Abstract:

This report presents an evaluation study of various models and features to predict the duration and fare of a taxi trip in New York City. We investigate the optimal model and hyper parameters for this problem. We show that Decision Trees based models, Gradient Boosting are effective in learning the predictive patterns from the given features. We examine the effective ways to represent the features and also study the importance of various features in our predictive models. I achieved lowest RMSE for duration prediction using the extra Gradient Boosting model and for the fare prediction problem using the Ensemble of Random Forest and Gradient Boosting model. Looking at results, we can claim that though the model was not completely successful in predicting the exact duration or fare, it can still be used as an effective tool to give an approximate estimate

1. Introduction:

In this analysis, we try to address some of the long standing problems in the transportation industry i.e. to predict the duration and fare at the beginning of the trip. With the advent of technology based cab services, this problem has become particularly important for the drivers and the customers. A good prediction mechanism can be instrumental for drivers in optimizing their returns, while also saving the customers from the uncertainties attached to a trip. In our analysis, we attempt to examine and understand the provided features to create a prediction model for this problem. We study the impact of various features and also attempt to find the best possible ways to leverage those features. We review the performance of various models and discuss the effectiveness and shortcomings for these models. In this report, we discuss four models: Random Forest, Gradient Boosting and an ensemble of Random Forest & Gradient Boosting. We evaluate these models based on the Root Mean Square Error (RMSE). We also discuss the importance of various features in our prediction algorithms.

2. Attribute Information:

We find:

- *vendor_id* only takes the values 1 or 2, presumably to differentiate two taxi companies. Vendor 1 has all of the trips beyond 24 hours, whereas vendor 2 has all of the (five) trips with more than six passengers and many more trips that approach the 24-hour limit.
- *pickup_datetime* and (in the training set) *dropoff_datetime* are combinations of date and time that we will have to re-format into a more useful shape
- *passenger_count* takes a median of 1 and a maximum of 9 in both data sets
- The *pickup/dropoff_longitute/latitute* describes the geographical coordinates where the meter was activate/deactivated.
- *store_and_fwd_flag* is a flag that indicates whether the trip data was sent immediately to the vendor ("N") or held in the memory of the taxi because there was no connection to the server ("Y"). Maybe there could be a correlation with certain geographical areas with bad reception?
- *trip_duration:* our target feature in the training data is measured in seconds.

3. Data Preprocessing:

TRIP_DURATION

#trip duration consists of min=1sec and maximum=3526282(980 hours).no one can travel 980 hours with taxi in a city the bill can be more and no one can travel in 1 sec so it consists of outliers in the target variable we can also assume that some people booked the taxi, after arriving the taxi there they cancelled the taxi trip from this we can get trip duration of 1 sec to 10 sec we can also remove that rows

#values are in seconds we can converte in to minutes to compete easily

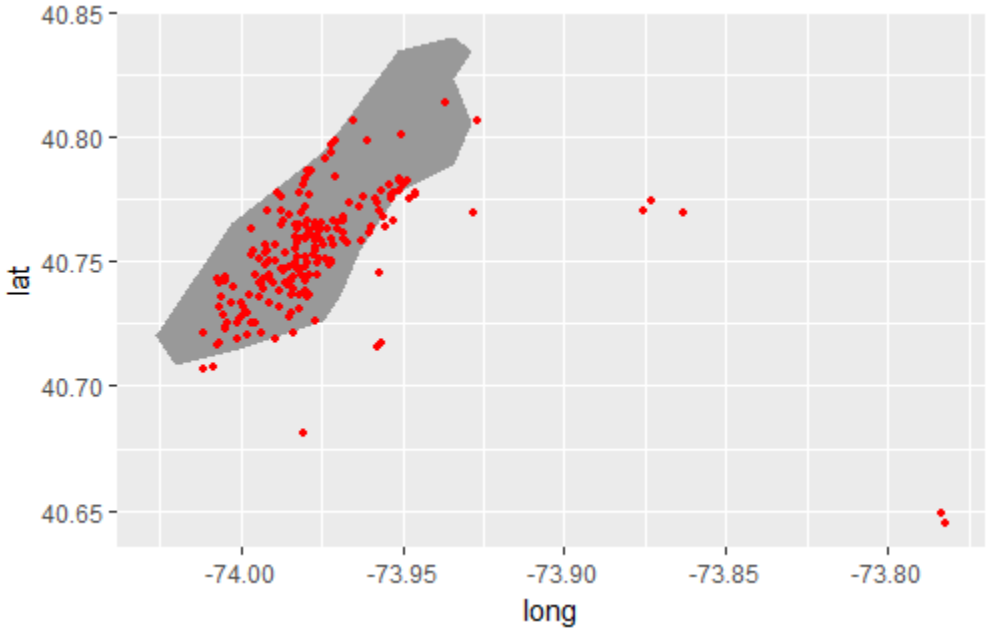
LONGITUDE AND LATITUDE

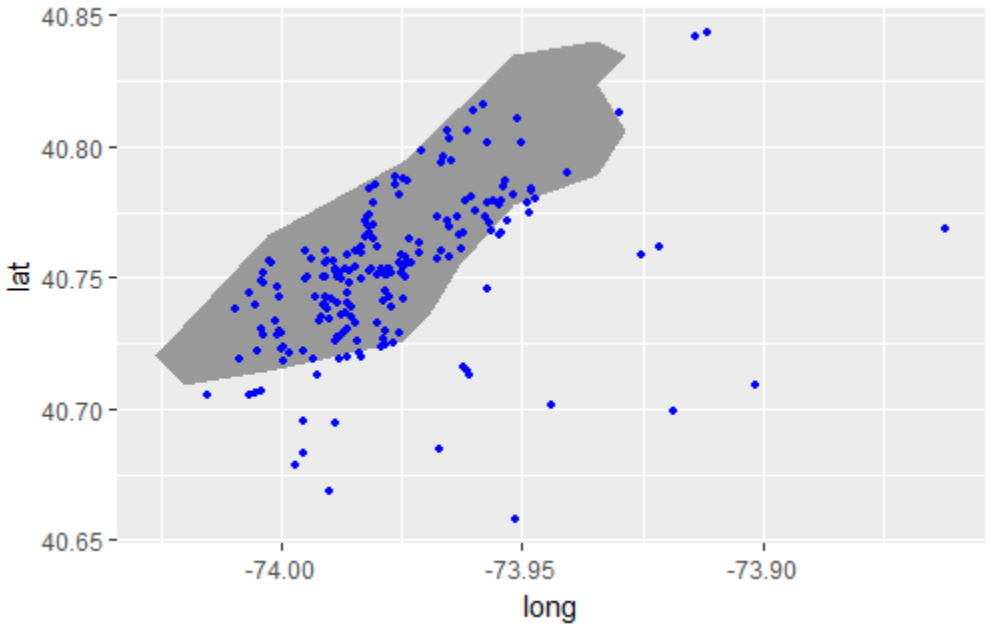
The coordinates of the NYC city is city_long_border = (-74.03, -73.75)

city_lat_border = (40.63, 40.85) the taxi cannot be drived out of the city so some coordinates are falling out- of these boarders so i removed those coordinates fall out of these boarders

###visualization for longitude and latitude

There is another insight here which is rather intuitive: trips to or from any of the airports (most prominently JFK) are unlikely to be very short. Thus, the a close distance of either pickup or dropoff to the airport could be a valuable predictor for longer trip_duration.after removing the outliers the pickupand dropoff are in same region



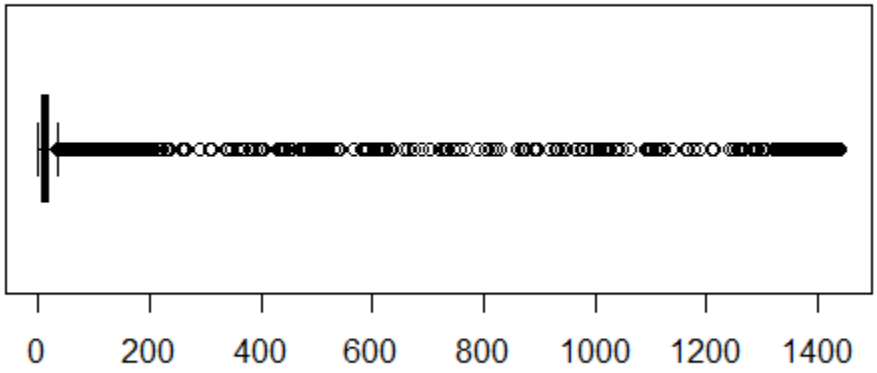


conversions:

- pickup_datetime should be in date format so by using ymd hms function I converted in to date format
- vendor_id should be factor

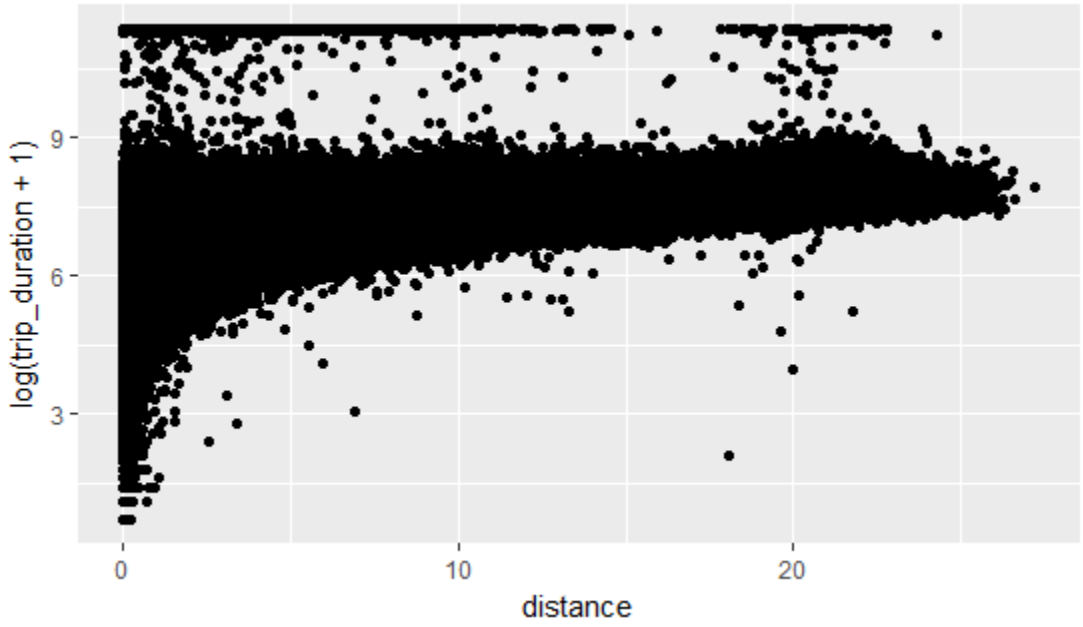
Dealing with missing values:

- While converting the pickup_datetime one value is missing in the data so I removed the data because we can't predict the date
- concatinating the dataset for feature engineering
- although it will be biased while doing any pre processing



feature engtneering:

- from summary of distance min=0 and max=33 km
- By using latitude and longitude we can calculate distance in kilometers
- a positive correlation between trip distance and duration. An interesting finding is that the variance of duration increases, as the trip distance increases



Holidays:

- from the sequence of the date we can predict the holidays present in the given datetime ,trip_duration can be less when compared to business days

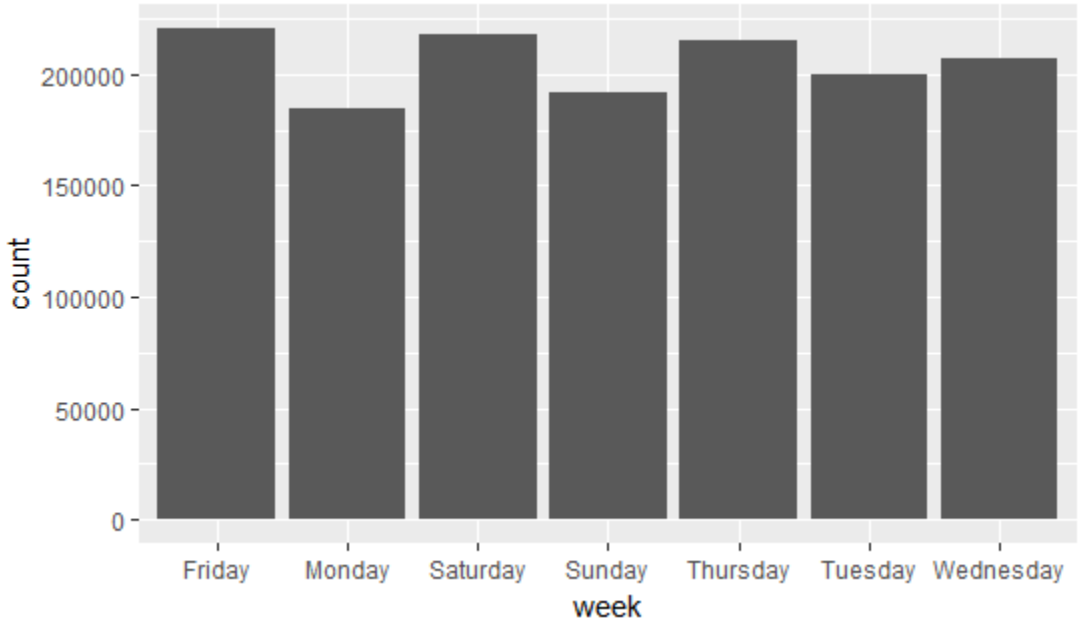
kmeans_minute:

- from datetime variable time is converted in to minutes to know at which minute taxi pickup is happened
- by doing cluster peak rides will fall in one cluster

Week:

- The day of the week can have a significant impact on the predictions because we expect the weekdays to be more congested than the weekends especially during the

day time. we can clearly see that the average pickup is higher on the weekdays than the weekends

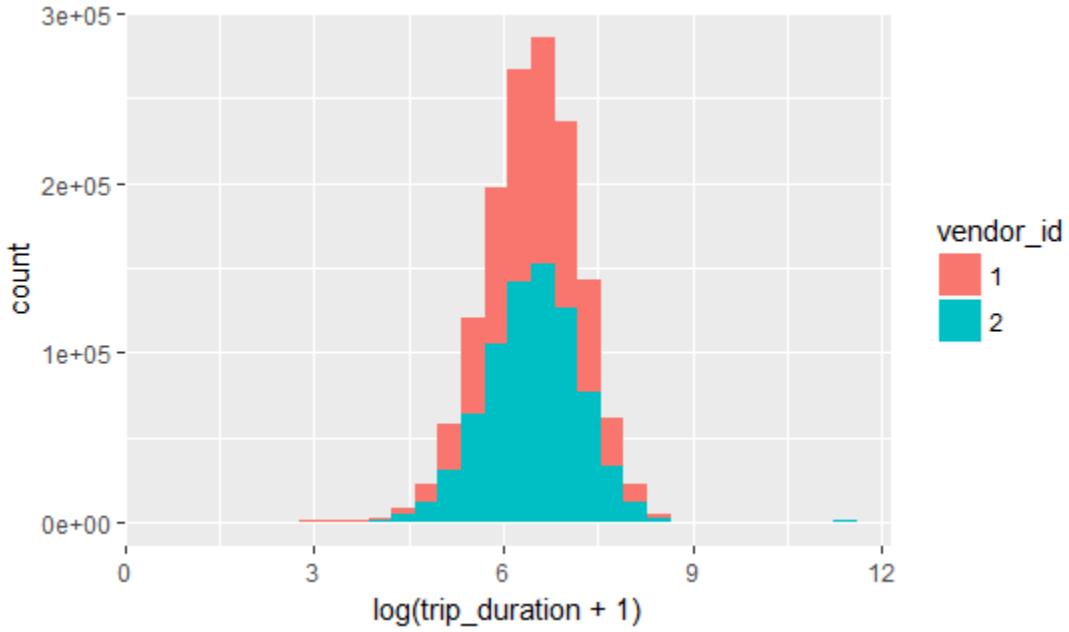


- Checking the seasonality for week and month and hour using cos and sin function

#visualization for vendor id

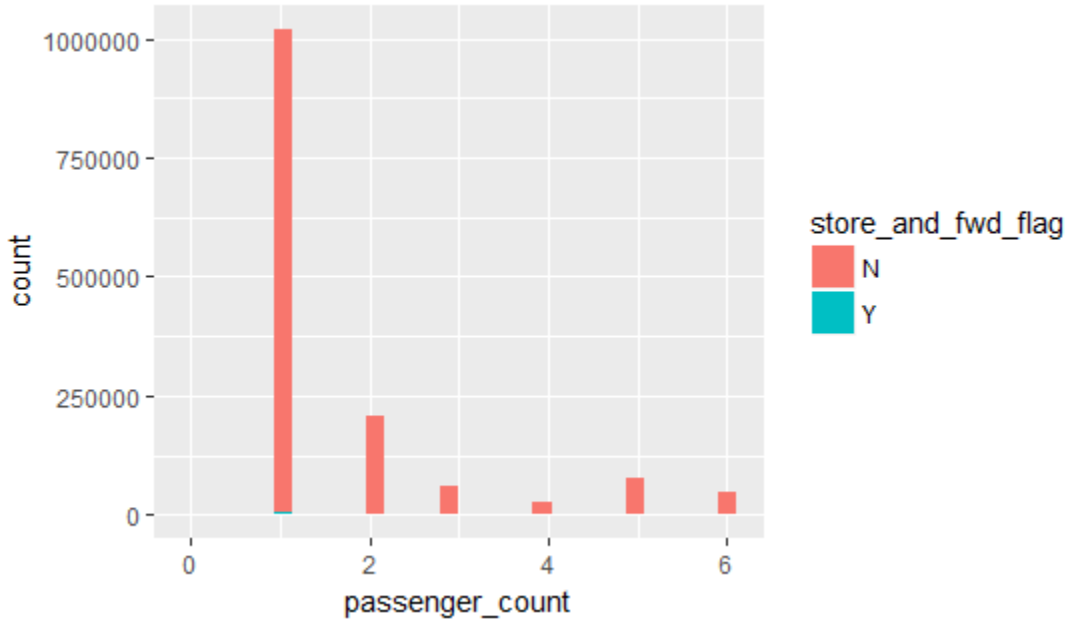
- vendor_id is in integer for so it should convert in to factor
- trip_duration is not normally distributed so log transformation is used on trip_duration
- trip_durations of about a minute might still be somehow possible, assuming that someone got into a taxi but then changed their mind before the taxi could move. Whether that should count as a "trip" is a different question.
- But trip durations of about 15 minutes (900 s) without any distance covered seem hardly possible. Unless they involve traffic jams

- It is also noteworthy that most trips in the less-than-a-minute-group were from vendor 1, whereas the 10-minute-group predominantly consists of vendor 2 taxis.

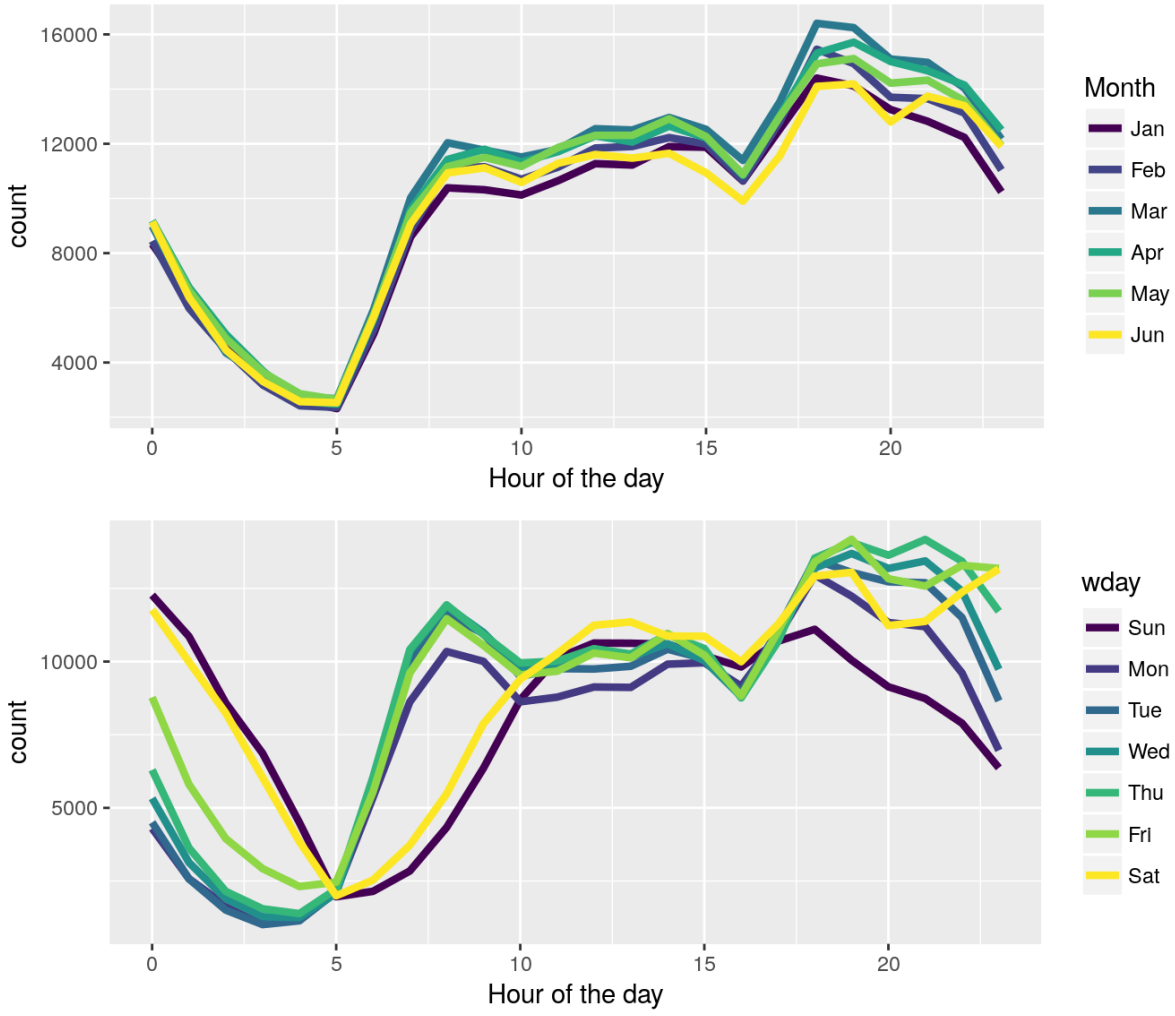


- *store_and_fwd_flag* is a flag that indicates whether the trip data was sent immediately to the vendor ("N") or held in the memory of the taxi because there was no connection to the server ("Y"). Maybe there could be a correlation with certain geographical areas with bad reception?

Visualization



The trip volume per hour of the day depends somewhat on the month and strongly on the day of the week:



Random Forests:

Random forest is a bagging technique which builds decision trees. Random forest is like bootstrapping algorithm with Decision tree (CART) model. Random forest tries to build multiple CART model with different sample and different initial variables. Here the hyper parameters tuned are ntree(# of trees) and mtry(# of random samples to take for each tree).

XGBOOST:

Xgboost is an ensamble technique which reduces variance and bias ,and herevhyper parameters tuned are nrounds,ntreads,maximum depth.

- Xgboost and randomforest can give feature selection plot