PROJECT REPORT

November 20, 2017

REPORT

1. Project Background and Description

New York City taxi rides paint a vibrant picture of life in the city. The millions of rides taken each month can provide insight into traffic patterns, road blockage, or large-scale events that attract many New Yorkers. With ridesharing apps gaining popularity, it is increasingly important for taxi companies to provide visibility to the ride duration, since the competing apps provide these metrics upfront. Predicting the duration of a ride can help passengers decide when is the optimal time to start their commute, or help drivers decide which of two potential rides will be more profitable. In order to predict duration data used includes pickup and drop-off coordinates, trip distance, start time, number of passengers, vendor id etc. Supervised learning algorithms such as linear regression, decision tree were used to predict duration of the travel time. The evaluation metric used for this model is root mean squared logarithmic error which is usually used when you don't want to penalize huge differences in the predicted and true values when both predicted and true values are huge numbers.

2. Literature Survey

The dataset has been taken from kaggle website [3]. Other works based on the same dataset that we used are as follows:

[1] Uses Linear Regression and Random Forest Regressor to predict the trip duration for the cab travelling in New York City.

[2] Uses one month of NYC taxi rides data. The model infers the possible paths for each trip and then estimates the link travel times by minimizing least squared difference between expected path travel times and the observed path travel times. In addition, they used the weather data of that particular season and what we observed was that even in winters when city witnesses heavy snowfall, the trip duration is not affected.

3. Dataset

The dataset and inputs were obtained from kaggle competition. The dataset has about 1.4 million records.

The following variables are given in the dataset:-

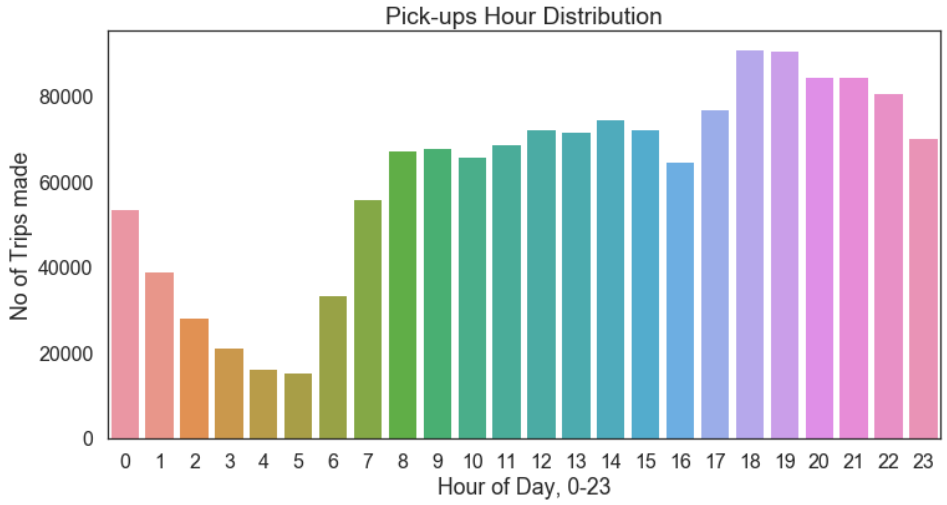
- id – a unique identifier for each trip.
- vendor_id – a code indicating the provider associated with the trip record.
- Pickup_datetime – date and time when the meter was engaged.
- dropoff_datetime – date and time when the meter was disengaged.
- passenger_count – the number of passengers in the vehicle (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.
- store_and_fwd_flag – this flag indicates whether the trip record was held in vehicle memory before sending to the vendor because the vehicle did not have a connection to the server - Y=store and forward; N=not a store and forward trip.
- trip_duration – duration of the trip in seconds.

The variable pickup and drop-off date time is used in a very efficient manner. With the help of pandas library we have extracted the weekday (Sunday=0, Monday=1 etc.), hours (24 hour clock) and month (January=0, February=1 etc.). We have the pickup and drop-off coordinates and from that we have calculated the radial distance using the harversian formulae.

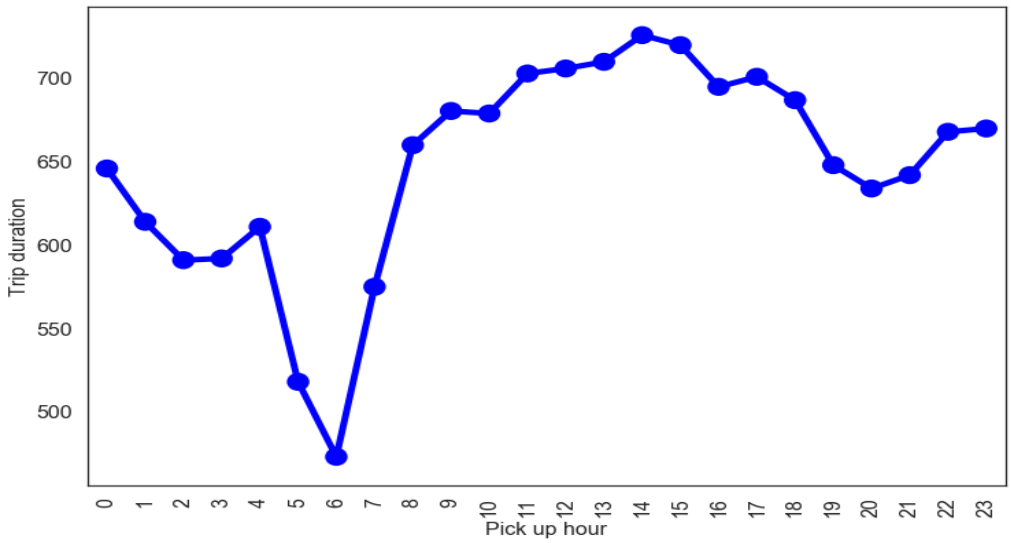
4. Exploratory Analysis

To have insight of the data we have performed some exploratory analysis on the data. The main objective of this is to see how different features behave while predicting the trip duration.

i. Time of the day

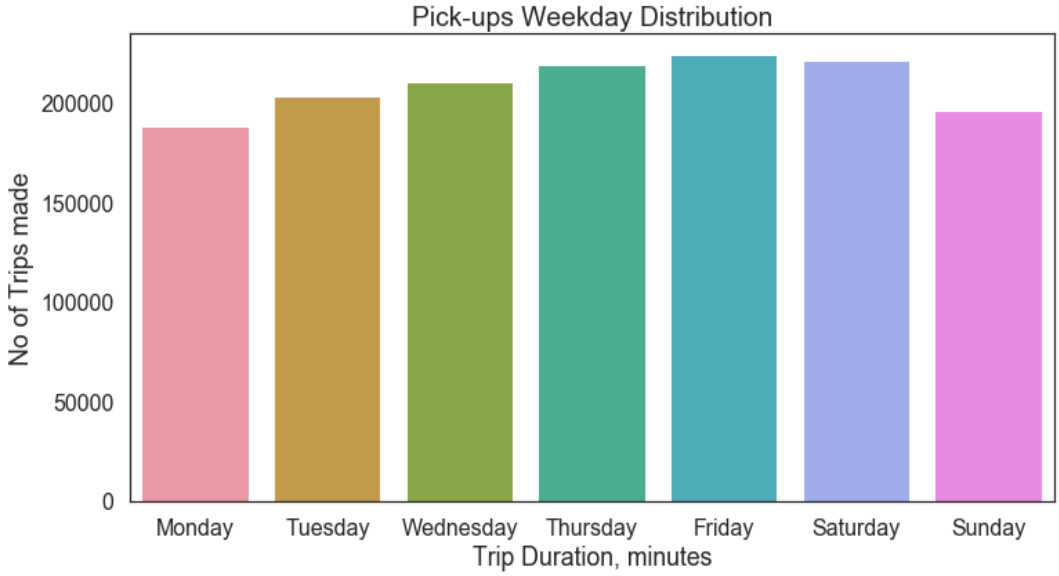


We see that how the no. of trips made vary during the day. It's very much obvious that during peak hours the no. of trips made is large when compared to non-peak hours especially during the midnight when no. of trips made decreases to a great extent.

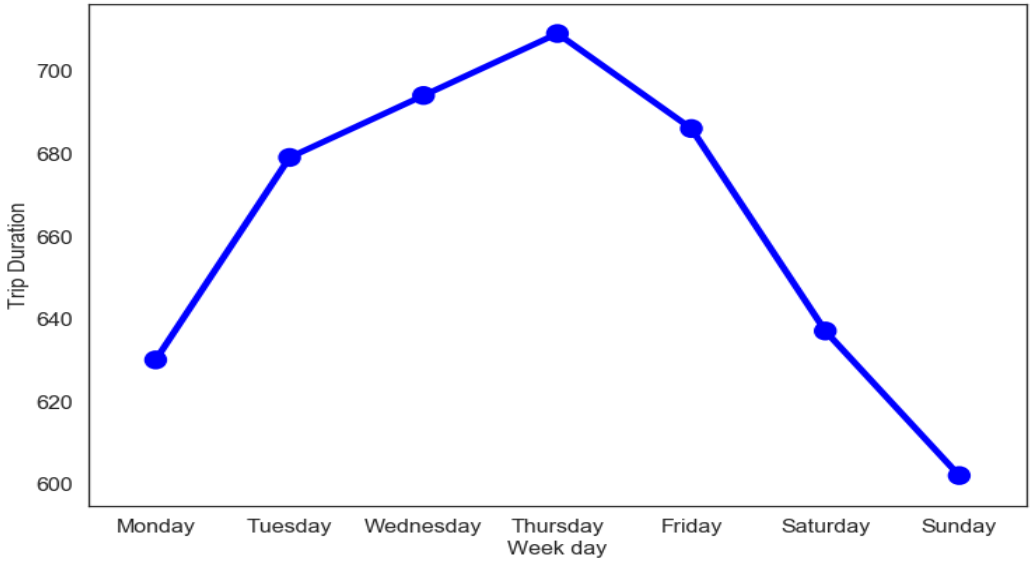


The above plot shows how the trip duration vary during the day. During peak hours we see that trip duration is time taking which suggests us that during peak hours, traffic on the road decreases the speed of the cab and hence the trip duration increases. While on the other hand during non-peak hours especially during midnight the trip duration is very less which is due to less traffic on the road.

ii. Day of the week



We see how the no. of trips made vary during the week. During weekends the no of trips made drops especially on Sunday, the no of rides made is very less when compared to other days of the week.



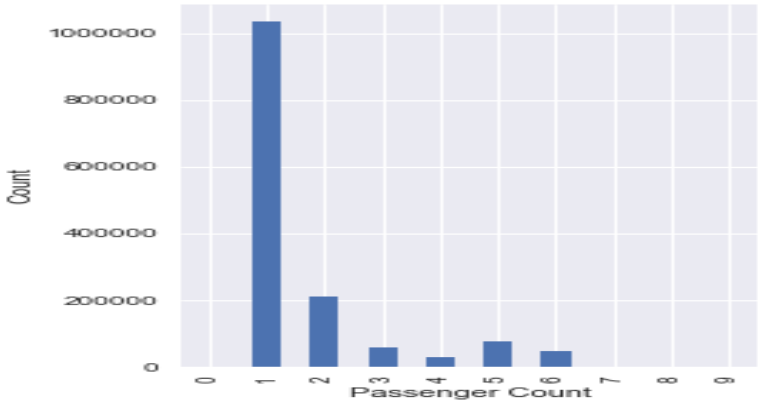
The above plot shows us how the trip duration vary across the week. During weekends the trip duration becomes very less which may be because the traffic on the road drops during weekends as people do not commute to their work during weekends so traffic on road drops which decreases the trip duration.

5. Methodology

We are predicting the duration of a NYC Taxi ride based on certain features. Since predicting the duration directly could lead to poor accuracy. So we are removing certain outliers from the data.

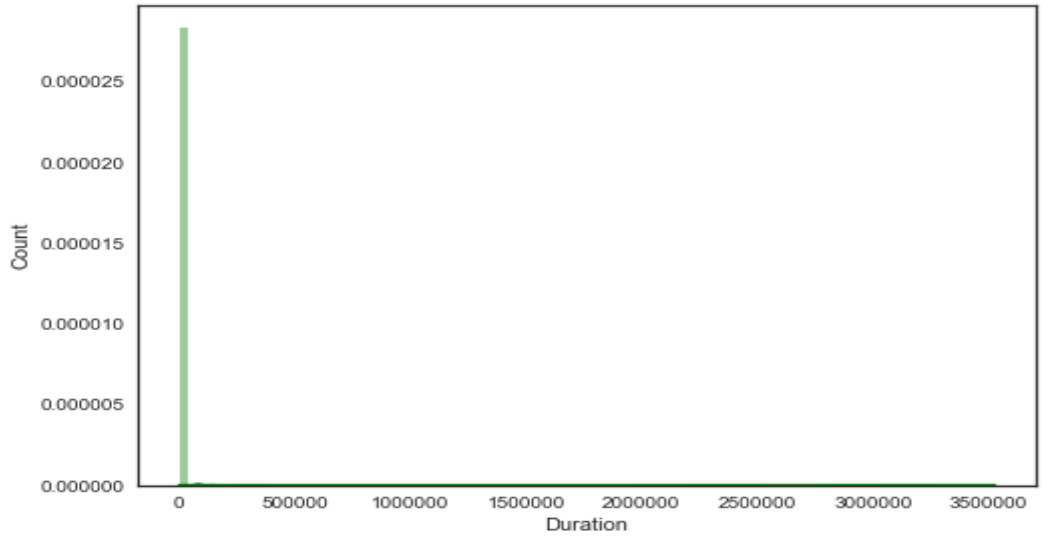
1. Exploratory analysis

i. Passenger Count



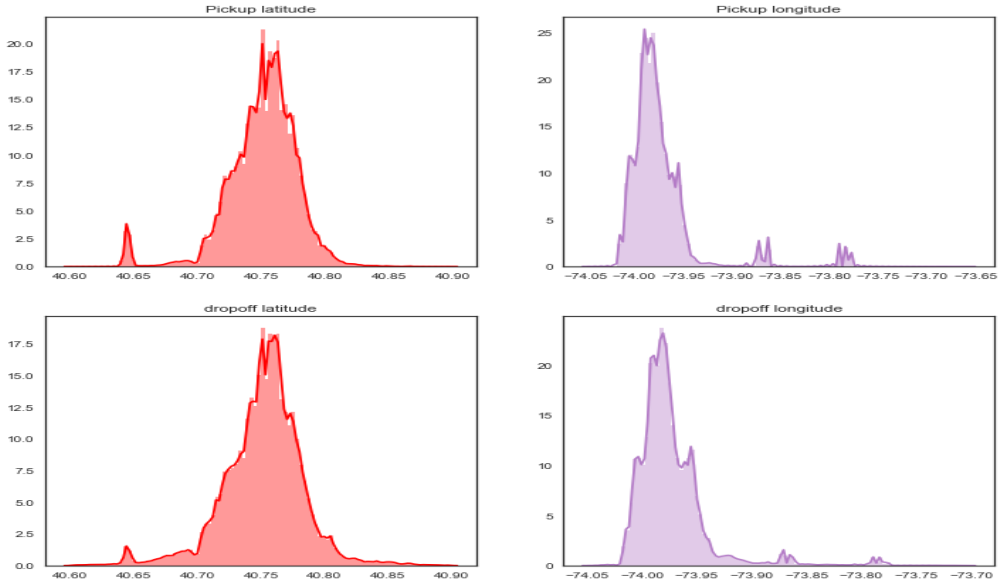
The above plot shows us the passenger count of each trip. Notice that apparently there were some rides with passenger count as 0. The main reason for this maybe that to complete the minimum ride for the day and to get incentives, the drivers do some fake rides. It is very important to remove such to get proper accuracy.

ii. Trip Duration



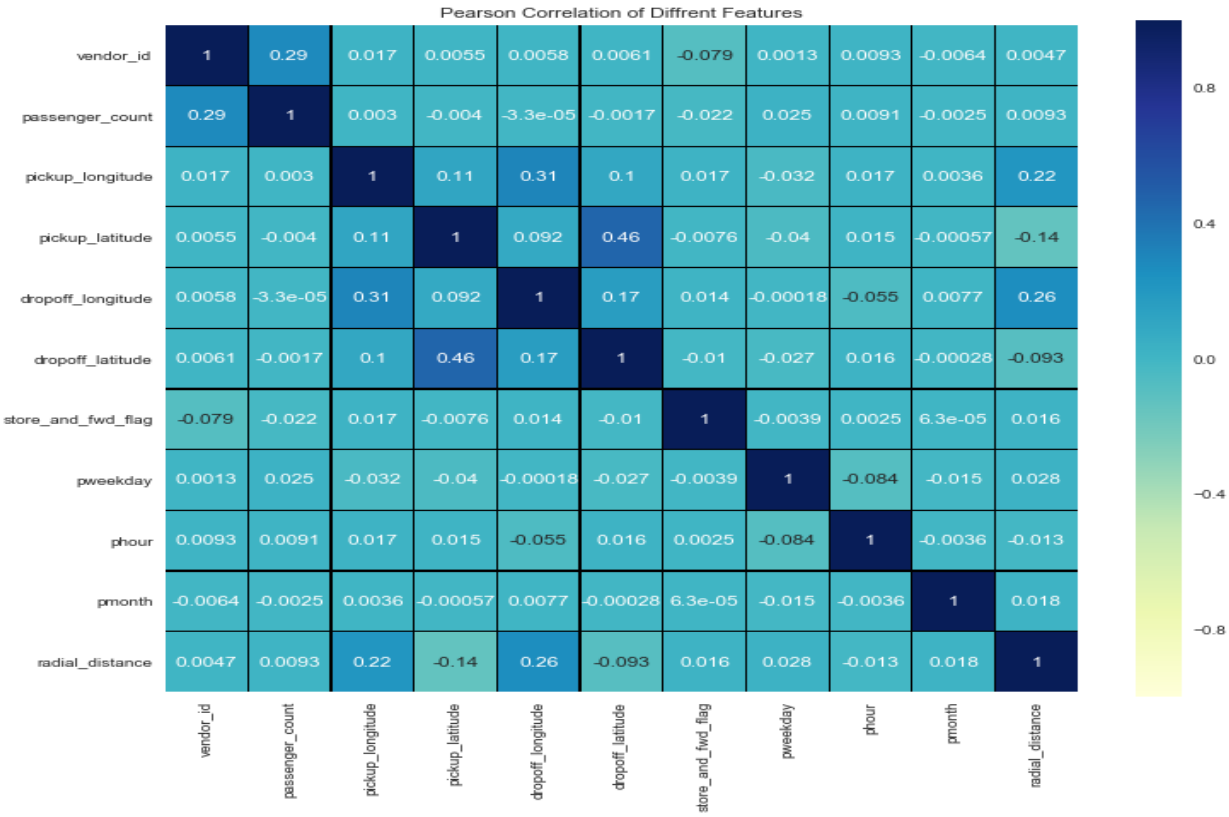
One might also be interested to view the number of trips made vary over time, since this could reveal not only apparent seasonality in the data and certain trends, but could point out any significant outliers. The above plot shows us the trip duration for each trip. We see that minimum trip duration is 1 second and maximum trip duration is around 35, 26,282 seconds which comes to around 979.5 hours so to have less noise in our data, trip duration has been capped to around 67 minutes.

iii. Pickup and Drop-off Coordinates



The above plot shows us the range of coordinates which is mostly used by passengers in the New York City so it is important that we remove such coordinates which is rarely used. This also suggests that these rides were taken in the outskirts of the city.

iv. Correlation



The above plot show us the correlation between the features. It suggests that correlation exists within the latitudes and longitudes and the distance which is quite obvious. We see that vendor id and passenger count are correlated as vendor id 2 is carrying 5 or 6 passengers most of the time.

2. Metrics

Root mean square logarithmic error (RMSLE) was used as a metric to measure our model's accuracy. It had the advantage of being convex and physically interpretable (it has the same unit as time for duration prediction). It is usually used when you don't want to penalize huge differences in the predicted and true values when both predicted and true values are huge numbers.

3. Predictive Model Prediction Results

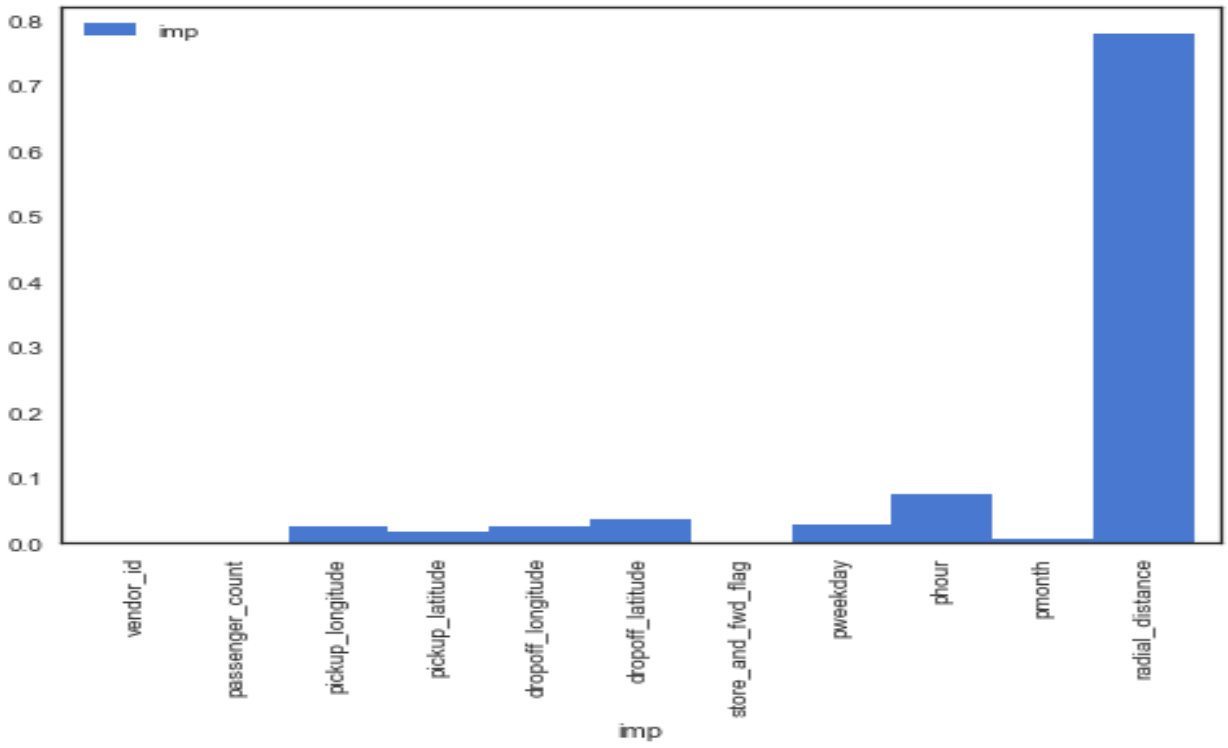
i. K Nearest Neighbor (KNN) Model

This is the first approach we tried. The target is predicted by local interpolation of the targets associated of the nearest neighbors in the training set. The parameters used here is n_neighbor which is the number of neighbor to use by default for kneighbor queries.This model performed better than we would we would expect with rmsle score of 0.452 which is very good.

ii. Decision Tree Model

This is the second approach we tried. **Decision tree learning** uses a **decision tree** (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). To avoid overfitting we used several parameters such as min_samples_split which is the minimum no. of sample which are required in a node to be considered for splitting. The rmsle score of 0.409 was obtained which is a clear improvement of the previous score of KNN model.

4. Feature Importance



The above plot shows which features have the greatest effect on trip duration. It would make logical sense that distance has the greatest affect. The further you travel, the longer it'll take. The rest of the features follow a similar logic in why it's ranked the way it is.

6. Conclusion

Considering what is and what is not accounted for in the models built in this study, their predicting results are fairly accurate. To further improve the prediction accuracy,

more variabilities need to be considered and modeled. We could also use the weather data for the city and try to increase the prediction accuracy. Also, modeling traffic and the effect of location in between pickup and drop-off points and the difference in driver's speed, will further explain us the dynamics of the city.

7. Reference and citations

[1] Fare and Duration Prediction: A Study of New York City Taxi Rides Christophoros Antoniades, Delara Fadavi, Antoine Foba Amon Jr. December 16, 2016

[2] Visual Exploration of Big Spatio-Temporal Urban Data: A Study of New York City Taxi Trips. Nivan Ferreira; Jorge Poco; Huy T. Vo; Juliana Freire; Cláudio T. Silva.

[3] Source of dataset:
https://www.kaggle.com/c/nyc-taxi-trip-duration

[4] Scikit-learn Cookbook by Trent Hauck