Decision Tree Regression Model to Predict Cab Rides

Vineet Kumar

Computer Science & Engineering Department
Jadavpur University

vntkumar8@gmail.com

Abstract

The Radio Cab Space is Heating up. Understanding the Cab demand and supply could definitely trigger the efficiency of the radio cab services. Getting the data of cab rides and using various ML models we can predict the best in favor of cab service providers. In this work, I present a few different models to predict the number of cab pickups that will occur at a specific time and location; these predictions could definitely boost the efficiency of cab service providers. I implemented decision tree regression model yielding coefficient of determination close to unity which is significant improvement over other regression models.

Introduction

Around the world, Cabs (Taxis) play a major role in urban transport system and present an alternative mode of travel to conventional public transportation systems. There are concerns from transportation planners about how cab services are distributed across the city in space. In order to effectively regulate the cab riding industry and plan for its effective integration into the citywide transportation system, it is necessary to understand the cab demand. In this work, various regression models have been experimented to predict the same.

Method

Data Handling

In India we dont have any publicly accesible repository to download cab ride data, in this project I used the US based data available by Google BigQuery[1]. The Data retrieval and processing technique is schematically shown below:-

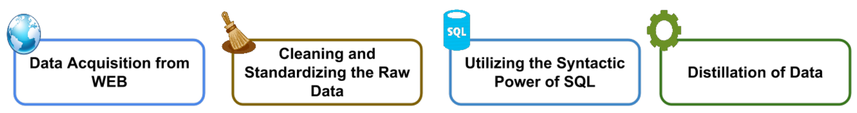

Figure 1: Data Retrieval, Conditioning, and Processing

- The raw cab ride data includes the information like date, time, and location of pickup and drop-off and many non significant fields like vendor id, vehicle number etc.
- Checked the validity and completeness of all data files. Eliminated parsing errors and fixed errant data values.
- Imported data from text files into SQL database using Python scripts. Flexibility of SQL queries allowed complex joins and merges to complete with ease. Data accessible from shell scripts as well as Python scripts.

Evaluation of Model

To evaluate the performance of any regression model, we need training data & test data. We split our *ready to use* data into two parts – training data and testing data. Also, in order to quantify the success of a model, requires an error metric.

- RMSD was my choice as it heavily penalizes predictions with a high deviation from the true value.

Regression Models

In this project I experimented with following regression models:-

- simpleAverage Model: simpleAverage model predicts the number of pickups on a test data point at a given location as the average number of pickups. This model works as a reference of improvement for other models
- Linear Least-Squares Regression: This model enables to explore the linear relationship in the dataset & feature set
- Decision Tree Model: Due to complex nature and outstanding capability of representing complex decision boundaries this model is perfect for our problem.

Results

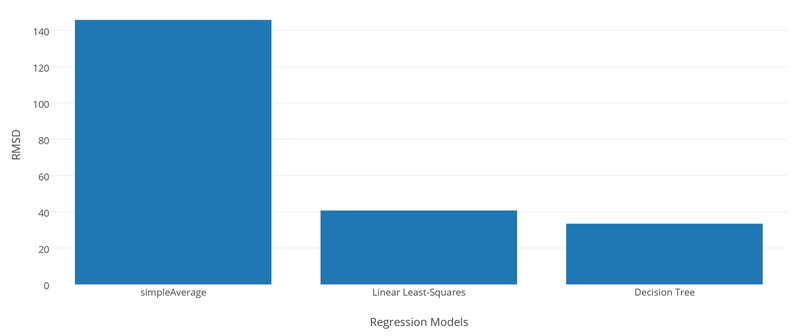

Figure 2: Variation of RMSD with various models

Model	RMSD
simpleAverage	142.62
Linear Least-Squares Regression	40.81
Decision Tree Regression	32.71

Table 1: Best results for each model

Table & Bar Chart summarises best results obtained by training the model using training data and subsequently testing on test data.

In order to visualize how well the models perform, we plot the true *versus* predicted number of pickups for each data point in the test set. The plots in figure[3] suggest that how well the linear regression and decision tree regression model performed. In the plots, the true value is a dotted stright line $y = x$ Most predictions lie close to the true values. The data points straddle the true value line evenly.

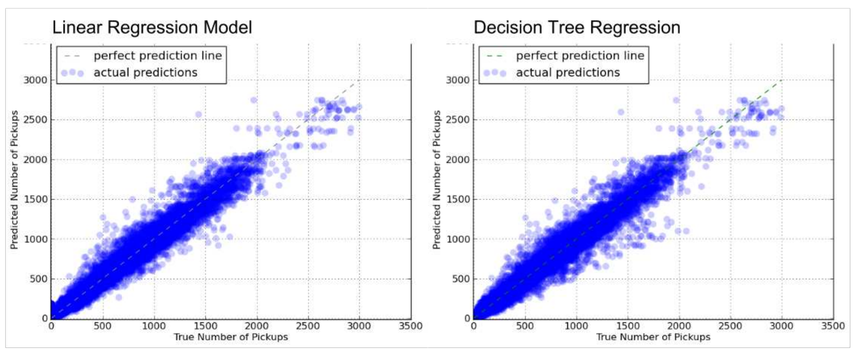

Figure 3: Predicted versus true number of pickups (left) Linear Regression (right) Decision Tree

The tree diagram in Figure[4] shows a subsection of the trained decision tree. Each non leaf node of the tree is a binary answer type question.

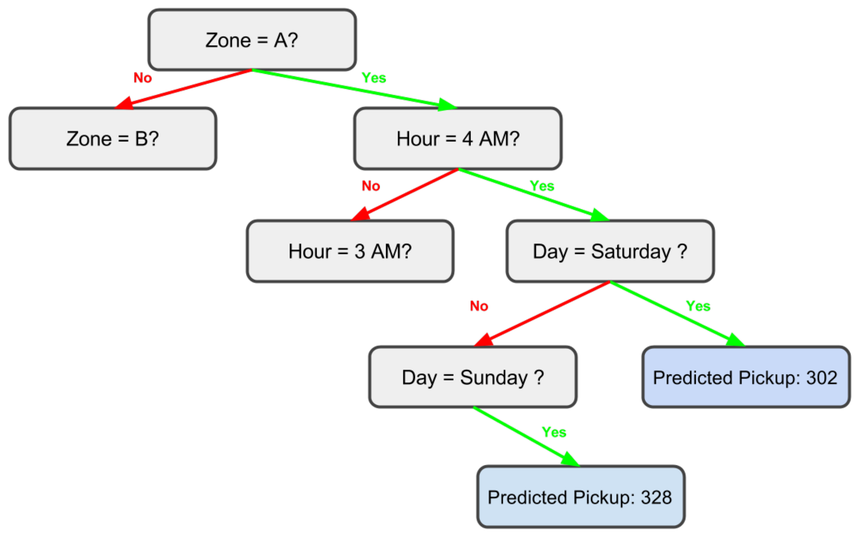

Figure 4: Shematic View of the final trained decision tree

Decision Tree model outperforms linear regression because paths within the decision tree are able to represent the combination of many features in Feature Set. It is because these features are highly dependent upon one another so therefore the decision tree regression performs best of all the models under consideration.

Conclusions

Cab Rides Predcting Model performed well. The decision tree regression model performed best, due to its outstanding ability to capture complex feature dependencies. The decision tree regression model achieved a value of 32.71 for RMSD and 0.9958 for coefficient of determination

References

[1] Google BigQuery:
https://cloud.google.com/bigquery

[2] Castro, Pablo Samuel, et. al, *Urban traffic modelling and prediction using large scale taxi GPS traces* Pervasive Computing. Springer Berlin Heidelberg, 2012. 57-72