Cab Fare Prediction

Venkatesh Kypa

1. Introduction
2. 1 Problem Statement

The objective of this project is to predict Cab Fare amount.

The Statement Given in the Problem is follows

You are a cab rental start-up company. You have successfully run the pilot project and now want o 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. 2 Data

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.

2. Methodology

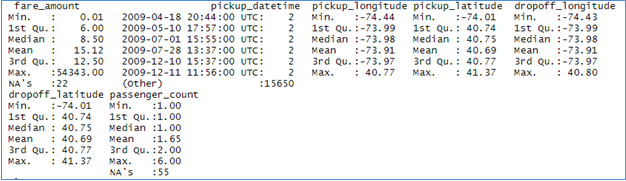
2.1 Pre-Processing

Data pre-processing is the first stage of any type of project. In this stage we get the feel of the data. For any machine learning problems input data will be structured in a way so as to get the meaningful insights from the model. For that data pre-processing is the first step and also the most important task.

We do this by looking at plots of independent variables vs target variables. If the data is messy, we try to improve it by sorting deleting extra rows and columns. This stage is called as Exploratory Data Analysis. This stage generally involves data cleaning, merging, sorting, looking for outlier analysis, looking for missing values in the data, imputing missing values if found by various methods such as mean, median, mode, KNN imputation, etc.

Further we will look into what Pre-Processing steps do this project was involved in.

Before we look into pre-processing just we see the summary of the data



2.1.1 Removing values which are not within desired range (outlier)

In this step we will remove values in each variable which are not within desired range and we will consider them as outliers depending upon basic understanding of all the variables.

In this given dataset

1. Fare amount which is currency cannot become zero or negative for any ride so we will remove entire rows where fare amount which is less than or equal to 0
2. Passenger count cannot be more than 6 since at maximum a car or cab can accommodate 6 or less 6 members and also there should be at least 1 passenger. Therefore we will remove entire rows where passenger count is more than 6 and less than 1
3. For Pickup longitude & Drop longitude which cannot exceed +180 and -180 therefore entire rows where longitude values not in the range of >180 and < -180
4. For Pickup latitude & Drop latitude which cannot exceed +90 and -90 therefore entire rows where longitude values not in the range of >90 and < -90
5. We also remove latitude and longitude values equal to 0 which says there is no pickup and dropping distances. Therefore we will remove entire rows where longitude & latitude values are equal to 0

2.1.2 Missing Values Analysis

In this step we look for missing values in the dataset like empty row column cell which was left after removing special characters and punctuation marks. Some missing values are in form of NA. Missing values left behind after outlier analysis; missing values can be in any form.   
Unfortunately, in this dataset we have found some missing values. Therefore, we will do some missing value analysis.

|  |  |  |  |
| --- | --- | --- | --- |
| S.No | Variables | Missing Values | Missing Percentage |
| 1 | passenger\_count | 55 | 0.3511684 |
| 2 | fare\_amount | 22 | 0.1404674 |
| 3 | pickup\_datetime | 1 | 0.006385 |
| 4 | pickup\_longitude | 0 | 0 |
| 5 | pickup\_latitude | 0 | 0 |
| 6 | dropoff\_longitude | 0 | 0 |
| 7 | dropoff\_latitude | 0 | 0 |

We will impute values for fare\_amount and passenger\_count as both of them have missing values 22 & 55 respectively.

We will drop 1 value(i.e entire row) in pickup\_datetime while changing it into date type

We will impute these missing values with the best approach i.e. with mode or mean or median or KNN technique

1. **For passenger count:**

Actual Value: 3

Mode: 1

Median: 1

KNN: 1.2 ~ 2

We will choose the KNN method here because because it imputes value closest to actual value. We will not use Mode method because whole variable will be more biased towards 1   
passenger\_count also passenger\_count has maximum value equals to 1

1. **For fare amount :**

Actual Value: 16.5

Mean Value : 6.5

KNN: 12.5

We will Choose KNN method here because it imputes value closest to actual value

After KNN Imputation we look at the passenger count values. Count should be integer values and accordingly we have to change the datatype

2.1.3 Outlier Analysis

We look for outlier in the dataset by plotting Boxplots. There are outliers present in the data. we have removed these outliers. This is how we done,

1. We replaced them with Nan values or we can say created missing values.
2. Then we imputed those missing values with KNN method.

Before we do outlier analysis first we convert longitude and latitude into distance travelled

* + 1. Feature Engineering

Here we split pickup\_datetime into several many values i.e. pickup\_month, pick\_year, pickup\_days, pickup\_hours. We will use this timestamp variable to create new variables.

After splitting one variable into many variables we also them into categorical variables

* + 1. Feature Selection

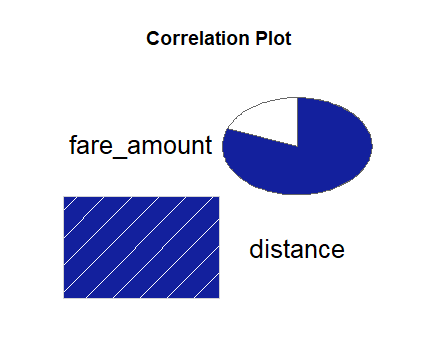
In this step we would allow only to pass relevant features to further steps. We remove irrelevant features from the dataset. We do this by some statistical techniques, like we look for features which will not be helpful in predicting the target variables. In this dataset we have to predict the fare\_amount.

1. Correlation analysis (For Numeric Variables)

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.

Correlation plot:



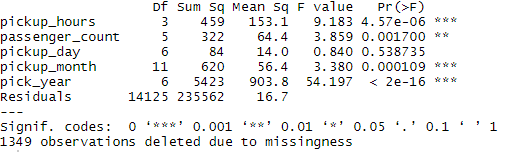
Plot shows that there is high correlation between distance & fare\_amount

1. Analysis of Variance(Anova) Test **(** for Categorical Variables)

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 different͘

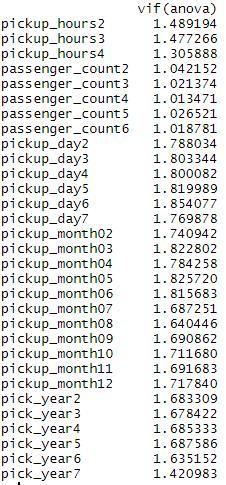
Anova table:



Looking at the table pickup\_day has p-value more than 0.05 therefore we will not consider pickup\_day variable in predicting fare\_amount

1. **Multicollinearity :**

In regression“multicollinearity”refers to predictors that are correlated with predictors. Multicollinearity occurs when our model includes multiple factors that are correlated not only just with the responsive variable but also with each other

* + 1.  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.

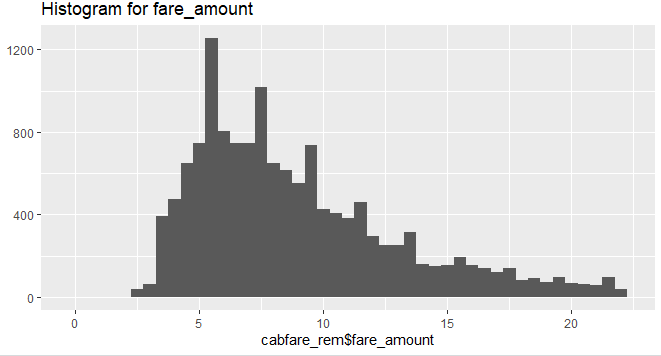
Ex: like heights, weights of people are different for different people and also those will not be in same scale therefore scaling or normalization help to arrive at common values for both the variables

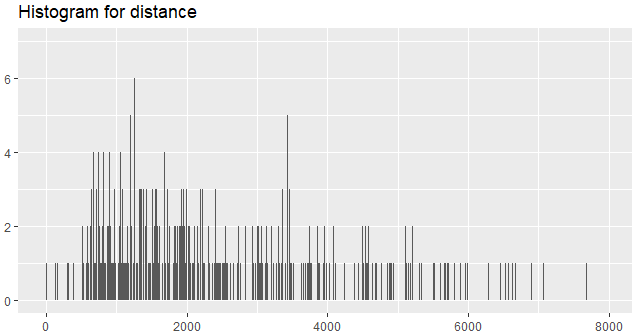
Normalization: Normalization refer to the dividing of a vector by its length. Normalization normalizes the data in the range of 0 to 1. It is generally used when we are planning to use distance method for our model development purpose such as KNN. Normalizing the data improves convergence of such algorithms. Normalisation 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 forstandardization.

Just check out the distribution for numeric variables

For fare\_amount & distance:





Since both the variable are skewed left therefore we go with normalization

3. Splitting Data into Train & Test Data

We have used sample() function in R & sklearn’s train\_test\_split() in python

80% will be train data and 20% we used as test data. Both train & test data we have split from the cabfare\_train dataset only for building and Identifying the best model and later we decided to apply on test.csv data

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 4 Regression Algorithms:

1. Multiple Linear Regression
2. Decision Tree Regression
3. Random Forest Regression
4. Gradient Boosting Technique
5. XG boost Regression

We use RMSE as error metric since it is a timestamp data

1. Multiple Linear Regression:

Performance metrics:

|  |  |  |  |
| --- | --- | --- | --- |
| MAE | MSE | RMSE | MAPE |
| 1.653 | 5.587 | 2.364 | 0.193 |

1. Decision Tree Regression:

|  |  |  |  |
| --- | --- | --- | --- |
| MAE | MSE | RMSE | MAPE |
| 1.807 | 6.279 | 2.506 | 0.209 |

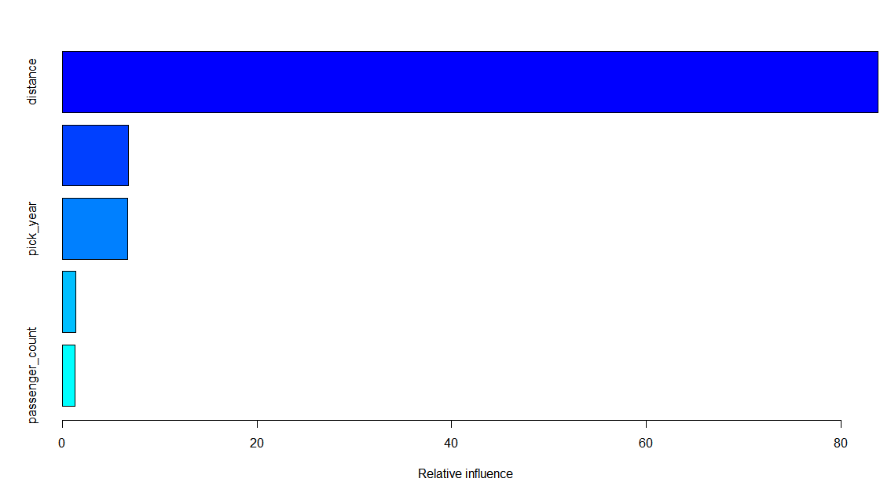
1. Random Forest Regression

|  |  |  |  |
| --- | --- | --- | --- |
| MAE | MSE | RMSE | MAPE |
| 1.987 | 6.998 | 2.645 | 0.251 |

1. Gradient Boosting Technique

|  |  |  |  |
| --- | --- | --- | --- |
| MAE | MSE | RMSE | MAPE |
| 1.640 | 5.445 | 2.333 | 0.190 |

Variable importance:



1. Improving Accuracy

For improving accuracy we have used XGBoost as an ensemble technique

|  |  |  |  |
| --- | --- | --- | --- |
| MAE | MSE | RMSE | MAPE |
| 1.637 | 5.509 | 2.347 | 0.188 |

Though we tried to improve accuracy using XG Boost algorithm we couldn’t achieve and clearly it shows the gradient boosting technique is the winner therefore we fill finalize the gradient boosting technique as the final model.

We will consider RMSE as error metric for this problem and let us list down RMSE for various techniques used so far

|  |  |  |
| --- | --- | --- |
| S.No | Model | RMSE |
| 1 | Multiple Linear Regression | 2.364 |
| 2 | Decision Tree Regression | 2.506 |
| 3 | Random Forest Regression | 2.645 |
| 4 | Gradient Boosting Method | 2.333 |
| 5 | XG Boost Regression | 2.347 |

1. Finalizing the Model and Predicting Test data

Now we finalized Gradient boosting method as our final model for this problem and now we will load test data given in the problem statement and try to predict the target class

We have trained Gradient boosting model on entire training dataset and used that model to predict on test data. Also, we have saved model for later use.