New York Taxi Fare Prediction

Cyril Cordor Matthew Sims Eung Keun Kim

December 11, 2018

Abstract

This report presents an analytical study of various predictive models and features in order to accurately predict taxi fares in New York City. The prediction of NYC taxi fares is important due to the rise of other ride sharing services, such as Lyft and Uber, that have become popular in the last decade. This research provides an exploratory analysis to determine which features are useful in predicting fare amount, and applies multiple interpretations of features to develop models using the following machine learning techniques: Linear Regression, Random Forest, and Gradient Boosting. The lowest root mean squared error (RMSE) of $3.40027 was achieved using Gradient Boosting with the Haversine distance formula. While these models developed did not achieve the most accurate results, they are useful in providing approximate estimates for those using taxi services in NYC.

1 Introduction

One of the most popular ways to travel around New York City is through the use of taxis, although rivaling taxi usage in New York are ride-sharing services such as Uber or Lyft. One of the major advantages of these ride-sharing services is that they provide estimates of how much the ride will cost before the trip takes place. This is not the case for users of taxi transportation who must wait until the end of their trip to determine the cost of the ride. This can often be frustrating because users may end up paying much more than they were expecting.

The use of predictive modeling to determine the cost of a taxi ride prior to the trip would be useful for those who prefer using a taxi service over the unregulated ride-sharing services. This project aims to develop a model based on the following information that is available prior to taxi trips:

Given Feature List
Fare Amount
Pick-up Datetime (Day of Week, Month, Year, and Hour)
Pick-up Location and Drop-off Location (Longitude / Latitude)
Number of Passengers

2 Related Work

The idea for this project came from a Kaggle competition where participants developed models to predict taxi fare for trips in New York City. Kaggle suggested as a baseline approach to develop a Linear Regression model that only considers the distance between the pick-up and drop-off points. This approach to predict taxi fare resulted in a RMSE of $5-8. The Linear Regression model provides a starting point to developing a new model that will substantially improve upon this RMSE.

As this was a competition on Kaggle, there are many discussions and articles on the topic, on exploratory analysis of the NYC taxi data, and creating predictive models for the fare amount by data science enthusiasts. One author, Aiswarya Ramachandran, details some of her findings from an initial exploration of the taxi data. She made interesting observations like the cluster of taxi pick-ups and drop-offs from the three airports that are close to NYC: LaGuardia (LGA), John F. Kennedy (JFK), and Newark International (EWR). Some of these observations were crucial in deciding what new features she needed to create for her predictive model[Ram16].

A metric tool considered in work related to predicting taxi fare is the rotation of the latitude/longitude coordinates to allow for the predictive model to better understand the traffic flow of different streets in Manhattan, one of the five boroughs in NYC. This transformation of the data was based upon the fact that the streets in Manhattan are aligned in a grid structure, and these researchers believe that the transformation would lead to better predictions in their Random Forest model[Ram16]. This rotational transformation was considered in the development of the predictive models in discussed in the Models section to see if it would lead to better results for trips all across NYC rather than just trips in Manhattan. The researchers suggest using a rotation angle of 36.1°, but this was only an estimate based on the data used in their research. A rotation angle of $\phi = 29°$ was used in the predictive models considered in this paper since this is the true angle that the Manhattan street grid is rotated from the north south axis [Rob06] The coordinates of given information are rotated in the following equations, where x is longitude and y is latitude.

$$x_{\text{rot}} = x\cos(\phi) - y\sin(\phi) \tag{1}$$

$$y_{\text{rot}} = x\sin(\phi) + y\cos(\phi) \tag{2}$$

3 Dataset

The dataset originally comes from the NYC Taxi and Limousine and it provides information about yellow taxi rides in New York City from 2010-2015. The data's features include: date/time of trip, pick-up/drop-off latitude, pick-up/drop-off longitude, and passenger count. Since the date/time of trips were represented into a single feature, this feature was separated into year, month, weekday, hour, and minute to facilitate analysis. The original dataset contained about 55 million rows, but only 300,000 data points were used due to

the limited resources of computation available. The data was then divided into a training set that consisted of 80% of the data and a validation set that consisted of 20% of the data.

While these features are useful for developing a predictive model, it was important to engineer new features based on the current ones available. The first feature added into the dataset was the estimated distance of the trip, but this was done using two different estimates.

The first estimate is the Haversine distance which calculates the distance between two points on a sphere:

$$d = \sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + \cos(\varphi_1)\cos(\varphi_2)\sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) \tag{3}$$

$$H(d) = 2R \cdot \tan^{-1}\left(\frac{\sqrt{d}}{\sqrt{1-d}}\right) \tag{4}$$

where φ_1, φ_2 are the pick-up and drop-off latitude coordinates, ϕ_1, ϕ_2 are the pick-up and drop-off longitude coordinates, and R is the radius of the Earth (3,959 miles).

The second estimate of the distance was the "Manhattan" distance also known as the taxicab metric. Using the Haversine function with "taxicab geometry," it calculates the absolute value of the sum of the vertical and horizontal distances between the two coordinates. The Manhattan distance may be more representative of an actual trip in NYC due to the grid-like layout of its streets.

Features based on trips that were going to or from the JFK and Newark Airports were also created because the NYC Taxi and Limousine Commission uses certain ways of calculating fares to these two destinations. The following information was found on the NYC Taxi and Limousine Commissions website [Com]:

- A metered fare is used for trips to/from the LaGuardia Airport.
- Trips to or from JFK and Manhattan have a flat fare of $52 plus tolls, 50-cent MTA State Surcharge, 30-cent Improvement Surcharge, and $4.50 rush hour surcharge. Rush hour is considered to be from 4PM to 8PM on weekdays.
- A metered fare is used for trips from JFK to other NYC destinations.
- There is a $17.50 Newark Surcharge along with a 30-cent Improvement Surcharge for trips to the Newark Airport

Using this information, features were created that indicate if a trip is to or from JFK Airport and Manhattan, and if a trip is to or from Newark since there is a flat fare for these trips. Also, another feature that indicates if a trip is to or from JFK Airport and Manhattan during rush hour was created to account for surcharges added on to these trips. The features were engineered by calculating a 2-mile Haversine radius within the latitude/longitude coordinates of the airports and Manhattan. An exploration of these different features was conducted to determine which ones can be used to predict fare amount.

4 Exploratory Analysis of the Data

From a preliminary analysis of the data, there appeared to be inaccuracies in the recording of trips. For some cases in the data, the number of passengers exceeded the capacity of a taxi. Also, some trips had fare amounts lower than the base fare of $2.50, so these data points were removed. Trips with unreasonable estimated distances were found as well and were removed in order to make sure that the predictive model was being trained on accurate data.

After feature engineering and data cleaning, the intuitions about the features relating to fare amount were checked to make sure there was justification for adding them into the model. In this section, the data is explored in order to both understand how the trips are distributed across New York and to determine which features should be included in the models developed to predict taxi fare. Provided below in Fig. 1, the pick-up and drop-off locations for all trips were plotted on a map of NYC in order to determine the distribution of where these taxi trips took place. The inspiration for these plots for exploratory analysis can be attributed to Albert Breemen, who created heat maps based on the pick-up and drop-off points [VB17].

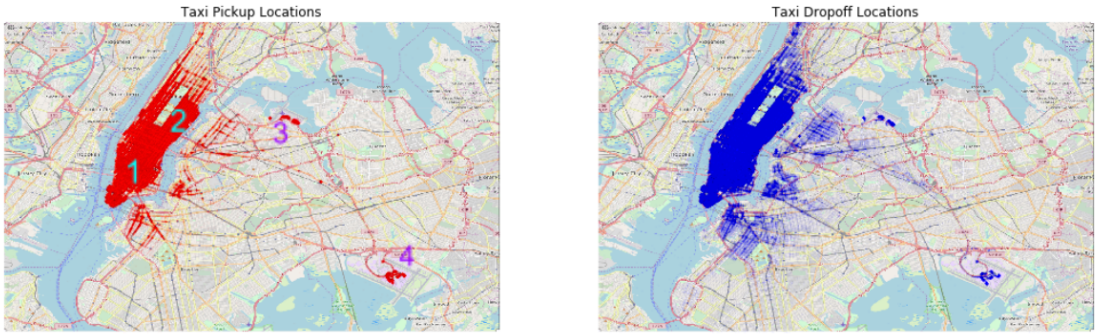

Figure 1: Visualization of Manhattan Map

The two maps display both the pick-up and drop-off locations of the main areas where trips took place. There appears to be a noticeable amount of trips that started and ended at LaGuardia Airport (3) and JFK Airport (4), which suggests adding features relating to airports would be beneficial to developing an accurate model. Despite trips to LaGuardia being a metered far, the amount of trips to and from this airport suggests that airport trips using yellow taxis are common. The trips to Newark Airport are not provided in the map since the airport is far from the general NYC area, but there were a substantial amount of these trips found within the data set. Also when analyzing the trips, it was found that most of them appear to take place in both the Manhattan and Bronx boroughs, which are represented by 1 and 2, respectively, in the left map. As explained in the Related Works section, the streets in the Manhattan borough are laid out in a grid-like structure, which means that it might be favorable to consider using the Manhattan distance formula rather than the direct Haversine distance, since most trips occur in this area. The use of both the Manhattan and the Haversine distances is explored within the Models section [CA16].

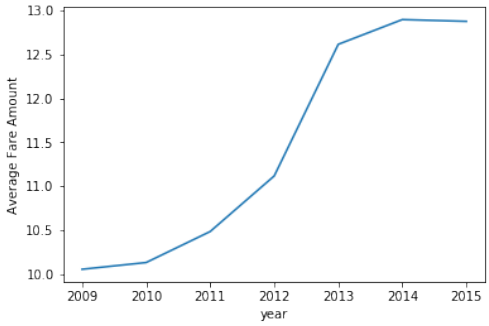

Figure 2: Increment of Average Fare from 2009 to 2015

When considering what measures of time (year, month, hour, and minute) to include as features, it was found that only two of measures seemed to have an apparent relationship with the fare: year and hour. As shown in Fig. 2, the average fare amount started at about $10.00 in 2009 and continued to increase slightly each year with the largest increase occurring between 2012 and 2013. Because of this increase, the use of year as a feature in our model to predict taxi fare seems appropriate.

As shown in the right plot in Fig. 3, there are particular hours when the average fare amount increases, which may be due to heavier traffic that cause trips to take longer than at other times. The fare amount dramatically spikes between 4AM and 6AM in UTC time (11PM to 1AM EST), and it also increases slightly from 1PM to 6PM (6AM to 1PM EST). The latter increase could be attributed to morning rush hour traffic, whereas the former is most likely due to the surcharge added onto fare after midnight.

Distance is clearly the most important feature in predicting the fare amount as it is the main metric taxis use to calculate the fare at the end of the ride. The two plots in Fig. 3 show the strong correlation between the fare amount and distance traveled. The plot of the mean distance traveled, calculated using the straight Haversine distance, regressed on the hour of day has an almost identical shape to the mean fare amount regressed on hour of the day. So, the intuition to prioritize distance traveled in the model was confirmed from the data exploration.

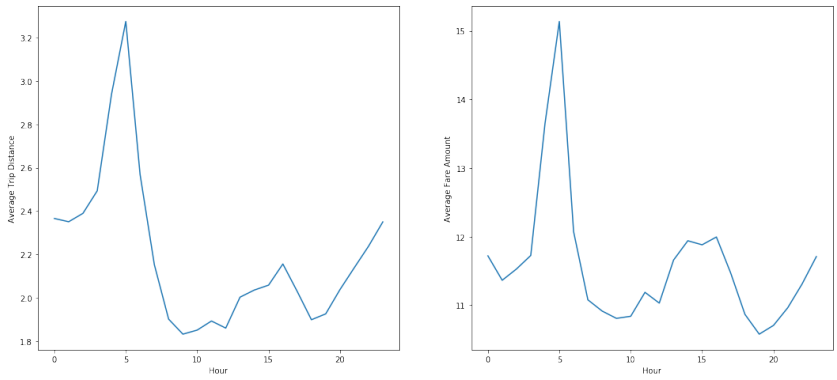

Figure 3: Visualization of Relationship between distance and fare

5 Models

As a means of comparing models, the root mean square error (RMSE) was the main performance metric:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \tag{5}$$

Linear Regression While the Kaggle competition provided a baseline model with an RMSE of \$5-\$8, it was believed that this could substantially be improved upon by developing a Linear Regression model that considered the features that were found to be related to fare amount in the Exploratory Analysis section. Linear Regression attempts to model the response, which in this case is fare amount, as a linear combination of the features:

$$Y = \beta_0 + \beta_1 x_1 + ... + \beta_d x_d + \epsilon \tag{6}$$

The estimated coefficients are calculated using a method known as Ordinary Least Squares. Using a Linear Regression model to make predictions, the RMSE was calculated to be \$4.41713 for the validation set, which is a slight improvement from the baseline provided by Kaggle. This was to be expected since more features were used in this model than in Kaggle's baseline regression model that only had one feature. One of the problems with using regression for predicting taxi fare is that it is highly unlikely that the fare can be calculated as a linear combination of some of the features, such as pick-up/drop-off latitude and longitude. Next, Random Forest Regression models were developed to both help make better predictions and also determine what features to use relating to pick-up and drop-off location along with the trip distance.

Random Forest Regression Random Forest Regression is a machine learning model that makes use of a technique known as bootstrapping. Multiple bootstrap samples are generated and only a certain number of features are included in these samples. These samples are used to train regression trees that are used to make predictions based on the data, and these predictions are averaged together in order to make a more accurate prediction. Since each tree is only trained using a certain amount of features, the correlation between these trees is reduced, which leads to a reduction in variance. Random Forest Regression was used to predict fare amount to account for the aspects of the pick-up and drop-off locations along with the time of day in which the trips occurred. There is often more traffic at certain hours of the day in different areas in NYC, and the Random Forest Model was thought to be able to account for these factors better than a Linear Regression model.

There were four Random Forest Regression models that were developed in total to compare the use of both of rotated coordinates and the use of the Haversine and Manhattan distances as features in the models. Each model considered different combinations of the coordinates and distances to determine the optimal combination. When these models were

developed, there were three hyper parameters that were tuned: the number of trees created, the number of features in each bootstrap sample, and the maximum depth of the trees. The two plots provided in Figs. 4 and 5 were used in order to determine the hyper parameters. The maximum depth of the tree that provided the lowest RMSE was found to be 13, the number of features per bootstrap sample was found to be half the number of total predictors, and the number of trees grown in the random forest that provided the lowest RMSE without significant changes in the RMSE was found to be 64. The maximum depth of the trees allows for the number of splits in the trees to be limited since regression trees are binary trees.

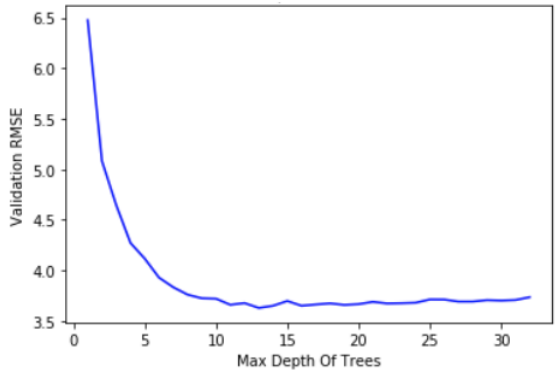

Figure 4: The Effect of Depth on validation RMSE

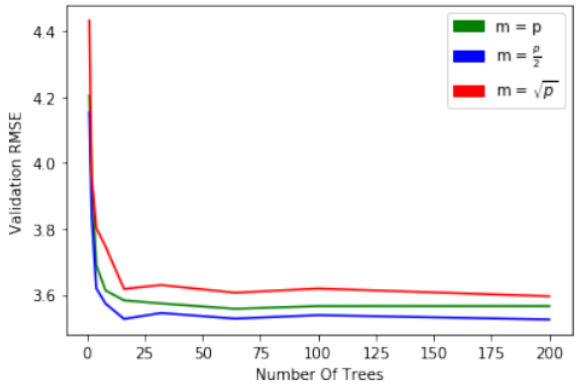

Figure 5: Maximum Number of Features in B-Strap Sample effects on validation RMSE

Using these hyper parameters in the Random Forest Regression models, four models were created. Provided below are the validation set RMSE from each of these Random Forest models:

Random Forest Model	RMSE
Haversine Distance / Original Coordinates	3.40555
Haversine Distance / Rotated Coordinates	3.48041
Manhattan Distance / Original Coordinates	3.58180
Manhattan Distance / Rotated Coordinates	3.58991

The combination of Haversine distance with the original coordinates provided the most accurate predictions for the validation set. This suggests that when developing a model that uses an ensemble of regression trees in order to make predictions, this combination of features should be used in order to make accurate predictions.

Gradient Boosting The last model created in order to make more accurate predictions of the taxi fare amount was developed using Gradient Boosting. Gradient Boosting works by constructing a forward stage-wise additive model through the use of gradient descent. Gradient descent is an optimization algorithm that is used to the find the minimum of an objective function. Unlike in Random Forest Regression where multiple regression trees are

created independently of each other, Gradient Boosting develops regression trees sequentially in order to allow for each model to correct for errors made by the previous tree. Since this algorithm uses gradient descent, it was important to determine the learning rate that provides the most accurate predictions. The learning rate that achieves the lowest RMSE was 0.1, as shown in Fig. 6. It is worth noting that the plot has been scaled to only include learning rates at or above 0.1 to provide a better visualization. When training the gradient boosting model, an RMSE of $3.40027, which is a slight improvement from the Random Forest model.

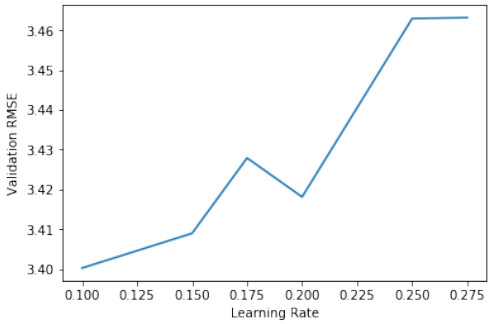

Figure 6: The Effect of the Learning Rate on RMSE

6 Conclusion and Future Works

Using Linear Regression, Random Forest, and Gradient Boosting models, the major task was to solve the problem of determining taxi fare in NYC based on information that is available prior to a trip. The rotational transformation of the pick-up/drop-off latitude and longitude coordinates was found to provide better predictions with the Random Forest model. Thus, the application of this same linear transformation with the Gradient Boosting model performed slightly better than Random Forest. The lowest RMSE obtained was $3.40027 which is about a 32% decrease of the $5 RMSE from the baseline approach. These results are relatively fruitful for investigating this type of machine learning task, but how to improve the performance of these models should be considered.

Intelligently crafting features to address certain surcharges and fixed fares for certain taxi trips is one possible improvement. The NYC Taxi and Limousine Commission provides these types of charges for trips to Westchester or Nassau County, according to their website. And, of course, the models could be improved by training a larger data set, though that would require more computational power than was available for this project. The most important problem is engineering the right metric for the distance between pick-up and drop-off locations. The exploratory analysis showed the strong correlation between any distance metric was investigated – whether the Haversine or Manhattan – and the hour of day and taxi fare. And so, this should be the focus of similar projects in the future.

References

[CA16] Antoine Foba Amon Jr. Christophoros Antoniades, Delara Fadavi. Fare and duration prediction: A study of new york city taxi rides. Available at http://cs229.stanford.edu/proj2016/report/AntoniadesFadaviFobaAmonJunior-NewYorkCityCabPricing-report.pdf, 2016.

[Com] NYC Taxi Limousine Commission. Taxicab rate of fare. http://www.nyc.gov/html/tlc/html/passenger/taxicabrate.shtml.

[Ram16] Aiswarya Ramachandran. Machine learning to predict taxi fare-part two: Predictive modelling. https://medium.com/analytics-vidhya/machine-learning-to-predict-taxi-fare-part-two-predictive-modelling-f80461a8072e, 2016.

[Rob06] Sam Roberts. City of angles. www.nytimes.com/2006/07/02/nyregion/thecity/02-gird.html, 2006.

[VB17] Albert Van Breemen. New York City taxi fare prediction playground competition [Kaggle kernel notebook]. https://www.kaggle.com/breemen/nyc-taxi-fare-data-exploration, 2017.