

**ABSTRACT**

The objective of our model is to predict the accurate trip duration of a taxi from one of the pickup locations to another drop-off location. In today’s fast-paced world, where everyone is short of time and is always in a hurry, everyone wants to know the exact duration to reach his/her destination to carry ahead of their plans. So, for their serenity, we already have million dollar startups such as Uber and Ola where we can track our trip duration. As a result of this, we proposed a technique in which every cab service provider can give exact trip duration to their customers taking into consideration the factors such as traffic, time and day of pickup. So, in our methodology, we propose a method to make predictions of trip duration, in which we have used several algorithms, tune the corresponding parameters of the algorithm by analyzing each parameter against RMSE and predict the trip duration. To make our prediction we used Random Forest, Decision Trees and Linear Regression. We improved the accuracy by tuning hyper-parameters and Random Forest gave the best accuracy. We also analyzed several data mining techniques to handle missing data, remove redundancy and resolve data conflicts.

**INTRODUCTION**

There are many possible methods of moving between two given points in a city; however, the taxi trip has found wide applications in urban cities when compared to any other mode of transport. It hence becomes very important

to analyze and predict trip duration between two points in the city when provided with the required set of parameters that affect the trip duration. For a good taxi service and its integration with the existing transportation system the project serves as a right means to comprehend the traffic system in New York City. For prediction purposes factors such as pick up latitude, pick up longitude, drop off latitude, drop off longitude etc. is considered. These geographical locations clubbed with other important factors such as pick up date, pick up time are used for the overall trip duration prediction. The primary focus of this project is in depth analysis of the factors associated with a taxi trip in NYC. The different algorithms used are: Linear Regression, Random Forest and Decision Trees.

1. LINEAR REGRESSION

It is a linear model that establishes the relationship between a dependent variable y (Target) and one or more independent variables denoted X (Inputs). Linear regression has been studied at great length, and there is a lot of literature on how your data must be structured to make best use of the model.

1. DECISION TREES

Decision Tree is a Supervised learning techniquethat can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation. It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.

1. RANDOM FOREST

The random forest approach is a bagging method where deep trees, fitted on bootstrap samples, are combined to produce an output with lower variance.Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of *combining multiple classifiers to solve a complex problem and to improve the performance of the model.* Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

However, random forests also use another trick to make the multiple fitted trees a bit less correlated with each other: when growing each tree, instead of only sampling over the observations in the dataset to generate a bootstrap sample, we also sample over features and keep only a random subset of them to build the tree. Sampling over features has indeed the effect that all trees do not look at the exact same information to make their decisions and, so, it reduces the correlation between the different returned outputs. Thus, the Random forest algorithm combines the concepts of bagging and random feature subspace selection to create more robust models.

**MATERIALS AND METHODOLOGY**

Materials that we have used include: Python software for coding and NYC Taxi Limousine Data. Our methodology involves the use of machine learning techniques such as: Linear Regression, Decision Trees and Random Forest Regression.

1. DATASET

We selected the following features: Trip Distance: Distance is an important factor for predicting the duration of a trip, as Distance = Speed/Time. Day of the week: Weekdays experience slow speed because of the daily routine of schools and offices, hence there is the need for this feature. Time of the day: Peak hours of offices and school start and end such as Morning 8 - 12 and evening 4-7 experience high traffic. Pick up and drop-off cluster: Route being travelled that is from one cluster to another is important to predict and identify that particular trip.

The columns that include are:

| **Table Header** | **Second Header** |
| --- | --- |
| 1. id | a unique identifier for each trip |
| 1. Vendor\_id | a code indicating the provider associated with the trip record |
| 1. Pickup\_datetime | date and time when the meter was engaged |
| 1. dropoff\_datetime | date and time when the meter was disengaged |
| 1. passenger\_count | the number of passengers in the vehicle (driver entered value) |
| 1. pickup\_longitude | the longitude where the meter was engaged |
| 1. pickup\_latitude | the latitude where the meter was engaged |
| 1. dropoff\_longitude | the longitude where the meter was disengaged |
| 1. dropoff\_latitude | the latitude where the meter was disengaged |
| 1. 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. |
| 1. trip\_duration | duration of the trip in seconds |

1. METHODOLOGY

We analyzed several data mining techniques to handle missing data, remove redundancy and resolve data conflicts. The Data Mining techniques are used to handle missing data. After analyzing we found that there were no missing data. So, In order to get rid of redundant data, we perform correlation analysis with the help of plots to check if the attributes are positively or negatively correlated if not redundant. To resolve data conflicts which we encountered for attributes such as time of pickup and drop-off, we converted the time to epoch format and worked on this epoch format to get our features. To train the model we used Random Forest Regressor.

**ANALYSIS**

In order to undergo the analysis part we examined several algorithms on regression and concluded that Random Forest is the well-suited Regression technique for our respective proposed model.

To begin with, we first examined the missing data, which was much less compared to the whole data so we decided to exclude the missing data.

By Exploratory Data Analysis we observed that the average speed is more from 0:00hrs to 5:00hrs and average speed is less during 16:00hrs to 20:00hrs in the evening.

We did correlation analysis to check for relation between two attributes which helped us to find the redundant data.

We trained our model using Linear Regression which gave us an accuracy of 75-78% and then we compared our model with a Random Forest Regressor and found that Random Forest was giving us more accurate results of 81- 83%.

In comparison to older techniques like Linear Regression our model gave a more accurate result by 6-7%.

Further, to improve the confidence we tuned the hyper-parameters such as number of trees and maximum depth for the Random Forest Algorithm.

With an increasing number of trees we observed that the Root Mean Square Error (RMSE) decreased rapidly to a healthy level.

**RESULTS**

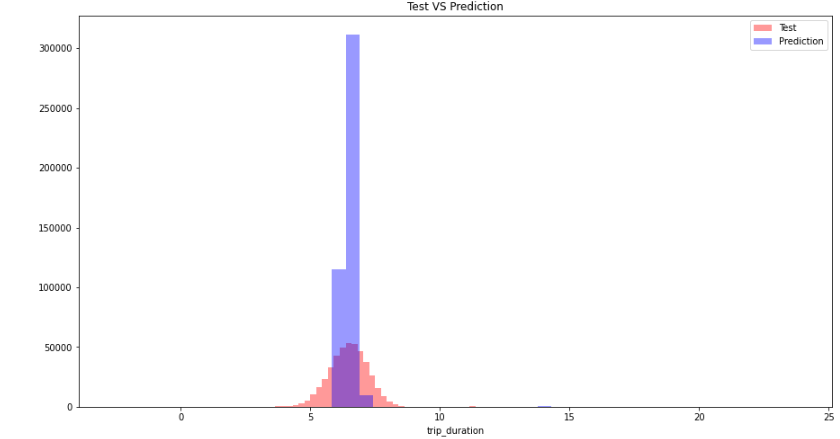
The proposed hierarchy of the workflow model was loading the data, Cleaning the data, Training the model, Making Predictions, Tuning the hyper Parameters to increase Confidence.

1. CLEANING THE DATA

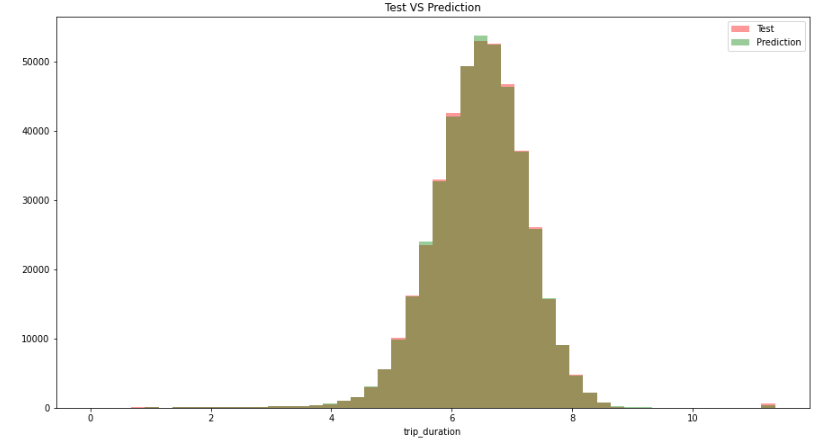
Cleaning the data involves eliminating the outliers and taking attributes required for feature extraction post Exploratory Data Analysis (EDA). To remove outliers some of the issues occurred are to make sure duration is greater than zero, ensure speed needs to be realistic (i.e.) speed needs to be between 6 and 140 mph, to make sure pickup and drop off locations are not random and belong to clusters close-by without loss of generality.

1. EXPLORATORY DATA ANALYSIS

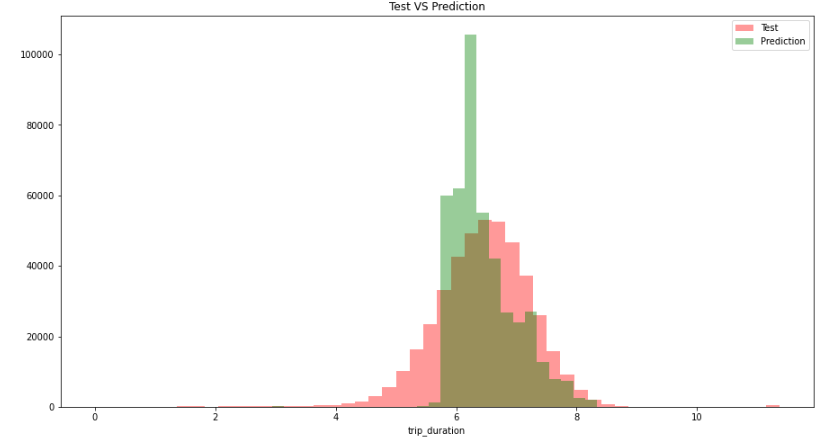
In statistics, exploratory data analysis is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. Data visualization is the graphic representation of data. It involves producing images that communicate relationships among the represented data to viewers of the images. This mapping establishes how data values will be represented visually, determining how and to what extent the property of a graphic mark, such as size or color will change to reflect changes in value of datum.



LINEAR REGRESSION



DECISION TREES



RANDOM FOREST

Visualizations show us how our model’s

predictions are close to Test Data. It is

evident that decision tree and Random

Forest is performing well.

C. TRAINING MODEL

To train the model we used Linear Regression and Random Forest Regression algorithms with 80-20 split of dataset for training and testing respectively. It gave an accuracy of 76-78 and 82-83 percent respectively, to improve the accuracy, tuning of

several hyper-parameters such as

number of trees and maximum depth

for a random forest algorithm.

**CONCLUSION**

* Observed which taxi service provider is most frequently used by the people of New York.
* Found out a few trips which were of duration 528 Hours to 972 Hours, possibly Outliers.
* Passenger count Analysis showed us that there were few trips with Zero Passengers and One trip with 7,8 and 9 passengers.
* Passenger count Analysis showed us that there were few trips with Zero Passengers and One trip with 7,8 and 9 passengers.
* Taxi giants such as UBER and OLA can use the same data for analyzing the trends that vary throughout the day in the city. This not only helps in better transport analysis but also helps the concerned authorities in planning traffic control and monitoring.

**FUTURE WORK**

As a part of the future work, the Multi-layer Perceptron model could be auto-tuned to further learn and determine which features need to be joined to detect numerous interactions between them as needed. Moreover, variabilities and quantities related to the various location features might also be computed in the upcoming research in order to localize the traffic-based effects on the taxi prediction coordinates. Speed limitations-based features could later be incorporated alongside to comprehend better analysis of the datasets. At last, enhancements to the K-Means Clustering algorithm could be provided by encompassing additional features such as distance to the closest metro station, number of bars and eateries in a given zone, etc. so as to exploit comparative qualities belonging to various zones. This would also ensure the rightful evaluation of various clusters in which each data point falls such that it fills in as an extra vital element for our models.

**REFERENCES**

* Analytics Vidhya
* Kaggle.com