

Ajay Darshan | Data scientist course | March 16, 2019

Cab fare prediction

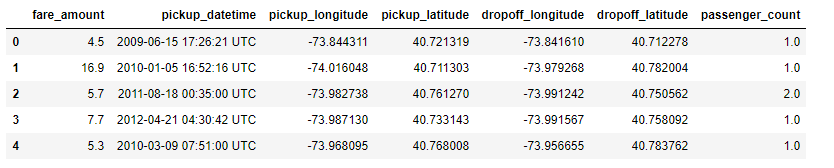
# Chapter 1 Introduction

## Problem statement

The objective of this project is to analyze historical data on cab fares and predict it. For this purpose, we will design a system that predicts the fare amount for a cab ride in the city.

## Data

Sample data is shown in below figure



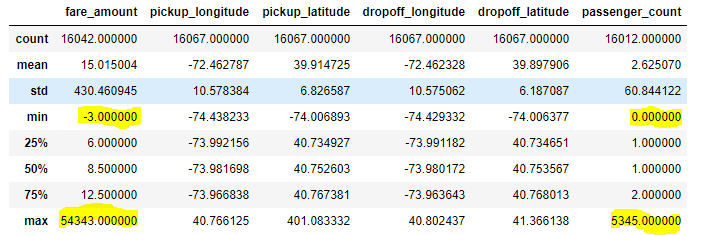
This data contains date and time of pickup, number of passengers and pickup and drop coordinates of latitude and longitude as independent variables. Using these variables fare amount is determined. As this involves prediction of fare amount, this is a regression problem.

# Chapter 2 Methodology

## Pre-Processing

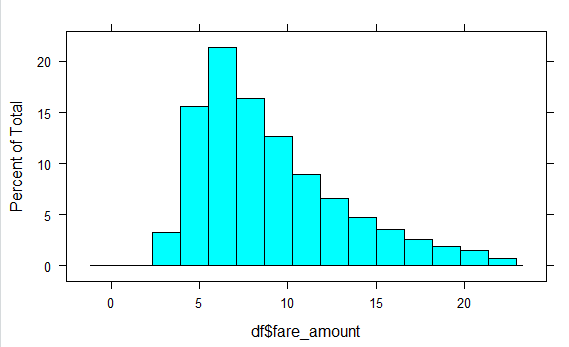
The first step in predictive analysis is pre-processing our data. This step is important as the well-prepared data gives better prediction on simpler model than bad data with complex model. It involves exploring data, cleaning data and visualizing it.

As a first step, all variables are converted into proper data types. Summary of numeric variables are shown in below figure.



As highlighted in above figure, we can see that fare amount is having negative values and maximum value is too high. Also, passenger count of zero and very high value doesn’t make sense. So as part of data cleaning we will remove rows having fare amount negative and passenger count having zero, fraction value and greater than 10. Here maximum passenger count of 10 is selected as maximum seating capacity of cabs which runs daily is 10-seater.

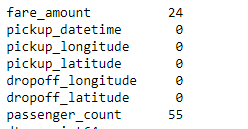
Below figure shows distribution of fare amount after cleaning data.



After this pickup hour, month, weekday and year are derived from pickup\_datetime variable. This is a part of feature engineering.

## Missing value analysis

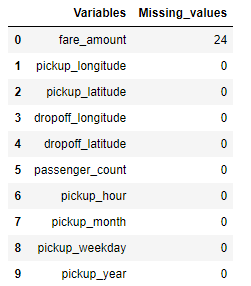
Next step in pre-processing is missing value analysis. Below figure shows number of missing values in data.



Here two variables have missing values fare\_amount and passenger\_count. Two different methods are employed for these two variables.

For passenger\_count, we cannot impute this value based on other independent variable, as these values cannot be influenced by other variables in real world. So, these rows having values missing are removed.

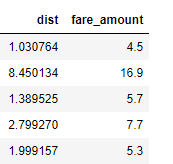
Below figure shows summary after performing above step.



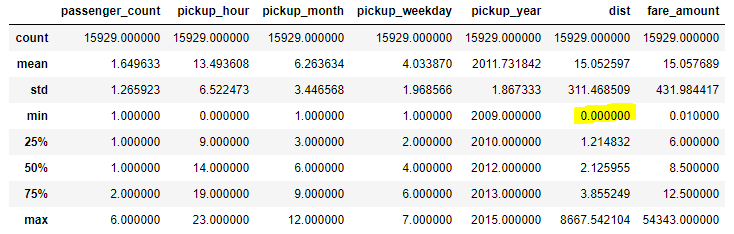
For fare\_amount, this value depends on other independent values. So, it is imputed using KNN imputation having k = 2.

## Feature engineering

In this part, total distance travelled is derived using pickup and drop co-ordinates of latitude and longitude. Distance is calculated using haversine method. Using distance, it is easy to explain relationship b/w distance and fare amount. Below figure shows, with increase in distance, fare amount also increases.

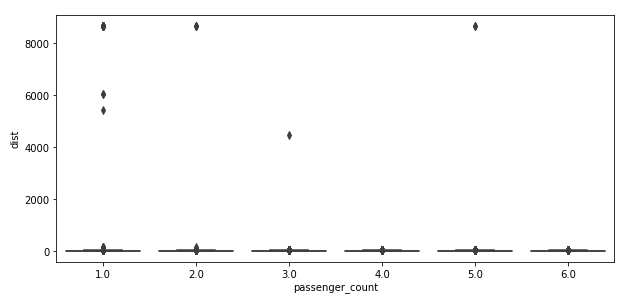


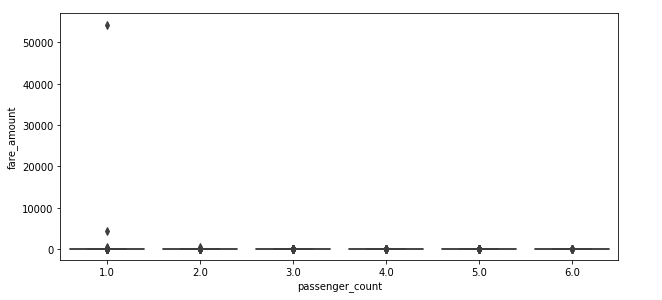
Summary of distance values are given below



As seen from figure, there are values having distance zero. We will remove all rows having distance zero, as this will not contribute any information to model. Higher values will be considered during outlier analysis. Here co-ordinate variables are dropped as both distance and co-ordinates carry same information.

## Outlier analysis

It is performed on two numeric variables distance and fare amount. Below figures show outliers in data.

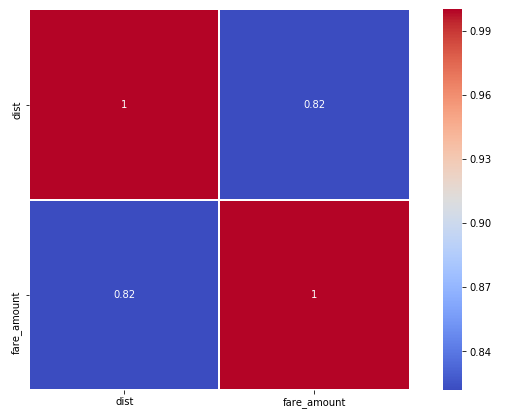


These outliers are imputed using KNN imputation with k=2.

## Feature selection

Feature selection is the process of selecting a subset of relevant features (predictors) for use in model construction. There is a possibility that many variables in our analysis are not important at all or may cause multicollinearity problem. Here, we use correlation analysis for numeric variables and ANOVA for categorical variables.

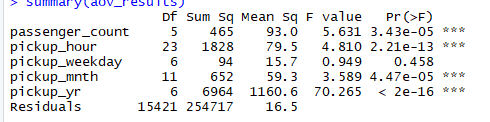
In correlation analysis, we remove variables, if they are highly correlated.



As seen from figure there is a high correlation between distance (independent variable) and fare amount (dependent variable). So, we will keep this variable.

ANOVA is applied on categorical variables v/s numeric fare amount.

Below table shows the summary of ANOVA



As seen from above figure, p value is greater than 0.05(95% confidence) for pickup weekday. We fail to reject NULL hypothesis. So pickup weekday and fare amount variables are independent

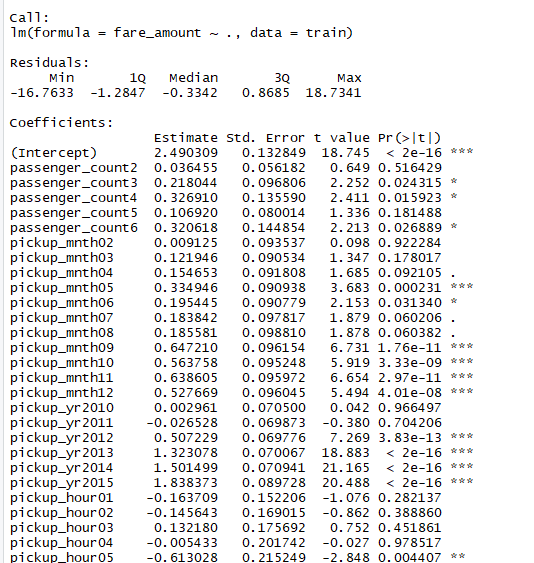
We will remove pickup weekday. Other independent variables are dependent on dependent variable.

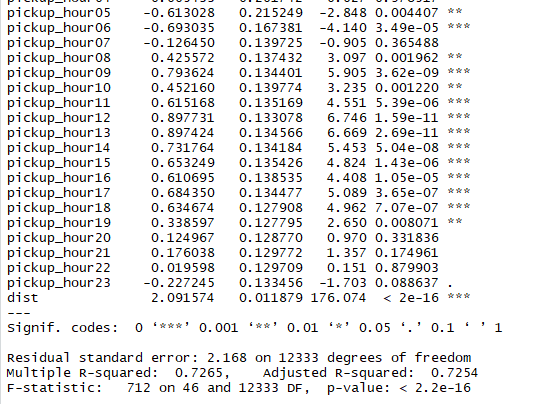
## Modeling

As this is regression problem, we will start from simple regression model and move towards complex ensemble models.

### Linear regression

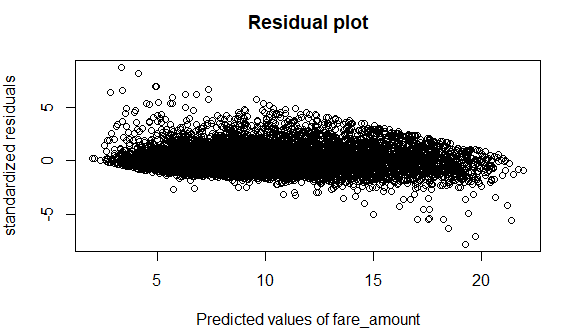
Below figure shows the summary of model applied on data



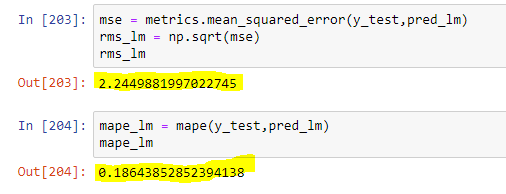


We can see R squared value is 0.7265, this model can explain 72.65% of data. Also, from F statistics we can conclude that full model is adequate.

Below figure shows residual plot



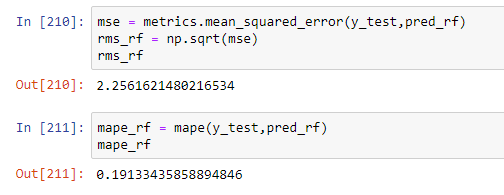
So, we can consider this model for evaluating with test data. After applying on test data, model performance is evaluated with RMSE, MAPE. Evaluation results are shown below,



RMSE is 2.24 and MAPE is 0.1864, accuracy of model is 81.6%. This model has performed fairly good on test data. But still we can consider other model and evaluate the results.

### Random forest

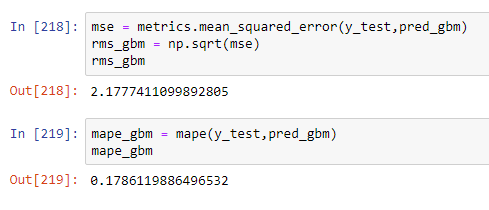
This model is applied on data and evaluated results are shown below



As seen from results above, RMSE is 2.25 and accuracy of model is 80.8%. This model has not performed better than Linear regression. So, we will check other models like GBM and XGBoost.

### Gradient boosting

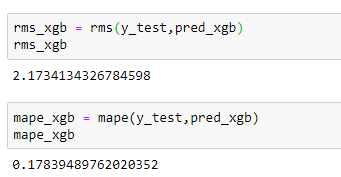
Results of this model on test data are given below



As seen from figure, this model gave better results than Linear regression and random forest. RMSE is 2.1774 and accuracy is 82.2%.

### XGBoost

Results of model is shown in below figure.



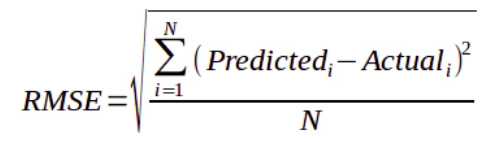
This model also outperformed Linear regression model and Random forest. It has RMSE of 2.1734 and accuracy of 82.1%

# Chapter 3 Conclusion

## Model Evaluation

Model evaluation on test data is necessary in selection of model. Here we will use error metrics RMS as primary measure for selecting the model. It is used because it provides measure of spread of the true y values about the predicted values. It will have same unit as that of value predicted, which helps us to understand the model easily. Here we can expect 68% of y values to be within one RMSE and 95% to be within two RMSE.

RMSE is calculated using below formula



## Model selection

By looking at selection criteria i.e. RMSE of all models, XGBoost and GBM has performed better than linear regression and random forest. Below figures shows distribution of errors for different models.

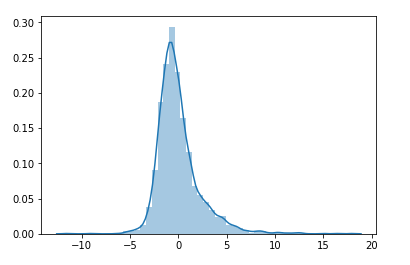


Figure: Distribution plot of (y\_test - pred) for Linear regression

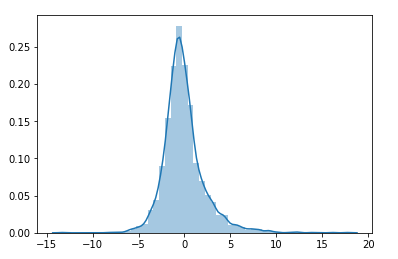


Figure: Distribution plot of (y\_test - pred) for Random forest

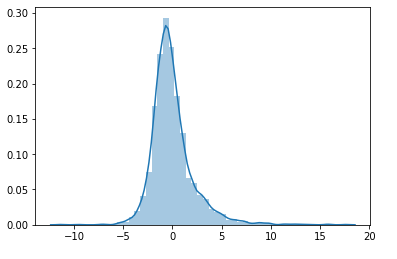


Figure: Distribution plot of (y\_test - pred) for Gradient boosting

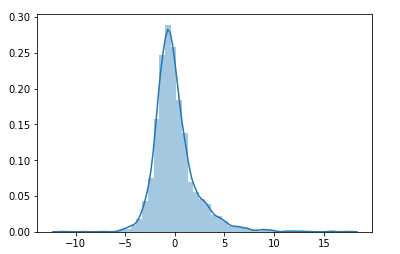


Figure: Distribution plot of (y\_test - pred) for XGBoost

As seen from figures above, errors are normally distributed. So, 68% of y values will be within one error distance and 95% within two error distance.

After considering all above factors, both Gradient boosting and XGBoost can be selected for prediction. In this project, prediction values of both models are submitted.