**EDWISOR**

Project: Prediction of   fare amount for a cab ride in the city.

Submitted by: Anurag Pratap Singh Chauhan

Date: 18/05/2019

**CONTENTS**

**1. INTRODUCTION**

1.1 Problem Statement

1.2 Data

**2. METHODOLOGY**

2.1 Pre-processing

2.1.1 Missing Value Analysis

2.1.2 Outlier Analysis

2.1.3 Conversion of Independent Variables

2.1.4 Feature Selection

2.1.5 Feature Scaling

2.2 Modeling

2.2.1 Model Selection

2.2.2 Evaluating Model Selection

2.2.3 Multiple Linear

2.2.4 Decision Tree

2.2.5 Random Forest

**3. CONCLUSION**

3.1 Model Evaluation

3.2 Model Selection

**REFRENCES**

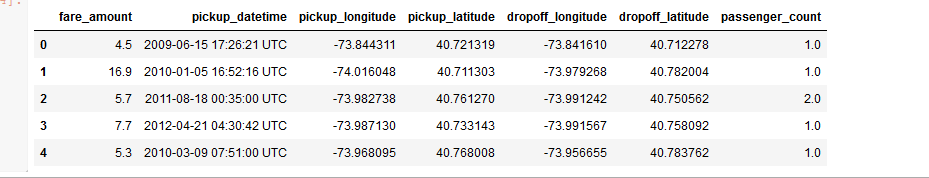
**1. INTRODUCTION**

1. **Problem Statement**

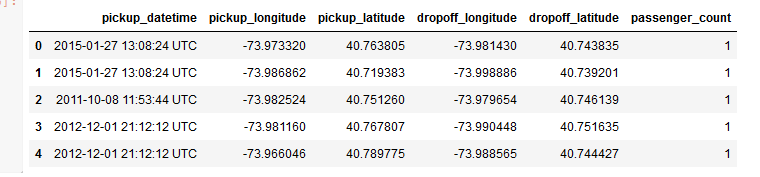
Our file “test\_cab.csv” contains the fare of the cab that is charged along with the pickup and drop latitudes and longitudes. Also we are given the date and time of the travel so need to develop a model that would help us predict the fares for future. We need to predict the cab fares for our test file which is “test.csv”.

**1.2 Data**

Our task is to build Regression model which will give the fare amount of the cab based on various factors such as distance , month of the year, week day, year. Given below is a sample of the data set that we are using to predict the cab fare.



And our test file for which the cab fare is to be predicted is as follows:



**Below are the variables we used to predict the count of cab fare as per our modified data:**

|  |  |
| --- | --- |
| **S.No** | **Variables** |
| **1** | Passenger Count |
| **2** | Pickup DateTime |
| **3** | Pickup Latitude |
| **4** | Pickup longitude |
| **5** | Dropoff latitude |
| **6** | Dropoff Longitude |

Keeping these variables in mind we are going to develop our model using various pre-processing techniques, predictive analysis and model evaluation to find the relationship between these variables and the target variable which is monthly rental bike counts.

**2.METHODOLOGY**

**2.1 Pre-processing**

Data preprocessing describes any type of processing performed on [raw data](https://searchdatamanagement.techtarget.com/definition/raw-data) to prepare it for another processing procedure. Commonly used as a preliminary [data mining](https://searchsqlserver.techtarget.com/definition/data-mining) practice, data preprocessing transforms the data into a format that will be more easily and effectively processed for the purpose of the user.

It 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 and look at all the distributions of the Numeric variables. Most analysis like regression, require the data to be normally distributed.

**2.1.1 Missing Value Analysis**

Missing values in data is a common phenomenon in real world problems. Knowing how to handle missing values effectively is a required step to reduce bias and to produce powerful models.

Below table illustrates that missing values are present in the data.

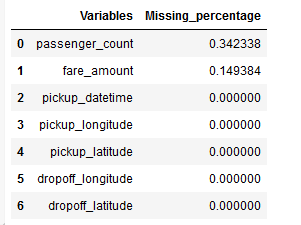
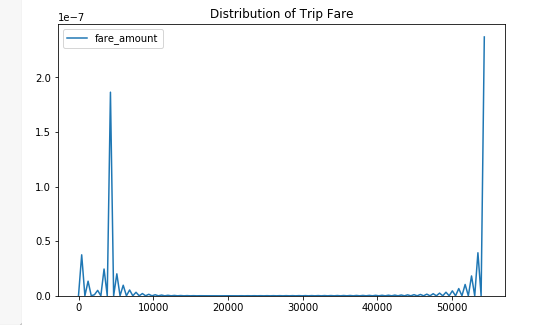


FIG 2.9: Missing Value analysis.

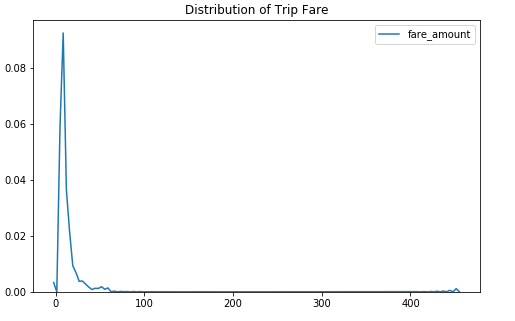
This missing data has to be taken care of since there are very high undulations in the data in few of the data sets, we would go for **median imputation** so that the median values take up the missing places.

**2.1.2 Outlier Analysis**

In statistics, an outlier is an observation point that is distant from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set. An outlier can cause serious problems in statistical analyses. Since most of the variables are categorical therefore we have removed the outliers taking different independent variables one at a time.

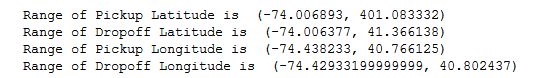


Above shown is the distribution of the trip fare, clearly we can see some outliers here, therefore we have removed the negative values from the data set. Also we have taken care of the fares greater than 500 by removing them.

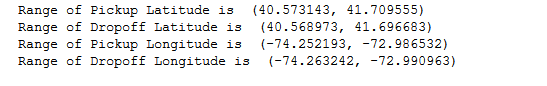
The following is the trip fare distribution after treating the outliers.

Apart from the fare amount pickup and drop-off latitudes and longitudes also have some incorrect data sets therefore we need to remove them too.

The figure below shows us the values of the pickup and drop-off latitudes and longitudes as follows

:

So after removing the outlying values which we have analyzed using the filters option in the “test.csv” file (by using the filters option) we have new range of the given latitudes and longitudes.

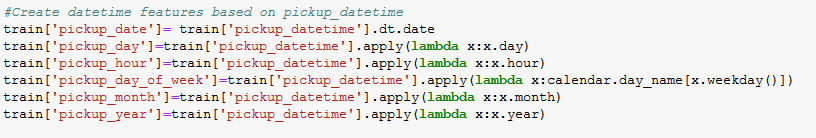


Passenger count also shows some very high values such as 5435 in many cases so we have removed the value greater than 6 as a cab cannot accommodate more than 6 people at a time. There were 6 such data sets therefore we can get rid of them.

**2.1.3 Conversion of independent factors**

**A) Pickup\_datetime to its components.**

The first step to analyze how the fares have changed over time, is to create features like hour, day of the week, day, month, year from pickup datetime. The code to extract these features is as below:



After conversion we can relate the relationship between various independent factors also we can relate these parameters with our dependent variable which is cab fare. For example, we have shown couple of plots below:

Fig: Number of trips vs Pickup Year

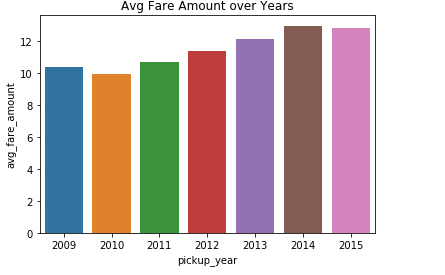


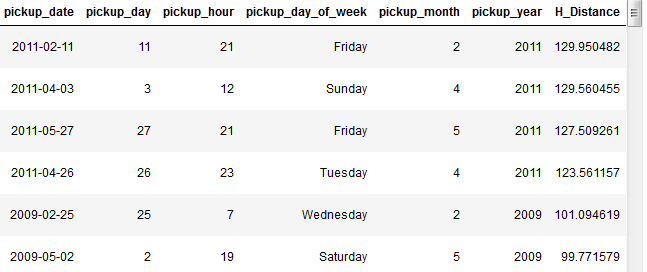
Fig: Average fare amount vs Pickup Year

**B) Pickup and drop-off latitudes and longitudes to Distance.**

Using the pickup and drop-off coordinates we calculate the trip distance in miles based on **Haversine Distance.**We will calculte the total distance travelled on the sphere using the haversine distance/ great circle distance.

Formulas used to calculate(phi = latitudes of the 2 pts, lambda = longitudes): haversine(θ) = sin²(θ/2)  
a = sin²(φB - φA/2) + cos φA \* cos φB \* sin²(λB - λA/2)  
c = 2 \* atan2( √a, √(1−a) )  
d = R \* c

After the above two conversions we have the following new variables in the data set:



**2.1.4 Feature Selection**

Before performing any type of modeling we need to assess the importance of each predictor variable in our analysis. There is a possibility that many variables in our analysis are not important at all to the problem of class prediction. There are several methods of doing it.

Machine learning works on a simple rule – if you put garbage in, you will only get garbage to come out. By garbage here, I mean noise in data.

This becomes even more important when the number of features are very large. You need not use every feature at your disposal for creating an algorithm. You can assist your algorithm by feeding in only those features that are really important. I have myself witnessed feature subsets giving better results than complete set of feature for the same algorithm or – “Sometimes, less is better!”.

We should consider the selection of feature for model based on below criteria

1. The relationship between two independent variable should be less and
2. The relationship between Independent and Target variables should be high.

Below figure illustrates that relationship between all numeric variables using Corrgram plot.

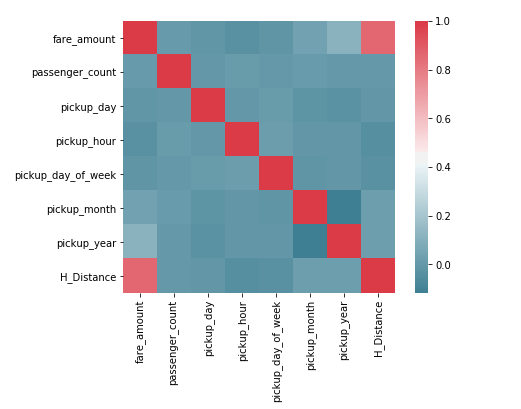


FIG 2.11 Correlation plot of numeric variables

The following inference can be made from the above correlation plot.

a) none of the dependent variables are negatively correlated with each other or with the dependent variable.

b) H\_distance is highly related with the fare\_amount.

c) there is no multi- co linearity between the independent variables. Hence we would not go for dimensional reduction we will consider all the variables in our analysis.

**2.1.5 Feature Scaling**

The word “normalization” is used informally in statistics, and so the term normalized data can have multiple meanings. In most cases, when you normalize data you eliminate the units of measurement for data, enabling you to more easily compare data from different places. Some of the more common ways to normalize data include:

[Rescaling data](http://www.statisticshowto.com/what-is-rescaling-data/) to have values between 0 and 1.  This is usually called feature scaling. One possible formula to achieve this is.  
[http://www.statisticshowto.com/wp-content/uploads/2015/11/normalize-data.png](http://www.statisticshowto.com/wp-content/uploads/2015/11/normalize-data.png)

But in our data set we have no such requirement as the data we have is in proper limits and we need to scale the data, as it could affect our model adversely.

**2.2.Modelling**

**2.2.1 Model Selection**

In out earlier stage of analysis we have come to understand that few variables like “Year”, “Haversine Distance” , “hour of the day”, “weekday” are going to play key role in model development , for model development dependent variable may fall under below categories

1. Nominal
2. Ordinal
3. Interval
4. Ratio

In our case dependent variable is interval so, the predictive analysis that we can perform is Regression Analysis

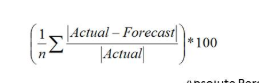
We will start our model building from Multiple Linear Regression

**2.2.2 Evaluating Regression Model**

The main concept of looking at what is called **residuals** or difference between our predictions f(x[I,]) and actual outcomes y[i].

We are using two methods to evaluating performance of our model.

1. **MAPE** : (Mean Absolute Percent Error) measures the size of the error in percentage terms. It is calculated as the average of the unsigned percentage error.



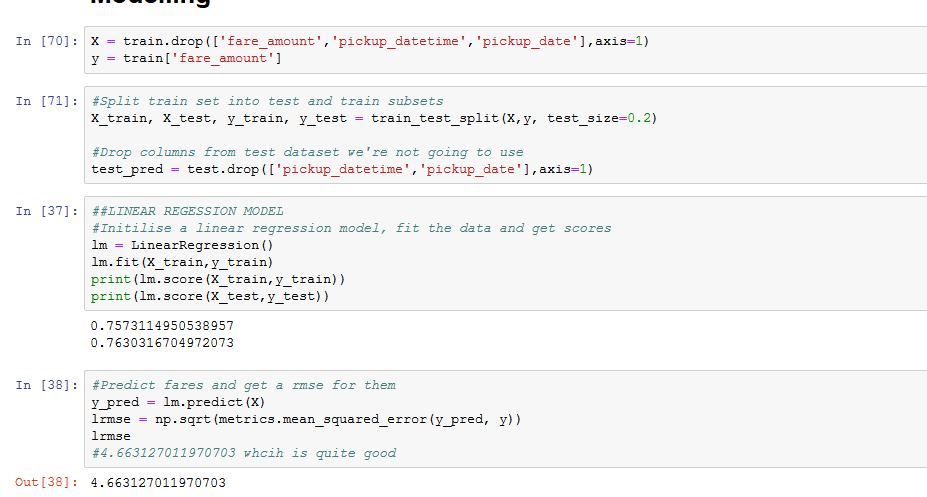
1. **RMSE :**(Root Mean Square Error) is a frequently used measure of the difference between values predicted by a model and the values actually observed from the environment that is being modelled. We are going to use RMSE for our model evaluation because our data is time based.



**2.2.3 Linear Regression**

[Multiple linear regression](http://www.statisticssolutions.com/academic-solutions/membership-resources/member-profile/data-analysis-plan-templates/data-analysis-plan-multiple-linear-regression/) is the most common form of linear regression analysis.  As a predictive analysis, the multiple linear regressions are used to explain the relationship between one continuous dependent variable and two or more independent variables.  The independent variables can be continuous or categorical.

As Linear regression will work well if multicollinearity between the Independent variables are less.

****

**FIG 3.7 : Development of Linear Regression Model.**

**2.2.4 Decision Tree**

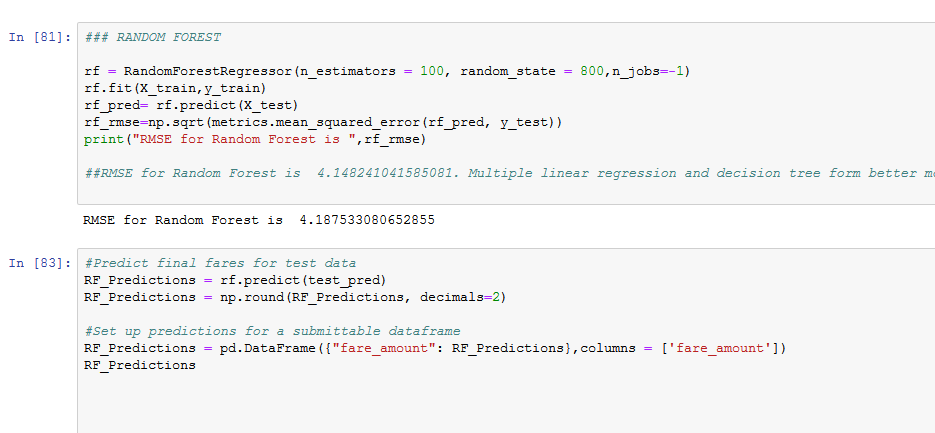
A tree has many analogies in real life, and turns out that it has influenced a wide area of **machine learning**, covering both **classification and regression**. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. As the name goes, it uses a tree-like model of decisions.



**FIG 3.1 Decision Tree Model**

**2.2.5 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 operate 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.



**FIG 3.5 Development of Random Forest**

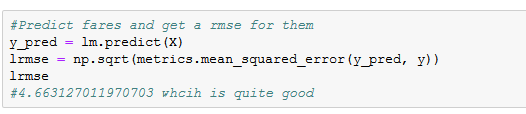
**3.CONCLUSION**

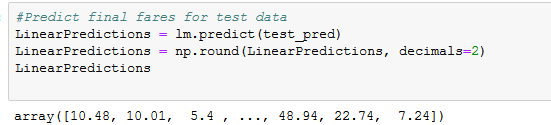
**3.1 Model Evaluation**

Now that we have a few models for predicting the target variable, we need to decide which one to choose. There are several criteria that exist for evaluating and comparing models. We can compare the models using Root Mean Squared Error because it is time based data and RMSE is suitable for such data types.

We have calculated the RMSE for all the three models and we will compare them to select the best model for cab rental prediction.

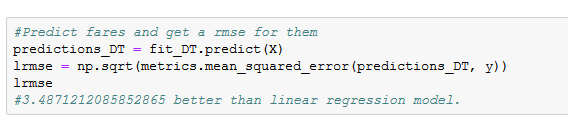
**a) Multiple Linear Regressions**

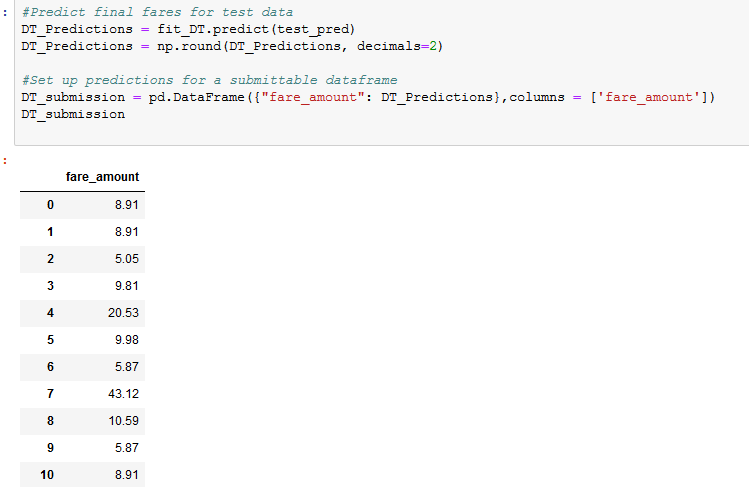
****

****

RMSE value for multiple linear regressions is 4.663 and the predicted values for our test data is shown in figure above.

**b) Decision Tree**

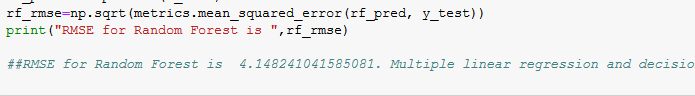
****



RMSE value for Decision tree is 3.48 and the predicted cab fare values for the test file are shown above.

Also Decision tree model has least RMSE value.

**c) Random Forest Regression**



****

RMSE value for Random Forest is 4.14 and the predicted cab fare values for the test file are shown above.

**3.2 Model Selection**

We can see that both models perform comparatively on average and therefore we can select any of the three models without any loss of information.

**References**

* [**WWW.EDWISOR.COM**](http://WWW.EDWISOR.COM)
* [**WWW.ANALTICSVIDHYA.COM**](http://WWW.ANALTICSVIDHYA.COM)
* **WWW.GOOGLE.COM**
* **WWW.YOUTUBE.COM**