NYC TAXI TIME PREDICTION

Prince kumar jha

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.

Problem statement :

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.

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, henceforth 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 traveled from one cluster to another is important to predict and identify that particular trip.

The columns that include are:

- **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**—T

his 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

STEPS INVOLVED :

➔Data Wrangling:

After loading the dataset we performed this method by cleaning, organizing, and transforming raw data into the desired format which makes us understand the data clearly. This process helped us to tackle the unwanted data, to produce accurate results, and to make better decisions.

➔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.

➔ 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 communication is achieved through the use of a systematic mapping between graphic marks and data values in the creation of the visualization. This mapping establishes how data values will

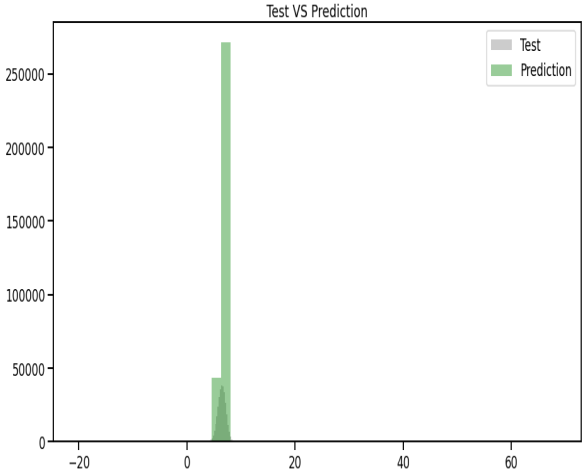
be represented visually, determining how and to what extent the property of a graphic mark, such as size or color will change to reflect changes in value of datum.

Model development :

1. Linear regression
2. Decision Tree
3. Random forest

➔ 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.

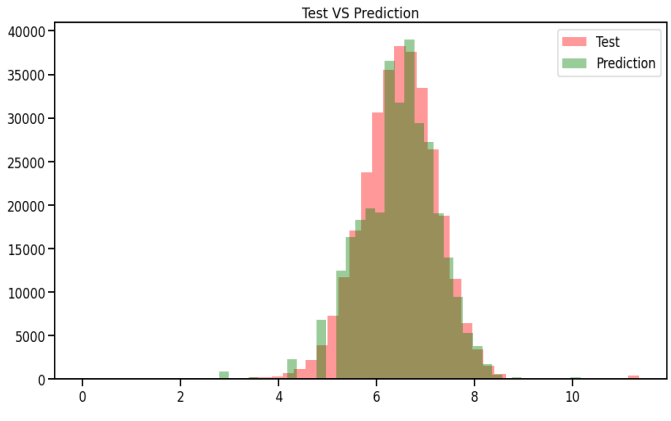


LINEAR REGRESSION

➔ DECISION TREES :

Decision trees can be used for classification as well as regression problems. The name itself suggests that it uses a flowchart like a tree structure to show the predictions that result from a series of feature-based splits. It starts with a root node and ends with a decision made by leaves.

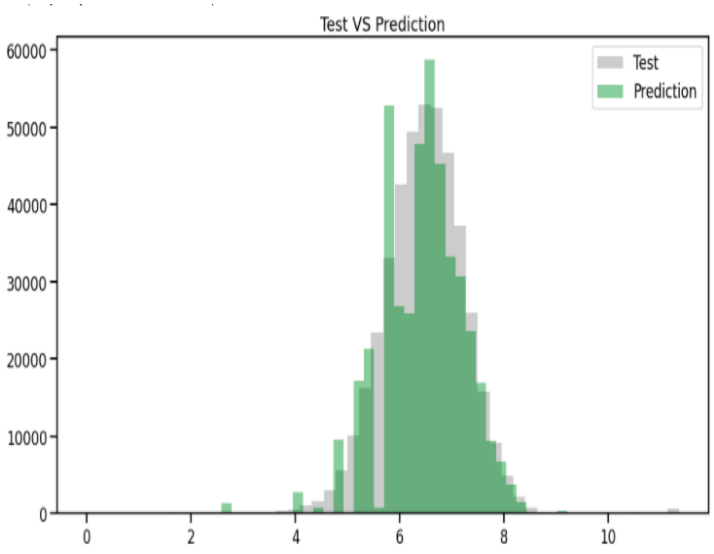
Decision trees are upside down which means the root is at the top and then this root is split into various several nodes. Decision trees are nothing but a bunch of if-else statements in layman terms. It checks if the condition is true and if it is then it goes to the next node attached to that decision.



DECISION TREES

➔ 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. 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.



RANDOM FOREST

➔ TRAINING & TESTING :

To train the model we use Linear Regression and Random Forest Regression algorithms with 80-20 split of dataset for training and testing respectively.

The results of our train & test are shown by the following tables . basically we apply two type of approach first approach is with PCA and second approach is without PCA

Ques- why do we use PCA?

Ans- we use PCA because our data is a high dimensional data and for better results we have to reduce its dimensionality and for this we use PCA which is very helpful for us.

Model evaluation result with PCA

Algorithms	Training Score	Validation Score	Cross Validation Score	R2-Score	RMSE
Linear Regression	0.0661	0.04564	0.06082	-10.221	--
Decision Tree	0.9506	0.9418	0.9406	0.9389	0.0306
Random Forest	0.9522	0.9454	0.9451	0.9424	0.0300

Model evaluation result without PCA

Algorithms	Training Score	Validation Score	Cross Validation Score	R2-Score	RMSE
Linear Regression	0.192	0.1766	0.1881	0.1766	--
Decision Tree	0.4672	0.4576	0.4469	0.4576	0.0884
Decision Tree with GridCV search	0.9091	0.5520	--	0.5130	0.0831
Random Forest	0.4805	0.4727	0.4702	0.4727	0.0876

CONCLUSION (with EDA):

- Observed which taxi service provider is most Frequently used by New Yorkers.
- Found out that the 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.
- 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.

CONCLUSION (with regression model):

- Apply Standard Scaling on the Dataset to Normalize the values.
- Further, Apply PCA to reduce dimensions, as you'll extract

features from our primary Date Time Feature. Those additional features might lead our model to suffer from “Curse of dimensionality” and could drastically affect performance.

- Pass the PCA Transformed data in our ML Regression Algorithm and Evaluate results.
- We can perform hyper tuning on our Algorithm to get the most out of it but
  Hyper Tuning consumes a lot of time and resources of the system depending upon how big the Data we have and what algorithm we're using. It will go through a number of Iterations and try to come up with the best possible value for us.
- We also applied some other types of algorithms like xgboost & elastic net but results are not acceptable so that's why I think that there is no need to show these types of algorithms .