**Project**

**CAB FARE PREDICTION**

**Submitted by**

**Vijay Jadhav**

**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.

**Procedure** – We could predict the fare price by following below steps these steps are necessary for approximate predictions.

1. **Business understanding :**

The given data is asking to predict the fare amount accurately.

1. **Data understanding**

Number of attributes:

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 ride.

fare\_amount – Fare of the given cab ride.

**Dependent Variable** : fare Amount remaining all others are Independent variables.

1. **Data preparation/Data Preprocessing**

- I have dropped pickup\_datetime as it will not have any impact on dependent variable.

- Missing Values :

In the given dataset it is found that there are lot of values which are missing. It is found in the following types:

• 1. Blank spaces : Which are converted to NA and NaN respectively for further operations

• 2. Zero Values : This is also converted to NA and Nan respectively prior further operations

• 3. Repeating Values : there are lots of repeating values in pickup\_longitude, pickup\_latitude, dropoff\_longitude and dropoff\_latitude

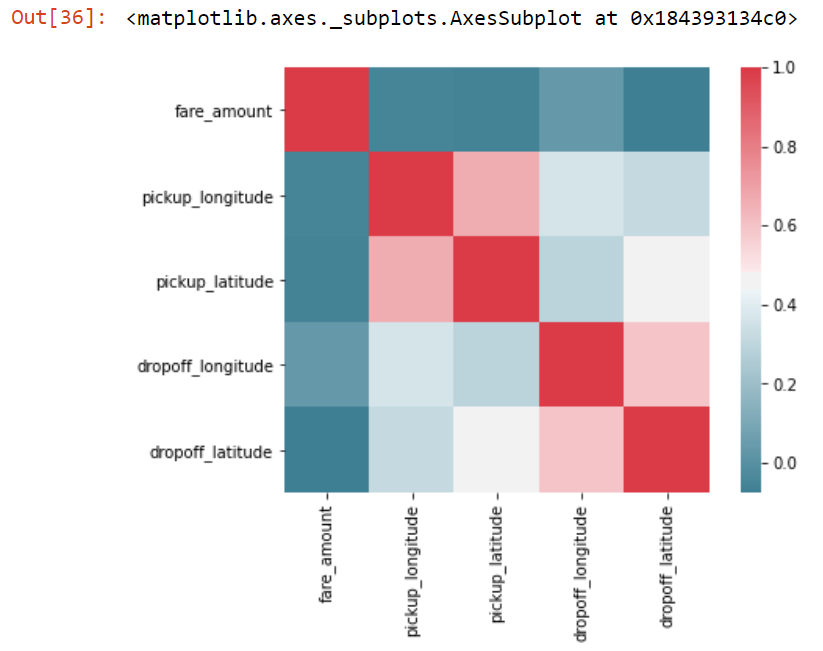
-Impute missing values with Median. Have demonstrated Median giving more close values then KNN Imputation and Mean.

-Outlier Analysis

Fare amount – Some observations are negative, highly impossible

Passenger Count – Have seen value in multiple of 100 a cab can carry maximum 8 passengers.

- Correlation Plot :



From the above plot it is found that most of the variables are highly correlated with each other, like fare amount is highly correlated with distance variable. All the dark grey charts represents that variables are highly correlated. And as there is no dark red charts, which represents negative correlation, it can be summarized that our dataset has strong or highly positive correlation between the variables.

1. **Modeling**

Next Step is to build a model check the Accuracy and error and compare that with different models like Decision Tree or Random Forest Classifier.

Linear regression : It is used to predict the value of variable Y based on one or more input predictor variables X. The goal of this method is to establish a linear relationship between the predictor variables and the response variable.

Coefficients : [-3.57402507e-05 -1.02419714e-01 -4.14493666e-03 -2.98061163e-02

-1.30904248e-01]

Intercepts : -1.9674121494453165

RMSE : 79.95683160904608

Decision Tree - is a supervised learning predictive model that uses a set of binary rules to calculate the target value/dependent variable. Decision trees are divided into three main parts this are :

• Root Node : performs the first split

• Terminal Nodes : that predict the outcome, these are also called leaf nodes

• Branches : arrows connecting nodes, showing the flow from root to other leaves.

#Accuracy: 56.95%

#Error : 43.05%

Random Forest Classifier :

It is a process where the machine follows an ensemble learning method for classification and regression that operates by developing a number of decision trees at training time and giving output as the class that is the mode of the classes of all the individual decision trees.

#Accuracy: 75%

#Error : 24.49%

1. **Evaluation**

We need to evaluate all above models based on MAPE (Mean Absolute Percentage Error) and Accuracy. RMSE is not used because we are not working with Timestamp value.

Mean Absolute Error (MAE) MAE or Mean Absolute Error, it is one of the error measures that is used to calculate the predictive performance of the model. In this project we will apply this measure to our models

Accuracy The second matrix to identify or compare for better model is Accuracy. It is the ratio of number of correct predictions to the total number of predictions made. Accuracy= number of correct predictions / Total predictions made It can also be calculated from MAE as Accuracy = 1- MAPE

1. **Deployment**: Deploying the model on cloud or using Flask or some other application.

**THANK YOU**