CAB FARE PREDICTION

(Project Report)

By

Parth Sarathi

Index:

1. 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.

2. Procedure

According to industry standards, the process of data analyzing involves 6 main steps and this process is abbreviated as CRISP-DM process, which is Cross Industry Process for Data Mining. The 6 main steps of CRISP-DM methodology are:

1. Business Understanding
2. Data Understanding
3. Data Pre-processing/ Data Preparation
4. Modeling
5. Evolution
6. Deployment

All the above steps are followed to develop the model.

2.1 Business Understanding:

It is important to understand the idea of business behind the data set. The given data set is asking to predict fare amount and it really becomes important to predict fare amount accurately, else, there might be a great loss to the revenue of the firm. Thus we have to make sure that the model is efficient.

2.2 Data Understanding:

Here, the given data is in the CSV format. It contains 7 variables and 16067 observations. The snapshot of the data is provided below.

	fare_amount	pickup_datetime	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
0	4.5	2009-06-15 17:26:21 UTC	-73.844311	40.721319	-73.841610	40.712278	1.0
1	16.9	2010-01-05 16:52:16 UTC	-74.016048	40.711303	-73.979268	40.782004	1.0
2	5.7	2011-08-18 00:35:00 UTC	-73.982738	40.761270	-73.991242	40.750562	2.0
3	7.7	2012-04-21 04:30:42 UTC	-73.987130	40.733143	-73.991567	40.758092	1.0
4	5.3	2010-03-09 07:51:00 UTC	-73.968095	40.768008	-73.956655	40.783762	1.0
...	...	...	...	...	...	...	...
16062	6.5	2014-12-12 07:41:00 UTC	-74.008820	40.718757	-73.998865	40.719987	1.0
16063	16.1	2009-07-13 07:58:00 UTC	-73.981310	40.781695	-74.014392	40.715527	2.0
16064	8.5	2009-11-11 11:19:07 UTC	-73.972507	40.753417	-73.979577	40.765495	1.0
16065	8.1	2010-05-11 23:53:00 UTC	-73.957027	40.765945	-73.981983	40.779560	1.0
16066	8.5	2011-12-14 06:24:33 UTC	-74.002111	40.729755	-73.983877	40.761975	NaN

16067 rows × 7 columns

The different variables of the data are:

fare_amount – Fare of the cab ride.

pickup_datetime – Timestamp value explaining the time of the ride.

pickup_longitude – Float value explaining the longitude location of the ride start.

pickup_latitude – Float value explaining the latitude location of the ride start.

dropoff_longitude – Float value explaining the longitude location of the ride end.

dropoff_latitude – Float value explaining the latitude location of the ride end.

passenger_count – Integer indicating number of passengers.

From the given Train data, it is understood that the variables pickup_datetime, pickup_longitude, pickup_latitude, dropoff_longitude, dropoff_latitude and passenger_count are independent and the Target variable is fare_amount.

2.3 Data Pre-processing / Data Preparation:

Data pre-processing involves transformation of raw data that into a format that helps to execute the model well. As, the data we usually we get are incomplete, inconsistent and also contains many errors. Thus, Data Preprocessing is a generic method to deal with such issues and get the data format that is easily understood by the machine and that helps developing the model in the best way. In this project I have followed various preprocessing techniques to rectify errors and issues in the data.

2.3.1 Missing Value Analysis

Missing value is the availability of incomplete observations in the dataset. This is occurred due to reasons like incomplete submission, wrong input, manual error, etc. These missing values effect the accuracy of the model, hence, it becomes very important to rectify these missing values.

All the 0's and blank spaces in the data is replaced by 'na' in R and 'nan' in python and also the variable are converted to required data-types accordingly.

In the given dataset there are few missing values present. The missing value and missing value percentage is as follows:

	Variables	count	Missing_percentage
0	pickup_longitude	313	1.951007
1	pickup_latitude	313	1.951007
2	dropoff_longitude	312	1.944773
3	dropoff_latitude	310	1.932307
4	passenger_count	113	0.704357
5	fare_amount	25	0.155831
6	pickup_datetime	1	0.006233

It is found from the above snapshot that there is no variable exceeding 30% of missing value. Hence, we need not eliminate any variable and proceed with imputation.

Missing Value Imputation

Missing value imputation is normally done by following methods. They are:

- Central Tendencies – by the help of Mean, Median, Mode.
- Distance based method like K-Nearest Neighbor (KNN) imputation.
- Prediction Method – It is based on predictive ML algorithms.

To use the best imputation method it is necessary to check which method predicts values which are closest to the actual data. This is done by taking a subset of data and making a note of it (Actual value). Then replacing that value with 'na' and applying available methods of imputation and noting down every value from the methods. The closest value to the actual value is chosen for imputing missing values. I my case I have chosen Median to impute the missing values.

	Variables	count	Missing_percentage
0	fare_amount	0	0.0
1	pickup_datetime	0	0.0
2	pickup_longitude	0	0.0
3	pickup_latitude	0	0.0
4	dropoff_longitude	0	0.0
5	dropoff_latitude	0	0.0
6	passenger_count	0	0.0

The above snapshot shows the missing values after imputation with Median.

2.3.2 Outlier Analysis:

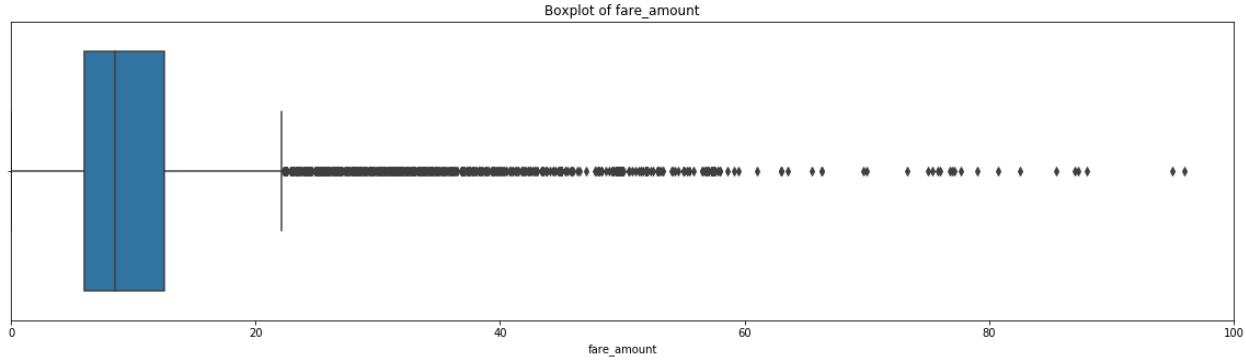
Outlier is an abnormal observation that stands or deviates away from other observations. This happens because of poor quality of data, manual error and it is correct but exceptional data. It can cause error in predicting the target variables. So it is important to rectify the outliers in the data.

In this dataset, I have found some irregular data those are considered as outliers. Before we start the outlier analysis, we make a copy of our dataset for further requirements. The outliers in each variable are rectified as follows.

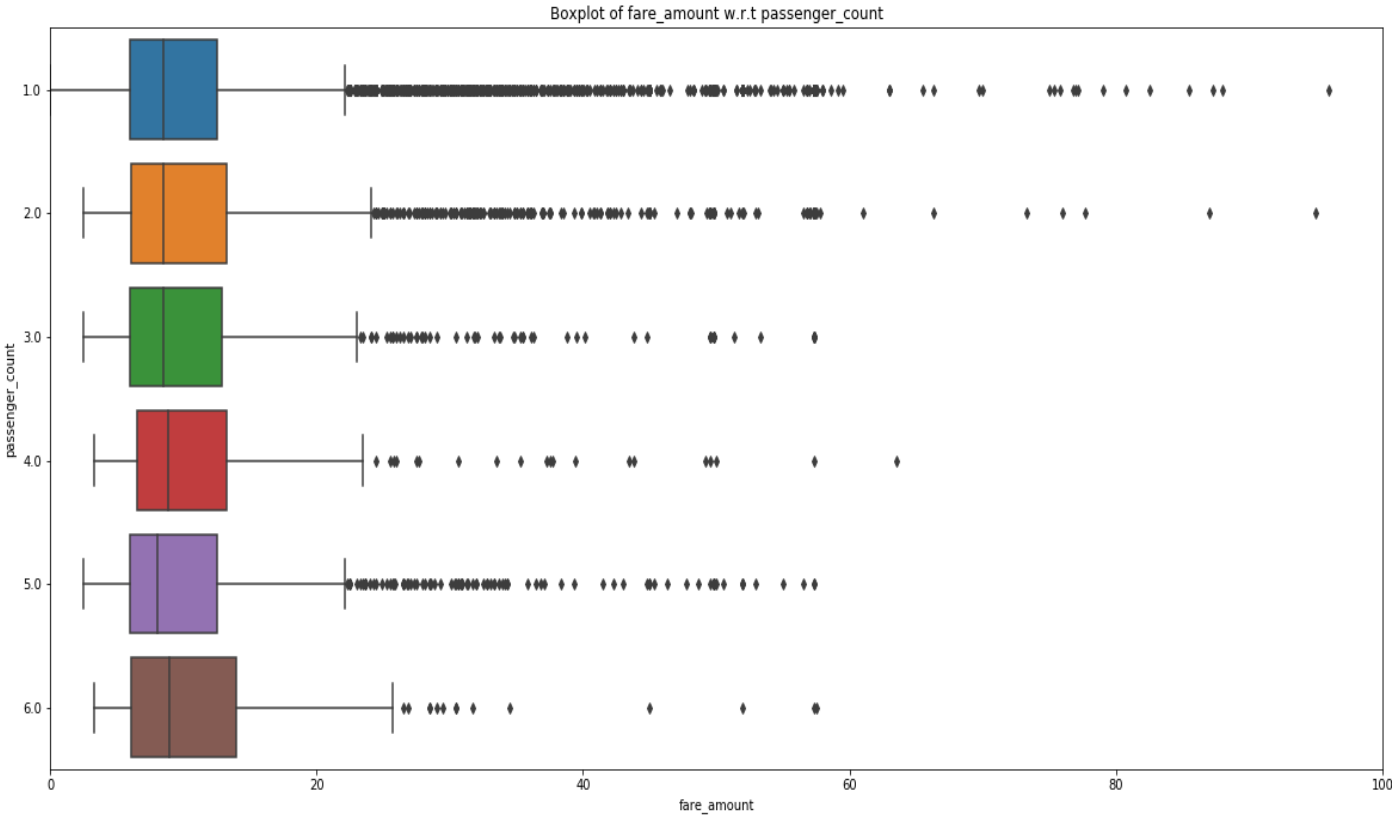
fare_amount – As we know that fare amount cannot be a negative value, hence, I dropped all the negative values from the fare_amount variable.

passenger_count – As it is a cab, passenger count cannot exceed 6. So, I have dropped all the observations containing passenger count more than 6.

Univariate Boxplots: Boxplot for target variable fare_amount.



Bivariate Boxplots: Boxplot for Numerical Variable Vs Categorical Variable



As we can see in the above boxplots, that there are outliers in the target variable. Imputing them may hamper the model and also as there are many observations, dropping them may cost in loss of information.

Hence By viewing all the observations I found that there were 2 observations where the fare amount is extremely high.

	fare_amount	pickup_datetime	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
1015	54343.0	2015-02-27 17:03:50 UTC	-74.003319	40.727455	-73.964470	40.764378	1.0
1072	4343.0	2012-01-15 20:42:04 UTC	-73.976309	40.751634	-74.014854	40.709044	1.0

I dropped the 2 observations.

For pickup_longitude, pickup_latitude, dropoff_longitude, dropoff_latitude, Outlier treatment is done after feature engineering step.

IQR Score:

Box plot use the IQR method to display data and outliers(shape of the data) but in order to be get a list of identified outlier, we will need to use the mathematical formula and retrieve the outlier data.

The interquartile range (IQR), also called the midspread or middle 50%, or technically H-spread, is a measure of statistical dispersion, being equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles, IQR = Q3 − Q1. In other words, the IQR is the first quartile subtracted from the third quartile; these quartiles can be clearly seen on a box plot on the data.

It is a measure of the dispersion similar to standard deviation or variance, but is much more robust against outliers.

In this project, I created a function “outlier_treatment” where I calculated the IQR value and then lower and upper bound is calculated and saved as minimum and maximum. Now, all the data in the list which is less than minimum and more than maximum is considered as outliers and is replaced by ‘nan’.

2.3.3 Feature Engineering:

Feature engineering is used to drive new features from existing features.

1. For ‘pickup_datetime’ variable

We will use this timestamp variable to create new variables such as ‘year’, ‘month’, ‘dayofweek’ and ‘hour’.

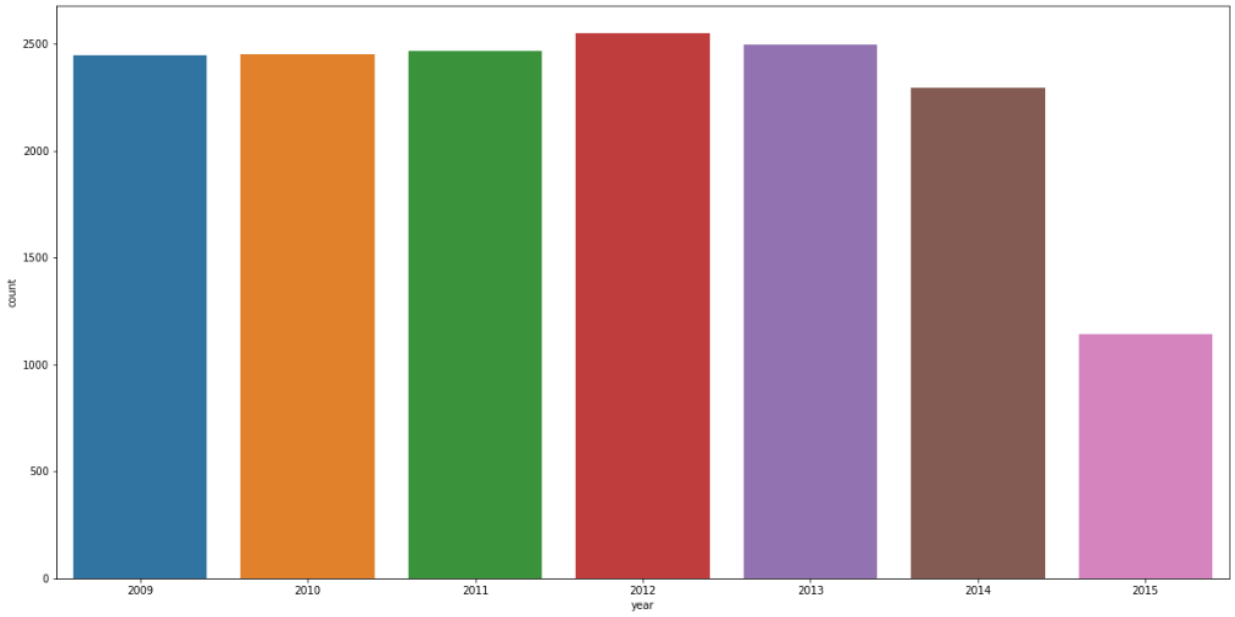
‘year’ will contain only years from pickup_datetime. For ex: 2009, 2010, 2011, etc.

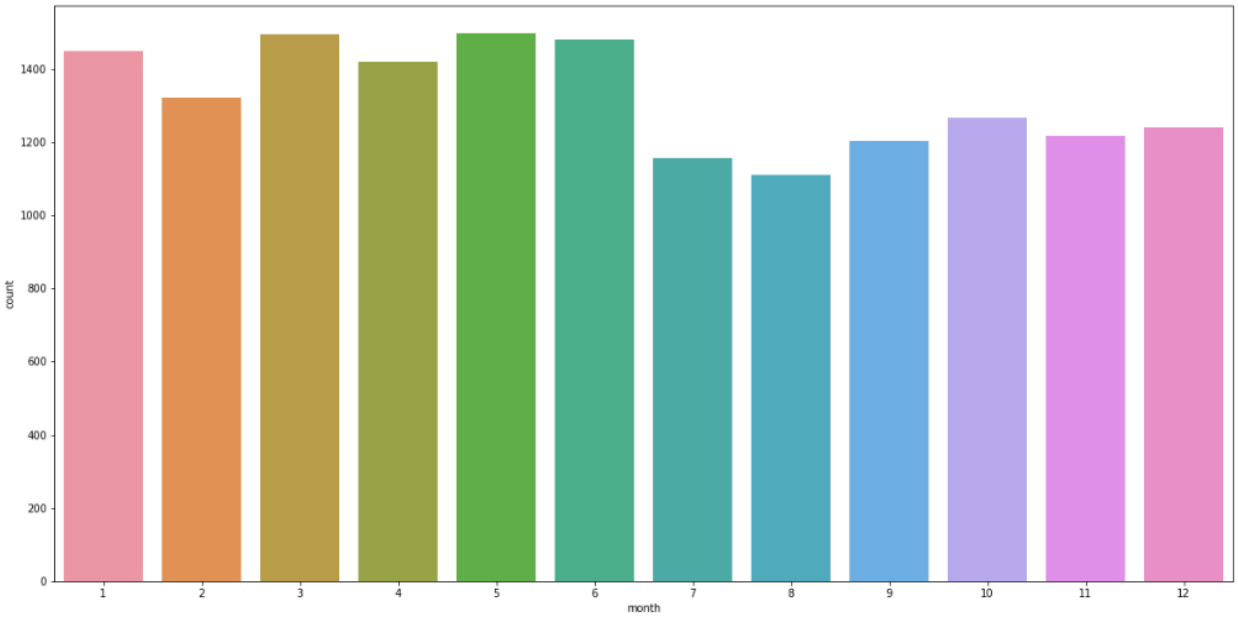
‘month’ will contain only months from ‘pickup_datetime’. For ex: 1 for January, 2 for February, 3 for March, and so on.

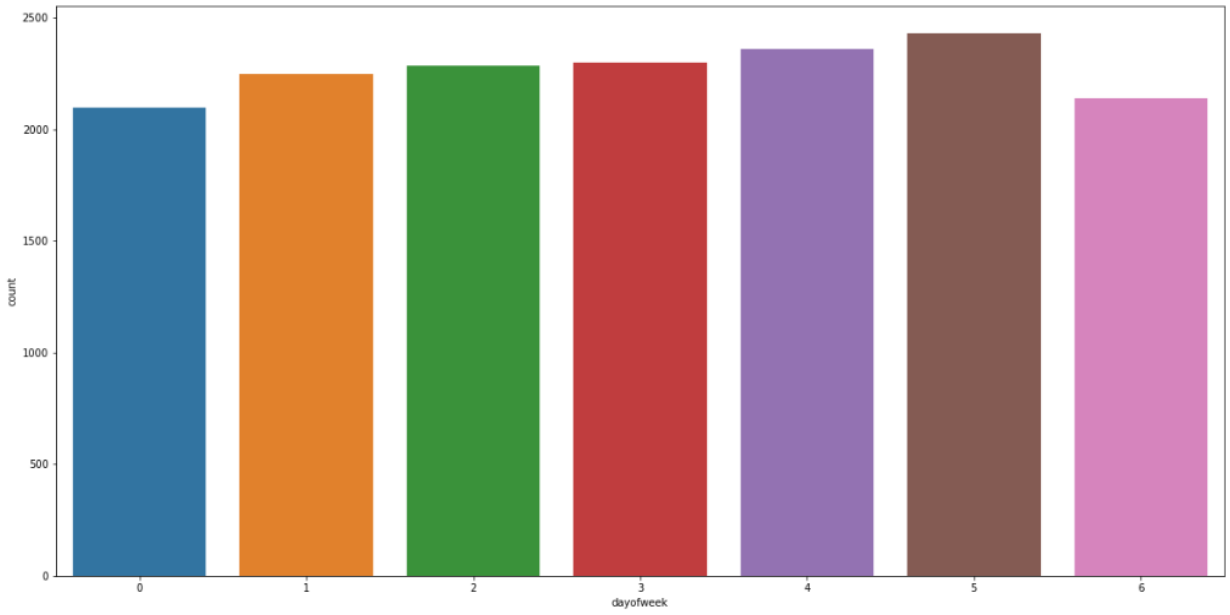
‘dayofweek’ will contain day in a week from ‘pickup_datetime’. For ex: 0 for Sunday, 1 for Monday, 2 for Tuesday and so on.

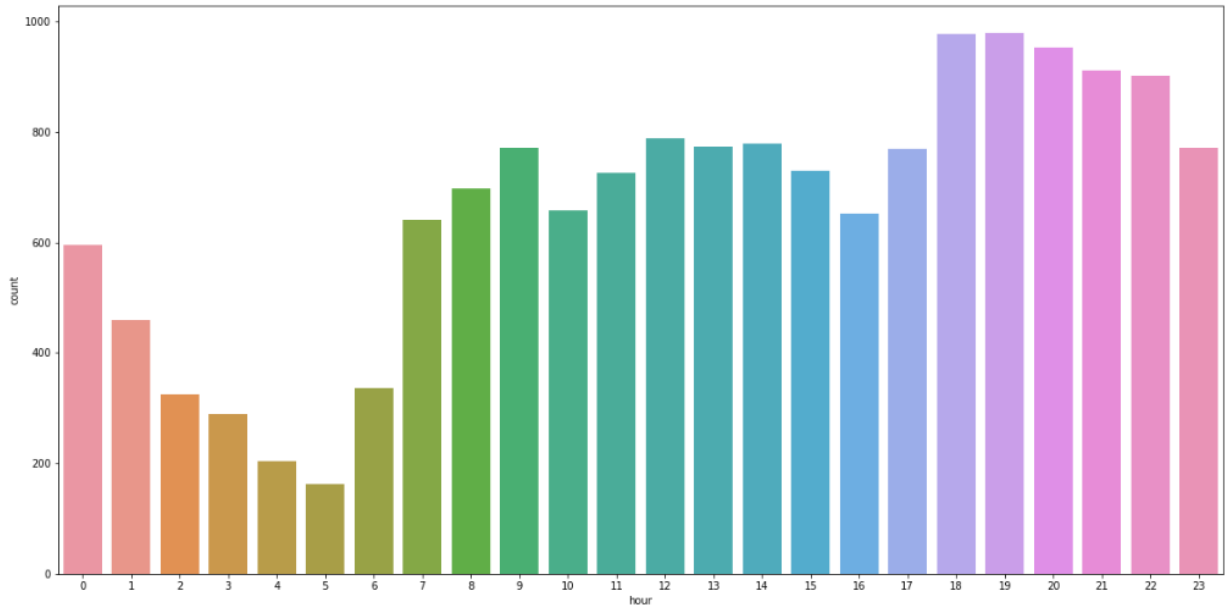
‘hour’ will contain only hours from ‘pickup_datetime’. For ex: 1, 2, 3, etc.

CountPlot of ‘year’, ‘month’, ‘dayofweek’ and ‘hour’ variables.




count
month
0
200
400
600
800
1000
1200
1400
1
2
3
4
5
6
7
8
9
10
11
12


count
dayofweek
0
500
1000
1500
2000
2500
0
1
2
3
4
5
6

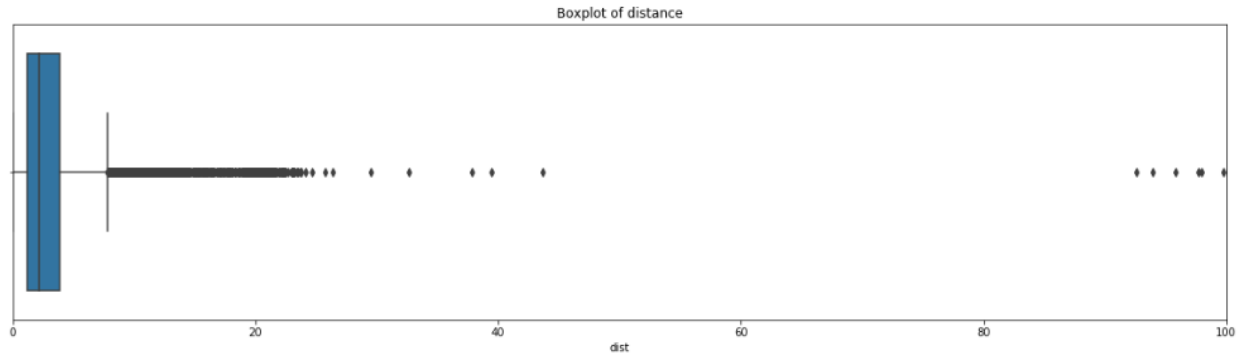


2. <u>For Latitudes and Longitudes variables:</u>

As we have latitudes and longitudes data for pickup and drop-off, we will find the distance the cab has travelled from pickup to drop-off location.

I used Haversine method to calculate distance and the new variable ‘dist’ is created.

Boxplot for new variable ‘dist’



As we see there are outliers in ‘dist’, we treat the outliers using the function ‘Outlier_treatment’ we created previously.

2.3.4 <u>Feature Selection:</u>

In this step we would allow only to pass relevant features to further steps. We remove irrelevant features from the dataset. We do this by some statistical techniques, like we look for features which will not be helpful in predicting the target variables. In this dataset we have to predict the fare_amount.

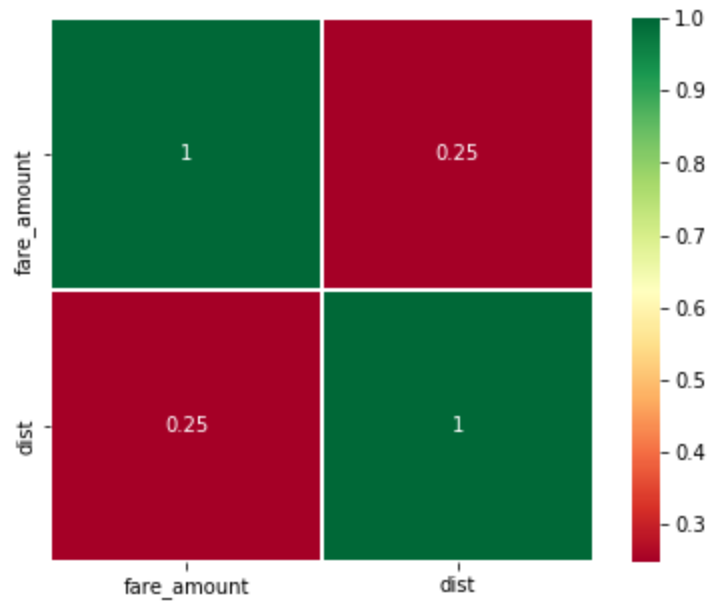
Further below are some types of test involved for feature selection:

1. Correlation analysis – This requires only numerical variables. Therefore, we will filter out only numerical variables and feed it to correlation analysis. We do this by plotting correlation plot for all numerical variables. There should be no correlation between independent variables but there should be high correlation between independent variable and dependent variable. So, we plot the correlation plot. we can see that in correlation plot faded colour like skin colour indicates that 2 variables are highly correlated with each other. As the colour fades correlation values increases. From below correlation plot we see that:

• 'fare_amount' and 'dist' are very highly correlated with each other.

• As fare_amount is the target variable and ‘dist’ is independent variable we will keep ‘dist’ because it will help to explain variation in fare_amount.

Correlation Plot:



2. Chi-Square test of independence – Unlike correlation analysis we will filter out only categorical variables and pass it to Chi-Square test. Chi-square test compares 2 categorical variables in a contingency table to see if they are related or not.

I. Assumption for chi-square test: Dependency between Independent variable and dependent variable should be high and there should be no dependency among independent variables.

II. Before proceeding to calculate chi-square statistic, we do the hypothesis testing:

Null hypothesis: 2 variables are independent.

Alternate hypothesis: 2 variables are not independent.

The interpretation of chi-square test:

I. For theoretical or excel sheet purpose: If chi-square statistics is greater than critical value then reject the null hypothesis saying that 2 variables are dependent and if it's less, then accept the null hypothesis saying that 2 variables are independent.

II. While programming: If p-value is less than 0.05 then we reject the null hypothesis saying that 2 variables are dependent and if p-value is

greater than 0.05 then we accept the null hypothesis saying that 2 variables are independent.

3. Analysis of Variance(Anova) Test –
   I. It is carried out to compare between each group in a categorical variable.
   II. ANOVA only lets us know the means for different groups are same or not. It doesn't help us identify which mean is different.

- Null Hypothesis: mean of all categories in a variable are same.

Alternate Hypothesis: mean of at least one category in a variable is different.

- If p-value is less than 0.05 then we reject the null hypothesis.
- And if p-value is greater than 0.05 then we accept the null hypothesis

Below is the Anova analysis table for each categorical variable.

	df	sum_sq	mean_sq	F	PR(>F)
C(passenger_count)	5.0	2.828216e+03	565.643267	5.055336	1.241770e-04
C(year)	6.0	2.034440e+04	3390.733858	30.304079	2.284210e-36
C(month)	11.0	6.459352e+03	587.213829	5.248119	2.535244e-08
C(dayofweek)	6.0	1.303777e+03	217.296135	1.942045	7.026184e-02
C(hour)	23.0	9.145410e+03	397.626514	3.553716	1.759060e-08
Residual	15802.0	1.768091e+06	111.890345	NaN	NaN

2.3.5 Feature Scaling:

Data Scaling methods are used when we want our variables in data to scaled on common ground. It is performed only on continuous variables.

• Normalization: Normalization refers to the dividing of a vector by its length. Normalization normalizes the data in the range of 0 to 1. It is generally used when we are planning to use distance method for our model development purpose such as KNN. Normalizing the data improves convergence of such algorithms. Normalization of data scales the data to a very small interval, where outliers can be loosed.

• Standardization: Standardization refers to the subtraction of mean from individual point and then dividing by its SD. Z is negative when the raw score is below the mean and Z is positive when above mean. When the data is distributed normally you should go for standardization.

Linear Models assume that the data you are feeding are related in a linear fashion, or can be measured with a linear distance metric. Also, our independent numerical variable 'dist' is not distributed normally so we had chosen normalization over standardization.

I have checked variance for each column in dataset before Normalization. High variance will affect the accuracy of the model. So, we want to normalize that variance.

Note: It is performed only on continuous variables.

2.4 Model Development

Our problem statement wants us to predict the fare_amount. This is a Regression problem. So, we are going to build regression models on training data and predict it on test data. In this project I have built models using 3 Regression Algorithms:

i) Linear Regression
ii) Decision Tree
iii) Random forest
iv) K – nearest neighbor(KNN)

1. Linear Regression:

Linear regression is a common Statistical Data Analysis technique. It is used to determine the extent to which there is a linear relationship between a dependent variable and one or more independent variables.

Here is the model summary,

OLS Regression Results

Dep. Variable:	fare_amount	R-squared:	0.074
Model:	OLS	Adj. R-squared:	0.073
Method:	Least Squares	F-statistic:	101.0
Date:	Wed, 05 Aug 2020	Prob (F-statistic):	2.40e-202
Time:	13:58:51	Log-Likelihood:	-47749.
No. Observations:	12683	AIC:	9.552e+04
Df Residuals:	12672	BIC:	9.560e+04
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
year	0.6204	0.051	12.272	0.000	0.521	0.720
month	0.1236	0.027	4.555	0.000	0.070	0.177
dayofweek	-0.0737	0.047	-1.551	0.121	-0.167	0.019
hour	-0.0452	0.014	-3.156	0.002	-0.073	-0.017
dist	12.8350	0.455	28.193	0.000	11.943	13.727
passenger_count_1	-1240.8579	101.722	-12.198	0.000	-1440.249	-1041.467
passenger_count_2	-1239.5935	101.721	-12.186	0.000	-1438.982	-1040.205
passenger_count_3	-1240.4099	101.719	-12.194	0.000	-1439.794	-1041.026
passenger_count_4	-1240.5865	101.725	-12.196	0.000	-1439.982	-1041.191
passenger_count_5	-1240.6892	101.706	-12.199	0.000	-1440.048	-1041.331
passenger_count_6	-1241.5740	101.789	-12.198	0.000	-1441.096	-1042.052

Omnibus:	23672.140	Durbin-Watson:	1.987
Prob(Omnibus):	0.000	Jarque-Bera (JB):	119007895.239
Skew:	13.630	Prob(JB):	0.00
Kurtosis:	476.767	Cond. No.	5.40e+06

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.4e+06. This might indicate that there are

Here, F-statistic explains the quality of the model. AIC is Akkaine Information Criterion , if we have multiple model with same accuracy then we need to refer this to choose the best model. T-statistic explains how much statistically significant the coefficient is. It is also used to calculate the P-value. If P-value is less than 0.05, we reject null hypothesis and say that the variable is significant. The R squared and the Adjusted R squared values show how much variance of the output variable is explained by the independent or input variables. Mean Absolute Percentage Error is calculated.

```
#Error - 47.99919382282029
#Accuracy - 52.00080617717971
```

2. Decision Tree:

Decision Trees are a non-parametric supervised learning method used for both classification and regression tasks. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

The below fit plot is shows the criteria that is used in developing the decision tree in python.

```
DecisionTreeRegressor(ccp_alpha=0.0, criterion='mse', max_depth=2,
                      max_features=None, max_leaf_nodes=None,
                      min_impurity_decrease=0.0, min_impurity_split=None,
                      min_samples_leaf=1, min_samples_split=2,
                      min_weight_fraction_leaf=0.0, presort='deprecated',
                      random_state=None, splitter='best')
```

To develop the model, I haven't provided any input argument of my choice except the depth as 2 to visualize the tree better. All other arguments in the model are default, in developing the model. After this the fit_DT is used to predict in test data and the error rate (MAPE) and accuracy is calculated.

```
#Error = 35.2988778277751
#Accuracy = 64.7011221722249
```

3. Random Forest:

Random forests are an ensemble learning method
for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of over fitting to their training set.

Like the decision tree, the below fit plot shows the criteria that is used in developing the random forest model in python.

```
RandomForestRegressor(bootstrap=True, ccp_alpha=0.0, criterion='mse',
                      max_depth=None, max_features='auto', max_leaf_nodes=None,
                      max_samples=None, min_impurity_decrease=0.0,
                      min_impurity_split=None, min_samples_leaf=1,
                      min_samples_split=2, min_weight_fraction_leaf=0.0,
                      n_estimators=10, n_jobs=None, oob_score=False,
                      random_state=None, verbose=0, warm_start=False)
```

RF_Predictions is used to predict in test data and the error rate (MAPE) and accuracy is calculated.

```
# Error = 27.48467609783952
#Accuracy = 72.51532390216048
```

4. K- nearest neighbor (KNN):

It finds the nearest neighbor and tries to predict target value. The method goes as, for the value of new point to be assigned, this value is assigned on the basis of how closely the point resembles the other points in the training set. After the model is run, the prediction fit is used to predict in test data. Finally, error and accuracy is calculated.

3. Evolution Of The Model

So, now we have developed few models for predicting the target variable, now the next step is to identify which one to choose for deployment. To decide these according to industry standards, we follow several criteria. Few among these are, calculating the error rate, and accuracy. MAPE is used in our project. RMSE is not used because we are not working with Timestamp value.

Model	**Error**	**Accuracy**
Linear Regression	47.99919382282029	52.00080617717971
Decision Tree	35.2988778277751	64.7011221722249
Random forest	27.48467609783952	72.51532390216048
KNN	33.7978	66.21

After comparison of the error and accuracy, we come to selection of the most effective model. From the values of error and accuracy it is found that Random forest gives better results as compared to other methods. So I prefer random forest model to be used for further processes.

THANK YOU

----------END----------