**TAXI FARE PREDICTION**

**A MINI PROJECT REPORT on DATA SCIENCE**

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF

**Master of Computer Applications**

**SUBMITTED TO:**

**Dayananda Sagar College of Engineering, Department of MCA, Bangalore**

**SUBMITTED BY:**

ANUSHA JM 1DS18MCA50

CHANDARMOULI 1DS18MCA54

MANASA KC 1DS18MCA63

MANU KUMAR D 1DS18MCA64

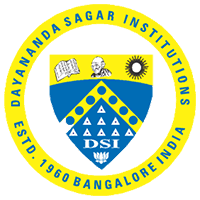
MANJUNATH R YADAWAD 1DS18MCA65

**NAME OF SUPERVISOR(S)**

Dr. Samitha Khaiyum, HOD- Dept. of MCA Mahendra Kumar B. Asst. Prof. Dept. of MCA

N G Rajeev ,SundarBabu R – Sr. Consultants, Enhance Knowledge Works

BillaVentakaSatyanarayana – Sr. Consultant and Technical Trainer



**Dayananda Sagar College of Engineering, Dept. of MCA, Bangalore**

**ACKNOWLEDGEMENT**

I would like to place on record my deep sense of gratitude to Prof. Dr. Samitha Khaiyum, HOD-Department Of MCA, Dayananda Sagar College of Engineering, India for his generous guidance, help and useful suggestions.

I express my sincere gratitude to Prof. Mahendra Kumar B, Dept. of MCA, Dayananda Sagar College of Engineering India, for his stimulating guidance, continuous encouragement and supervision throughout the course of present work.

**Signature(s) of Students**

**CERTIFICATE**

I hereby certify that the work which is being presented in the MCA, Mini Project Report entitled **“Taxi fare Prediction”,** in partial fulfillment of the requirements for the award of the **MCA** and submitted to the **Department of MCA, Dayananda Sagar College of Engineering, Bangalore** is an authentic record of my own work carried out during a period **July 2019**

The matter presented in this Project Report has not been submitted by me for the award of any other degree elsewhere.

**Signature of Student (S):**

This is to certify that the above statement made by the student(s) is correct to the best of my knowledge.

**Signature of Supervisor**

**Date:**

**Head of Department**

**MCA**

**TABLE OF CONTENTS**

**Contentspage no.**

Introduction 5

Defining the Problem 5

Collection of Data 5

Prepare the Data 5-7

Splitting the Data into Training and Testing 7

Algorithm Selection 8

Training the Algorithm with Data for Machine 8-9

Evaluate Test Data 9

Parameter Tuning 10

Start using your Model 12

References13

**Introduction:**

Taxi rate prediction deals about predicting the fare of a taxi trip. In this challenge we are given a training set of Taxi trips in the train data and few records in the test data. The goal of this challenge is to predict the fare of a taxi trip given information about the pickup and drop off locations, the pickup date time and number of passengers travelling.

In any analytics project 80% of the time and effort is spent on data cleaning, exploratory analysis and deriving new features. In this post, we aim to clean the data, visualize the relationship between variables and also figure out new features that are better predictors of taxi fare.

STEP 1: DEFINING THE PROBLEM

* Taxi rate prediction deals about predicting the fare of a taxi trip
* The goal of this challenge is to predict the fare of a taxi trip given information about the pickup and drop off locations, the pickup date time and number of passengers travelling.
* we aim to clean the data, visualize the relationship between variables and also figure out new features that are better predictors of taxi fare

STEP 2: COLLECTION OF DATA

- The data was collected from WHO and United Nations website.

- <https://www.kaggle.com/kumarajarshi/life-expectancy-who>

STEP 3: PREPARE THE DATA

* First you must upload the data to the jupyter notebook for analysis.



* Check are there any null rows or columns, if any remove the rows or fill the columns with the mean value of that particular column.
* ***Import all the packages and functions that are necessary for the analysis process.***



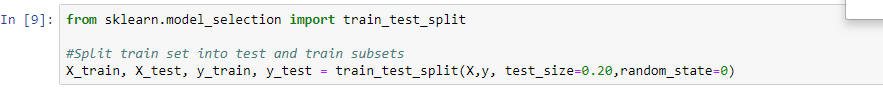
* We can analyze the data quality using scatter plots.

STEP 4: SPLITTING THE DATA INTO TRAINING AND TESTING

1. Here in this step we split the data into two separate tables training and testing respectively.
2. Training table is used for analysis and alteration purpose, whereas testing table is used to

test the data in the final step.

1. We will also split the table data into 80% and 20%.



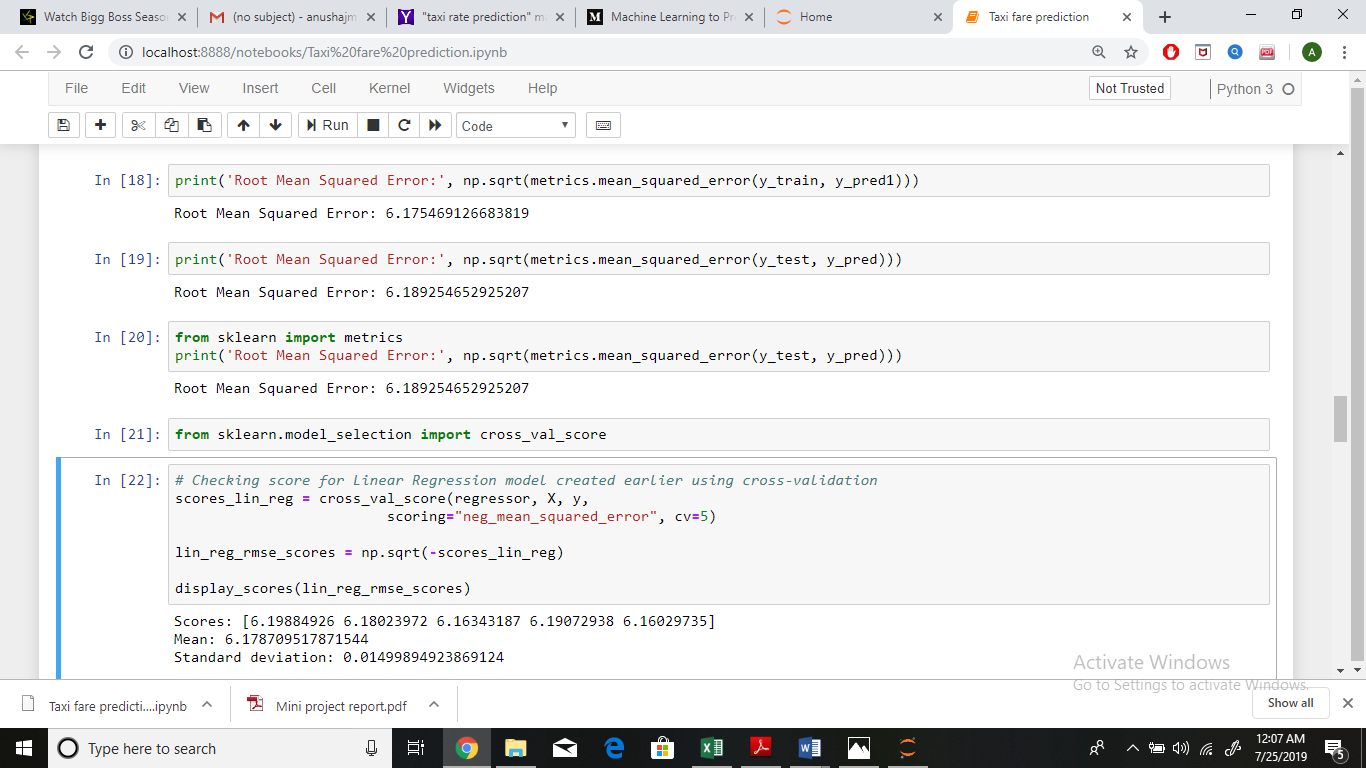
STEP 5: ALGORITHM SELECTION:

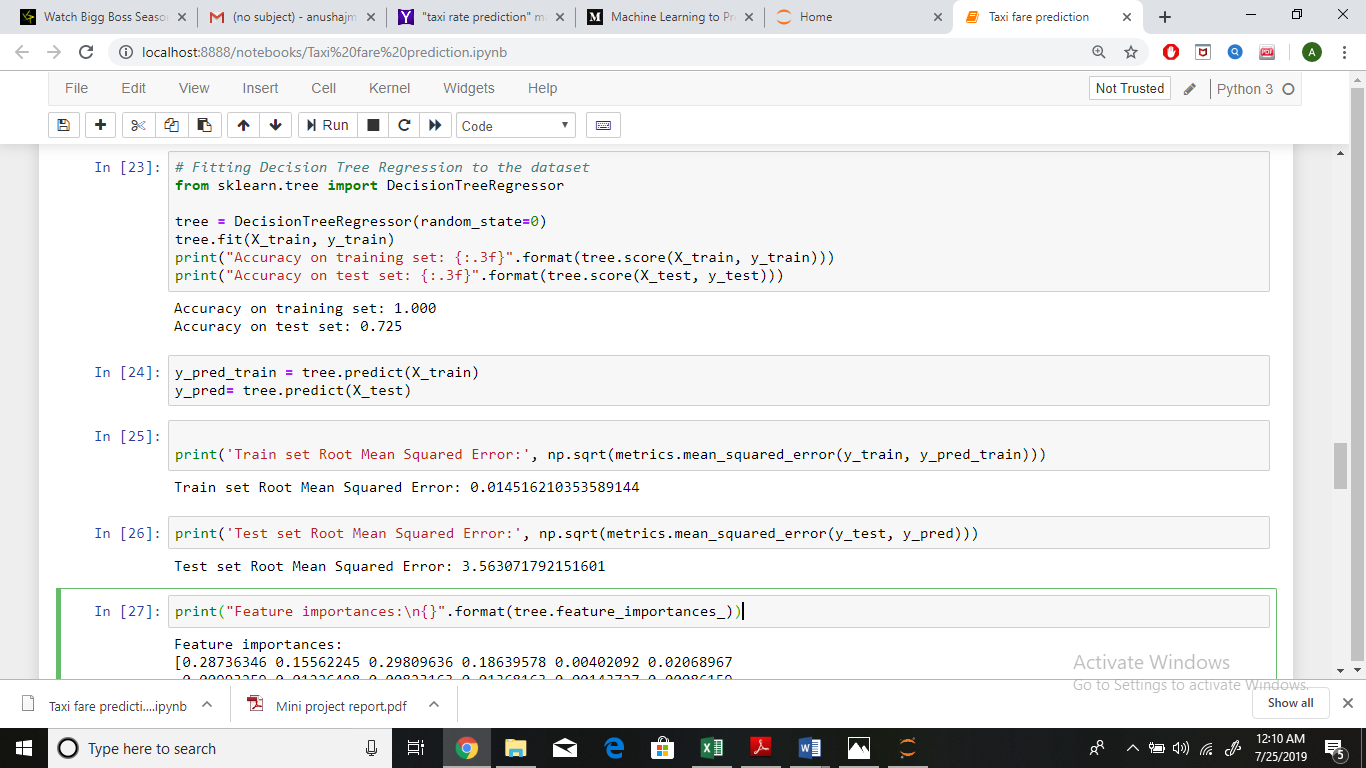
Here we select the appropriate algorithm/model that is necessary for the analysis purpose, we have selected the following models for processing the dataset.

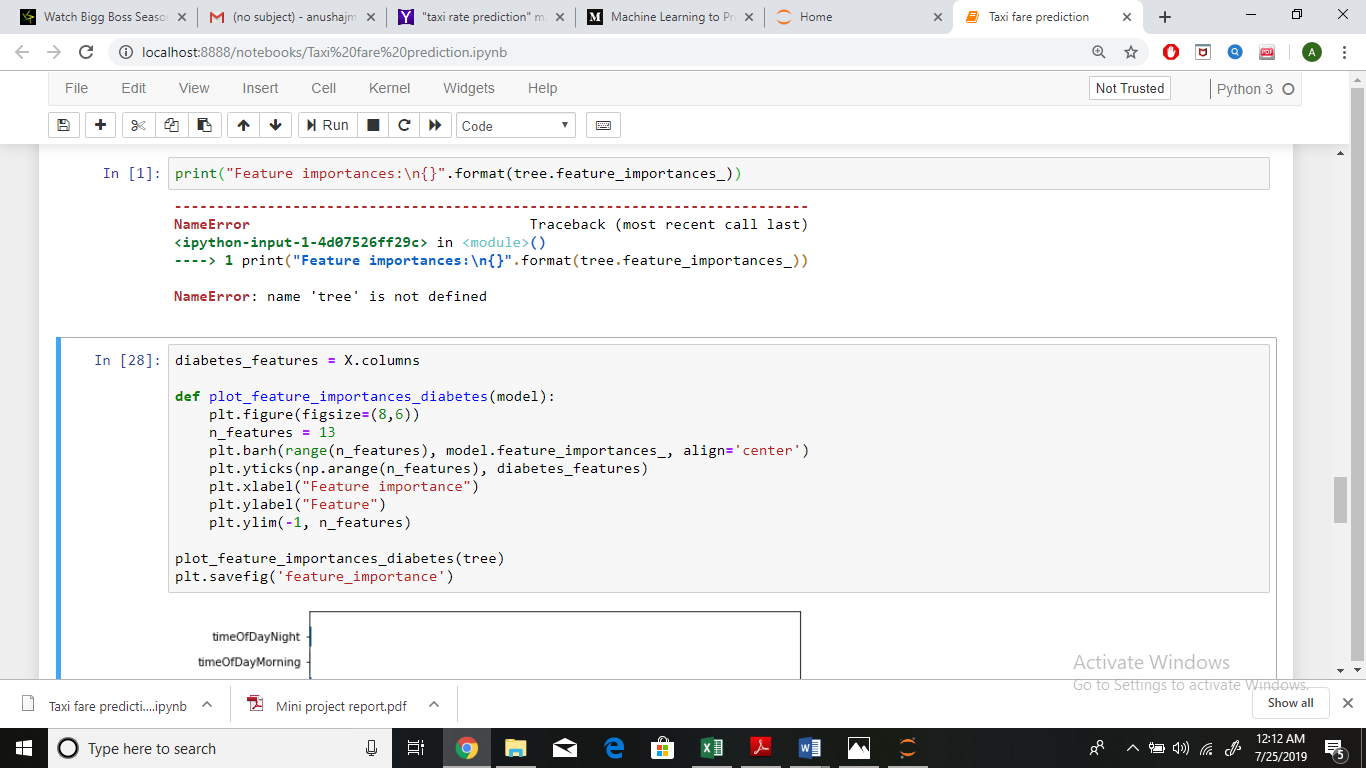
1. Linear Regression
2. Linear Regression with polynomial features
3. Decision Tree
4. Random Forest

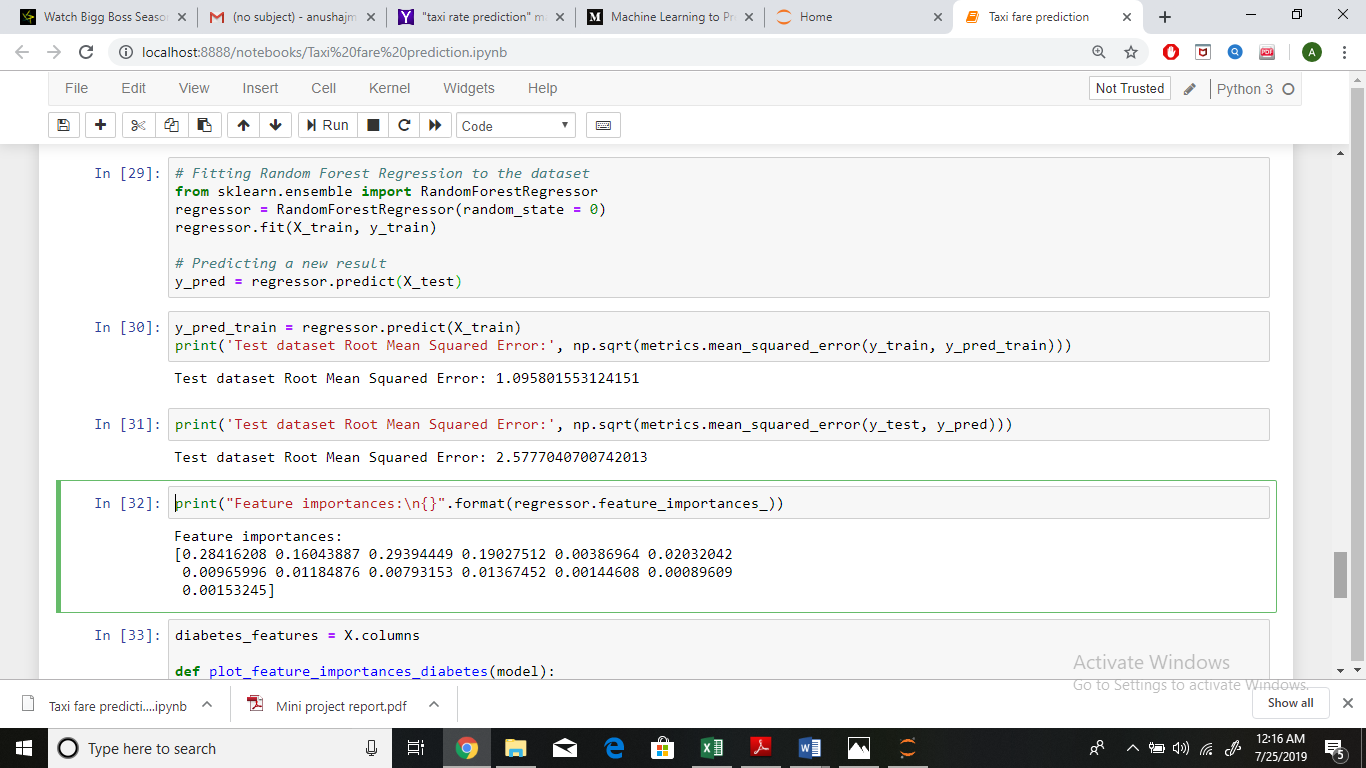
STEP 6: TRAINING THE ALGORITHM WITH DATA FOR MACHINE

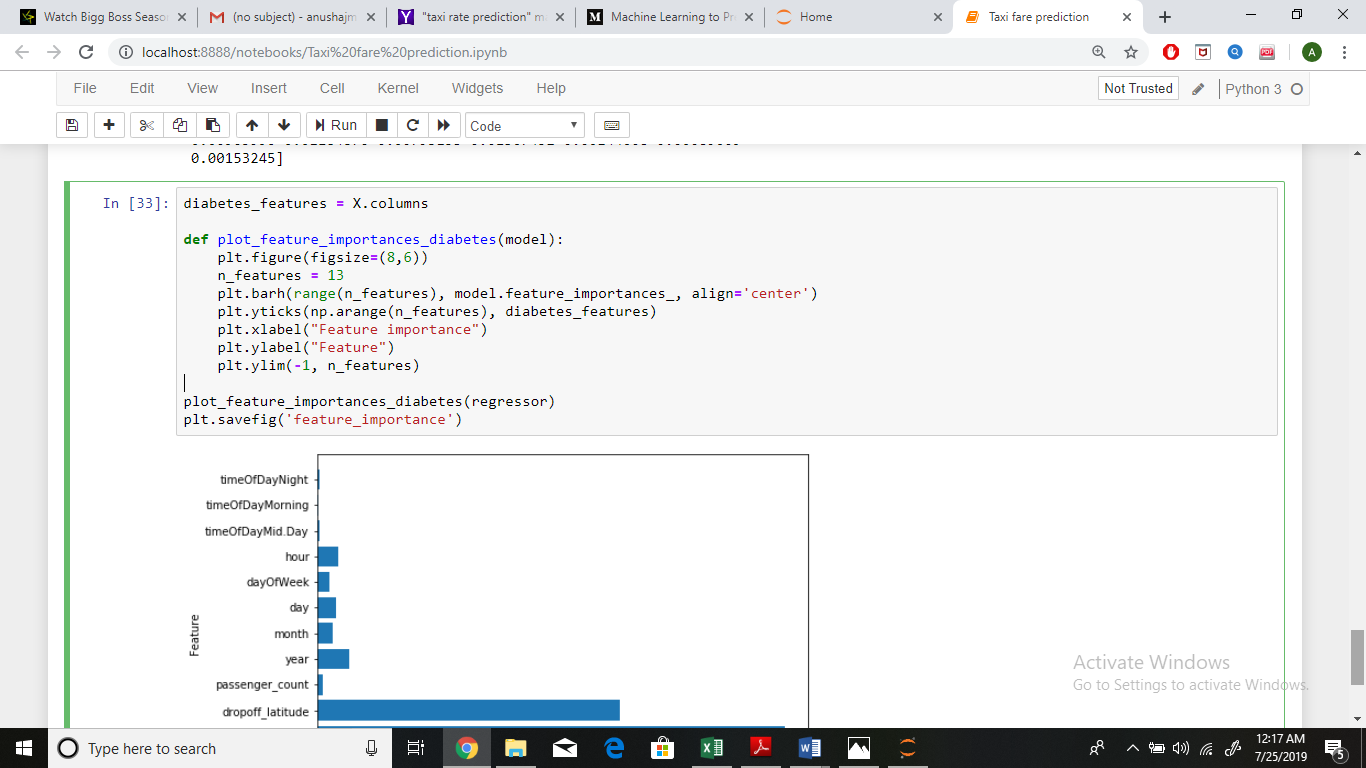
1. Here the data is divided into xtrain,ytrain,xtest and ytest where 80% of data is taken as train data and remaining 20% of data is taken as test data.
2. Then if there are any nulls in the rows or column, remove those and clean the data and
3. Also if you have any high range values take the mean of those and reduce the range of those values.
4. ***Then the algorithm is trained with 80% cleaned data of xtrain and ytrain***

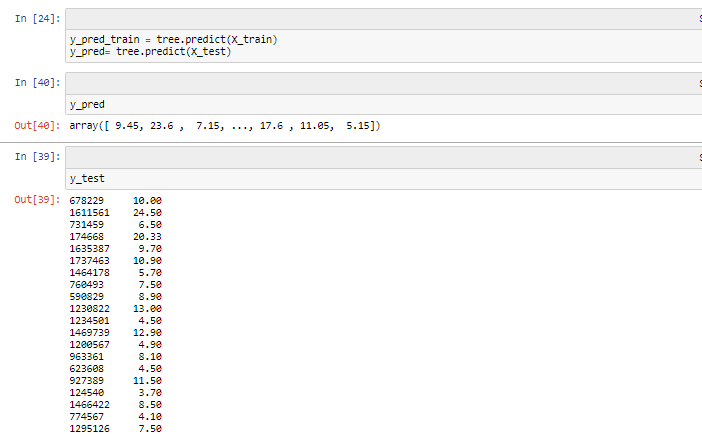


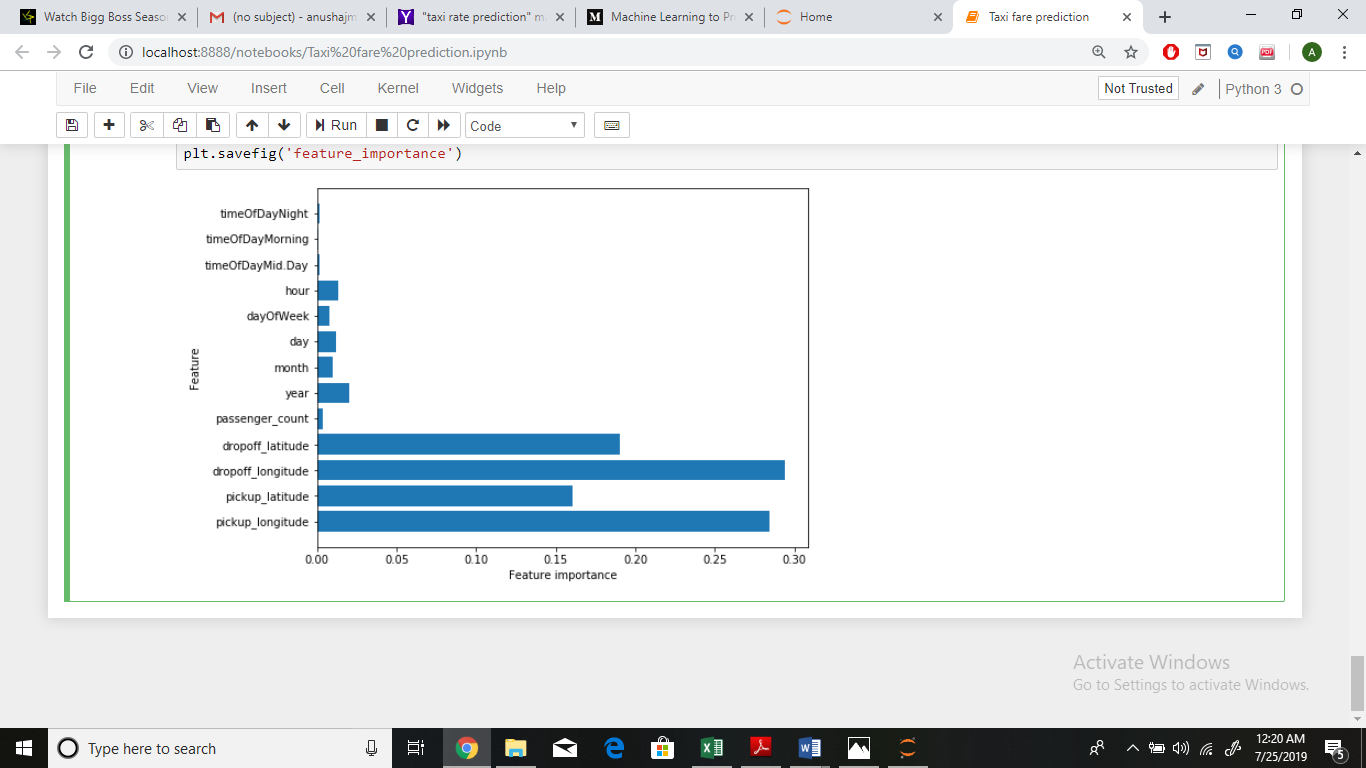




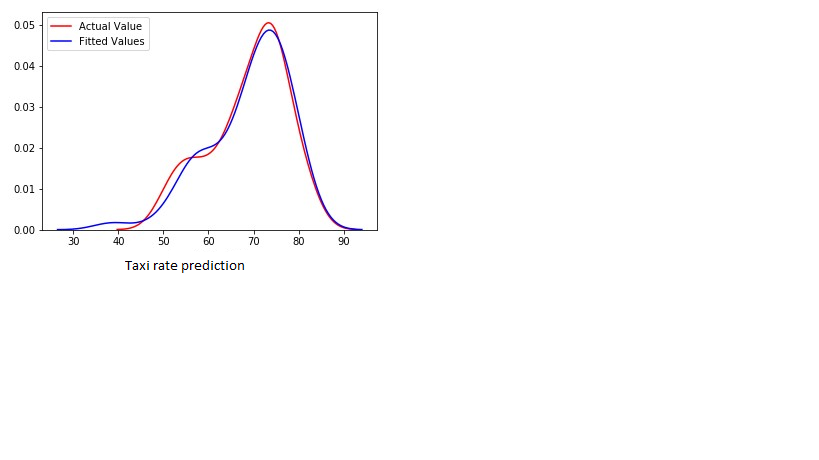


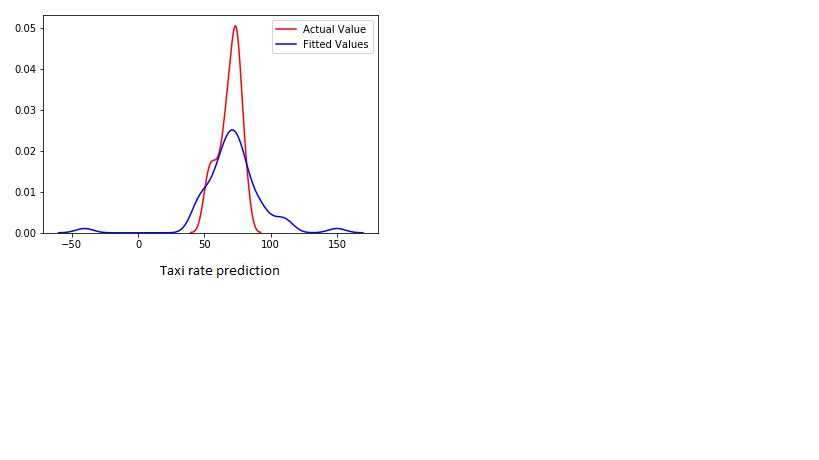


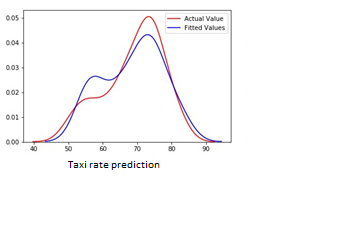


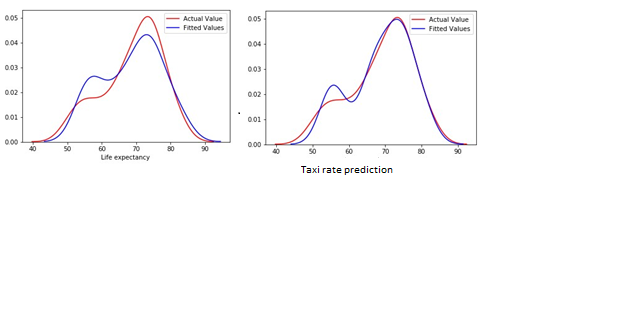


STEP 7: EVALUATE TEST DATA

* 1. After the algorithm is trained with 80% of data, the algorithm is to be tested with remaining 20% of data.
  2. Before that check there are any nulls in the test data, if present remove those nulls and
  3. Obtain clean data. Then evaluate the algorithm with test data for predictions.







STEP 8: PARAMETER TUNING

* 1. A tuning is to be done for algorithm in order to control the behavior of the algorithm.
  2. There are many tuning methods available, here we applied Random Forest.
  3. Random Forest is suitable for Taxi fare prediction because its most widely used predictive model.

STEP 9: START USING YOUR MODEL

* 1. After the algorithm is trained with test data, it gives the prediction that is it gives the accuracy of our model.
  2. If the prediction is below 70% then the model is failed, there may be a mistake in
  3. Choosing the data, cleaning the data or methods etc.
  4. If the prediction is above 70% then the model is good and ready for usage.

**REFERENCES:**

1. https://www.eurekalert.org/pub\_releases/2019-06/imi-ld062719.php

2. https://priceonomics.com/whats-the-difference-between-data-science-and/

3. https://www.analyticsindiamag.com/a-day-in-the-life-of-fractal-analytics-techie-who- wants-to-marry-economics-data-science/