**PROJECT REPORT**

**CAB FARE**

**BY**

**DIBYANSHU KUMAR**

**10-04-2019**

CONTENTS:

1 Problem Statement . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

2 Data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

3 EDA . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 5

4 Data Pre Processing . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6

5 Missing Value Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 6

6 Outlier Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 8

7 Feature Selection . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10

8 Feature Scaling . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 11

9 Univariate Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 12

10 Bivariate Analysis. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 16

11 Inferences from Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 19

12 Modeling . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 20

**INTRODUCTION**

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.

Variables Information:

* pickup\_datetime : timestamp value indicating when the cab ride started.
* pickup\_longitude - float for longitude coordinate of where the cab ride started.
* pickup\_latitude - float for latitude coordinate of where the cab ride started.
* dropoff\_longitude - float for longitude coordinate of where the cab ride ended.
* dropoff\_latitude - float for latitude coordinate of where the cab ride ended.
* passenger\_count - an integer indicating the number of passengers in the cab

Data Sets:

* Training data:

Shape: (16067, 7)

Info:

fare\_amount 16043 non-null object

pickup\_datetime 16067 non-null object

pickup\_longitude 16067 non-null float64

pickup\_latitude 16067 non-null float64

dropoff\_longitude 16067 non-null float64

dropoff\_latitude 16067 non-null float64

passenger\_count 16012 non-null float64

* Test data:

Shape: (9914, 6)

Info:

pickup\_datetime 9914 non-null object

pickup\_longitude 9914 non-null float64

pickup\_latitude 9914 non-null float64

dropoff\_longitude 9914 non-null float64

dropoff\_latitude 9914 non-null float64

passenger\_count 9914 non-null int64

Type of problem:

We have to predict fare\_amount for our test data and since our target variable is a countnuous

variable therefore this a Regression problem.

DATA PRE PROCESSING

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in

certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a

proven method of resolving such issues.

In this project we would use preprocessing technique such as missing value analysis, outlier

analysis, feature selection and feature scaling.

MISSING VALUE ANALYSIS:

We have seen in the data info that this data consists of missing values. We have to figure out

what is to be done with the missing data.

Methods used to impute missing data:

* Mean method : In this method we impute the missing values with the mean of the data in

the column.

* Median method : In this method we impute the missing values with the median value of

the column.

* KNN method : In this method we impute the missing values based on the mean of the

closest neighbours.

Here are the variables which contain missing data:

**Variables count Percentage**

|  |  |  |  |
| --- | --- | --- | --- |
| **0** | passenger\_count | 55 | 0.342317 |
| **1** | fare\_amount | 25 | 0.155598 |
| **2** | dropoff\_latitude | 0 | 0.000000 |
| **3** | dropoff\_longitude | 0 | 0.000000 |
| **4** | pickup\_latitude | 0 | 0.000000 |
| **5** | pickup\_longitude | 0 | 0.000000 |
| **6** | pickup\_datetime | 0 | 0.000000 |

For the given dataset we can see that we have only two variables with missing data and too the

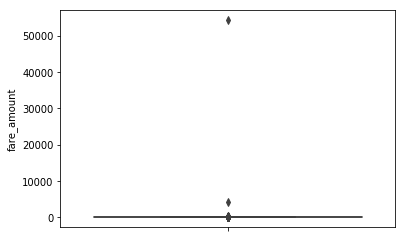
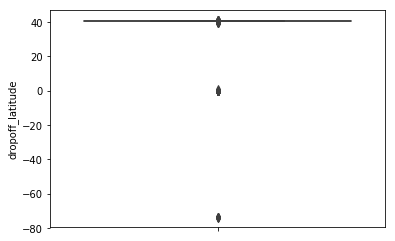
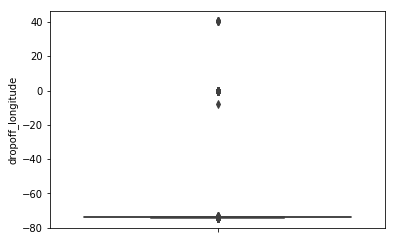
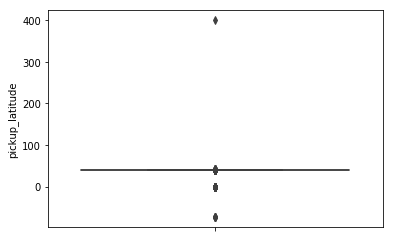
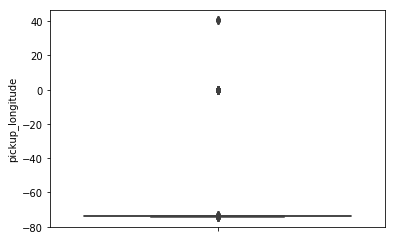
percentage of missing value is very less . So for this case we would be removing the samples

which have missing value.

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. Also we will be using some machine learning models which requires variables to be normalized thus outliers should be taken care of since it could make the distribution skewed.

**Box Plot** is the best method to find the outliers. For this data set we have plotted some boxplot for the continuous data as shown:



From the boxplots we could see that the range of the variables are very less. Below is the count

of outliers in each variable.

fare\_amount :1394

pickup\_longitude :1110

pickup\_latitude :786

dropoff\_longitude :1170

dropoff\_latitude :1002

Method to fix outliers:

To fix the outliers the values which are more than the maximum value are replaced by

maximum value and the values which are less than minimum value are replaced by the

minimum value.

FEATURE SELECTION:

Feature selection is a process where we select those features which contribute most to our prediction variable or output in which we are interested in. Having irrelevant features in our data can decrease the accuracy of the models and make our model learn based on irrelevant features.

Before feature selection we have added one more variable which is ‘distance’ in km

and it calculated by the latitude and the longitude of the points in the data.

Following is the mathematical way to calculate distance between two geographical points.

R = 6373.0

dlon = radians(lon2) - radians(lon1)

dlat = radians(lat2) - radians(lat1)

a = (sin(dlat/2))\*\*2 + cos(radians(lat1)) \* cos(radians(lat2)) \* (sin(dlon/2))\*\*2

c = 2 \* atan2(sqrt(a), sqrt(1-a))

distance = R \* c

VARIANCE INFLATION FACTOR:

To check the multi-collinearity in our data we have calculated VIF for each variable

fare\_amount 2.925835e+00

pickup\_longitude 1.399047e+00

pickup\_latitude 1.556862e+00

dropoff\_longitude 1.441208e+00

dropoff\_latitude 1.567970e+00

passenger\_count 1.000658e+00

distance 2.759796e+00

Since the VIF for every given variable is quite appropriate(<4) so we would be taking all the

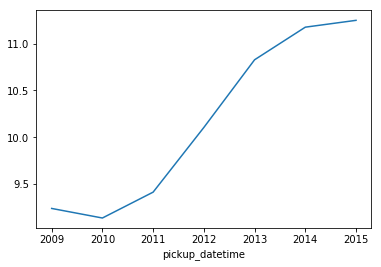
variables in our model.

EXPLORATORY DATA ANALYSIS:

In statistics, exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods.

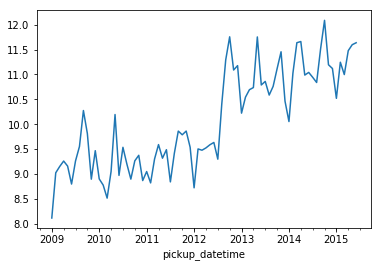
For this problem we have a variable which is ‘pickup\_datetime’ . This variable is converted into date time index to make out some inference from the data.

* Firstly we have to see that how the fare amount varies for every year.



Here is a plot of average fare amount from year 2009 to 2015 . We can see that after year2009 there is a slightly decrease in the fare amount but after year 2010 the fare gradually increased.

* Now we would also see , how the fare amount varies monthly.



Here we see that in between years the graph is not constant thus for months , there is

not a constant increase or decrease in fare amount but it overall it increases with the

year.

* Below is the data for no of customers each year:

number of customers in year 2009 is 2469

number of customers in year 2010 is 2491

number of customers in year 2011 is 2496

number of customers in year 2012 is 2573

number of customers in year 2013 is 2508

number of customers in year 2014 is 2304

number of customers in year 2015 is 1146

Also let see the average fare every year:

2009-12-31 9.240871

2010-12-31 9.138840

2011-12-31 9.414864

2012-12-31 10.104858

2013-12-31 10.826750

2014-12-31 11.174609

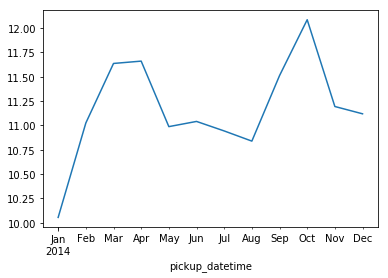
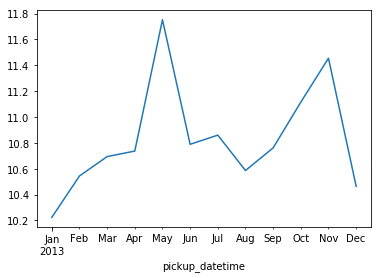
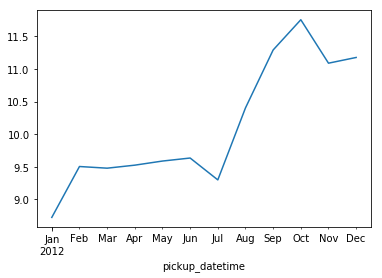
2015-12-31 11.247504

We can see that the no of customers are maximum in the year 2012. Till 2012 the no of

customer increased but after 2012 there is a gradual decrease in no of customers taking

this cab service. At the same time the fare has increased gradually from 2012 to 2015.

* Now , let see how the fare varies in a particular year.



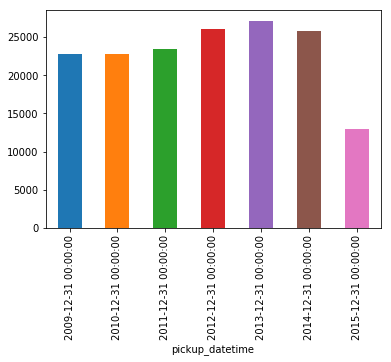
Here are the graphs of year 2012 ,2013 and 2014 . Although the fare amount does no vary in a similar way but we can see that for all the cases common points are:

1) The fare goes down during the month of July and August.

2) Again the fair rises in the month of October and November.

The fare can also increase if people are travelling more distance by cab.

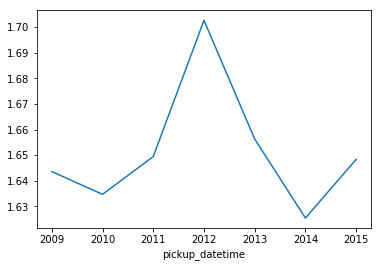
* Earning of the cab company in different years:



So we can see that the earning of the company increased till the year 2013 , and is

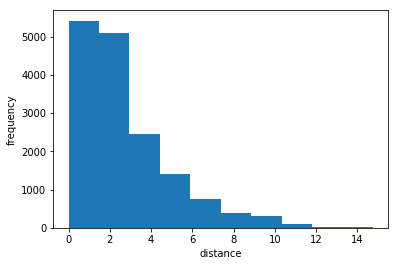
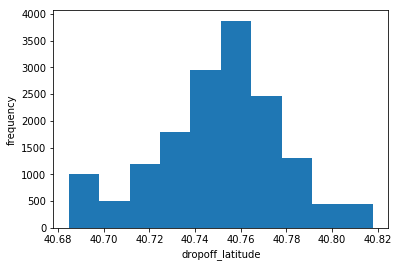
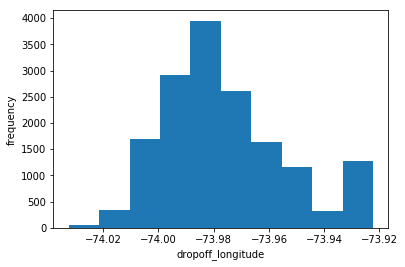
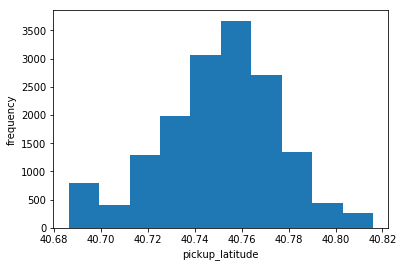
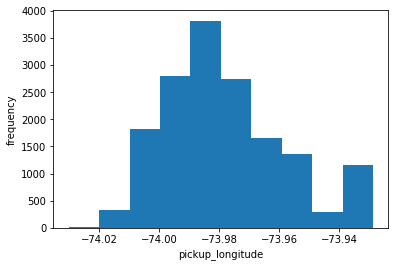
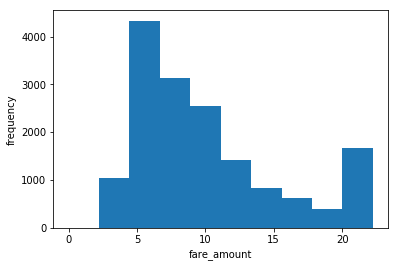
maximum in 2013 but after that it decreased. The earning in 2015 is very low since for this year we have half year data only.

* Relation between passenger\_count and year:



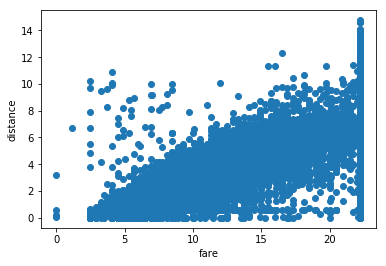
The average number of passenger increased in the year 2012.

* Now we would see the distribution of different variables:



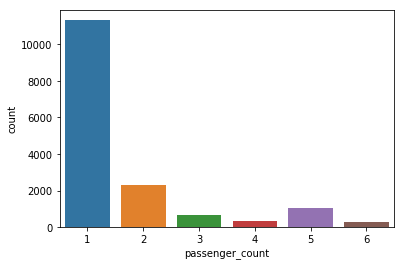
We could see that most of our variables are normally distributed. For fare amount we see that there is rise at 20 which means that there is a fixed price for certain distance range.

* Scatter plot of fare vs distance:



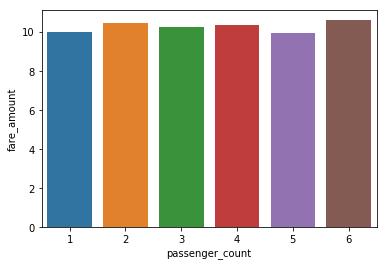
A linear relation can be seen between distance and fare which is quite obvious. There are very few cases in which fare is less for large distance.

* Passenger count:



From the bar graph its very clear that passenger\_count 1 is more frequent than others.

* Relation between passenger count and fare amount



Although the fare is slightly more if passenger count is 6 but there is not a significant

change in fare amount with the increase in passenger count.

FEATURE SCALING:

**Feature scaling** is a method used to standardize the range of independent variables or

features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step. Since the range of values of raw data varies widely, in some machine learning algorithms, objective functions will not work properly without normalization. For example, the majority of classifiers calculate the distance between two points by the Euclidean distance. If one of the features has a broad range of values, the distance will be governed by this particular feature. Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance. Since most of our data is uniformly distributed we will use Standardisation for feature scaling.

Standardisation = (ith data – mean of the column) / (standard deviation of the column)

MODELLING

This is the final phase of our project where we would build some machine learning models and will train our model on the data for future predictions. We would consider different machine learning algorithms to check which gives the best result.

Linear Regression:

Linear regression is perhaps one of the most well known and well understood algorithms in statistics and machine learning. Linear regression is a **linear model**, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).

Results for Linear Regression:

Root Mean Squared Error For Test data = 0.5887080535042504

R^2 Score = 0.65496046532007

Mean Absolute percentage Error For Test data = 1.4011466291521695

mean of accuracies of 10 folds = 0.657389058434841

standard deviation of accuracies of 10 folds = 0.02013918934757513

Ridge Linear Regression:

Ridge and Lasso regression are some of the simple techniques to reduce model complexity and prevent over-fitting which may result from simple linear regression. Here we add a parameter alpha which decreases the complexity of model. With the increase of alpha the complexity decreases.

Results for Ridge Linear Regression:

R^2 Score for alpha=1 is 0.6549619866408769

R^2 Score for alpha=10 is 0.6549753195598951

R^2 Score for alpha=100 is 0.655073673035268

R^2 Score for alpha=1000 is 0.653056907651409

We can see that there is not a significant change in accuracy in above models. We will now be using Decision tree and Random Forest.

Decision Tree Algorithm:

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.

Results for Decision Tree:

Root Mean Squared Error For Test data = 0.6980801117591496

R^2 Score = 0.5148462017314859

Mean Absolute percentage Error For Test data = 2.273001694684463

mean of accuracies of 10 folds = 0.5540969878962182

standard deviation of accuracies of 10 folds = 0.029312488171128658

We can see that the accuracy decreased may be it is because of overfitting. We would now

use Random forest to increase the accuracy and decrease overfitting.

Random Forest Algorithm:

Random Forest algorithm is a supervised algorithm. We can see it from its name, which is to create a forest by some way and make it random. There is a direct relationship between the number of trees in the forest and the results it can get , the larger the number of trees, the more accurate the result.

Results for Random Forest:

Root Mean Squared Error For Test data = 0.5187789056855896

R^2 Score = 0.7320624863903773

Mean Absolute percentage Error For Test data = 1.563972677051762

We can see that the root mean square error is least in Random Forest.

Hypertuning:

In random forest we will now vary the parameters and we’ll check what are the best parameters.

parameters = {

'bootstrap': [True],

'max\_depth': [80, 90, 100, 110],

'max\_features': [2, 3],

'n\_estimators': [100,500,800]

}

These are the parameters for which we will find the best value.

After Hypertuning the best parameters we got are:

{

'bootstrap': True,

'max\_depth': 100,

'max\_features': 2,

'n\_estimators': 500

}