Technical Documentation

NYC Taxi Time Prediction

By- Rajvee Sharma

Abstract

The objective of our model is to predict the accurate trip duration of a taxi from one pickup location to another drop off location. In today's fast paced world where everyone is short and always in a hurry, people want to reach their destinations within the particular period of time. So, they preferred taxi transport which is actually cheaper in cost and safer. So, for this serenity, we already have million dollar start-ups like Uber, Ola where we can track our trip duration. So, in this methodology, we propose methods to make predictions of trip duration and for this we have used several algorithms, tune the corresponding parameter by analysing each and every parameter against RMSE and predict the trip duration. To make our predictions we used Random Forest, Decision Tree, Linear Regression. We improved the accuracy by tuning hyper parameters and Random Forest gave the best result. Also observed some data mining techniques to handle missing data, data redundancy and resolve data conflicts.

Introduction

There are many possible methods of moving between two points in the city, however taxi trips have found wider application in urban cities. Hence, it becomes very important to predict taxi trip time duration between two points in the city. For a good taxi service it's an integration with the existing transportation system. The project serves as the right means to comprehend the traffic system in New York city. For prediction purposes such as pickup latitude, pickup longitude, drop-off latitude, drop-off longitude is considered. The primary focus of this project is in depth analysis of the factors associated with taxi trips in NYC. The different uses are Random Forest, Decision Tree, Linear Regression.

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.

2. Decision Tree

Decision Trees (DTs) are a non parametric supervised learning method used for classification and regression. 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.

3. 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.

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, Random Forest Regression 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 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
id	A unique identifier for each trip
Vendor_id	A code indicating the provider associated with each trip
Pickup_datetime	The date and time when meter was engaged

Dropoff_datetime	The date and time when meter was disengaged
Pickup_latitude	The latitude when meter was engaged
Dropoff_latitude	The latitude when meter was disengaged
Pickup_longitude	The longitude when meter was engaged
Dropoff_longitude	The longitude when meter was disengaged
Store_and_fwd_flag	The flag indicates whether the trip record was held in vehicle memory before sending to the vendor.
Trip_duration	Duration of trip in seconds

METHODOLOGY

We analysed several data mining techniques to handle missing data, remove redundancy and resolve data conflicts. The Data Mining techniques are used to handle missing data.

Analysis

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 Repressor which gave us an accuracy of 75-78% and then we compared our model with a Random Forest Repressor 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.

A) 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.

A.EXPLORATORY DATA ANALYSIS

In statistics, exploratory data analysis is an approach to analysing 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 modelling 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. To train the model we used Linear Regression and Random Forest Regression algorithm 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 New Yorkers.
- 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.
- Taxi giants such as UBER and OLA can use the same data for analyzing the trends that vary throughout the day in the city.

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