New York City Taxi Fare Prediction

Milestone: Project report

Group 11

Kevin Kurias Saibu

Saujanya Sanjay Zemse

845-263-0285

857-241-0499

saibu.k@northeastern.edu

zemse.s@northeastern.edu

Percentage of Effort Contributed by Student 1: 50%

Percentage of Effort Contributed by Student 2: 50%

Signature of Student 1: Kevin Kurias Saibu

Signature of Student 2: Saujanya Sanjay Zemse

Submission Date: 04/23/2021

NEW YORK CITY TAXI FARE PREDICTION

PROBLEM SETTING

New York City is the busiest and happening city in the US, people are always thriving to achieve something, trying to get rich and famous, that is why it is also called the city of dreams. Empire State of mind is a song by Jay-Z which describes New York perfectly, I would recommend you guys to listen if you have not already. New York City also has three airports which increase the taxi traffic, going to and fro from the airports. Sometimes it happens that taxi drivers are not reasonable and charge increased amounts, predicting the fare would put this practice on hold. We faced multiple challenges while exploring the data like data no being in proper format, useful features were derived and not already present, the data had more than 55 million records which was computationally impossible for us to implement. Taxicabs in New York City come in two varieties: yellow and green; they are widely recognizable symbols of the city. Taxis painted yellow (medallion taxis) can pick up passengers anywhere in the five boroughs. Those painted apple green (street hail livery vehicles, commonly known as "boro taxis"), which began to appear in August 2013, are allowed to pick up passengers in Upper Manhattan, the Bronx, Brooklyn, Queens (excluding LaGuardia Airport and John F. Kennedy International Airport), and Staten Island. Both types have the same fare structure. Taxicabs are operated by private companies and licensed by the New York City Taxi and Limousine Commission (TLC). It also oversees over 40,000 other for-hire vehicles, including "black cars", commuter vans, and ambulettes. The average hourly gross income for a medallion driver in 2015 was $30.41, not including tips, according to the 2016 TLC Factbook. Boro (green) taxi drivers made $20.63, also excluding tips. Evening hours are typically more lucrative. Medallion Taxis collectively make between 300,000 and 400,000 trips a day, while Green Taxis, which hit the road in New York City in 2013, do almost 50,000 daily on average. While the TLC does not regulate specific shifts, the morning shift for a medallion taxi typically begins at 5:30 a.m., and the evening shift often starts at 5:15 p.m. Trips peak Friday evenings for medallion cabs, and Saturday night for Green Taxis.

PROBLEM DEFINITION

We usually have the question that how is the fare already calculated when we book any taxi service online? This project would help you answer this question, and other questions such as what factor influences fare the most, do fare change according to the time of the day or day of the week, etc. Predicting the fares is a regression task so we are going to use multiple regression models trained on the data for the prediction of fares. Given the multiple features in the data we need to implement a regression model which would successfully predict taxi fare and with minimum error possible. Most of the online taxi services implement this method to show the estimated fares before the trip to the customer. This helps customer decide if the trip is affordable for him/her or not. Through this project we trying to find out reasons why taxi fare would change for same distance, major factors which affect the fare etc.

DATA SOURCES

Data was obtained from competition conducted by Kaggle, Google Cloud and Coursera. It contained more than 55 million records. We took a sample of the whole data for better analysis and faster computing as it took 4-5 minutes to just load the data in our Jupyter Notebooks. The dataset included features as date and time, fare amount, pickup and drop off coordinates and passenger count. Following is the link from which we got our dataset. Kaggle, a subsidiary of Google LLC, is an online community of data scientists and machine learning practitioners. Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges. Kaggle got its start in 2010 by offering machine learning competitions and now also offers a public data platform, a cloud-based workbench for data science, and Artificial Intelligence education. Its key personnel were Anthony Goldbloom and Jeremy Howard. Nicholas Gruen was founding chair succeeded by Max Levchin. Equity was raised in 2011 valuing the company at $25 million. On 8 March 2017, Google announced that they were acquiring Kaggle. Following is the link of the dataset and competition:

https://www.kaggle.com/c/new-york-city-taxi-fare-prediction/data

DATA DESCRIPTION

Data was obtained from competition conducted by Kaggle, Google Cloud and Coursera, it contains more than 55 million records of taxi hired by the people of New York City. It contained many outliers and values which made no sense, so we had to perform data cleaning task to remove the outliers and covert the data into proper format to predict the fares. Dataset contained features such as date and time, fare amount, pickup and drop-off coordinates (longitude and latitudes) and passenger count. We found out that the mean of fare is 8.07$ while maximum fare in a single trip is 9400$ and minimum fare amount for single trip is -3$ which is impossible as fare cannot be negative. Getting a negative fare means no amount paid so it doesn't make any sense. Passenger count analysis shows us that max passengers travelled in a taxi for a single trip were 2000 people, this value is not possible as more than 5 people cannot fit into a taxi. So, we found out that passenger count variable also contains outliers and unusual values which cannot be explained which we would need to discard. After data cleaning task we finally got 450k clean data on which we performed exploration and visualization.

DATA EXPLORATION

We found out that date and time is in UTC for some reason, so we converted it to EST. We used pythons datetime library to convert the date and time into hour, day, month and year, and stored them in different columns for better analysis. We calculated distance between source and destination using Haversine distance formula. This formula gives great-circle distance between two points – that is, the shortest distance over the earth's surface. Using this formula, we calculated the distance between pickup coordinates and drop-off coordinates, in miles.

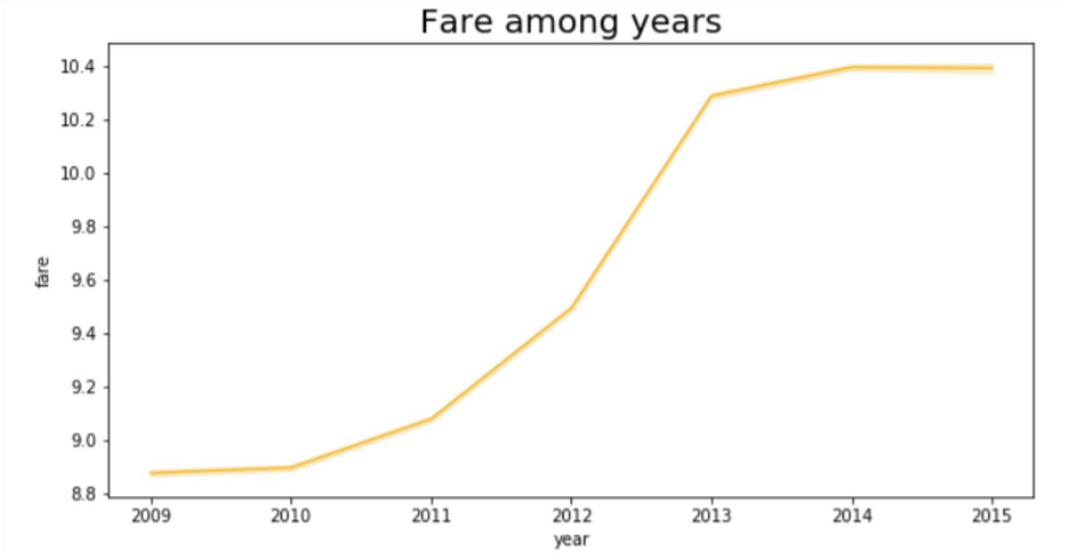
Haversine formula:

$$a = \sin^2(\Delta\varphi/2) + \cos\varphi_1 \cdot \cos\varphi_2 \cdot \sin^2(\Delta\lambda/2)$$
$$c = 2 \cdot \text{atan2}(\sqrt{a}, \sqrt{(1-a)})$$
$$d = R \cdot c$$

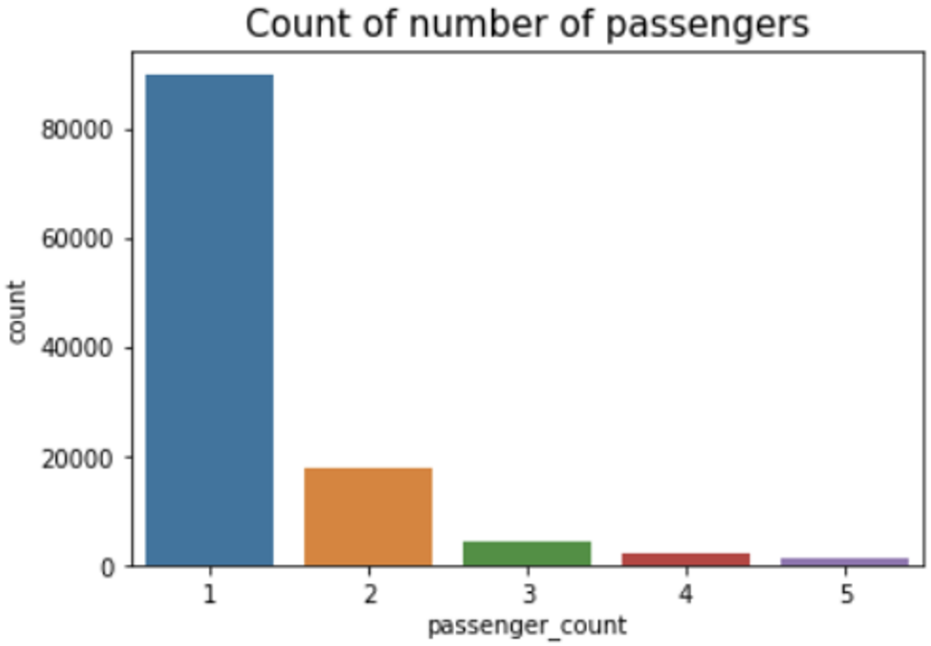
where φ *is latitude,* λ *is longitude,* R *is earth's radius (mean radius = 6,371km); note that angles need to be in radians to pass to trig functions!*

In order to remove outliers, we deleted records containing fare greater than 30$, data contained some co-ordinates which were of outside New York city, so we removed them too. We limited the distance to under 30 miles and passenger count to 5 as more than 5 people can't fit into a taxi. And at last, we took sample of the complete data as the dataset was very large and it would be computationally impossible for us to work on such a large dataset. Taking the sample was justified as we were doing regression i.e., predicting the fare so there is no problem of imbalance data which would affect the prediction of the model. We found out that max fare amount was 9500$ which is too high for New York City so the dataset contained outliers and unusual values if trained on them model would give inaccurate results, so we had to remove all the outliers. Through domain knowledge we found out the co-ordinates of New York City, we deleted all the values in the dataset which contained co-ordinates outside of New York City. We thought that main city is not more than 30 miles, so to get more accurate results we limited the distance to 30 miles. Passenger count is also one of the most crucial factor in determining the fare. More number of passengers require a larger taxi and thus fare increases. So, for normal travel we considered number of passengers to be less than 5 as more than 5 people cannot fit in a taxi. Our data contained values from 1 to 9. Having more than 5 people, we thought would be impractical for a normal taxi, so we neglected those records. We used multiple bar plots and box plots to find out outliers.

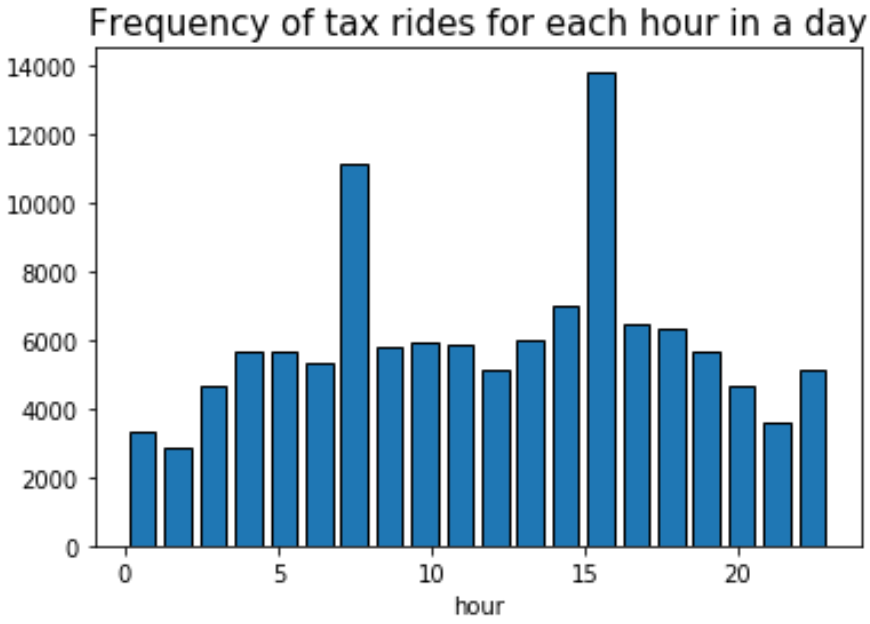
Even though the primary aim is to predict the taxi fare for each ride, exploratory data analysis on the data must be performed to find out hidden patterns and trends in the data. It is also crucial to find out the nature of taxi fare and all factors influencing the fare in a ride. In order to perform the aforementioned tasks various visualizations were generated from the data.



We used this line graph to showcase the change of taxi fare as the years went by. We can see that there was sharp rise in taxi fares from 2011,2012,2013, this could be because of increase in fuel prices or taxi union workers increasing the base fare of taxi.

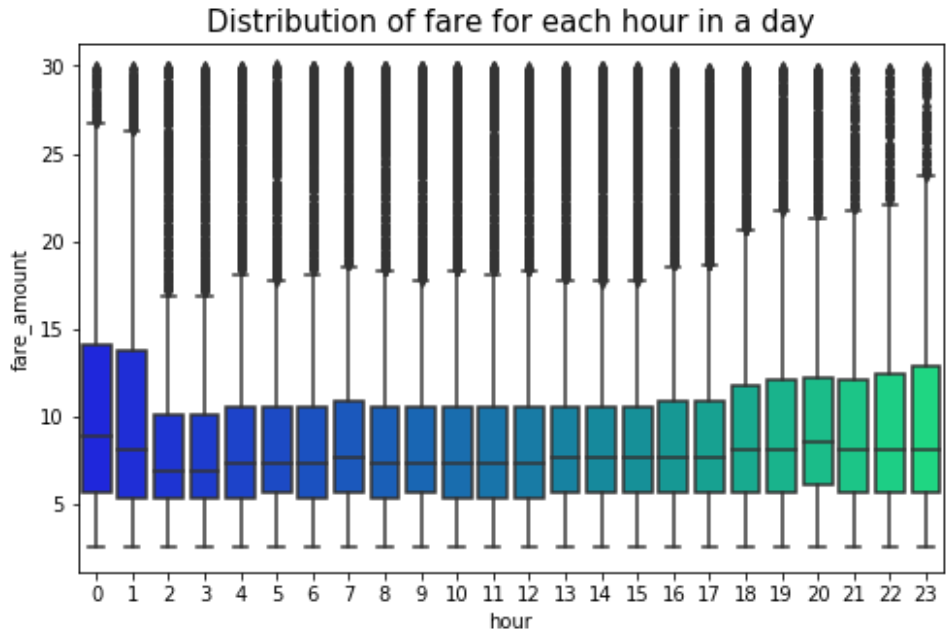


Through the above bar plot, we visualized the count of passengers travelling in taxis. We can clearly see that single passengers travel the most as it is specified by the high count.

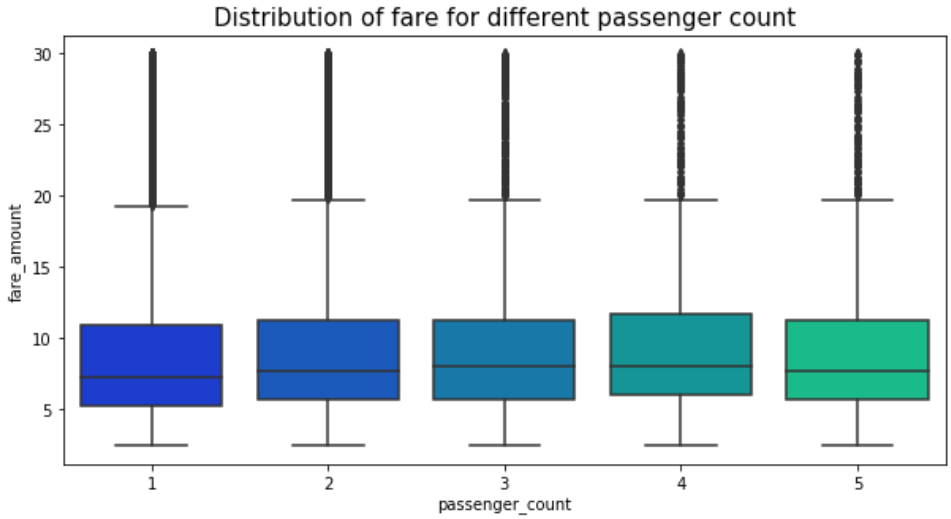


Newyork city is also known as the city that never sleeps. Even though there is a significant amount of traffic caused by the yellow cabs at every hour, from the above bar chart we can infer that the majority of rides takes place during morning around 7am to 9 am and during evening around 3pm to 5pm. This might be because of the working professionals making use of these cabs for their daily commute.

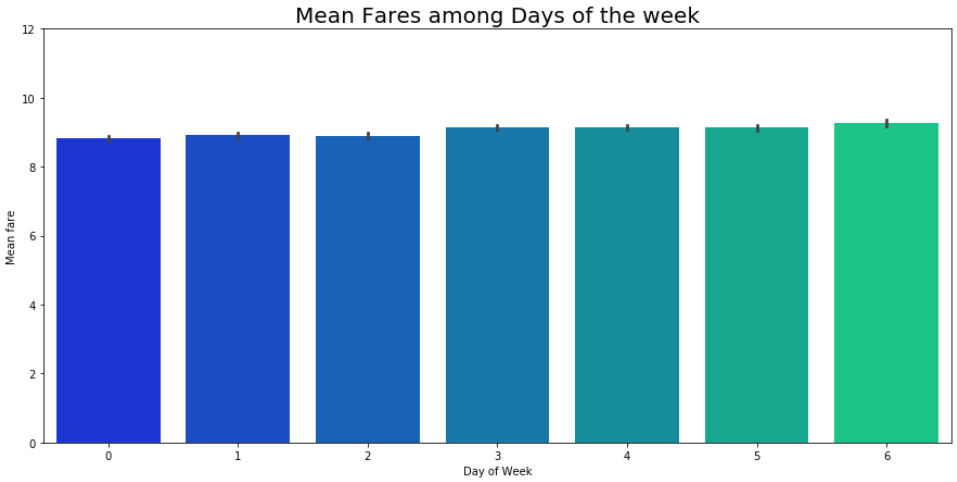
As already mentioned, the total fare of a taxi ride depends on various factors like distance, time of the day, day of week, destination etc. Each of these factors affects the fare in different ways. The influence of all these factors can be discussed as follows.



The above side by side boxplots represents the distribution of taxi fare on each hour of the day. It can be easily noted that the fare is significantly higher during late nights and past midnight compared to that of daytime. Since there is no factual explanation to this, this might be because the drivers charge more for tourists who are exploring the nightlife.

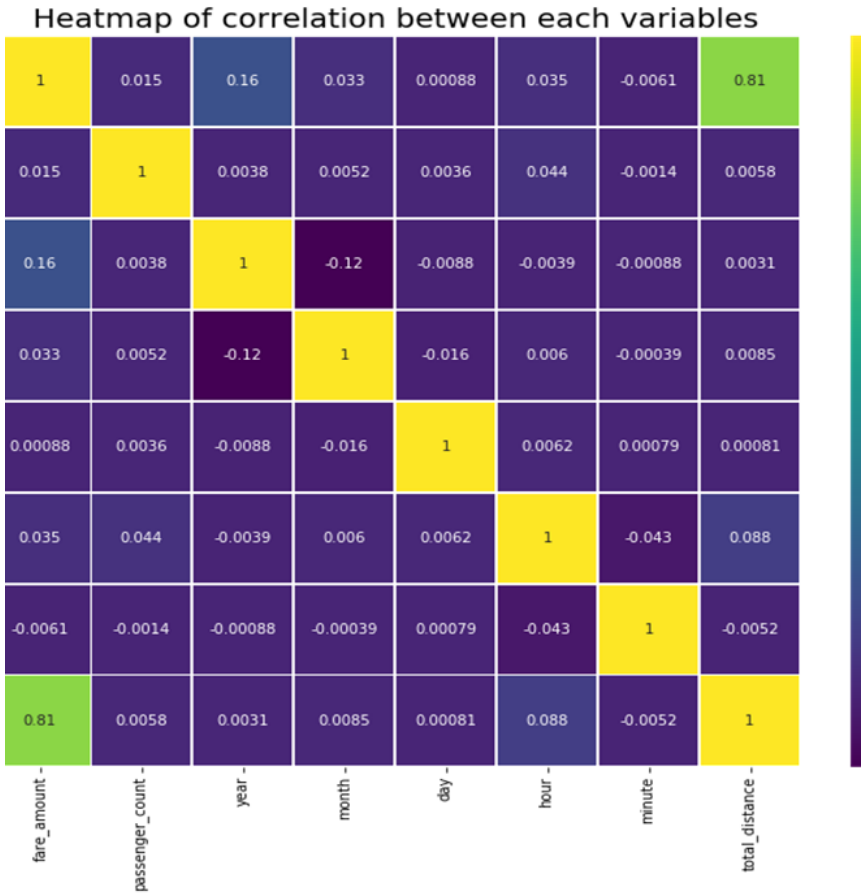


The above side by side boxplot explains the distribution of fare on different passenger counts. By analyzing the above plot, we can conclude that there is no significant effect on fare caused by number of passengers. But it can be noticed that there is a slight increase in the fare for rides including four passengers.



The influence of day of week on taxi fare can be analyzed from the above bar graph. Like number of passengers there is no significant difference in the mean fare for each day. Even though we can notice that there is a slight increase in fare during weekends. It can be assumed that this is because of the huge tourist flow in Newyork city.

One of the perfect methods to identify features which affects the fare directly is by means of creating a correlation plot which shows the correlation between each predictor and predictors to target variables. The correlation plot of this data can be depicted as,



From the correlation plot it can be noted that there is a high positive correlation between total distance and the fare. All the other predictors show only a little positive correlation. While it is a domain knowledge that there won't be any influence on fare by minute and year, those features can be neglected when training a model.

DATA MINING MODELS

In order to achieve the final goal of this project which is to predict the total fare of a taxi ride when certain features are given, certain machine learning algorithms has to be trained using the available data. Taxi fare is a continuous variable and prediction of it makes it a regression problem. In order to train and validate each machine learning algorithm the data has been splitted into eighty percent training data and the remaining twenty percent as test data. There was no scope for standardizing the data since all of the predictors were in similar range. Similarly, there was no scope for dimension reduction since some of the predictors were neglected through domain knowledge and based on the insights retrived during exploratory data analysis. This includes minute and number of passengers. Several regression algorithms from multiple linear regressor to 4-layer artificial neural network was trained to analyze and compare their performance on the predicting the taxi fare. To compare the performance of each model metrics like RMSE and R-Squared value were calculated. Different algorithms trained for this project are,

1. Multiple Linear Regressor
2. Lasso Regressor
3. Bayesian Ridge Regressor
4. K-Nearest Neighbors Regressor
5. Random Forest Regressor
6. Gradient Boosting Regressor
7. Xgboost Regressor
8. Adaptive boosting Regressor
9. 4 Layer Artificial Neural Networks.

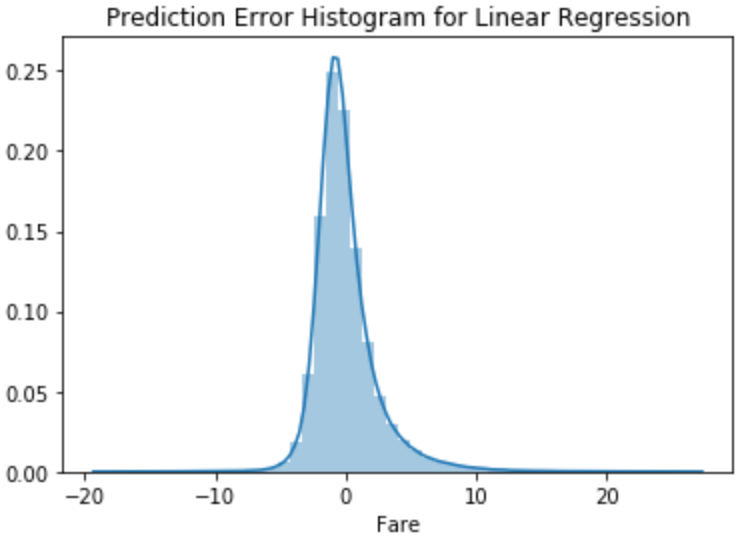
The performance of each model in predicting the taxifare can be discussed as follows,

Multiple Linear Regressor:

The multiple linear model is trained on the training data and the performance metrics on the test data can be represented as,

R-Squared	0.6791
RMSE	2.493
MSE	6.217

The distribution of prediction error of Linear regressor can be represented as,

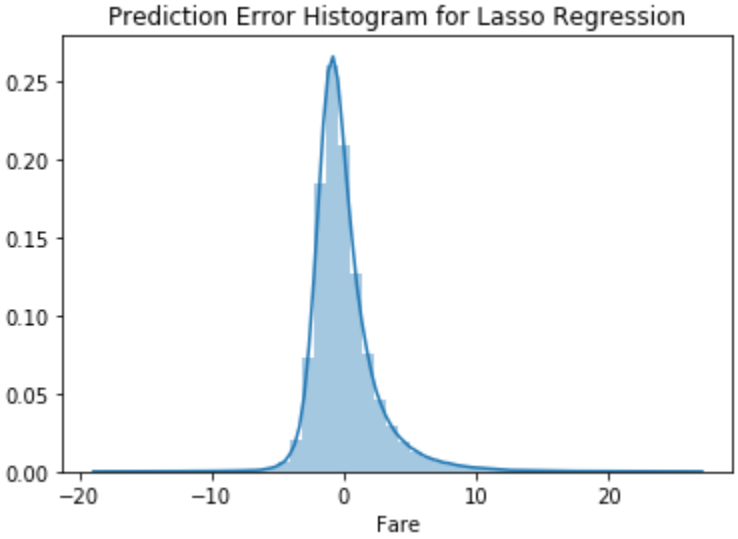


Lasso Regressor:

The previous model is modified by adding a parameter alpha which penalizes the sum of absolute values of the weight. The best alpha value is selected by means of hyperparameter optimization. The performance metrics of Lasso regressor on the test data is as follows,

R-Squared	0.679
RMSE	2.493
MSE	6.217

The distribution of prediction error can be represented as,

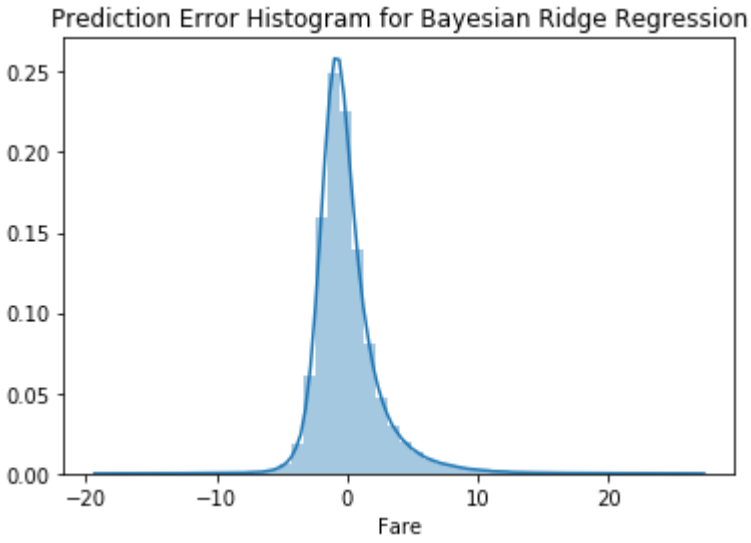


Bayesian Ridge Regressor:

The above model is further modified as a bayesian ridge regressor, The primary difference is that in ridge regression the cost function is altered by adding a penalty equivalent to square of the magnitude of the coefficients. The performance metrics of this model on the test data is as follows,

R-Squared	0.679
RMSE	2.493
MSE	6.217

The distribution of prediction error can be illustrated as,

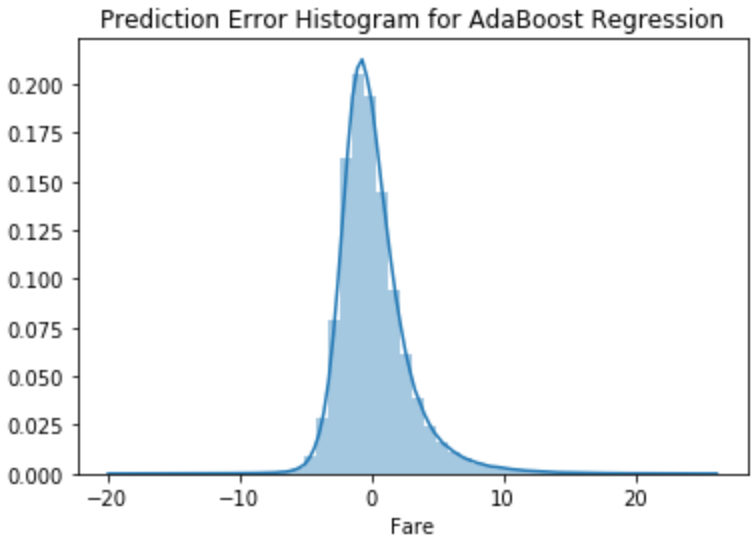


K Nearest Neighbours Regressor:

Even though KNN is not ideal for very large datasets because of its tendency to follow the noise in the data, we wanted to see how it performs in predicting taxi fare. The model is trained with different values of K and all the models gave similar response in the test data. The performance metrics on the test data obtained from the model with K=10 is as follows,

R-Squared	0.6366
RMSE	2.6534
MSE	7.0408

The distribution of prediction error can be illustrated as,



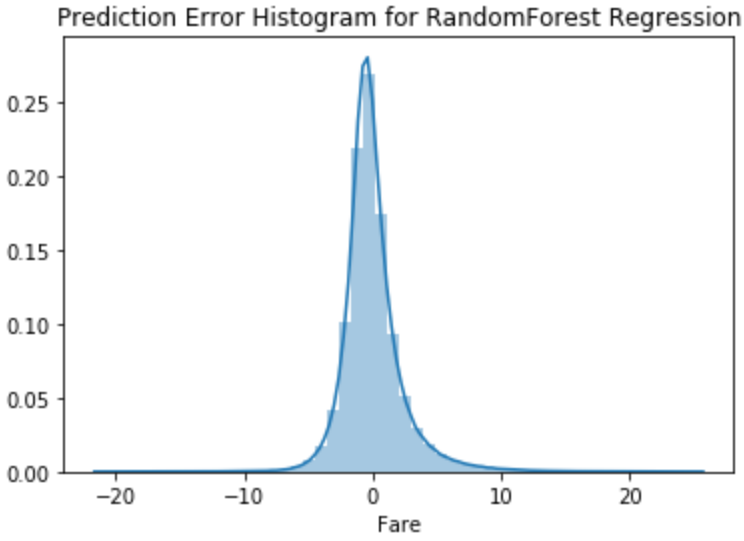
Large datasets usually perform well on ensemble methods. Taking this point in mind the data is trained on various bagging and boosting algorithms and the corresponding metrics were noted.

Random Forest Regressor:

Random Forest regressor is one of the bagging algorithm which follows a bootstrap approach.This algorithm helps in minimizing the noise and bias variance.A random forest is a meta estimator that fits a number of classifying decision trees on various subsamples of data to improve the predictive accuracy and to control over fitting. The algorithm is trained on the training data and the performance metrics on the test data is as follows,

R-Squared	0.7106
RMSE	2.36798
MSE	5.607

The distribution of prediction error obtained from random forest regressor can be illustrated as,

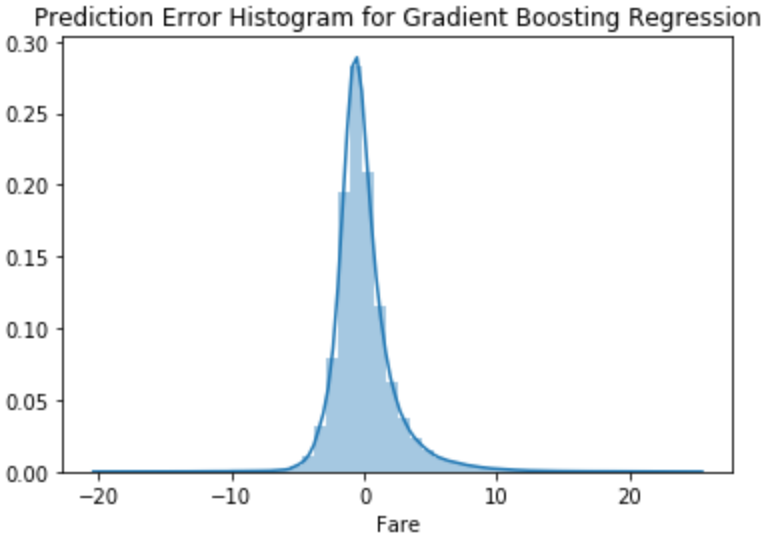


Gradient Boosting Regressor:

Gradient boosting regressor is the basic boosting technique which can be used for both classification and predictions. Boosting is a method of converting weak learners into strong learners. In boosting, each new tree is fitted on a modified version of the original dataset. Gradient boosting trains many models in a gradual,additive and sequential manner which helps in increasing the efficiency. The performance metrics obtained from the test data after training the gradient boosting regressor is,

R-Squared	0.7185
RMSE	2.3355
MSE	5.458

The distribution of prediction error on the test data can be illustrated as,

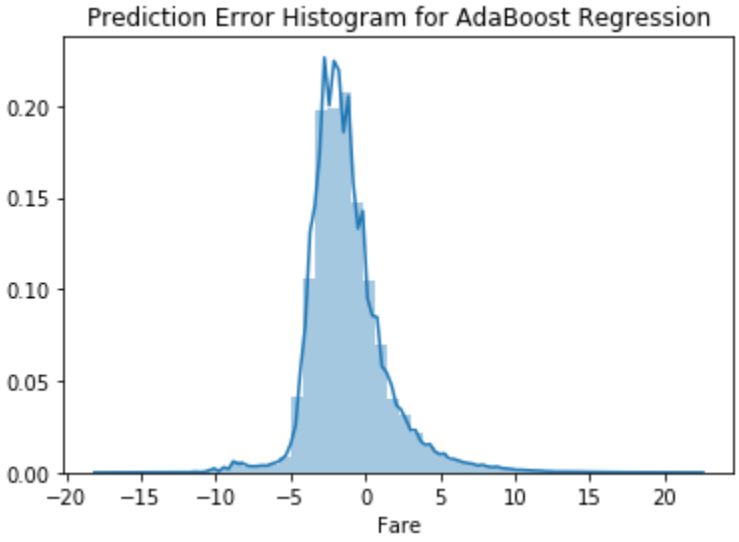


Adaptive Gradient Boosting Regressor:

Adaboost Regressor is another ensemble method in which the weights are assigned to each instance, with higher weights to incorrectly classified instances. The adaboost model is trained on the training dataset and the performance metrics on the test data is as follows,

R-Squared value	0.5651
RMSE	2.9021
MSE	8.427

The distribution of prediction error can be illustrated as,

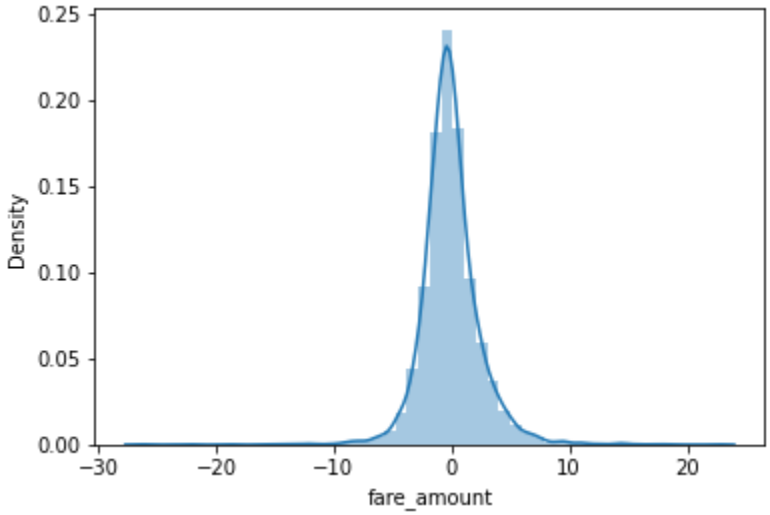


Xgboost Regressor:

Extreme gradient boosting algorithm is one of the advanced ensemble method which showed promising results in both classification and regression. Hyperparameter tuning is performed for this model with the help of RandomizedSearchCV to make a total of 25 fit on the training data with different combinations of parameters. The performance metrics after tuning is,

R-Squared	0.7576
RMSE	2.5569
MSE	6.537

The distribution of prediction error obtained from xgboost regressor can be illustrated as,



Artificial Neural Network:

A neural network with four hidden layers is created with the help of Keras and Tensorflow. Relu activation function is used in the firest 4 layers and linear activation function is used in the output layer since the aim is to predict a continous variable. Adam optimization is used to change the weights. Dropout layers were added to avoid overfitting by assigning some noise to the activation function of certain nodes. The performance metrics obtained from the neural network is as follows,

R-Squared	0.668
RMSE	2.532
MSE	5.85

PERFORMANCE EVALUATION

The performance metrics of each model that has been trained can be summarized as,

Machine learning algorithm	R-Squared	RMSE	MSE
Multiple Linear Regressor	0.67	2.493	6.217
Lasso Regressor	0.67	2.493	6.217
Ridge Regressor	0.67	2.493	6.217
KNN Regressor	0.6366	2.6534	7.0408
Random Forest Regressor	0.710	2.36798	5.607
Gradient boosting Regressor	**0.718**	**2.3355**	**5.458**
Adaboost Regressor	0.566	2.9021	8.427
Xgboost Regressor	0.7576	2.5569	6.537
Artificial Neural Network	0.668	2.532	5.85

It can be seen that almost all the models gave similar results on the test data with only slight difference between each other. The multiple linear regressor and its modifiers exhibits same values of RMSE on the test data. KNN Regressor and adaboost regressor appears to be the worst performing models with RMSE on test data above 2.6. It can be also noted that much better results were obtained from some ensemble method among which gradient boosting algorithm outperformed all other promising models like random forest regressor and artificial neural network.

PROJECT RESULTS:

Since the primary aim of this project was to predict the fare of a taxi ride when some features are given, almost all models performed well in predicting the fare. It can be noted that the distribution of each model's prediction error resembles a smooth bell curve which depicts that the model is performing well in predicting the fare of new rides. Among all models trained, gradient boosting regressor outperformed all other promising models by giving a low **RMSE value of 2.33.** The residuals distribution curve of gradient boosting regressor is a very fine bell curve.

FUTURE UPDATIONS AND IMPROVISATIONS:

This project has a wide scope for future updating. Due to the lack of good computing power and since the data was too large with 5 million records, additional fine tuning on each model couldn't be done. Some of the future updating that can be done in the project are as follows,

- Taking notable destinations like airports, famous tourist attractions etch into consideration which could influence the total performance of the model. So far only the total distance of a ride is considered. For two different destinations the fare might be different even if the distance is same in both cases.
- Promising results would have been obtained from neural networks if proper hyperparameter tuning is performed to find out the optimum number of nodes that can be present in each hidden layer. This can be performed with the help of Keras auto tuner.
- Various unsupervised machine learning algorithms like K means clustering can be performed on the data to figure out rush hours and the busiest areas within the city.
- The best model can be deployed in any cloud-based platforms like AWS Sagemaker, Heroku etc to make it more user friendly so that a user can easily get benefitted from this project.

IMPACTS OF THE PROJECT OUTCOME:

At the end of road, the primary objective of this project is to give the user an idea about taxi fare when he/she wants to hire an NYC yellow cab. Since NYC is a business and tourist hub with an average flow of 66 million tourists each year, each person is susceptible to cheating and fraudulence. The developed model can be used to predict the fare when certain inputs like approximate distance, day, number of passengers, hour etc. This can be used by tourists and working professionals moving to city to stay away from fraudulence. Further fine tuning and modifications can be done on the model to improve its predictive efficiency.