*Cab Fare Prediction*

Ankush Saha

# 

Table of Contents

[1 Introduction 3](#_Toc14815828)

[1.1 Problem Statement 3](#_Toc14815829)

[1.2 Data 3](#_Toc14815830)

[2 Methodology 4](#_Toc14815831)

[2.1 Preparing Data 4](#_Toc14815832)

[2.2 Data Pre-Processing 4](#_Toc14815833)

[2.2.1 Missing Value Analysis 4](#_Toc14815834)

[2.2.2 Zero Value Analysis 5](#_Toc14815835)

[2.2.3 Outlier Analysis 5](#_Toc14815836)

[2.2.4 Feature Selection 5](#_Toc14815837)

[2.2.5 Sampling 5](#_Toc14815838)

[3 Model development 6](#_Toc14815839)

[3.1 Linear Regression 6](#_Toc14815840)

[3.2 Decision Tree 7](#_Toc14815841)

[3.3 Random Forest 7](#_Toc14815842)

[4 Model Evaluation 8](#_Toc14815843)

[4.1 RMSE 8](#_Toc14815844)

[5 Model Selection 9](#_Toc14815845)

[6 Steps To run the code 9](#_Toc14815846)

[6.1 In R 9](#_Toc14815847)

[6.2 In Python 10](#_Toc14815848)

[7 Summary 11](#_Toc14815849)

[8 Visualization 12](#_Toc14815850)

[8.1 Avg. fare Amount year wise 12](#_Toc14815851)

[8.2 Avg. fare Amount weekday wise 13](#_Toc14815852)

[8.3 count of fare year wise 14](#_Toc14815853)

[8.4 count of rides weekday wise 15](#_Toc14815854)

# 1 Introduction

## 1.1 Problem Statement

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

## 1.2 Data

Our task is to build a regression model which will predict the cab fare based on given pickup date & time, pickup point, drop-off location, passenger count. Given below is a sample of the data set that we are using to predict cab fare :

**Train\_cab:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| fare\_amount | pickup\_datetime | pickup\_longitude | pickup\_latitude | dropoff\_longitude | dropoff\_latitude | passenger\_count |
| 4.5 | 2009-06-15 17:26:21 UTC | -73.844311 | 40.721319 | -73.84161 | 40.712278 | 1 |
| 16.9 | 2010-01-05 16:52:16 UTC | -74.016048 | 40.711303 | -73.979268 | 40.782004 | 1 |
| 5.7 | 2011-08-18 00:35:00 UTC | -73.982738 | 40.76127 | -73.991242 | 40.750562 | 2 |
| 7.7 | 2012-04-21 04:30:42 UTC | -73.98713 | 40.733143 | -73.991567 | 40.758092 | 1 |
| 5.3 | 2010-03-09 07:51:00 UTC | -73.968095 | 40.768008 | -73.956655 | 40.783762 | 1 |
| 12.1 | 2011-01-06 09:50:45 UTC | -74.000964 | 40.73163 | -73.972892 | 40.758233 | 1 |
| 7.5 | 2012-11-20 20:35:00 UTC | -73.980002 | 40.751662 | -73.973802 | 40.764842 | 1 |
| 16.5 | 2012-01-04 17:22:00 UTC | -73.9513 | 40.774138 | -73.990095 | 40.751048 | 1 |

**Test:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| pickup\_datetime | pickup\_longitude | pickup\_latitude | dropoff\_longitude | dropoff\_latitude | passenger\_count |
| 2015-01-27 13:08:24 UTC | -73.97332001 | 40.76380539 | -73.98143005 | 40.74383545 | 1 |
| 2015-01-27 13:08:24 UTC | -73.98686218 | 40.71938324 | -73.99888611 | 40.73920059 | 1 |
| 2011-10-08 11:53:44 UTC | -73.982524 | 40.75126 | -73.979654 | 40.746139 | 1 |
| 2012-12-01 21:12:12 UTC | -73.98116 | 40.767807 | -73.990448 | 40.751635 | 1 |
| 2012-12-01 21:12:12 UTC | -73.966046 | 40.789775 | -73.988565 | 40.744427 | 1 |
| 2012-12-01 21:12:12 UTC | -73.960983 | 40.765547 | -73.979177 | 40.740053 | 1 |
| 2011-10-06 12:10:20 UTC | -73.949013 | 40.773204 | -73.959622 | 40.770893 | 1 |
| 2011-10-06 12:10:20 UTC | -73.777282 | 40.646636 | -73.985083 | 40.759368 | 1 |

# 2 Methodology

Any predictive modeling requires that we look at the data before we start modeling. First we prepare the data-set and we prepare the Data-set to feed into our model. In data preparation we perform actions like missing value analysis, outlier analysis, feature scaling, feature sampling.

## 2.1 Preparing Data

Before any kind of analysis we prepare the data. In this case we take the target variable fare\_amount from 1st column to last column for ease of making model

## 2.2 Data Pre-Processing

### **2.2.1 Missing Value Analysis**

Here we check if there is any missing value cell in the data set or not. If there is any missing value cell in the data set it may affect the model, and our model may not be efficient enough to predict correct value. So we need to either empty or impute those empty cells. If the number of empty cells in a column is more than 30 percent we can drop that variable as it may not add much value in our model. And if it is less than 30 percent we can impute them with either of any basic statistical process (i.e. mean, median, mode) or we can use KNN imputation method which one is best fit.

After our analysis we found out that In our data set there are 0.4% missing data which we can easily drop. So we have drop the missing value observations.

### **2.2.2 Zero Value Analysis**

In our dataset there are some zero values that does not make any sense. Like 0 longitude latitude does not make any sense as this location is of Ghana, and in some observations the passenger count is 0 which also does not makes any sense. So we have removed those observation as the number was very less.

### **2.2.3 Outlier Analysis**

Sometimes in a column of data set there are some values which do not comply with general behavior of other data. These data are called outliers. These values may manipulate the behavior of the data set. And our model may not work accurately. To tackle this kind of situation we can either delete the row or the impute the values with mean, median, mode or KNN imputation.

In our dataset we could not use boxplot analysis, as it would not give any sensible data. So using some common sense we have deleted some outliers like there was a time-date stamp of value ‘43‘ which makes no sense. And in a cab there cannot be more than 6 passengers if it is a shuttle. So we have removed those observation which has more than 6 passengers. And we have also removed the longitude, latitude points which are more than 360. And we also have removed the observation with fare amount ‘430-‘. After removing them the size of the data set was (15586X7).

### **2.2.4 Feature Selection**

There might be some variables who are highly dependent to each other. So keeping both of them in our data set may create some partiality towards some features. So we can remove one of the variables. To check the correlation of the variables we can plot them in a graph and we can visually see that how much they are dependent to each other.

In our dataset we could not delete any columns as all of them are important for prediction. What we can do is we can make them useful data for our model. Like pickup longitude latitude and drop-off longitude latitude. We can extract the distance between them and use the distance to feed in our model. And we can drop four columns of longitude and latitude, and we can drop those observation which has 0 distance.

### **2.2.5 Sampling**

First we need to divide our data set into two parts. First part on which we will develop our model and second part on which we will test our model how accurate it will predict.

So for sampling we have used random sampling technique which will randomly separate data into train and test data set for given size and for our data set we have taken 70% for train data and 30% for test data

After sampling train data size is (10799 X 4)

And test data size is (4629 X 4)

# 3 Model development

For prediction we need to develop a model by which we will predict the cab fare between two points. There are different types of model building technique available. For different kind of data set different techniques work accurately. So we will develop three models using three different techniques and will check which model is working more accurately, and we will feed our test data set to the best model and develop our prediction.

The three techniques we are goanna use are

1. Linear regression
2. Decision tree
3. Random Forest

## 3.1 Linear Regression

In statistics, the linear regression model is used to predict some value using values of other dependent variables. Behind the algorithm it calculates the relation between the dependent variables with independent variables and calculate co-efficient for those dependent variables. Then while prediction it put the values of the dependent variables in the formula and calculate the dependent variable

First it makes a formula depending on the independent variables to calculate the dependent variable. The formula is as shown below

Where y = predicted value

= coefficient of the nth independent variable.

It also calculate std. Errors, t value and p value.

Std error provides estimated error associated with the estimations.

t-value is (estimation)/(std. Error).

The absolute value of t-value must be less than std error.

The p-value is a probability value. The p-value will be always between 0 and 1.

Lower p-value indicates statistically significant effect of predictor on dependent variable.

## 3.2 Decision Tree

In [computer science](https://en.wikipedia.org/wiki/Computer_science), Decision tree learning uses a [decision tree](https://en.wikipedia.org/wiki/Decision_tree) (as a [predictive model](https://en.wikipedia.org/wiki/Predictive_modelling)) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). It is one of the predictive modeling approaches used in [statistics](https://en.wikipedia.org/wiki/Statistics), [data mining](https://en.wikipedia.org/wiki/Data_mining) and [machine learning](https://en.wikipedia.org/wiki/Machine_learning). Tree models where the target variable can take a discrete set of values are called classification trees; in these tree structures, [leaves](https://en.wikipedia.org/wiki/Leaf_node) represent class labels and branches represent [conjunctions](https://en.wikipedia.org/wiki/Logical_conjunction) of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically [real numbers](https://en.wikipedia.org/wiki/Real_numbers)) are called regression trees.

## 3.3 Random Forest

Random forests or random decision forests are an [ensemble learning](https://en.wikipedia.org/wiki/Ensemble_learning) method for [classification](https://en.wikipedia.org/wiki/Statistical_classification), [regression](https://en.wikipedia.org/wiki/Regression_analysis) and other tasks that operates by constructing a multitude of [decision trees](https://en.wikipedia.org/wiki/Decision_tree_learning) at training time and outputting the class that is the [mode](https://en.wikipedia.org/wiki/Mode_(statistics)) of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of [overfitting](https://en.wikipedia.org/wiki/Overfitting) to their [training set](https://en.wikipedia.org/wiki/Test_set).

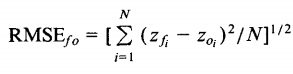
# 4 Model Evaluation

For Model Evaluation We use various techniques. We are building a regression model. For regression model we use methods like MAPE, MAE, RMSE.

In our dataset there is a column of time and date so it can be considered as time series analysis. And for time series Evaluation we RMSE is appropriate.

## 4.1 RMSE

Root Mean Square Error (RMSE) is the standard deviation of the [residuals](https://www.statisticshowto.datasciencecentral.com/residual/) (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the [line of best fit](https://www.statisticshowto.datasciencecentral.com/line-of-best-fit/). Root mean square error is commonly used in climatology, forecasting, and [regression analysis](https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/regression-analysis/) to verify experimental results.



In R RMSE for linear regression = 798.039, for decision tree = 799.841, and for Random forest = 90.24

In Python for linear regression = 13.28, for decision tree = 64.15, and for Random forest = 197.99

# 5 Model Selection

From all of the test we can clearly say that Random Forest is the best fit for the model in R and linear regression is best fit for the model in python. As its RMSE is very low from the other models.

So we will use Random Forest model for predicting cab fare of our test data when we will use R and we will use linear regression when we will use python.

# 6 Steps To run the code

Below are the instruction to keep in mind to run the code.

1. Download the files and keep them in a folder and get th file path.
2. In R you have to press ctrl+Enter to execute a line of code. You can run multiple lines together by selecting the line and pressing ctrl+Enter.
3. In python you have to press shift+Enter to run a cell. You have to run a cell each at a time. The python program is written in Jupyter notebook.

Please follow the Below Steps to predict the test data. Before performing the steps.

## 6.1 In R

1. Open RStudio and open the file: “Project\_2.R”.
2. First run line 1 to 13 to install all the required library.
3. in 15th line of command you have to give your file path between the quotation ex: “C:/Users/Ankush Saha/Documents/Project\_2”.
4. Then run line 18&19 to load the train data and take a backup of it. ***You can run line 19 after every step to take the updated backup.***
5. Run line 23 to 31 to adjusting the columns.
6. Now you can test the data set for missing values by running line 37 to 43. And we can remove the missing values in line 46 & 47.
7. And also we can test the data set for zero value by running line 51 to 57. And we can remove the zero values in line 59 to 64.
8. Run 68th to 75th line to remove outliers.
9. Run line 88 to 94 to change the longitude and latitude to make it meaningful distance and make the dataset feed able
10. For dividing the data set into train and test data run line no. 98 to 101.
11. Now to develop Linear Regression model rum 107th line. It will save the model in **lm\_model.**
12. And to get the predicted data run 108th line.
13. Now get the model evaluation in 112th line.
14. You can also develop develop Decision tree and Random forest models. From this file. run those lines as instructed in 12th and 13th point.
15. ***In any of these steps if you have done anything wrong just run line 20 to get the last backup. It will avoid rerunning the whole code from start, and it will save time.***
16. Now to test it on our test data we first have to load the test file and prepare it to feed it into the model. To achieve this run 131st line and ***take a backup by running line 132.***
17. Now as we have made the model to feed a proper data set, we have make the test data feed able, and by running line 134 to 179, we can achieve this.
18. Now run line 182 to 187 to predict the values and make it usable.
19. Now finally run line 190 to save the data set in csv format in your hard-disk.

## 6.2 In Python

1. Open Jupyter notebook and open “Project\_1.ipynb” from the folder.
2. Run cell 1 to install required libraries.
3. Run cell 2 by giving your file location ex: “C:/Users/Ankush Saha/Documents/Project\_2”. to set your location.
4. Run cell 3 & 4 to load the train data set and take a backup. ***You can run cell 4 after every step to take the updated backup.***
5. You can check your data by running cell 5.
6. Run cell 6 to prepare your data.
7. You can run cell 7 to check missing value.
8. Run cell 8 to drop the missing values
9. Run cell 9 to check Zero values in dataset.
10. To remove zero values and outliers run cell 10 and 11.
11. Run cell 12 to convert pickup date-time to proper form.
12. Run Cell 13 & 14 to convert longitude latitude to distance and add them to dataset.
13. For dividing the data set into train and test data run cell 15.
14. We need to define function RMSE for model evaluation and we can achieve it by running cell 16.
15. Now to develop Linear Regression model run cell 17. It will save the model in **model** and the output of the cell will be RMSE for that model
16. To develop Decision Tree model run cell 18. It will save the model in **model2** and the output of the cell will be RMSE for that model
17. To develop Random Forest model run cell 19 & 20. It will save the model in **model3** and the output of the cell will be RMSE for that model
18. ***In any of these steps if you have done anything wrong just run cell 4 to get the last backup by commenting the first line uncommenting the second line. It will avoid rerunning the whole code from start, and it will save time.***
19. Now to test it on our test data we first have to load the test file and prepare it to feed it into the model. To achieve this run cell 21 and do the missing value analysis in cell 22 & 23.
20. Run cell 24, 25 & 26 to prepare a backup data to feed into the model in proper form.
21. Now run cell 27, 28, 29 & 30 to predict the values and make it usable.
22. Now finally run cell 31 to save the data set in csv format in your hard-disk.

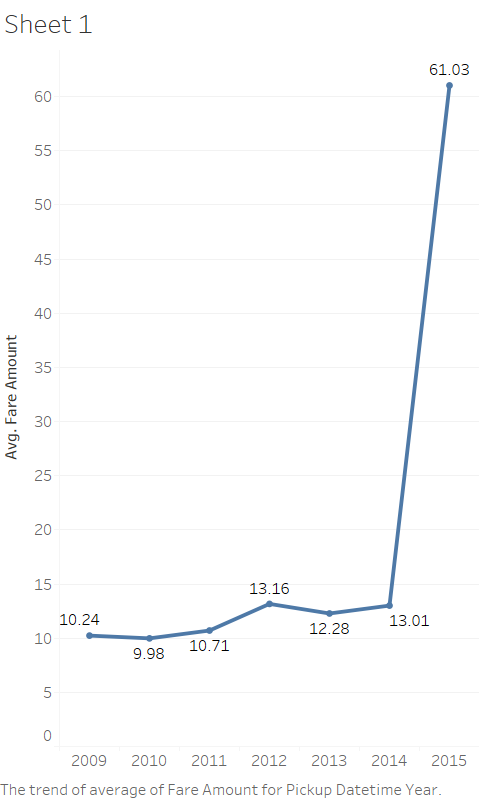
# 7 Summary

In this project we have built a model which will predict cab fare between two points of city. It does not give quite accurate result as the dataset with which we have trained our model is not accurate. Like in train data set it is given for 5427 kms fare was $8.5 in 2012 but in 2011 for 127 kms fare is $11.30. There might be some other factors on which the fare depends like last ride’s cancellation charge, Car Model e.t.c.

We can revise our train dataset and make them correct then after building model on that dataset our model’s RMSE would be very low and it will predict quite accurately.

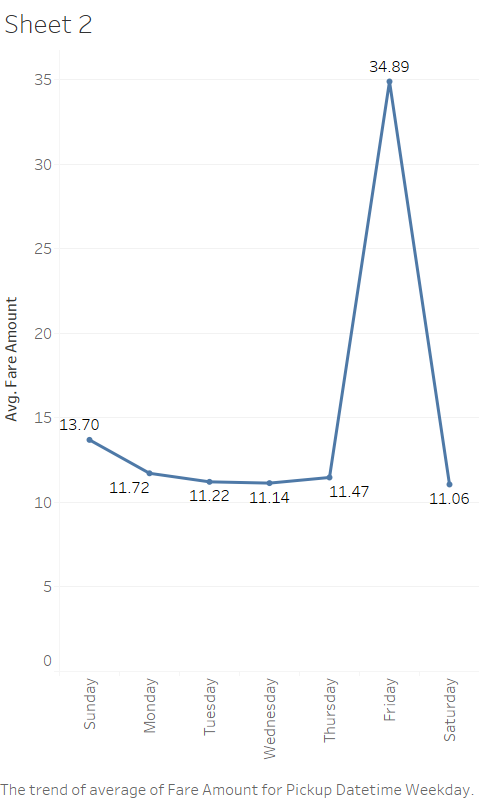
# 8 Visualization

## 8.1 Avg. fare Amount year wise



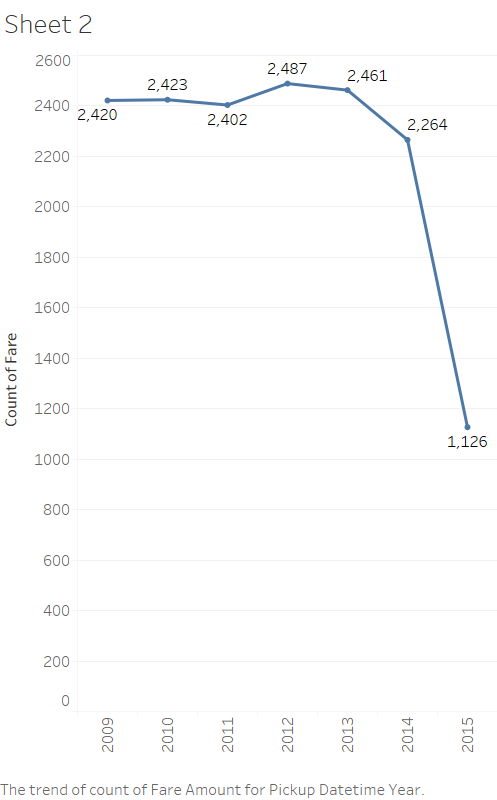
We can clearly see that the average fare amount 2015 was much higher than prior years.

## 8.2 Avg. fare Amount weekday wise



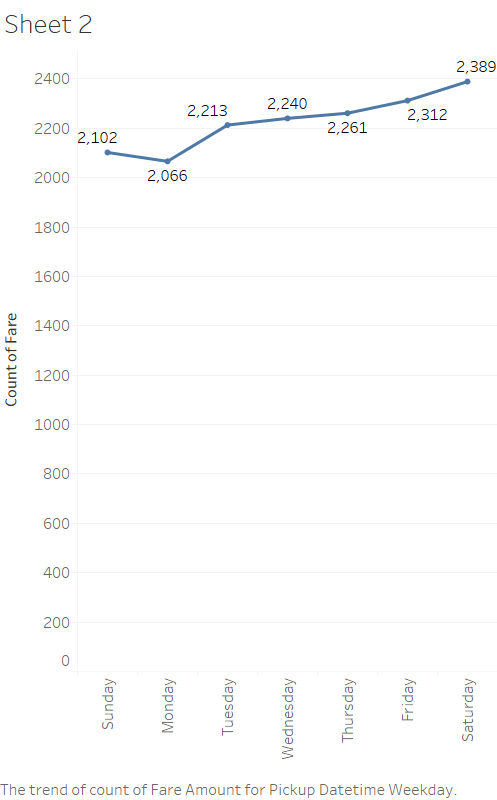
Here we can clearly see that the rate rides on Fridays and it goes down from Sunday to Wednesday.

## 8.3 count of fare year wise



As the avg. fare has increased drastically in 2015 the no. of rides has reduced drastically. We have to control our fare to maintain avg. no. of rides.

## 8.4 count of rides weekday wise



People tends to book more cabs in weekend. So we can use surge charge during weekend to make profit.