NYC Taxi Trip Time Prediction

Ashik Kumar
Data science trainees,
AlmaBetter, Bangalore

Abstract:

New York City taxi rides form the core of the traffic in the city of New York. The many rides taken every day by New Yorkers in the busy city can give us a great idea of traffic times, road blockages, and so on. Predicting the duration of a taxi trip is very important since a user would always like to know precisely how much time it would require of him to travel from one place to another. Prediction of duration and price of trips can help users to plan their trips properly, thus keeping potential margins for traffic congestions. It can also help drivers to determine the correct route which in-turn will take lesser time as accordingly.

This data includes pickup and drop-off point coordinates, the distance of the trip, start time, number of passengers, and a rate code belonging to the different classes of cabs available such that the rate applied is based on a regular or airport basis. Hereafter, we applied LINEAR REGRESSION and DECISION TREE to find out which one of them provides better accuracy and relationships between real-time variables

1.Problem Statement

The New York City Taxi Fare prediction challenge is a *supervised regression* machine learning task. Given pickup and drop-off locations, the pickup timestamp, and the passenger count, the objective is to predict the fare of the taxi ride. this problem isn't 100% reflective of those in industry, but it does present a realistic dataset and task on which we can hone our machine learning skills.

Some of the important attributes of the dataset are discussed below:

- **Id**: which provides a unique identification to a trip.
- **vendor id**: a unique code which gets assigned to the different cab companies.
- **pickup datetime**: starting statistics of the pickup.
- **dropoff datetime**: ending statistics of the pickup.
- **passenger count**: passengers travelling in a particular trip.
- **pickup longitude**: longitudinal location of the pickup.
- **pickup latitude**: latitudinal location of the pickup.
- **dropoff longitude**: longitudinal location of the drop off.
- **dropoff latitude**: latitudinal location of the drop off.
- **store and fwd flag**: a code to identify whether the data is stored on the device and then gets forwarded to the database.
- **trip duration**: the total time of the trip in seconds.

2. Introduction

New York City is one of the highly advanced cities of the world with extensive use of taxi services. Along with a vast population, the requirement of commonly available transportation serves the common purpose as it provides a very large transportation system. New York facilitates one of the largest subway systems in the world and comprises various green and yellow cabs which approximately count around 13,000 taxis. Most of the population of New York depends upon public transport, and it has been estimated that 54 percent of the people do not own a car or a personal vehicle. As a matter of fact, it accounts for almost 200 million taxi trips per year.

Successful prediction of the taxi trip duration would eventually be very useful in the future to make better taxi trip duration predictions applicable to multiple cities.

3. Steps involved:

- **Exploratory Data Analysis**
  After loading the dataset, we performed this method by comparing our target variable that is trip_duration with other independent variables. This process helped us figuring out various aspects and relationships among the target and the independent variables. It gave us a better idea of which feature behaves in which manner compared to the target variable.

- **Null values Treatment**
  Fortunately, in this dataset we do not have any missing values which is great.

- **Encoding of categorical columns**
  We used One Hot Encoding to produce binary integers of 0 and 1 to encode our categorical features because categorical features that are in string format cannot be understood by the machine and needs to be converted to numerical format.

- **Standardization of features**
  Our main motive through this step was to scale our data into a uniform format that would allow us to utilize the data in a better way while performing fitting and applying different algorithms to it.
  The basic goal was to enforce a level of consistency or uniformity to certain practices or operations within the selected environment.

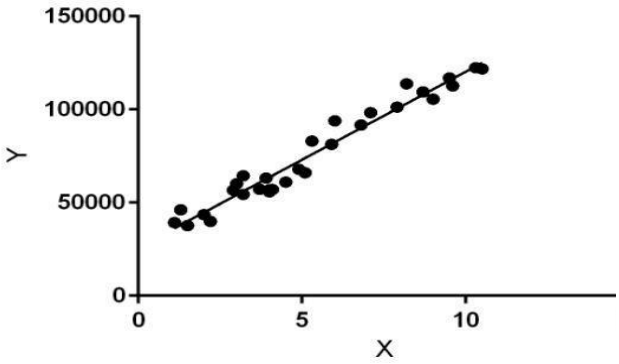
- **Fitting different models**
  For modelling we tried various classification algorithms like:

1. **Linear Regression**
2. **Decision Tree**

4. Algorithms:

1. Linear Regression:

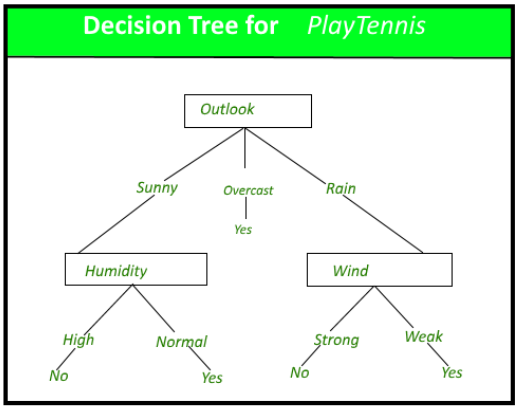
Linear Regression is a machine learning algorithm based on **supervised learning**. It performs a **regression task**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering, and the number of independent variables being used.



Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.
In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model.

2. Decision Tree:

Decision tree is the most powerful and popular tool for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label.



7.2. Model performance:

Model can be evaluated by various metrics such as:

1. **Confusion Matrix**-

The confusion matrix is a table that summarizes how successful the classification modelis at predicting examples belonging to various classes. One axis of the confusion matrix is the label that the model predicted, and the other axis is the actual label.

2. **Precision/Recall**-

Precision is the ratio of correct positive predictions to the overall number of positive predictions : TP/TP+FP

Recall is the ratio of correct positive predictions to the overall number of positive examples in the set: TP/FN+TP

3. **Accuracy**-

Accuracy is given by the number of correctly classified examples divided by the total number
of classified examples. In terms of the confusion matrix, it is given by: TP+TN/TP+TN+FP+FN

4. **Area under ROC Curve(AUC)-**
   ROC curves use a combination of the true positive rate (the proportion of positive examples predicted correctly, defined exactly as recall) and false positive rate (the proportion of negative examples predicted incorrectly) to build up a summary picture of the classification performance.

5. Conclusion:

We are successfully able to implement both algorithms on the New York City Taxi Trip Duration dataset and able to draw certain conclusions from several inferences. After implementing both algorithms, we come across that DECISION TREE is better than LINEAR REGRESSION as it shows a slightly better accuracy than the latter one. This in turn helps to conclude that the DECISION TREE Model is more efficient and reliable in predicting the taxi trip duration as compared to LINEAR REGRESSION.

References-

1. MachineLearningMastery
2. GeeksforGeeks
3. Analytics Vidhya