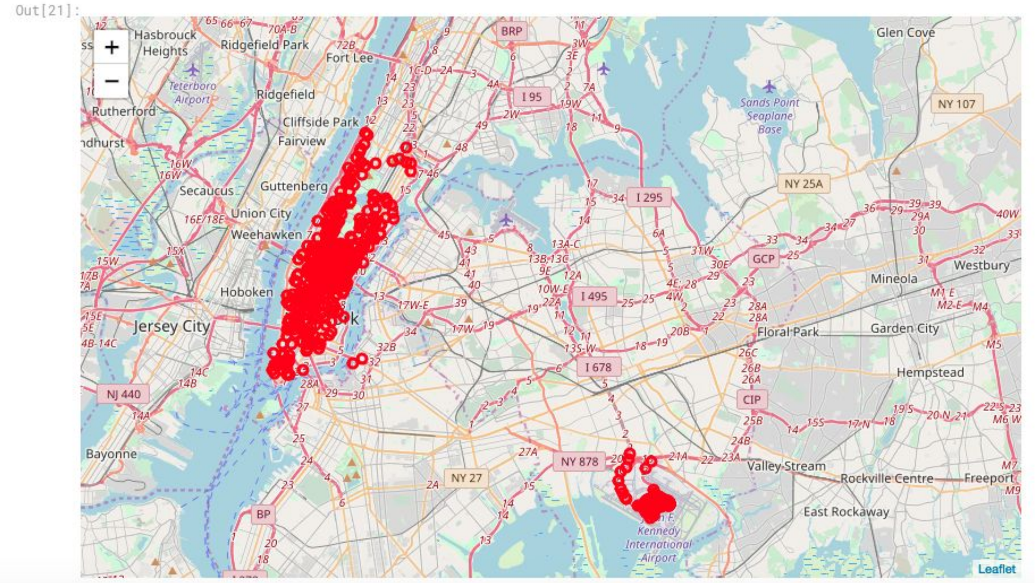
Peak and Night hour:

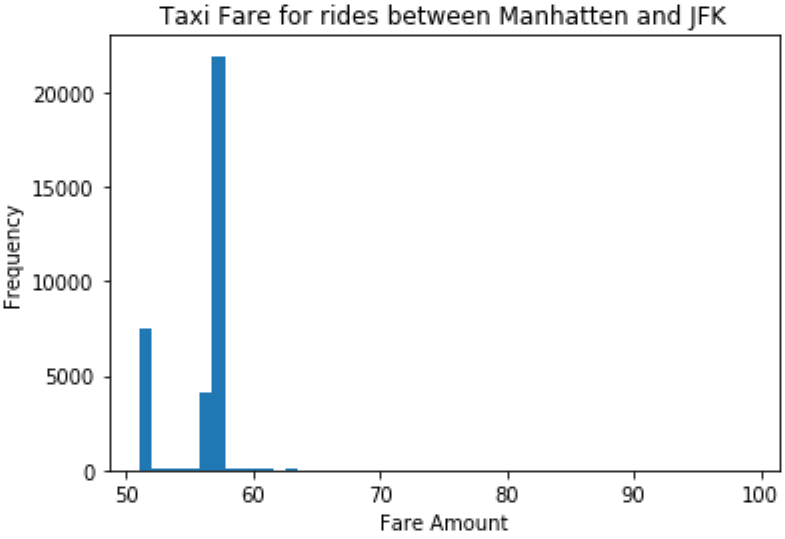
From the time of the day information, two new feature is developed for *peak* if time of the ride is between 4:00 pm and 8:00 pm. Similarly a new feature called *night_hour* is developed if ride time is between 8:00 pm and 6:00 am. Time of the day feature is then dropped afterwards before modeling along with the pickup-datetime column.

Airport JFK-Manhattan rides:

All rides between JFK airport and Manhattan have fixed charge of 52.0, so a new feature, JFK_airport, is added if pickup or dropoff location is within the boundary of JFK or Manhattan and minimum set fare.

Rides with pickup/dropoff between Manhattan and JFK airport

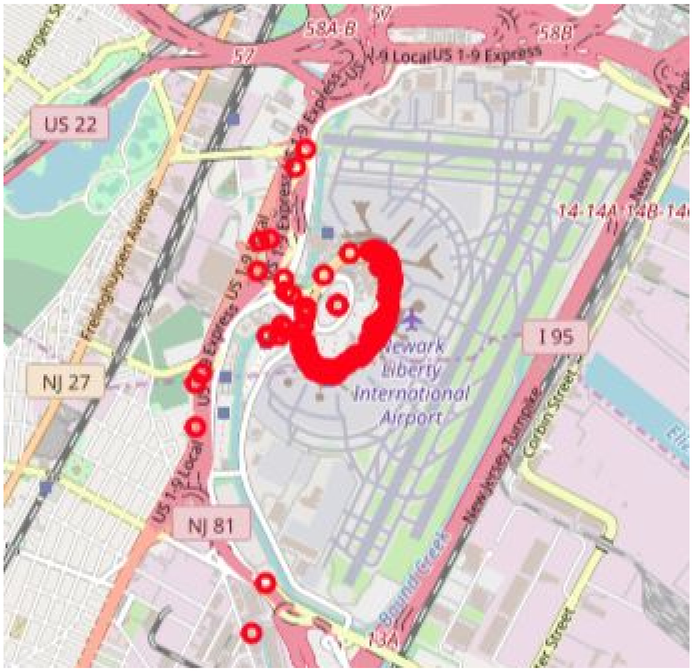




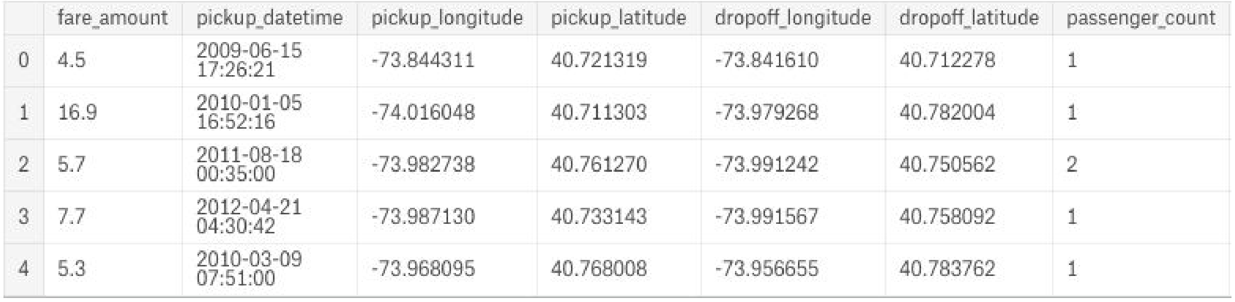
Newark Airport Drop off rides:

If the ride has drop off location as Newark airport then NEW_airport feature column is set to 1 otherwise set as 0. In the figure below, the number of such rides are plotted.

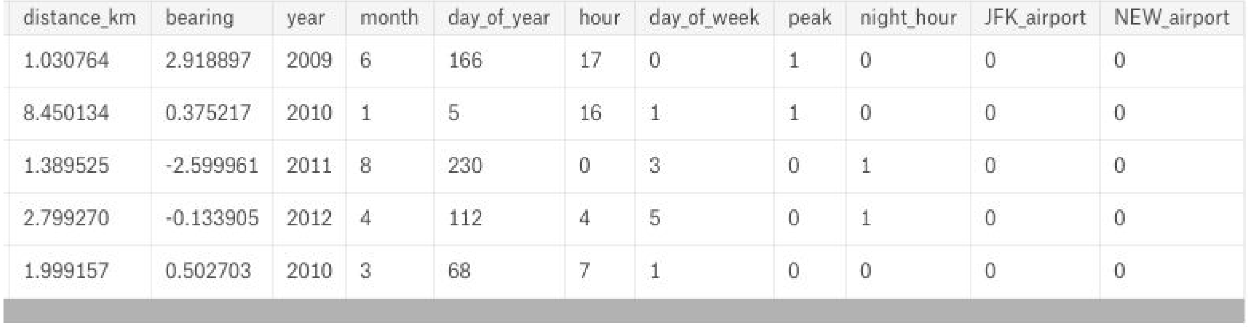
Rides with drop off location as Newark Airport



First few lines of data after the cleanup and new features columns added looks like this.

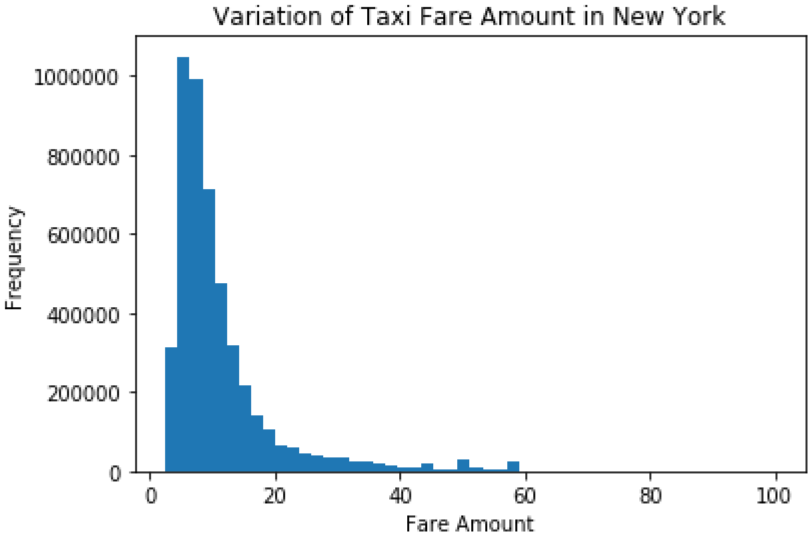


	fare_amount	pickup_datetime	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
0	4.5	2009-06-15 17:26:21	-73.844311	40.721319	-73.841610	40.712278	1
1	16.9	2010-01-05 16:52:16	-74.016048	40.711303	-73.979268	40.782004	1
2	5.7	2011-08-18 00:35:00	-73.982738	40.761270	-73.991242	40.750562	2
3	7.7	2012-04-21 04:30:42	-73.987130	40.733143	-73.991567	40.758092	1
4	5.3	2010-03-09 07:51:00	-73.968095	40.768008	-73.956655	40.783762	1

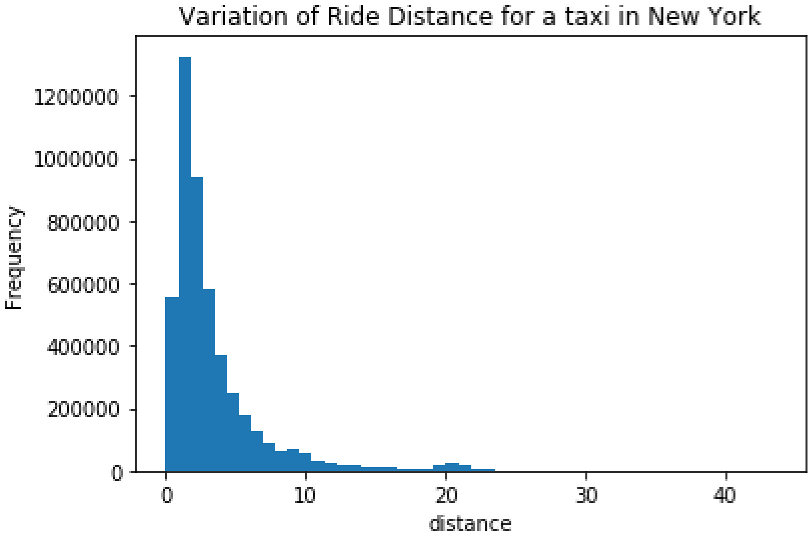


distance_km	bearing	year	month	day_of_year	hour	day_of_week	peak	night_hour	JFK_airport	NEW_airport
1.030764	2.918897	2009	6	166	17	0	1	0	0	0
8.450134	0.375217	2010	1	5	16	1	1	0	0	0
1.389525	-2.599961	2011	8	230	0	3	0	1	0	0
2.799270	-0.133905	2012	4	112	4	5	0	1	0	0
1.999157	0.502703	2010	3	68	7	1	0	0	0	0

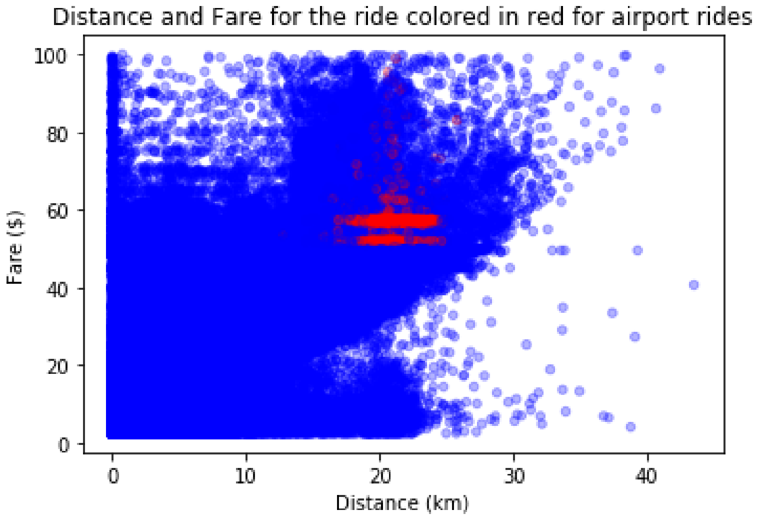
Now let us plot the new features of the data. From the plot of variation of taxi fare, it can be noticed that large number of features are around the mean value of $11. There are also large number of rides which are short distance with minimum fare of $2.5.



Frequency of the taxi ride distance is plotted in the plot below, which indicates that most number of rides have distance between 0-15 km.



In the plot below, variation of fare with ride distance is plotted and rides to-from JFK airport-Manhattan airport are colored red. Except for the rides with near zero distance and high fares, most of the data shows linear relationship between fare and distance. Rides with short distance and high fares, could be due to tolls paid, traffic jams and waiting period.

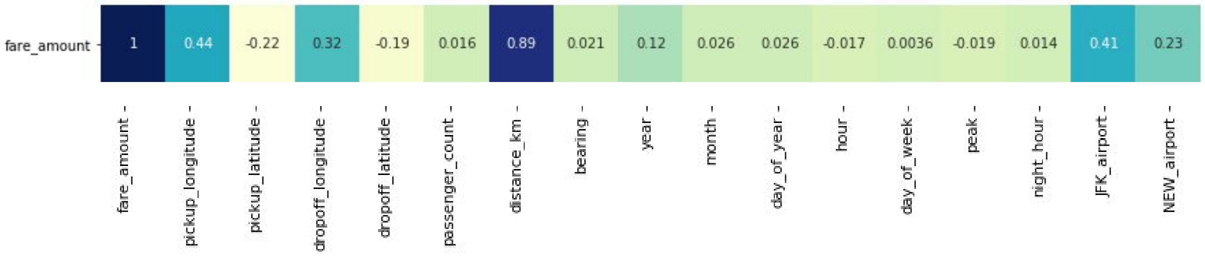


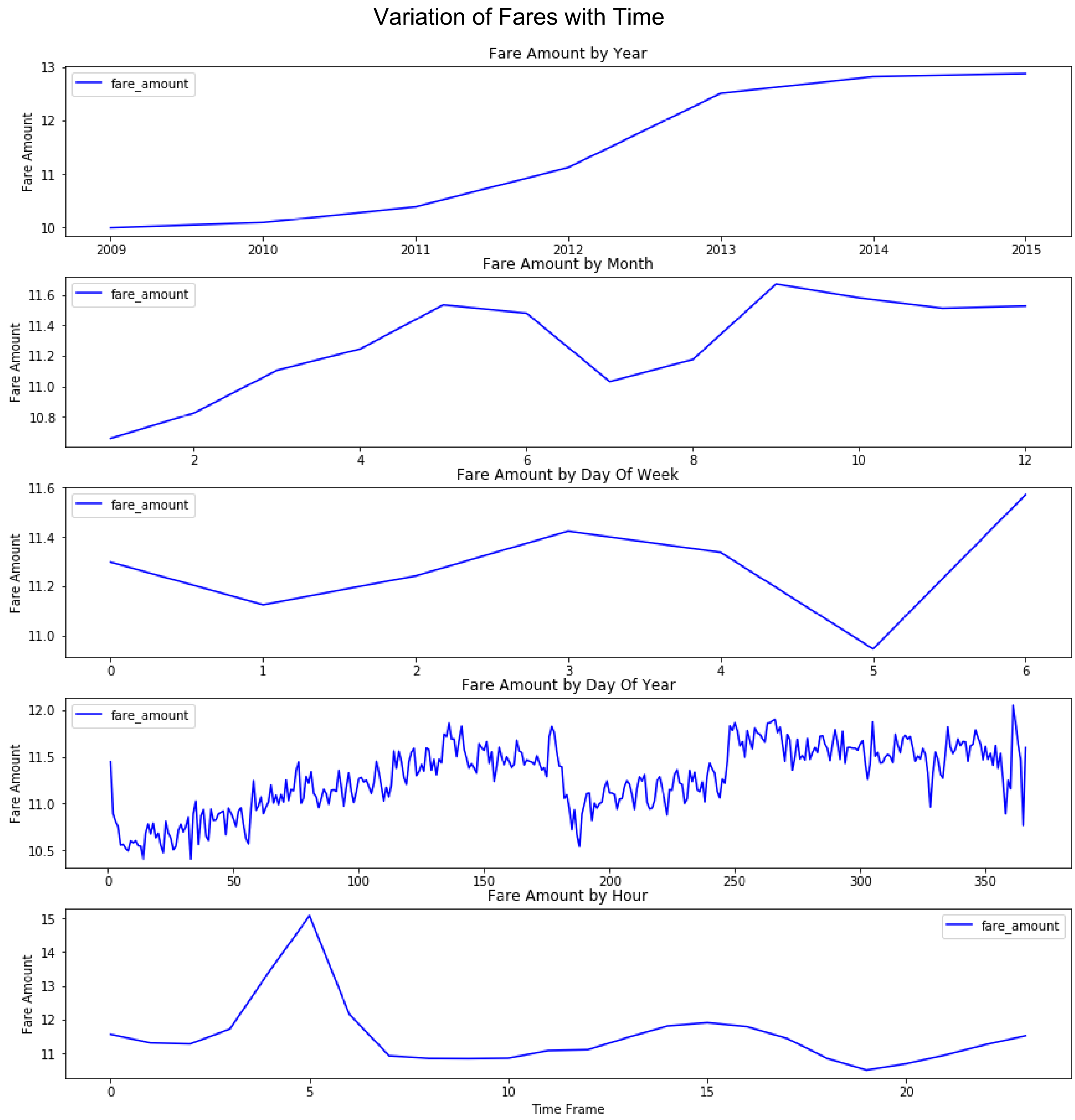
Looking at the graphs showing variation of fares with time, it can be concluded that taxi fare increases by $3 over the period of six years. So it is one of the variable which influences the taxi fare amount.

From the variation of fare with month, day of the year and day of the week, it can be concluded that changes in are are negligible hence these variables can be neglected.

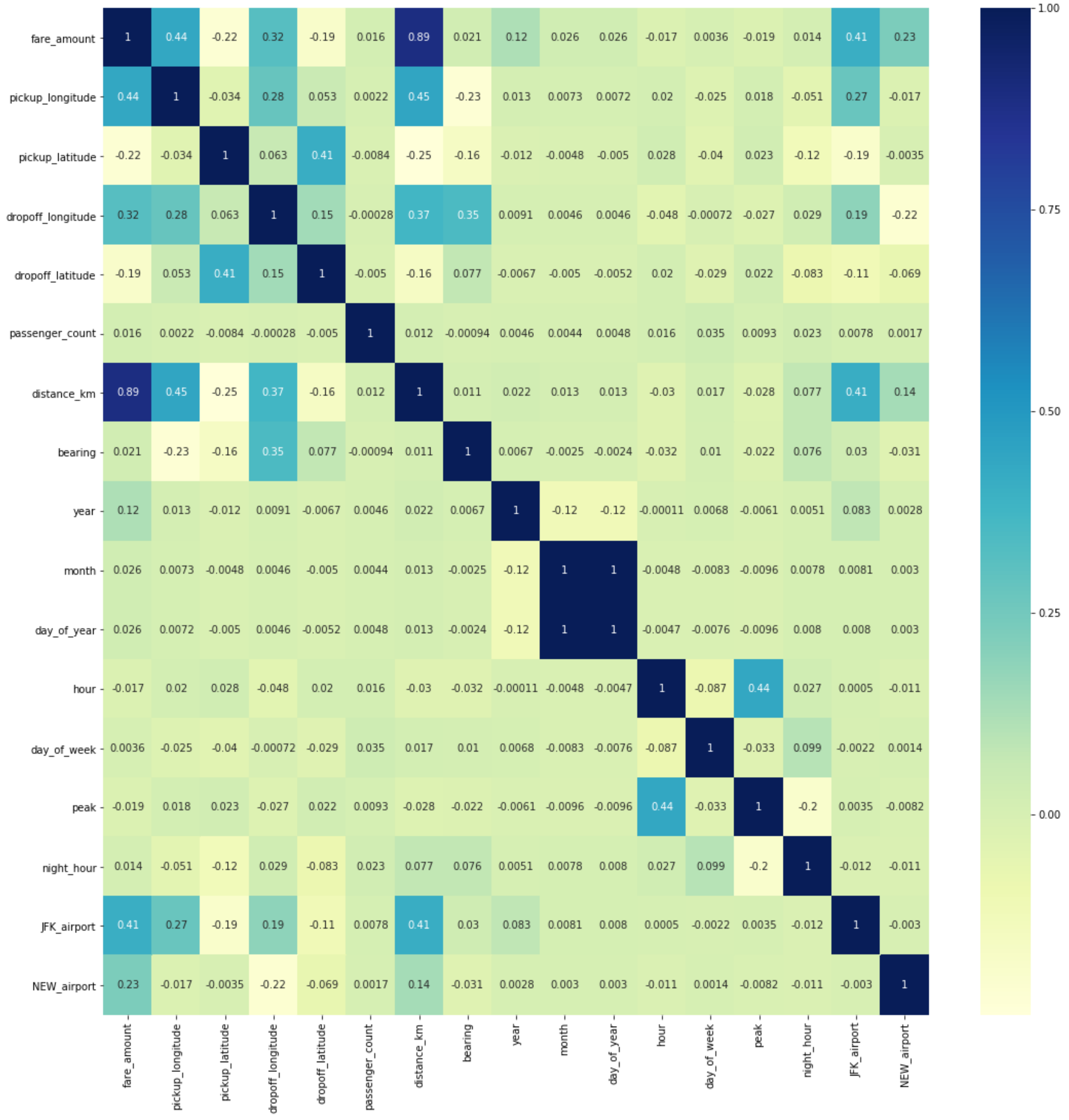
In addition, from Pearson's correlation matrix scatter plot below, it can be seen that following variables correlates strongly with taxi fare in the following order. Other variables have very weak correlation with taxi fare.

1. Distance (0.89)
2. Pickup Longitude (0.44)
3. JFK Airport (0.41)
4. Dropoff Longitude (0.32)
5. NEW Airport (0.23)
6. Year (0.12)
7. Pickup Latitude (-0.22)




Variation of Fares with Time
Fare Amount by Year
fare_amount
Fare Amount
10
11
12
13
2009
2010
2011
2012
2013
2014
2015
Fare Amount by Month
fare_amount
Fare Amount
10.8
11.0
11.2
11.4
11.6
2
4
6
8
10
12
Fare Amount by Day Of Week
fare_amount
Fare Amount
11.0
11.2
11.4
11.6
0
1
2
3
4
5
6
Fare Amount by Day Of Year
fare_amount
Fare Amount
10.5
11.0
11.5
12.0
0
50
100
150
200
250
300
350
Fare Amount by Hour
fare_amount
Fare Amount
11
12
13
14
15
0
5
10
15
20
Time Frame

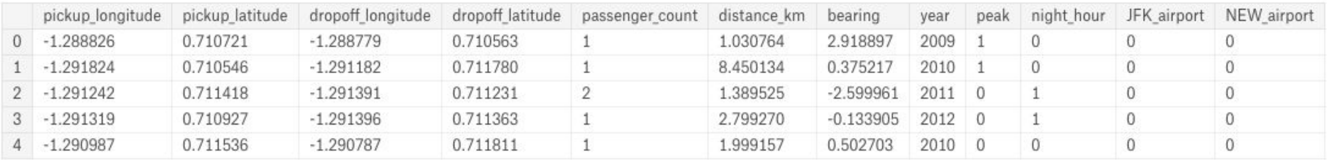
Pearson Correlation Matrix and Scatter Plot



Preparing Data for Model Development:

- Dropping features of least importance:
  - Before development of the baseline model, data of location which is in degrees is converted to radians which is conventional in mathematical modeling.
  - Features of pickup_datetime and hour are dropped as they are split or converted in to other variables such as hour and peak etc.

- ○ Other features such as day of the year, day of the month, month are dropped as they do not influence the fares significantly.
- ○ After dropping features the data looks like as shown below.



	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count	distance_km	bearing	year	peak	night_hour	JFK_airport	NEW_airport
0	-1.288826	0.710721	-1.288779	0.710563	1	1.030764	2.918897	2009	1	0	0	0
1	-1.291824	0.710546	-1.291182	0.711780	1	8.450134	0.375217	2010	1	0	0	0
2	-1.291242	0.711418	-1.291391	0.711231	2	1.389525	-2.599961	2011	0	1	0	0
3	-1.291319	0.710927	-1.291396	0.711363	1	2.799270	-0.133905	2012	0	1	0	0
4	-1.290987	0.711536	-1.290787	0.711811	1	1.999157	0.502703	2010	0	0	0	0

- Scaling the data: Furthermore the data is scaled using MinMaxScalar in sklearn library.
- Holdout data for evaluation: Data is split with sklearn train_test_split and 10% of the data is reserved for evaluation of model.
- Remaining data is split into training (80%) and validation data (20%).

Implementation

Baseline Model Implementation and Evaluation

As discussed earlier, baseline model is a multivariate linear regression model. Linear regression library from Scikit-learn is used for the fitting the linear regression model.

Deep Learning Model Implementation:

For the deep learning model development Keras with Tensorflow is used. Model has 4 layers with, 256 neurons in 1st layer and 128 neurons in second and 64 neurons in the third layer followed by 1 neuron in the output prediction layer.

Each layer use batch normalization and dropout percentage of 20%. Dropout layers drop some of the neurons during training to reduce overfitting. Each layer uses ‘relu’ as activation function.

For optimization of neural network, mean_squared_error is set as loss function and optimized with ‘Ada’. Batch size is set to speedup the optimization with batch size set to 2000. With batch size specified in the model, neural network solver updates the weights based on the sum of batch size of training set.

To reduce overfitting L2 regularization is used for third layer and L1 regularization for the last layer. Kernel regularization helps to detect and eliminate the features with least importances. This reduces overfitting of the model.

To avoid overfitting, some part of the dataset termed as validation data set is used to validate the model. If error on validation data increases in 5 consecutive iterations then neural network training is stopped.

Refinement

A deep neural network regression model has several hyper parameters. These include optimizer function choice, learning rate, epochs, batch size, number of layers and neurons in each layer.

In the refinement stage, the grid search method is used for optimization of dropout percentage and batch size of the deep neural network. Batch size is varied from 250 to 750 for the given neural network with 5 million data points. Running grid search is very expensive for the large dataset.

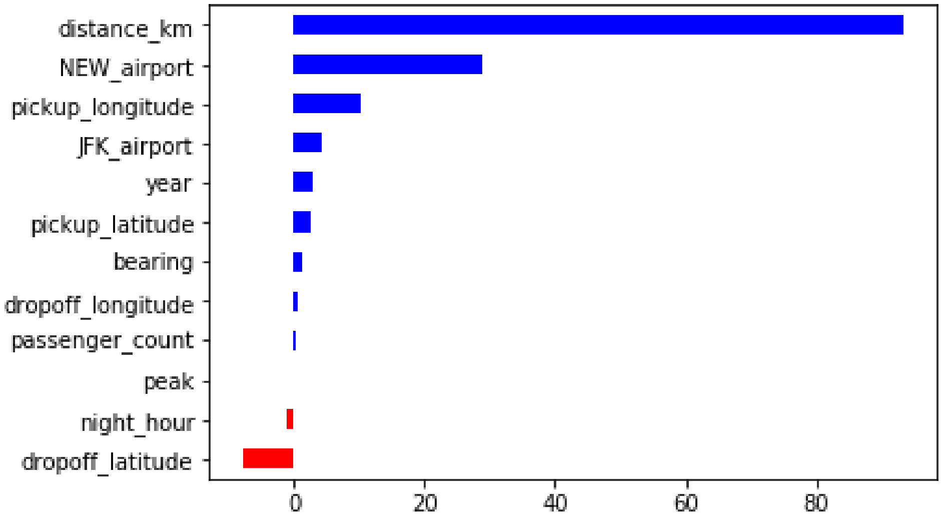
Results

Model Evaluation and Validation

Baseline Model Results

In the image below, coefficients of the linear regression model are plotted.

- Coefficient of the variable Distance_km is largest, followed by the NEW_Airport, Pickup_Longitude, Dropoff_Latitude, JFK_airport variables.
- Coefficients of the other variables, Year, Pickup_Latitude, Bearing, Dropoff_Longitude, Passenger_Count, Peak_Hour, Night_Hour are negligible.

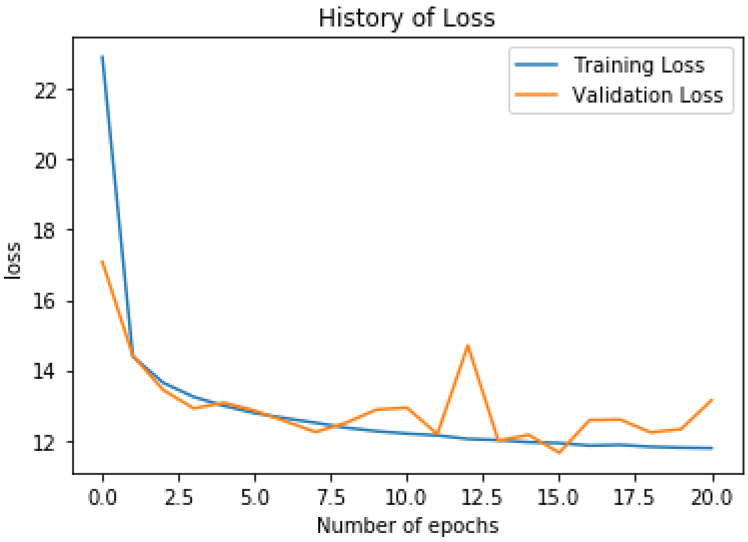


Variables such as Bearing, Passenger Count, Pickup Latitude, Drop-off Longitude play a small role on fare amount. Dropoff Latitude and Night_hour have negative coefficients.

Baseline model is then evaluated with holdout dataset. For the holdout dataset, **the baseline model predictions have RMS error of $3.93. and R^2 score of 0.82.**

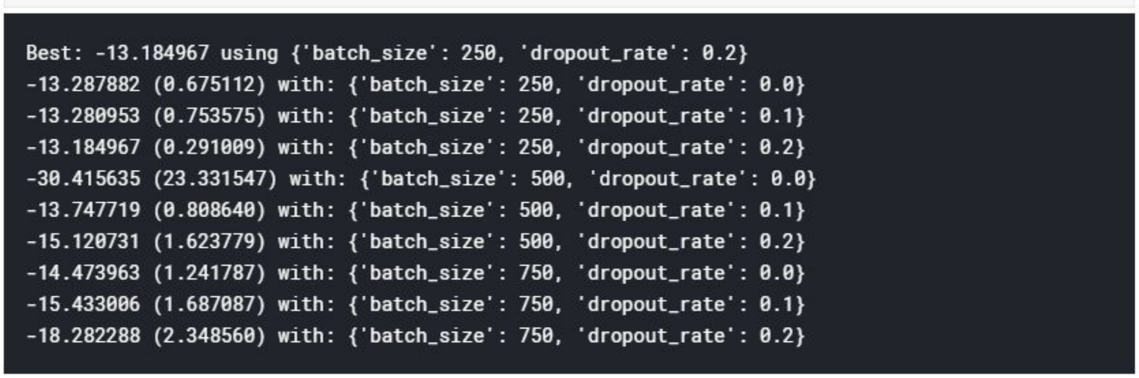
Deep Learning Model Results

History of the loss for training and validation dataset is shown in the Fig. History of loss function. After the validation loss increases after 5 iterations, optimization of neural network is stopped.



After the training is stopped, the model is then used to predict the results for the holdout dataset. Holdout dataset is not seen by the neural network model. For this holdout has a RMS of $3.4.

Following is the result of optimal batch size for best results is returned by the grid search method. Optimal batch size is 250 and dropout rate of 0.2.



```
Best: -13.184967 using {'batch_size': 250, 'dropout_rate': 0.2}
-13.287882 (0.675112) with: {'batch_size': 250, 'dropout_rate': 0.0}
-13.280953 (0.753575) with: {'batch_size': 250, 'dropout_rate': 0.1}
-13.184967 (0.291009) with: {'batch_size': 250, 'dropout_rate': 0.2}
-30.415635 (23.331547) with: {'batch_size': 500, 'dropout_rate': 0.0}
-13.747719 (0.808640) with: {'batch_size': 500, 'dropout_rate': 0.1}
-15.120731 (1.623779) with: {'batch_size': 500, 'dropout_rate': 0.2}
-14.473963 (1.241787) with: {'batch_size': 750, 'dropout_rate': 0.0}
-15.433006 (1.687087) with: {'batch_size': 750, 'dropout_rate': 0.1}
-18.282288 (2.348560) with: {'batch_size': 750, 'dropout_rate': 0.2}
```

Justification

For the same holdout dataset, baseline model RMSE of $3.92 and deep neural network model has RMSE of $3.5. RMSE value of deep learning model is lower than the baseline model with linear regression.

Conclusions

Reflection

It can be seen from the earlier analysis, distance is single most important feature influencing the taxi fare amount. However dataset does not provide any cues on actual travelled distance during a ride. In absence of data, distance used in this project is Haversine distance together with bearing.

In reality, taxi fare is based on driving distance through a city between two locations which can involve wait time due to signal, traffic, customer delays and payment to tolls. To implement this in the code, requires the support of paid GPS services such as Google or Bing maps. Though possible, performing travel distance calculations for millions of lines of data is more complicated and time consuming using GPS maps.

NY Taxi charges 50 cents per 60 seconds in slow traffic or when the vehicle is stopped. However data of idle time is also not available in the data which can not come from the GPS. Taxis in New York also report tolls paid in final bill, however this information is not shared here in the dataset. All of this is also reflected in the data where for a large number of rides distance is small but fare is large.

Improvement

Improvement could be made in the form of data collection such as obtaining record of idle time, distance on road travelled in city during a ride and tolls paid during each travel.

These new features will help improve the accuracy of the model which affect the taxi fares significantly.

References:

1. https://www.kaggle.com/c/new-york-city-taxi-fare-prediction
2. https://datasmart.ash.harvard.edu/news/article/case-study-new-york-city-taxis-596
3. http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

4. http://www.aei.org/publication/chart-of-the-day-creative-destruction-the-uber-effect-and-the-slow-death-of-the-nyc-yellow-taxi/
5. https://datasmart.ash.harvard.edu/news/article/analytics-city-government
6. https://stackoverflow.com/questions/27928/calculate-distance-between-two-latitude-longitude-points-haversine-formula
7. http://mathforum.org/library/drmath/view/55417.html
8. https://www.kaggle.com/nicapotato/taxi-rides-time-analysis-and-oof-lgbm
9. https://machinelearningmastery.com/grid-search-hyperparameters-deep-learning-models-python-keras/