

Cab Rental Prediction

Pranav Tewari | 28 March 2019

Chapter 1

Introduction

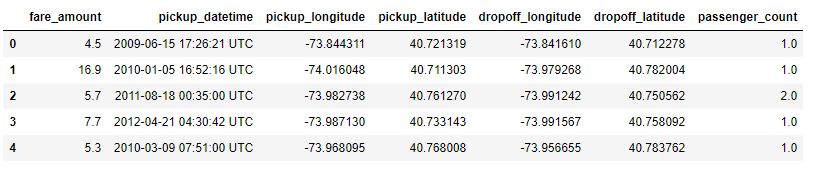
* 1. Problem Statement

You are a cab rental start-up company. You have successfully run the pilot project and now want to launch your cab service across the country. You have collected the historical data from your pilot project and now have a requirement to apply analytics for fare prediction. You need to design a system that predicts the fare amount for cab ride in the city.

* 1. Data

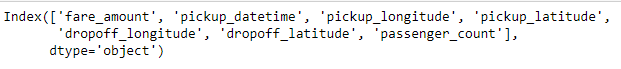
Below is the sample of data set which we are using to predict:

Table 1.1: Sample Data (Column: 1-5)



The table below shows the columns of the data set using which we have to predict the values.

Table 1.2: Columns of Data



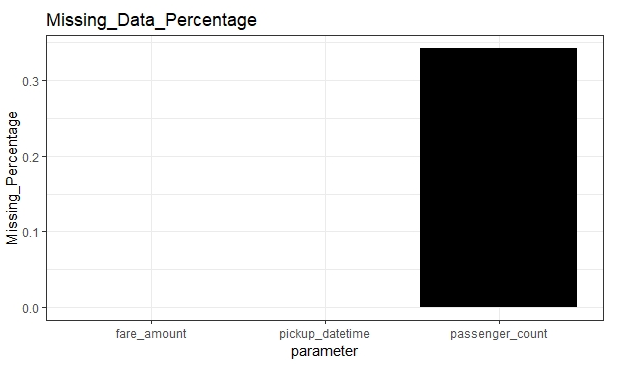
Chapter 2

Methodology

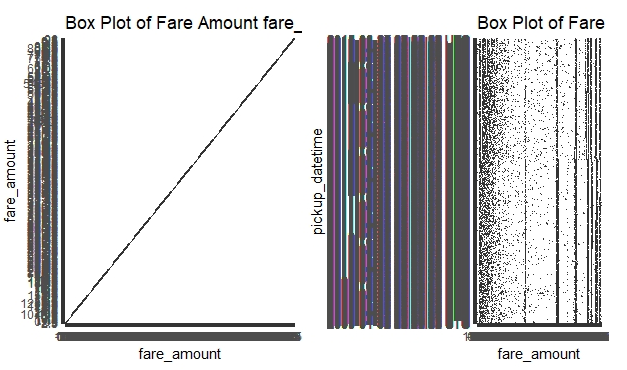
* 1. Pre Processing

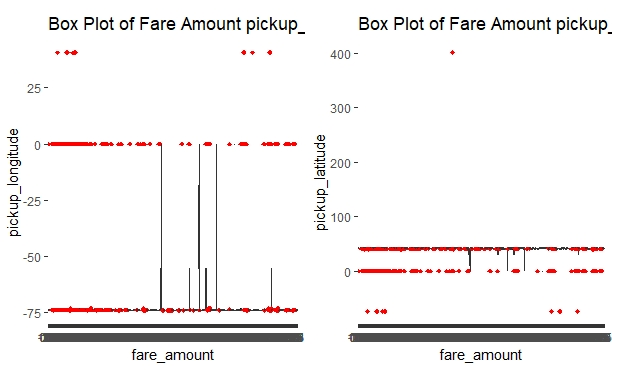
Any predictive modelling requires that we look at the data before we start modelling. However, in data mining terms looking at data refers to so much more than looking. Looking at data refers to exploring the data, cleaning the data as well as visualizing the data through graphs and plots. This is often called as **Exploratory Data Analysis**. To start this process we will first try to look at all the probability distributions of the variables. Most analysis like regression, requires data to be normally distributed. We can visualize that in a glance by looking at the probability distribution or probability density functions of the variable.

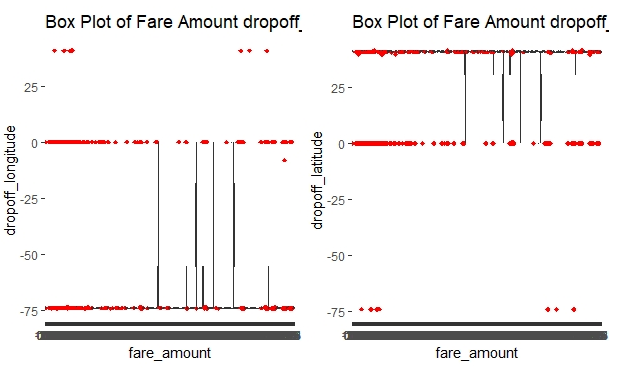
Below figures are Bar Plot of the Missing Values in the Data set.

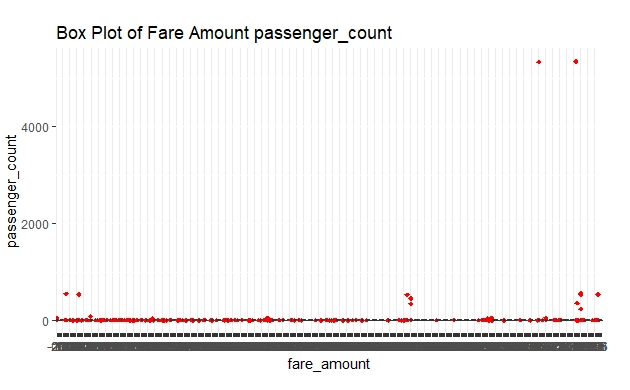


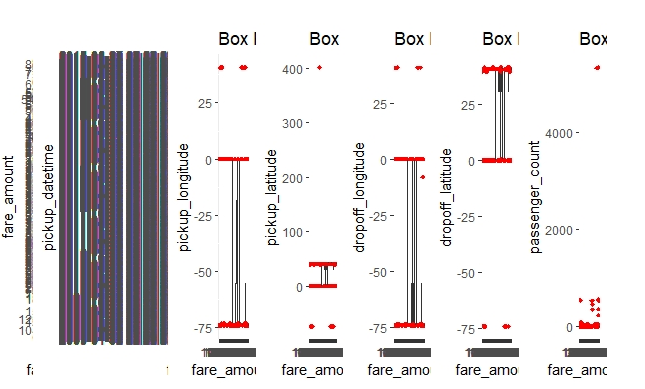
Below figures are box Plot of the Variables against the count.





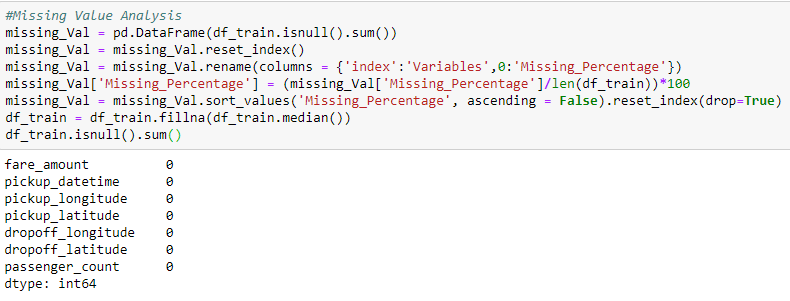






* + 1. Null Value Analysis

It is one of the Pre Processing steps in which the missing values in the variables are imputed with either of the methods i.e., (Mean, Median, Mode, KNN). The values are imputed only if they are missing up to 30% of the total data in the variable else it is left out as the variable does not carry meaningful details to predict the outcome.

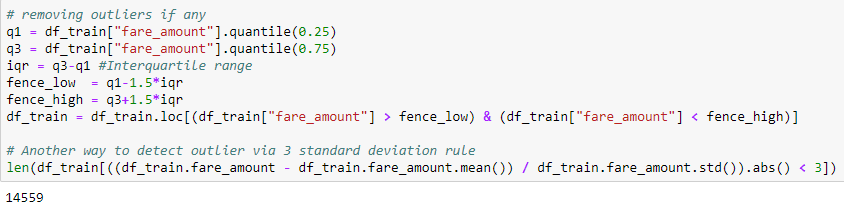


* + 1. Outlier Analysis

One of the step of Pre Processing is outlier analysis where we can see the variables having values which are varying from other values very widely.

These values are those which can be rejected or substituted using KNN.

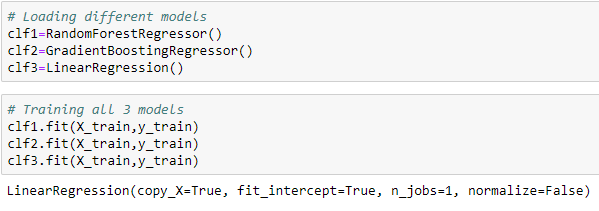
But herein case we have removed the outlier data from our data set to make it normally distributed.



* 1. Modeling
     1. Model selection

Model selection can depend upon the independent variable, in our case its **Fare Amount.** Here we can try and test multiple models and the one which best suits or gives the best result on train model.

You can start your model with the most simplest to the most complex and can try all these models simultaneously.



If the dependent variable is nominal the only predictive analysis that we can perform is **Classification,** and if the dependent variable is interval or ratio, the normal way is to do **regression** analysis, or classification after binning, but here in our case we do both classification and Linear Regression.

* + 1. Classification

Classification for the prediction of count cannot be done as count of anything is not a categorical value. Though the dependent variables are categorical but predicting count may not give the result expected and may have to categorize the count as well depending upon the dependent variables selected which will lead to a very hectic task.

Chapter 3

Conclusion

* 1. Model Evaluation

Now that we have few models for predicting the target variable, we need to decide which one to choose.

There are several criteria that exits for evaluating and comparing models. We can compare the models using any of the following criteria:

1. Predictive Performance
2. Interpretability
3. Computational Efficiency

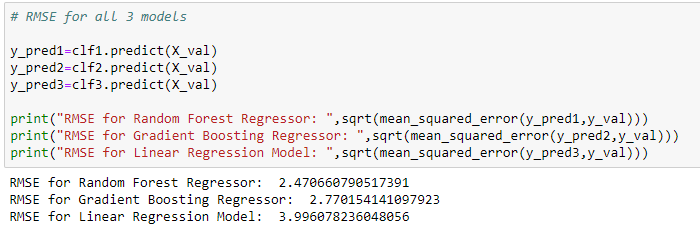
In our case i.e., predicting Fare Amount, the latter 2, Interpretability and computational Efficiency, do not hold much significance. Therefore we will use predictive performance as criteria to compare and evaluate models.

Predictive performance can be measured by comparing predictions of the models with real values of the target variables, and calculating some average error measures.

* + 1. Root Mean Square Error (RMSE)

RMSE is one of the error measures used to calculate the predictive performance of the model. We will apply this measure to our model.

Mean Absolute percentage error (MAPE) measures the deviation from the actual data in terms of percentage and consider both the negative and positive errors which cancels out each other. Hence we use RMSE which is more accurate and by squaring the errors they become positive and does not cancels each other and stays in existence till the end of commutation thus adding more accuracy to the result.



* 1. Model Selection

We can see that Random Forest is comparatively performing better than linear regression and Gradient Boosting hence we will use that model to avoid any loss of information.

* 1. Answer to Problem Statement

Here we can divide the range of fare depending upon the kilometers travelled by the customer so according here the customer can travel within the city and the rental is between 5-17 (currency specific).

Chapter 4

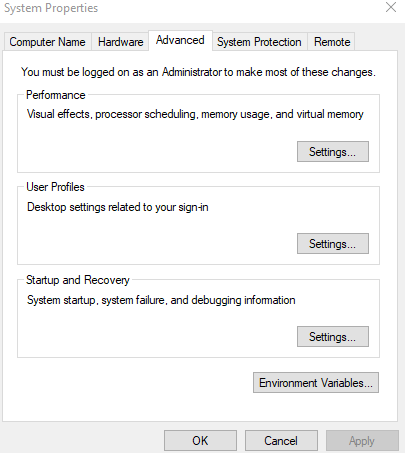
Deployment

* 1. Steps to deploy model on DOS

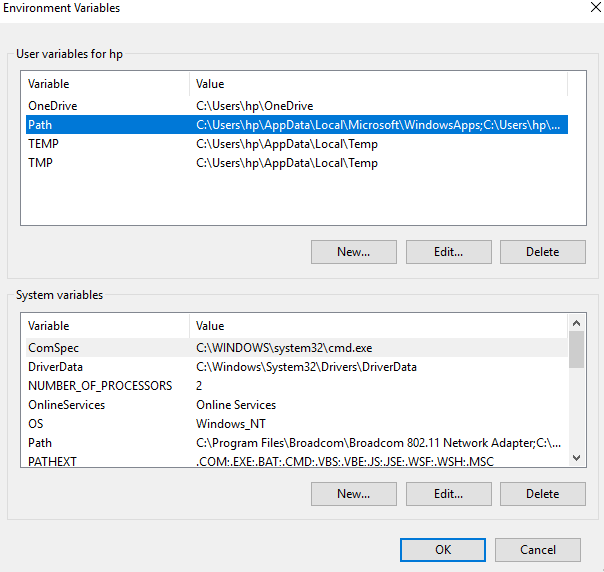
A .py file is created and attached with the other files which can run on DOS format.

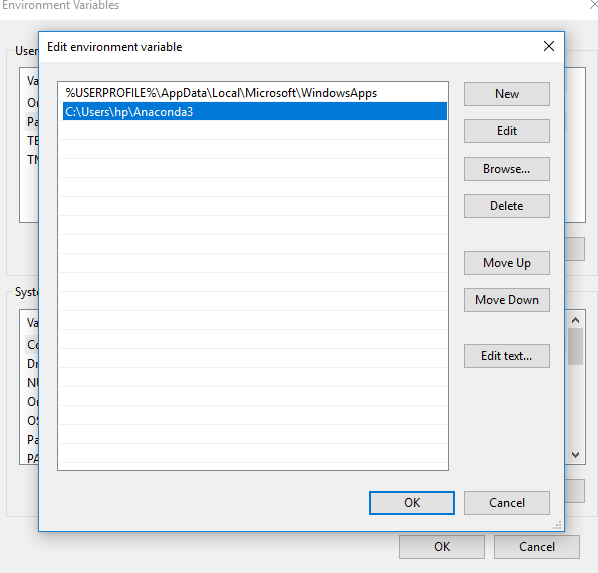
1. First of all open the path in which python has been installed

Go to control panel open edit the system environment variables

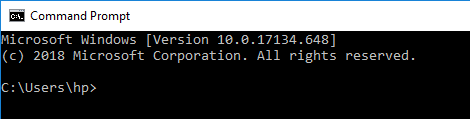


1. Go to advanced tab and click on Environment Variables and click on variable path there and then click on edit

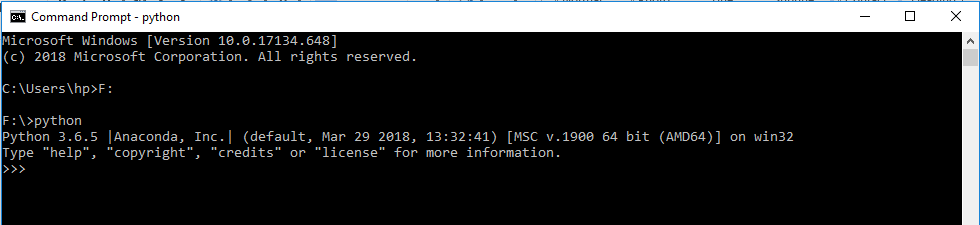




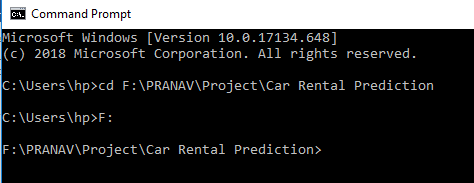
1. Paste the path copied having the python.
2. Click on OK and then open CMD on your system.



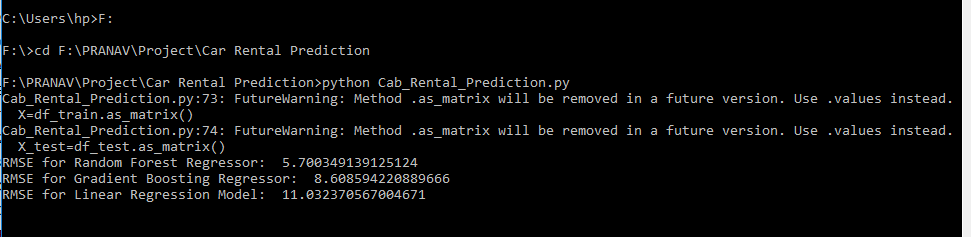
1. To check python commands are working on it enter **PYTHON/ Rscript for R**

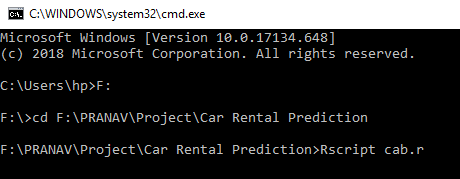


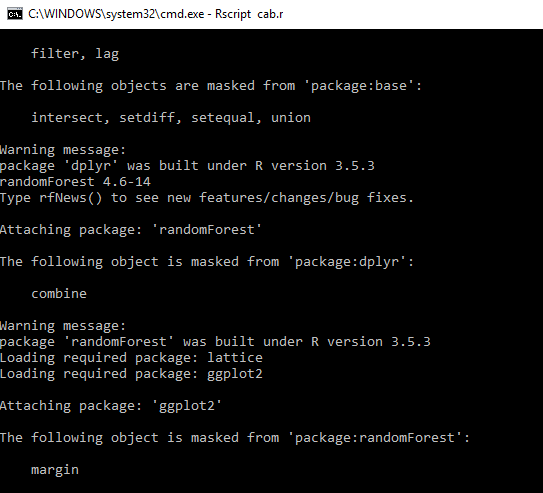
1. Enter the drive name having your **.py/.R** file as shown below



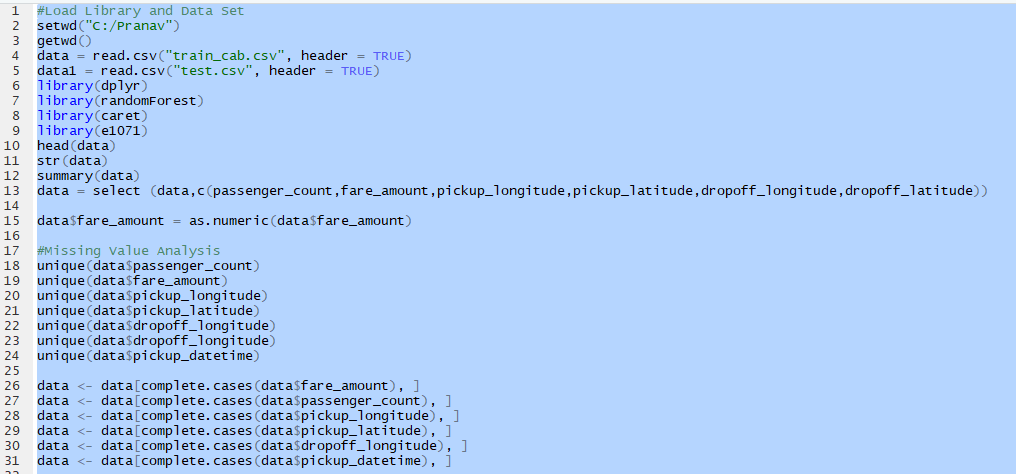
1. Enter the name of file after **python for python/Rscript for R** as entered below the code will start running

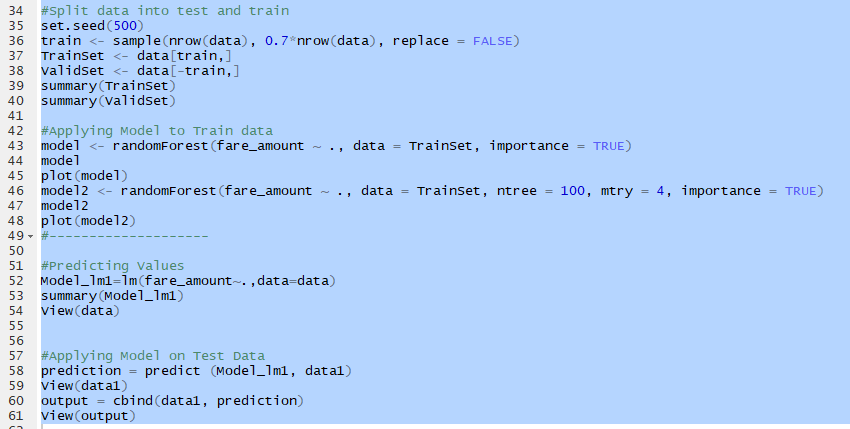






Appendix A – R code





Appendix B – Python Code

