National Institute of Technology, Karnataka Surathkal

Data Warehousing & Mining

CO461

Trip Duration Prediction

Author:
Khursheed 14CO255
Pradhumn 14CO232

Supervisor:
Dr.M Venkatesan

November 15, 2017

Summary

The purpose of this modelling and work represented in this report is to accurately predict the trip duration of taxi's from one pickup location to other dropoff location. With today's fast growing world of billion dollar startup's such as Uber and Ola, every customer wants to know the exact duration to reach his/her destination to carry ahead their plans. As a result goal of every cab service provider is to get exact duration to their customers taking into consideration the factors such as traffic, time and day of pickup. To solve the above problem we propose a method to make predictions of duration, we will use 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 use RandomForest Regressor, LinearSVR and LinearRegression. We improved accuracy by tuning hyperparameters and RandomForest gave best accuracy of 83 percent. Implementation and Presentation can be obtained by browsing hyperlinks.

1 Approach Used

We used/analyzed several data mining techniques to handle missing data, remove redundancy and resolve data conflicts. We used the NYC Limousine OpenData. We used the travel details of the month of January in the year 2015 to carry ahead feature extraction and prediction.

1.1 Data Mining techniques

1. To handle missing data we did use binning, but on analysis we found that the data missing was very less in number (i.e) only few tuples of around 1 percentage of total tuples were missing so we preferred removing missing data.

2. To remove data redundancy we did correlation analysis by using plots to check if the two attributes are positively or negatively co-related if not redundant.

3. To resolve data conflicts which we encountered for attributes such as time of pickup and dropoff, we converted the time to epoch format and worked on this epoch format to get our features.

4. To cluster dropoff and pickup locations we used KMeans algorithm as KMmeans tries to group based solely on euclidean distance between objects we will get back clusters of locations that are close to each other, also as locatiosn are not spread across the world and confined NYC KMeans does a decent job here.

5. To train the model we used RandomForestRegressor because let us consider an example, Let's suppose you are trying to predict income. The predictor variables that are available are education, age, and city. Now in a linear regression model, you have an equation with these three attributes. Fine. You'd expect higher degrees of education, higher "age" and larger cities to be associated with higher income. But what about a PhD who is 40 years old and living in Scranton Pennsylvania? Is he likely to earn more than a BS holder who is 35 and living in Upper West SIde NYC? Maybe not. Maybe education totally loses its predictive power in a city like Scranton? Maybe age is a very ineffective, weak variable in a city like NYC? This is where decision trees are handy. The tree can split by city and you get to use a different set of variables for each city. Maybe Age will be a strong second-level split variable in Scranton, but it might not feature at all in the NYC branch of the tree. Education may be a stronger variable in NYC. Applying similar analogy to our data set where in we need to split based on day of the week and pick up , RandomForest comes handy.

1.2 Data Set

We used NYC Taxi Limousine OpenData for the year 2015 in the month of January. We selected the following features:

1. **Trip Distance:** Distance is an important factor for predicting the duration of trip, as Distance = Speed/Time

2. **Day of the week:** Weekdays experience slow speed because of daily routine of schools and offices, hence forth the need for this feature.

3. **Time of the day:** Peak hours of offices and school start and end such as Morning 8 - 12 and evening 4 to 7 experience high traffic.

4. **Pick up and dropoff cluster:** Route being travelled that is from one cluster to another is important to predict and identify that particular trip.

2 Results & Discussions

The pipeline of the work flow was:

- Loading the data
- Cleaning the data
- Training the model
- Making Predictions
- Tuning the hyper Parameters to increase Confidence

Cleaning the data involved removing the outliers and getting attributes required for feature extraction post Exploratory Data Analysis(EDA). To remove outliers some of the issues addressed are:

1. Make sure duration is greater than zero.
2. Ensure speed is realistic (i.e) speed must be between 6 and 140 mph.
3. Make sure pickup and drop off locations are not random and belong to clusters close-by without loss of generality.

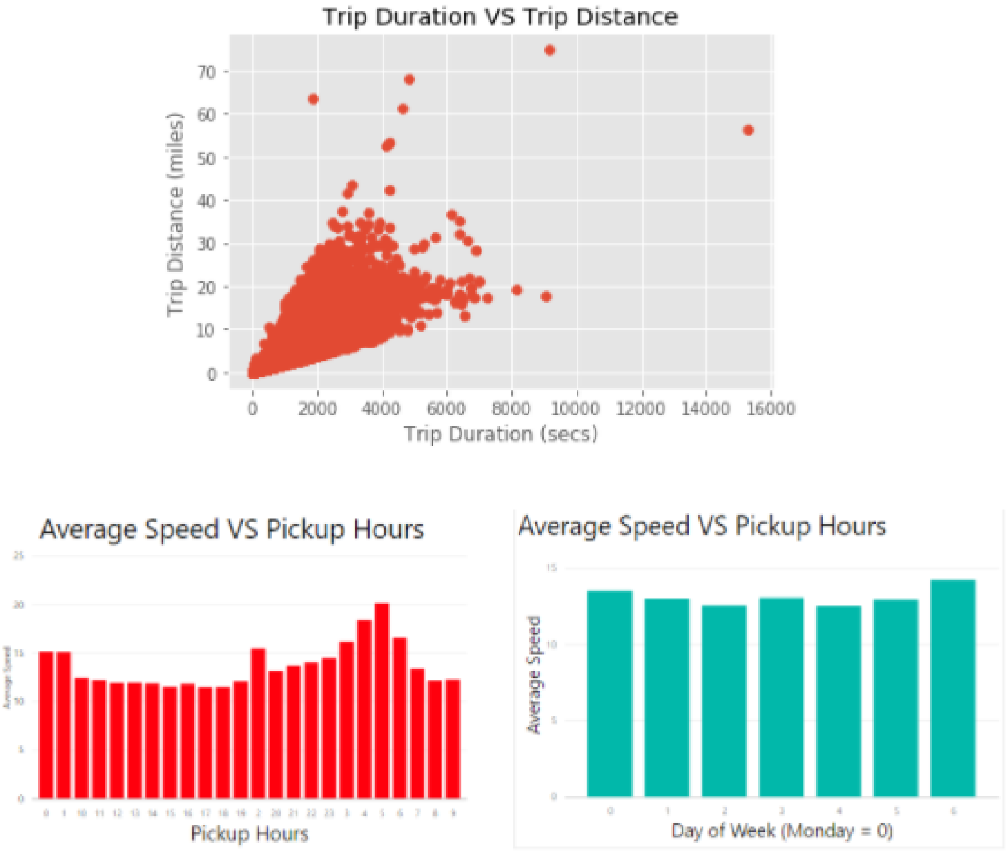

Figure 1: Exploratory Data Analysis(EDA)

EDA as shown in Figure 1 made us draw the following conclusions:

1. The average speed is more during the time 00hrs - 05hrs in the morning
2. The average speed is less during the time 16hrs - 20hrs int he evening.
3. There exists a positive co-variance/co-relation between trip distance and trip duration.

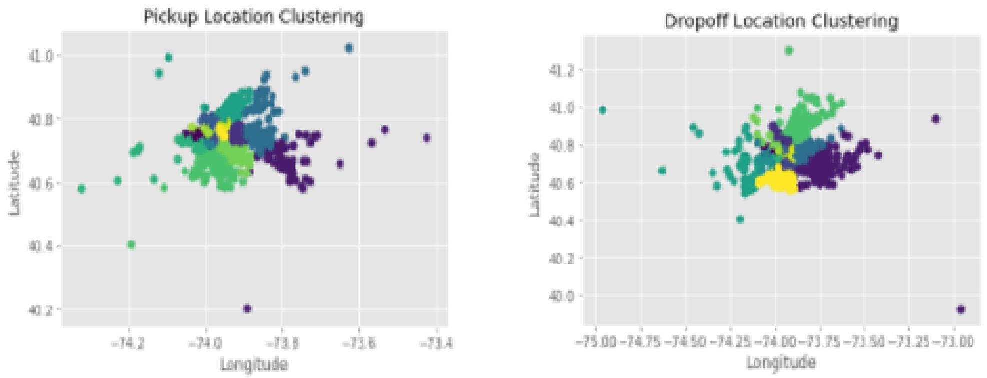

Figure 2: Pickup & Dropoff clusters using KMeans

Getting features involved clustering the pick up and drop off locations as shown in Figure 2 and having the one hot encoding of pickup and dropoff clusters as feature along with one hot encoding of other features such as time and day of pickup wherein these were selected as features after EDA done above.

Now to train the model we used Random Forest Regression algorithm with 80-20 split of dataset for training and testing respectively. It gave an accuracy of 82-83 percentage.

To improve the accuracy, tuning of several hyper-parameters such as number of trees and maximum depth for random forest algorithm was done and Figure 3 is the result of analysis.

With respect to the Figure 3 increasing the number of trees as observed above the RMSE decreased to good levels until it reached the elbow point of value greater than 20.

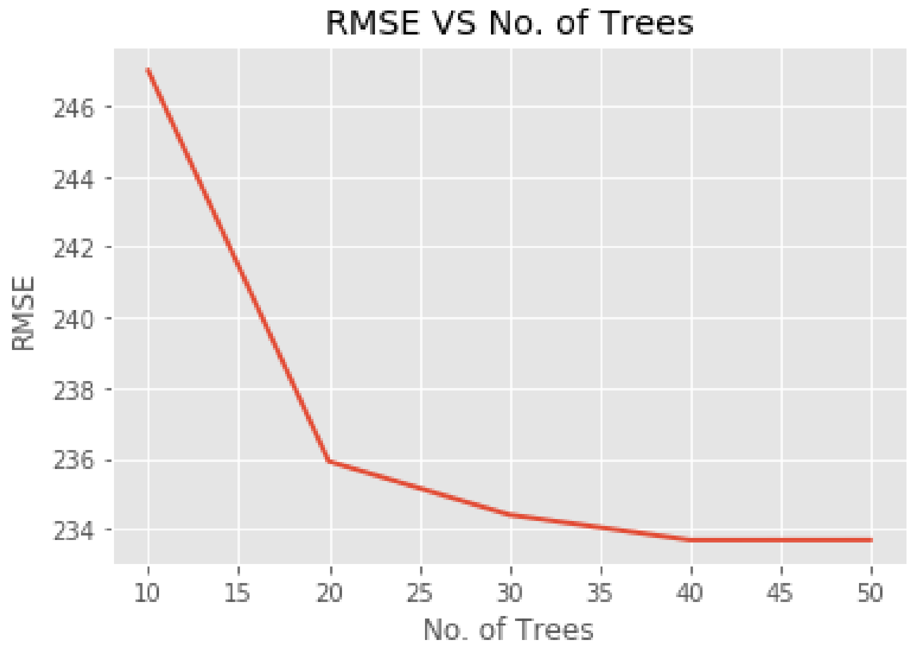

Figure 3: Tuning of hyper-parameters

3 Conclusions

Compared all the other algorithsm such as Linear regression(accuracy: 78 percent) and its variants, Random Forest(accuracy: 83 percent) gives the best result. However a more realistic approach to solve the problem statement would be to get dynamic data or real data via Cab service provider's API. This would help us get the traffic at that time and will provide accurate results. We aim to carry this work ahead using dynamic data sets via API's, getting real data and using other algorithms such as Stochastic gradient descent to train the model and make predictions.

References

[1] *Documentation on Machine Learning algorithms*, scikit-learn.org.

[2] *Fare Prediction for NYC Taxi ride's*, University of California, San Diego.