New York Taxi Fare Prediction

Contents

Domain and Context:

Advent of technology in field of transportation especially in the taxi industry have created a several challenges to cab industry. Electronic dispatchment has replaced the traditional VHF radio system. these mobile data terminals are installed in each vehicle typically provides information on GPS localization and distance. This information collected over a period of time will be serving as dataset to analyse and predict the model for future.

Problem:

To improve the efficiency of electronic taxi dispatching systems it is important predict the fare. Given the information on historical rides and fare in centralized system, we need to analyse and predict the near accuracy fare using the optimal model.

Solution :

Supervised learning algorithms will be applied for prediction and unsupervised algorithm will be applied to get insights.

Datasets :

The dataset for this exercise is compiled by NYC Taxi and Limousine Commission , sample of the dataset is available in Kaggle repository

Data fields

- id - a unique identifier for each trip
- 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

Proposed Solution:

1. Load libraries :Identification ofML libraries Panda, numpy, scikit and tensor flow
2. Load dataset : We downloaded the dataset from kaggle which is of 5GB.
   Sample of 1million records(100 MB) has been taken for EDA.

3. Summarize Data: Describing the data/ Correlating the data.
4. Visualisation Data: Plotting graphs for data visualisation and analysis using Matplot and Tableau.
5. Data Cleansing: Identifying null values and replacing with mean and
   Identification of outliers through box plot and removing it.
6. Feature Selection: In addition to the existing 4 columns, we are deriving additional columns viz pick hour, day of the week and month from Pick up date/time column. Demand is derived by grouping lat/long for a time interval of every 15 mins.

7. Build Prediction Model:

   - To build a prediction Model following Algorithm will be used.
     1. Linear Regression
     2. Naive Bayes'
     3. Neural N/w
     4. Random Forest
   - Compare Algorithms: Comparing the precision and recall rates of various algorithms.
   - Improve Accuracy: Hyper tuning of parameters using GridSearchCV.
   - Cross validation
   - Ensemble Technique – Bagging, Boosting and Stacking.

8. Finalize Model
9. Predictions on validation dataset
10. Create standalone model on entire training dataset

Evaluation Metrics:

Following Key Performance Indicator would be used for assessing the performance of different prediction models.

1. Confusion Matrix:
   - Accuracy
   - Precision
   - Negative Predicted value
   - Recall
   - Specificity

2. Area under ROC curve:

3. RMSE

Exploratory Data Analysis

1. We extracted 1 million records from 5GB dataset.
2. Findings based on the summary statistics:

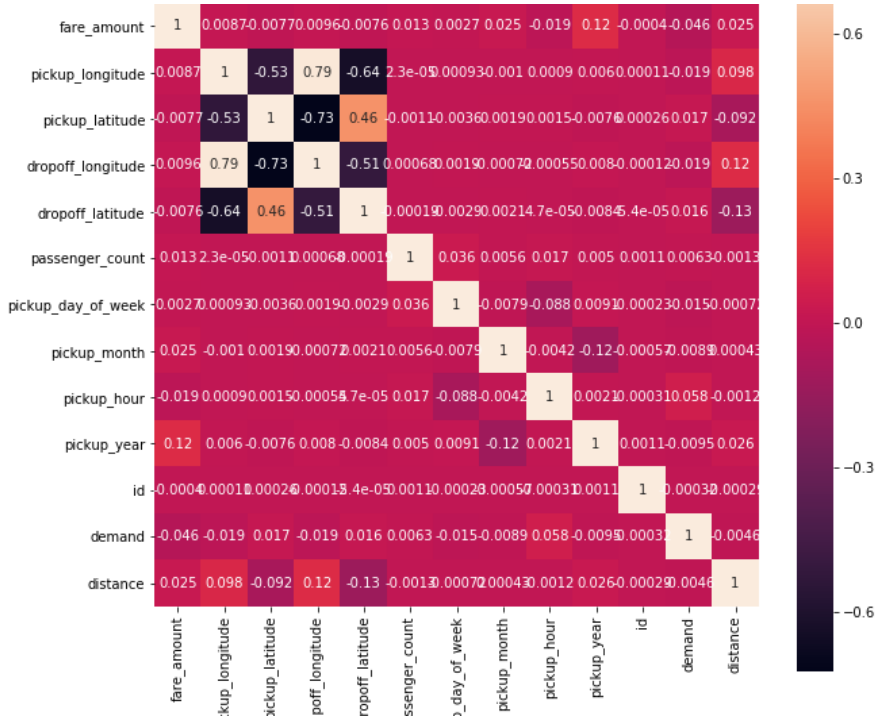
    - Fair price is usually in the range of 6 and 12$.
    - Max Fair price seems to be 500$, which is an outlier.
    - Based on the pickup and drop of lat/long information most of the trips are off short duration trips.
    - Most of the rides are taken by single passengers.
    - Trip distance is mostly between 1 and 5 KMs.
    - There are few discrepancies in data where it shows 200 passengers travelled on a trip and also a distance of 22k KMs covered for one case, which is not valid.

Findings based on Correlation Matrix:

As most of the rides are of short distance, except pickup lat/long and dropoff lat/lat coordinates are correlated, except above none are correlated.

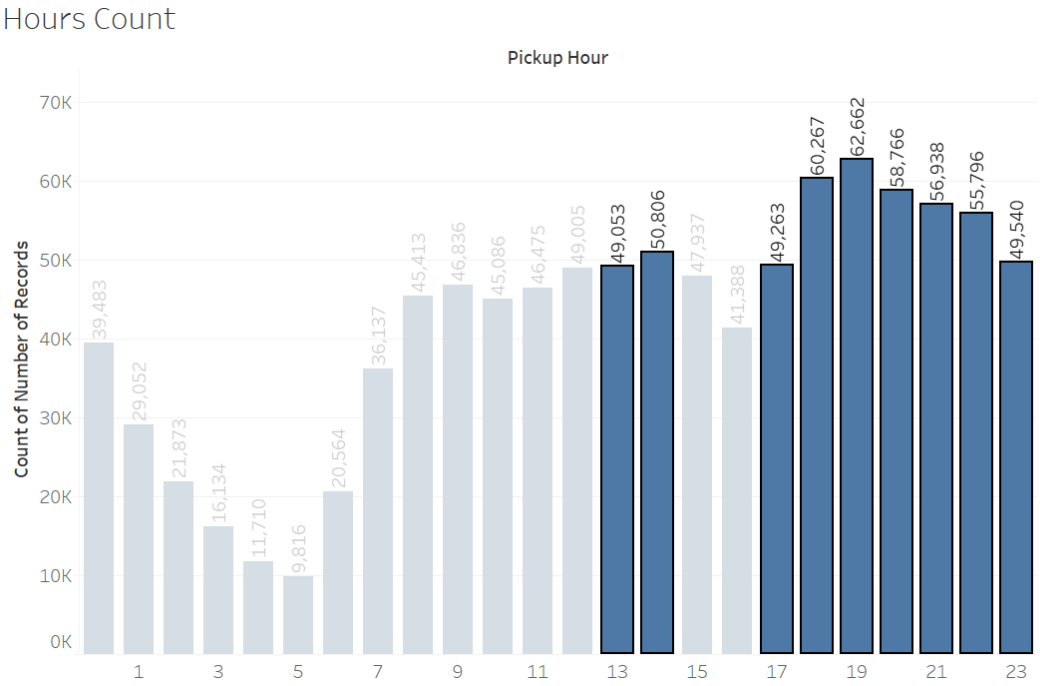
	count	mean	std	min	25%	50%	75%	max
fare_amount	1000000.0	11.348079	9.822090	-44.900000	6.000000	8.500000	12.500000	500.000000
pickup_longitude	1000000.0	-72.526588	12.057929	-3377.680000	-73.990000	-73.980000	-73.970000	2522.270000
pickup_latitude	1000000.0	39.928901	7.626147	-3116.290000	40.730000	40.750000	40.770000	2621.630000
dropoff_longitude	999990.0	-72.527860	11.324494	-3383.296608	-73.991385	-73.980135	-73.963654	45.581619
dropoff_latitude	999990.0	39.919954	8.201418	-3114.338567	40.734046	40.753166	40.768129	1651.553433
passenger_count	1000000.0	1.684924	1.323911	0.000000	1.000000	1.000000	2.000000	208.000000
pickup_day_of_week	1000000.0	3.039856	1.949970	0.000000	1.000000	3.000000	5.000000	6.000000
pickup_month	1000000.0	6.267875	3.436243	1.000000	3.000000	6.000000	9.000000	12.000000
pickup_hour	1000000.0	13.509477	6.513840	0.000000	9.000000	14.000000	19.000000	23.000000
pickup_year	1000000.0	2011.741106	1.860754	2009.000000	2010.000000	2012.000000	2013.000000	2015.000000
id	1000000.0	499999.500000	288675.278933	0.000000	249999.750000	499999.500000	749999.250000	999999.000000
demand	1000000.0	1.178272	0.446575	1.000000	1.000000	1.000000	1.000000	6.000000
distance	999990.0	24.987805	482.911957	0.000000	1.576720	2.764460	5.039720	22139.911397

Correlation matrix

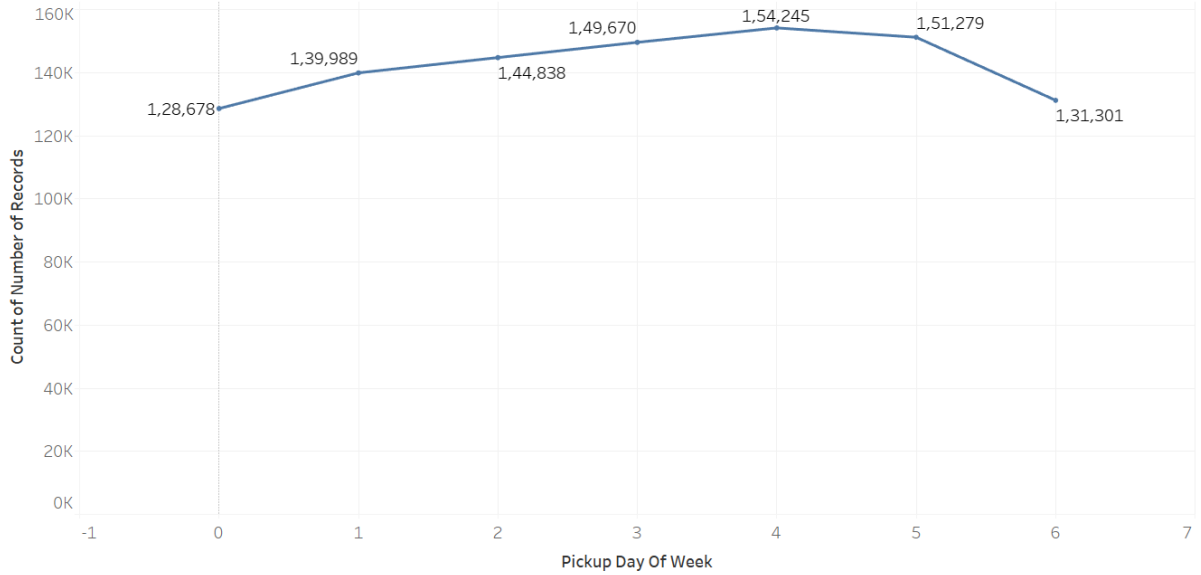


Exploratory Visualization

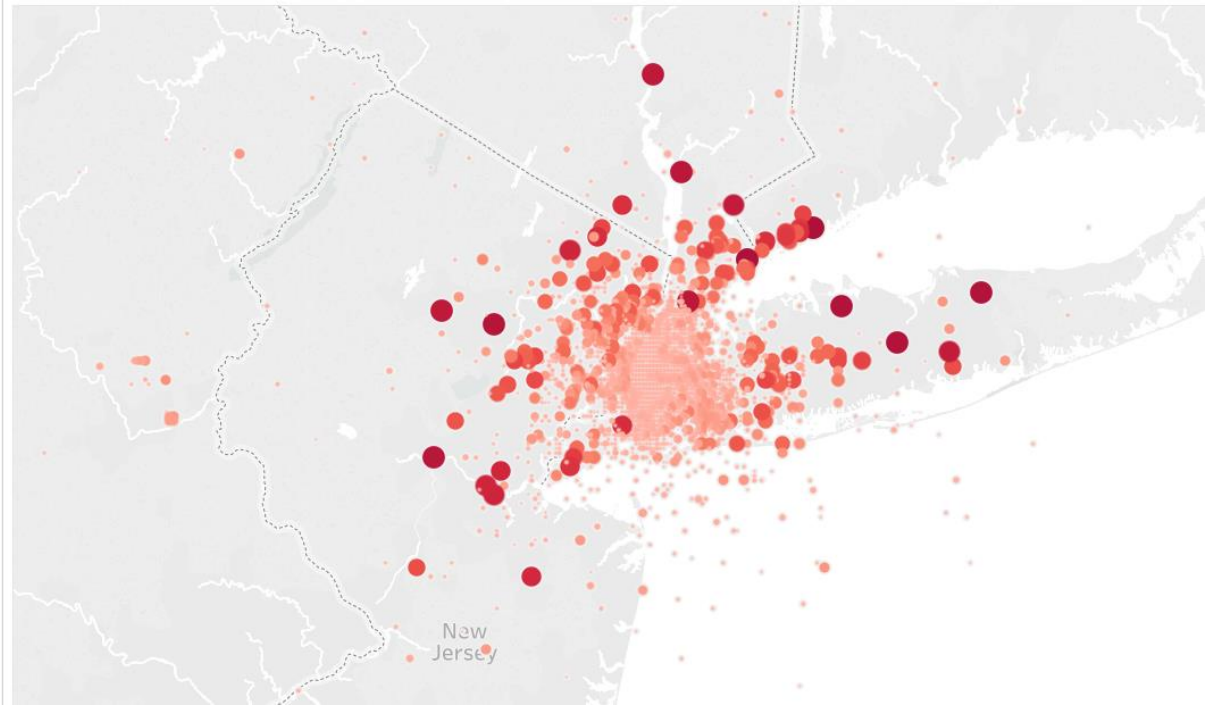
- Peak hours of taxi rides is between 5 and 8 PM then gradually declines till 12 PM.
- Long distance rides are usually between 3 and 5 AM and fare amount is usually higher.
- Fare amount is higher if pickup location is outside the city.

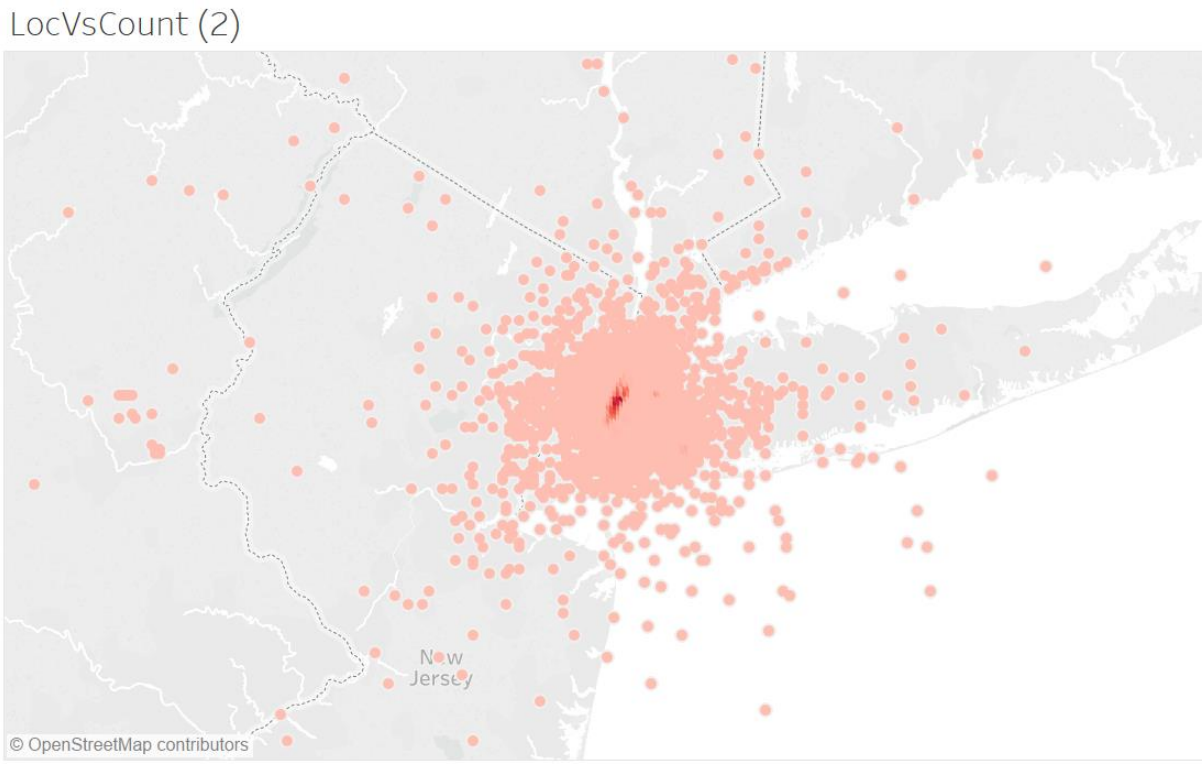


Week count



LocVsFare





Summary of Initial Findings

- Pickup hour, lat/long and distance play a major role in determining the taxi price. Above is used for building the prediction model.
- We have used linear regression in building a prediction model since fare amount is a continuous data.

Challenges

As there are close to 1 million records in the dataset , it is becoming tedious to process huge set of data.

Next steps

- In order to improve the accuracy of the model , neural networks (tensorflow) has been identified as best option to determine the fare price.
- Spark will be used in conjuction with neural networks to process huge amount of data.

Appendix

Codebase link : https://github.com/sreelakshminarayanan/nyctaxianalysis

Tableaue public profile:

https://public.tableau.com/profile/prabakar5597#!/vizhome/Capstone_NYC_Taxi/