CAB RENTAL PREDICTION

Shivam Agarwal

12/05/2019

Contents

1. **Introduction 3**
   1. Problem Description 3
   2. Problem Statement 3
   3. Data 3
   4. Performance Metric 4
2. **Methodology 5**
   1. Exploratory Data Analysis 5
      1. Data Visualization 5
         1. Univariate Analysis 5
         2. Bivariate Analysis 8
      2. Data Preparing and Cleaning 11
         1. Missing Value Analysis 11
         2. Outlier Analysis 12
         3. Feature Engineering 12
         4. Feature Selection 13
      3. Modelling 13
         1. Linear Regression 13
         2. Decision Tree 14
         3. Random Forest 14
3. **Conclusion 15**
   1. Model Evaluation 15
      1. Root Mean Square Value 15
   2. Model Selection 16

**References 17**

Chapter 1

**Introduction**

* 1. **Problem Description**

In this project we have to predict the fare of the cab per trip using the pickup and drop-off longitudes and latitudes, number of passengers at the different time around the year.

**1.2. 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 a cab ride

in the city.

**1.3. Data**

The data is a Time-Series data but instead we will approach it as Regression Problem. Our task is to build a regression model which will predict the fare amount per trip using different latitudes and longitudes and some other variables.

Table 1.1: Cab Rental Attribute (Columns 1-7)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| fare\_amount | pickup\_datetime | pickup\_longitude | pickup\_latitude | dropoff\_longitude | dropoff\_latitude | passenger\_count |
| -3.00 | 2013-08-30 08:57:10 UTC | -73.995062 | 40.740755 | -73.995885 | 40.741357 | 4.0 |
| -2.90 | 2010-03-09 23:37:10 UTC | -73.789450 | 40.643498 | -73.788665 | 40.641952 | 1.0 |
| -2.50 | 2015-03-22 05:14:27 UTC | -74.000031 | 40.720631 | -73.999809 | 40.720539 | 1.0 |
| 0.00 | 2010-02-15 14:26:01 UTC | -73.987115 | 40.738808 | -74.005911 | 40.713960 | 1.0 |
| 0.01 | 2015-05-01 15:38:41 UTC | -73.939041 | 40.713963 | -73.941673 | 40.713997 | 1.0 |

As we can see in the table above, we have we have the following 6 variables, using which we have to correctly predict the ‘fare\_amount” on each trip of the cab.

Table 1.2: Predictor Variables

|  |  |
| --- | --- |
| S.No. | Predictors |
| 1. | pickup\_datetime |
| 2. | pickup\_longitude |
| 3. | pickup\_latitude |
| 4. | dropoff\_longitude |
| 5. | dropoff\_latitude |
| 6. | passenger\_count |

**1.4 Performance Metric**

RMSE : Root Mean Square Error (RMSE) is the standard deviation of the residuals (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.

Also, since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. So, RMSE becomes more useful when large errors are particularly undesirable. So, Roost Mean Square value seems like a perfect choice for our problem at hand.

Chapter 2

**Methodology**

**2.1 Exploratory Data Analysis**

Exploratory Data Analysis (EDA) is the first step in our data analysis process. We do this by taking a broad look at patterns, trends, outliers, unexpected results and so on in our existing data, using visual and quantitative methods to get a sense of the story this tells. To start with this process, we will first have a look at univariate analysis like plotting Box plot and whiskers for individual features, Histogram plots, Bar plots and Kernel Density Estimation for the same for the same. Then we will proceed to Multivariate analysis like Bar and Histogram and Bar plot using group-by function and Pivot table for the features with respect to the target variable.

**2.1.1. Data Visualization**

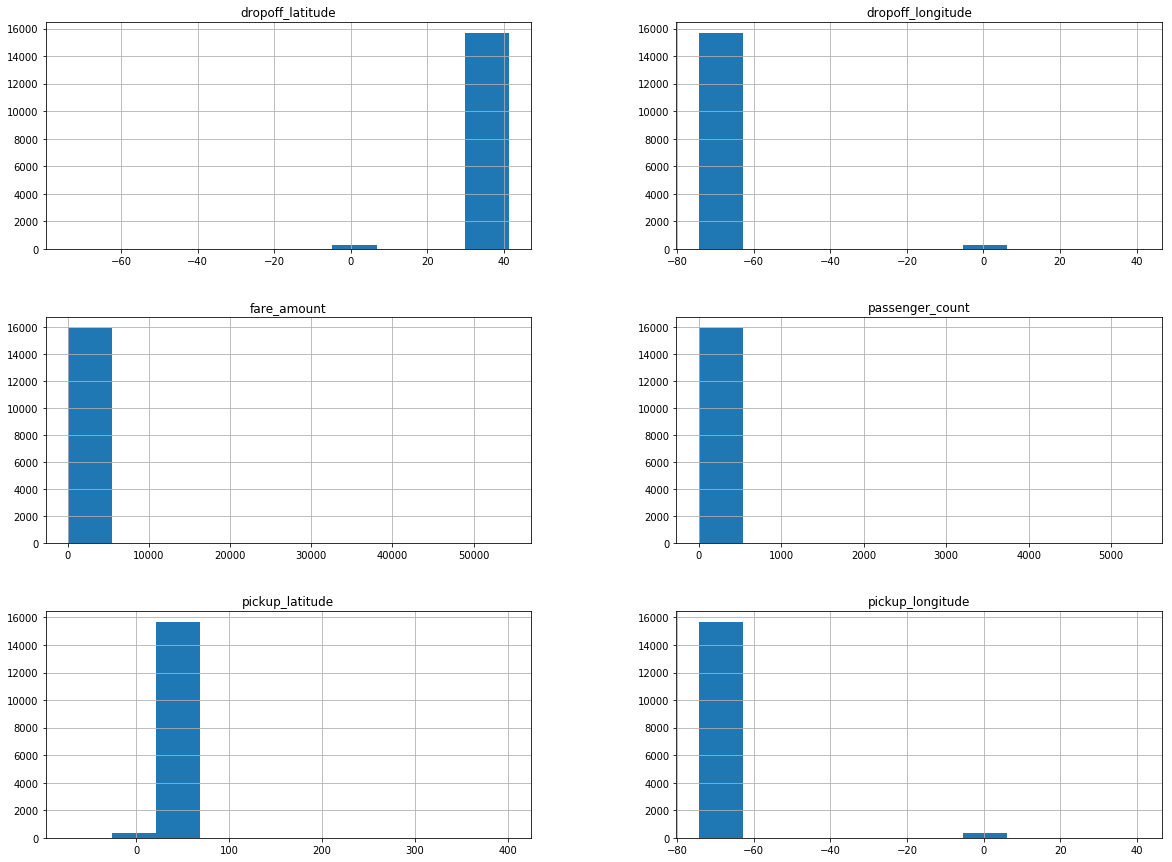
Data visualisation helps us to get better insights of the data. By visualising data, we can identify an area that need attention or improvement and also clarifies which factors influence target behaviour and how the resources are used by the target.

**2.1.1.1 Univariate Analysis**

Univariate analysis is the simplest form of data analysis where the data being analysed contains only one variable. Since it's a single variable it doesn’t deal with causes or relationships. The main purpose of univariate analysis is to describe the data and find patterns that exist within it.

So, Let’s have a look at histogram plot, to identify the characteristic of the features and the data.

Figure 2.1 Histogram plot for distribution of features in the data



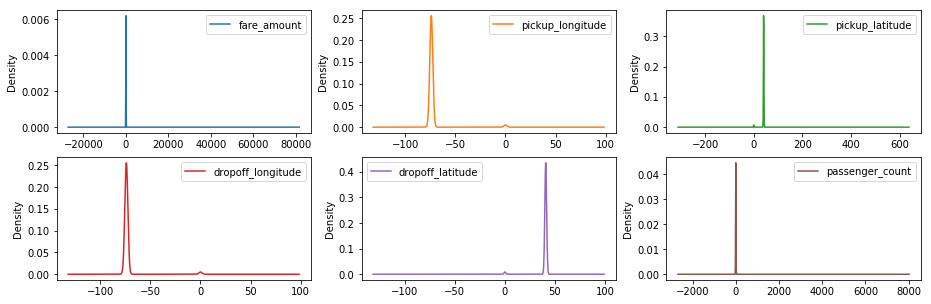
Histograms are constructed by binning the data and counting the number of observations in each bin. The objective of plotting Histogram plot is usually to visualise the shape of the distribution. The number of bins needs to be large enough to reveal interesting features and small enough not to be too noisy.

From the above histogram plot, we can clearly observe that the entire variable distributed over a large number of values. Also if observed properly, It is worth noting the following points:

1. Most latitude and longitude are near to (40.7128°, - 74.0060°).

2. A passenger count ranges from 1 to 5345.

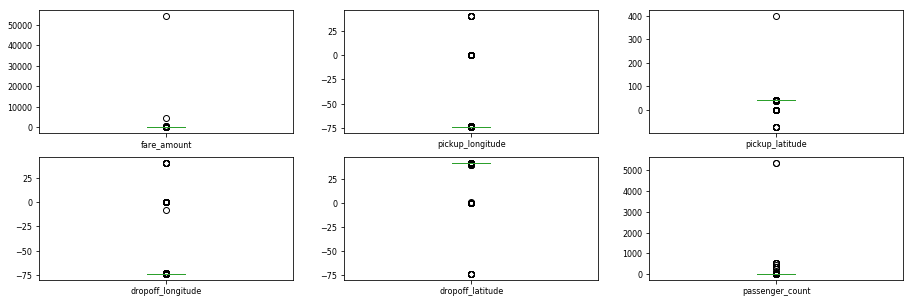
Figure 2.2: KDE plot for distribution of features in the data.



A Density Plot visualises the distribution of data over a continuous interval or time period. Density plots can be thought of as plots of smoothed histograms. An advantage Density Plots have over Histograms is that they're better at determining the distribution shape because they're not affected by the number of bins used.

So, looking at the above density plot, we can observe that some of the features follow Gaussian distribution. Few of the features seems to follow Gaussian distribution at first sight but they either have long tail at the left or right or they are either jiggered at the end.

Figure 2.3: Box plot of features in the data.



From the above Box and whisker plots, we can observe that all the features contain outliers. It is also evident from the above plot that none of the features are symmetric to the median and it can easily be interpreted that none of the features follow symmetric distribution. Also, it can also be observed that every feature have a very narrow range between minimum and maximum value.

**2.1.1.2 Bivariate Analysis**

Bivariate analysis refers to the analysis of bivariate data. It is one of the simplest forms of statistical analysis, used to find out if there is a relationship between two sets of values. It usually involves one predictor variable and one target variable.

So, let’s have a look at the Histogram and Bar Plots to understand the Employee behaviour better.

Figure 2.4: Bar graph showing the number of passengers after removing unrealistic data.

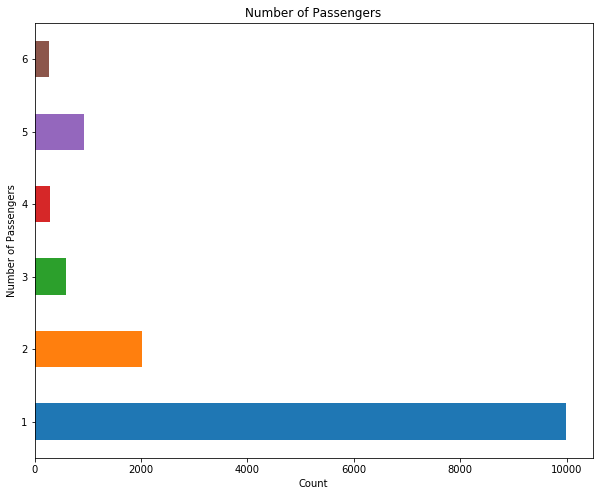
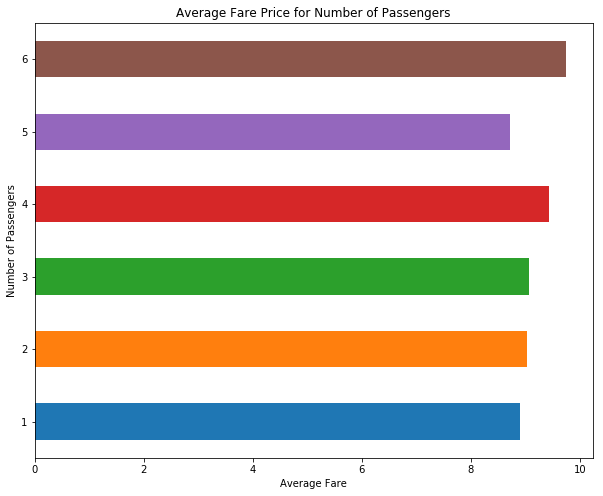
****

Figure 2.5: Bar graph showing average fare price for number of passenger



The average fair price for number of passengers is almost same.

Figure 2.6:- Bar graph showing total number of passenger on day of the week

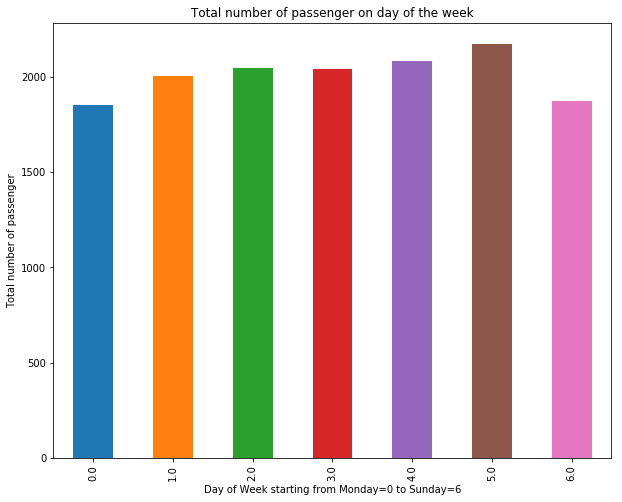
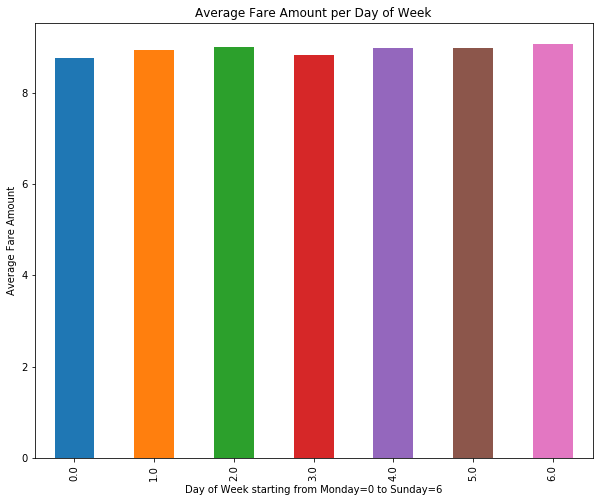


Figure 2.7:- Bar Graph showing average fare amount per day of week



**2.1.2 Data Preparation and Cleaning**

**2.1.2.1 Missing Value Analysis**

One of the most common problems I have faced in Data Cleaning/Exploratory Data Analysis is handling the missing values. Firstly, there is no good way to deal with missing data. But still missing value analysis helps address several concerns caused by incomplete data. If cases with missing values are systematically different from cases without missing values, the results can be misleading. Also, missing data may reduce the precision of calculated statistics because there is less information than originally planned. Another concern is that the assumptions behind many statistical procedures are based on complete cases, and missing values can complicate the theory required.

So, in our data, there are plenty of missing values available in different variables. So, after computing the percentage of missing data that is available to us in the dataset, it accounts to around 3 – 4 % of the data. In this project we delete the missing values in place of imputation as we have very less data missing, another reasons the features like longitude and latitude we cannot replace them with mean, median or KNN imputation.

**2.1.2.2 Outlier Analysis**

In statistics, an outlier is an observation point that is distant from other observations. In layman terms; we can say that an outlier is something which is separated from the crowd. Also, Outlier analysis is very important because they affect the mean and median which in turn affects the error (absolute and mean) in any data set. When we plot the error we might get big deviations if outliers are in the data set.

From the box and whisker plot we observe that we have a very narrow range of data between the minimum and maximum value. So here we simply not are removing outliers.

In case of latitudes and longitude we observe the data are from New York City. So we have removed the values which is outside New York City.

In case of passenger count the minimum and maximum values starts from -1 to 3.5 which is a very small range. So we decided to keep all the values ranges from 1 to 6 (passengers).

**2.1.2.3 Feature Engineering**

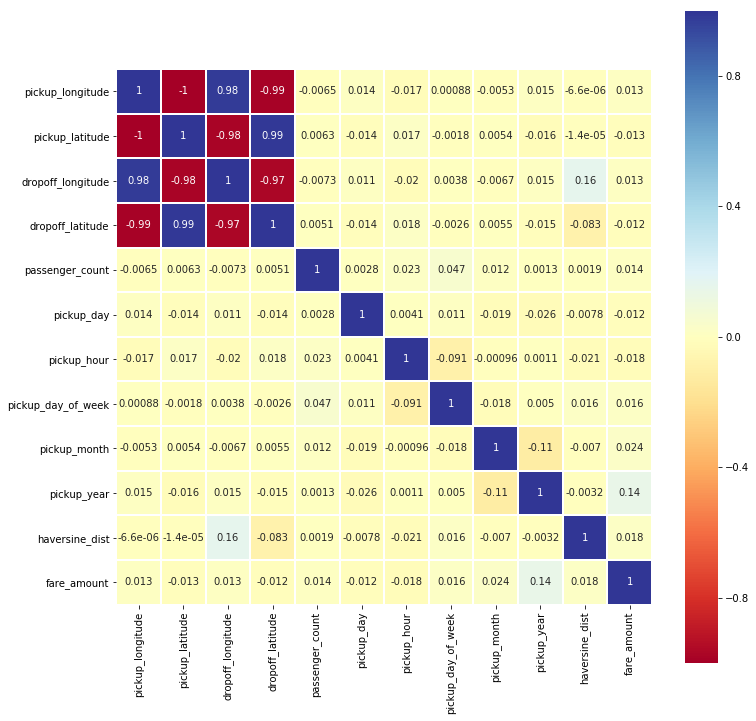
Feature engineering is the process of using domain knowledge of the data to create features that make machine learning algorithms work. Feature engineering is fundamental to the application of machine learning, and is both difficult and expensive.

Here we have datetime column but it is not possible to use datetime variable for modelling so we create more variable from that column like day, month, year, and hour.

We have also calculate a haversine distance in KM by using the pickup and drop-off longitudes and latitudes to find the distance of travel to accurate predict the fare amount.

**2.1.2.4 Feature Selection**

As we know we have only continuous variable se we go for filter method correlation analysis for feature selection and remove the highly correlated variables.



From the above heat map we observe that, latitudes and longitudes are highly correlates so we drop the drop-off longitude and drop-off latitude.

**2.1.3 Modelling**

**2.1.3.2 Multiple Linear Regression**

Multiple linear regression is the most common form of linear regression analysis. As a predictive analysis, the multiple linear regression is used to explain the relationship between one continuous dependent variable and two or more independent variables.

**Linear Regression RMSE: 4.0688726023890345**

**2.1.3.2 Decision Tree**

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.

The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data

.So, after we implement Decision Tree Regression on our data, we get the following results:

**Decision Tree RMSE: 2.271276793236065**

**(RMSE, when max\_depth=6)**

**2.1.3.3 Random Forest**

Random Forest Regression or Regression Trees are known to be very unstable, in other words, a small change in our data may drastically change your model. The Random Forest uses this instability as an advantage through bagging resulting on a very stable model. So, after we implement Random Forest Regression or Regression Trees on our data, we get the following results:

**depth - 2000 -- n\_estimators - 10 RMSE : 2.2150410784896244**

Chapter 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 any of the following criteria:

1. Predictive Performance

2. Interpretability

3. Computational Efficiency

In our case of Employee Absenteeism, the latter two, Interpretability and Computation Efficiency, do not hold much significance. Therefore we will use Predictive performance as the 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 measure.

**3.1.1 Root Mean Square Error**

Root Mean Square Error (RMSE) is the standard deviation of the residuals (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.

Also, Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. So, RMSE becomes more useful when large errors are particularly undesirable. So, Roost Mean Square value seems like a perfect choice for our problem at hand.

**3.2 Model Selection**

We saw both the model Decision tree and Random Forest perform comparatively on RMSE (Root Mean Square Error) ,Although Decision Tree gives the most promising result.

Table 3.1: Model Performance table

|  | **Model** | **RMSE** |
| --- | --- | --- |
| **0** | Decision Tree Regression | 2.2710 |
| **1** | Random Forest Regression | 2.2150 |
| **2** | Linear Regression | 4.0688 |
|  |  |  |

So , it is obvious from above model performance table, both Decision Tree and Random Forest perform good but we go with Random Forest.

**References**