**Cab Fare Prediction**

***Tanishka***

***JULY 2019***

# Contents

1. **[Introduction](#_bookmark0) 3**
   1. [Problem Statement](#_bookmark1) 3
   2. [Data](#_bookmark2) 3
2. **[Methodology](#_bookmark3) 4**
   1. [Pre Processing](#_bookmark4) 4
      1. Variable identification 4
      2. Data Cleaning 4
      3. Feature Engineering 4
      4. Missing Value Analysis and Treatment 4
      5. Visualization 4
3. Univariate Analysis 4
4. Bivariate Analysis 4
   * 1. Outlier Analysis and Treatment 10
     2. Feature Selection 10
5. Correlation 4
6. Feature Importance 4
   * 1. Feature Scaling 10
   1. Modelling 12
      1. Model Selection 10
7. Decision Tree 4
8. Random Forest 4
9. Simple Linear Regression 4
10. Multiple Linear Regression 4
    * 1. Visualizing models 12
11. Prediction Plots
12. **Conclusion 15**
    1. Model Evaluation 15
       1. Mean Absolute Error (MAE) 15
       2. Mean Squared Error (MSE) 15
       3. Mean Squared Error (RMSE) 15
       4. R2 15
       5. Model Score 15
    2. Model Selection 16
    3. Execution 20

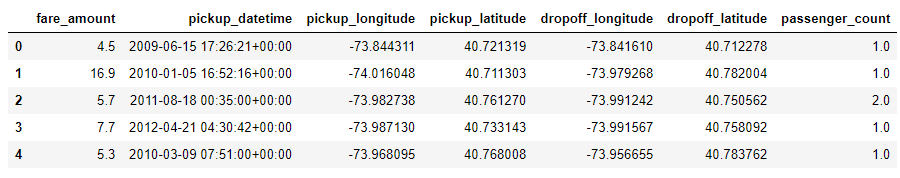
# *Chapter 1*

## *Introduction*

### *Problem Statement:* The objective of this Project is to predict the fare of the Cab rental in the city. This Fare prediction takes distance, date/time and other factors in account from historical data which was gathered from the pilot project for the same. We would be building a model that can successfully predict the fare of rentals on relevant factors.

### *Data:* As the dataset given has dependent and independent values, it will come under supervise Machine learning. Our task is to build Regression models which will help us predicting the fare for our cab which depends on the factors provided. Given below is a sample of the data set that we are using for our prediction. This dataset contains 07 variables in which 6 are independent variables and 1 (Fare\_amount) is dependent variable.

|  |  |
| --- | --- |
| **Variable** | **Explanation** |
| fare\_amount | Float amount of the ride. |
| 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. |



# *Chapter 2*

## *Methodology*

### *2.1 Pre Processing:* Before we proceeding to create our model on top of the provided data. It is necessary to do Exploratory Data Analysis. EDA is very first and necessary step to take before proceeding further. As the result depends on the data, EDA makes sure the quality of input data is high which will lead to high quality results. We can perform EDA as follows: a) Variable Identification:.

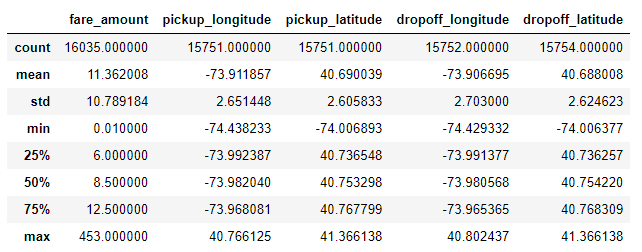
#### **2.2.1 Variable Identification:** In Order to understand the data, we need to first, Identifying Predictor (Input) and Target (output) variables. Then, Identifying the data type and category of the variables

##### Types of Variable: Our Target Variable is ‘fare\_amount’, and Predictor variables are (pickup\_datetime, pickup\_longitude,pickup\_latitude,dropoff\_longitude,dropoff\_latitude,passenger\_count) .

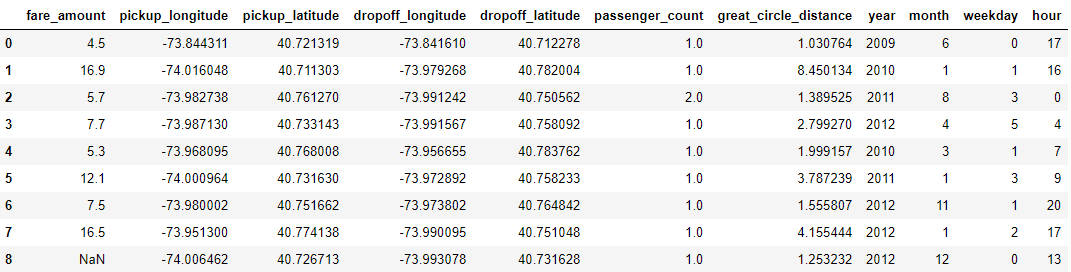
##### Data Types: Character(passenger\_count), Numeric(fare\_amount) ,factor( pickup\_longitude,pickup\_latitude,dropoff\_longitude,dropoff\_latitude), datetime(pickup\_datetime ). We have converted the data as per our requirement NOTE: I made passenger\_count to category after making it to Int and cleaning it.

##### Variable Categories: Categorical (passenger\_count), Continuous (pickup\_datetime,pickup\_longitude,pickup\_latitude,dropoff\_longitude,dropoff\_latitude) Before I made passanger\_count to category , our data looked like below After I made passanger\_count to category our data looks like

#### **2.2.2 Data Cleaning:** We can clearly observe from Summary in R and Describe Function in Python that Passenger counts of maximum values is very high and Pickup\drop off longitude and latitude is not under 90 and 180 which is not possible as per geographical information. Passenger count is too high as we know cab can accommodate max 8 passenger if consider its SUV. Distance is also extremely high as per regular cab which roams within the city. So, to proceed further we are, Keeping fare\_amount under 100 (as during visualization I realized the distribution of data is under 60 and after that just tail is stretched), Pickup\drop off longitude and latitude under under 90 and 180, passenger\_count under 8. I have not dropped observation for not but just imputed it with NA. After cleaning our data looks like



#### **2.2.3 Feature Engineering:** Before we proceed , 1) I have split pick\_up datetime into hours, day, month and year and dropped the main variable pickup\_datetime. This will help us understanding our data more efficiently. 2) I have calculated distance based on our Pickup\drop off longitude and latitude using great\_circle\_distance function. After Feature engineering our data looks like this



#### **2.2.4 Missing values treatment:** Missing values occur when no data value is stored for the variable in an observation. Missing values are a common occurrence, and you need to have a strategy for treating them. A missing value can signify a number of different things in your data. Perhaps the data was not available or not applicable or the event did not happen. It could be that the person who entered the data did not know the right value, or missed filling in. Typically, ignore the missing values, or exclude any records containing missing values, or replace missing values with the mean, or infer missing values from existing values. We check for missing values in our data and came to know we have missing data in almost every variable

I have deleted observations with Missing Values in R coding since distribution of missing values are same across the different variable and as a another try imputed using mean, median , KNN and other fitting formulas in Python.

#### **2.2.5 Visualization: E**xploring Variables one by one to understand central tendency, spread of the variable, distribution of each category, association and disassociation between variables at a predefined significance level.

As we see below our target variable data is under 0 to 60 and further that tail is stretched.

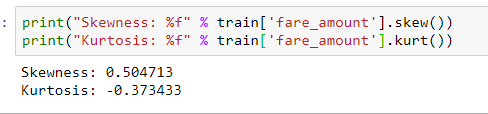
##### ***2.2.5.1 Univariate Analysis:*** Checking the distribution of individual variables As We see below our target variable data is under 0 to 60 and further that tail is stretched As per the below figure , we can understand our city is under 20 to 40 latitude, and passenger frequency of 1 is higher.

##### ***2.2.5.2 Bi-variate Analysis:*** We are checking relation of variables with each other to understand the relationship of variables with our target variable Fig 1.1 – bar pot for distribution of year, month, weekday and hour against fare\_amount Fig 1.2 lm plot for distance Vs fare\_amount Fig 1.3 lm plot for abs\_lat Vs fare\_amount Fig 1.4 lm plot for abs\_longi Vs fare\_amount

#### ***2.2.6 Outlier treatment:*** An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. Outliers can drastically change the results of the data analysis and statistical modelling. There are numerous unfavourable impacts of outliers in the data set. It increases the error variance and reduces the power of statistical tests. If the outliers are non-randomly distributed, they can decrease normality. They can also impact the basic assumption of Regression, ANOVA and other statistical model assumptions. In our data, we can clearly observe from Summary in R and Describe Function in Python that Passenger counts of maximum values is very high and Pickup\drop off longitude and latitude is not under 90 and 180. Hence I have marked passenger count more than 8 to 2(to avoid outliers) as a cab can only accommodate 8 people at the max and kept Pickup\drop off longitude under 90 and 180 as per geographical information. I have also minimized Distance to 500km (but my model was highly skewed so further minimize to 100).

#### ***2.2.7 Feature Selection:*** We have converted Pickup \ drop off latitude and longitude as absolute location points and from these variables we have extracted the total distance travelled. From Pick date and Time extracted Year, Month, day, Hours. Here is some graphical representation of the same A picture containing writing implement, pencil, stationary Description generated with very high confidence **Correlation Analysis :** We make heat map to understand the co relation of contiguous variable. A heatmap is a graphical representation of data where the individual values contained in a matrix are represented as colors. Here each numerical variable’s correlation is mapped with each other’s in a matrix which has been plotted in the following heatmap. **Feature Importance:** The concept is really straightforward: We measure the importance of a feature by calculating the increase in the model’s prediction error after permuting the feature. A feature is “important” if shuffling its values increases the model error, because in this case the model relied on the feature for the prediction. A feature is “unimportant” if shuffling its values leaves the model error unchanged, because in this case the model ignored the feature for the prediction. Checking via Tree: Here , we can see the the importance of distance is extremely high. So, instead of deleting all other variables , I am going to create out model with two inputs one with distance only and one with all the variable including distance.

#### ***2.8 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. Normalization also called Min-Max scaling. It is the process of reducing unwanted variation either within or between variables. Normalization brings all of the variables into proportion with one another. It transforms data into a range between 0 and 1. All our continuous variables are already normalized except the target and the distance which we took out from logi/lati variable which we prefer not to scale because its variation is spread quite widely and after scaling, the difference between the number is diminishing. Checking for Skewness and Kurtosis: Skewness is usually described as a measure of a dataset’s symmetry – or lack of symmetry. A perfectly symmetrical data set will have a skewness of 0. If the skewness is between -0.5 and 0.5, the data are fairly symmetrical If the skewness is between -1 and – 0.5 or between 0.5 and 1, the data are moderately skewed. If the skewness is less than -1 or greater than 1, the data are highly skeweness.



Our data is little skewed and sample does not look Gaussian. Skewed data messes up the predictive model and it affects the regression intercept, coefficients associated with the model.

So, to reduce the Skewness I am log transforming our data. After log transform our data looks like below.

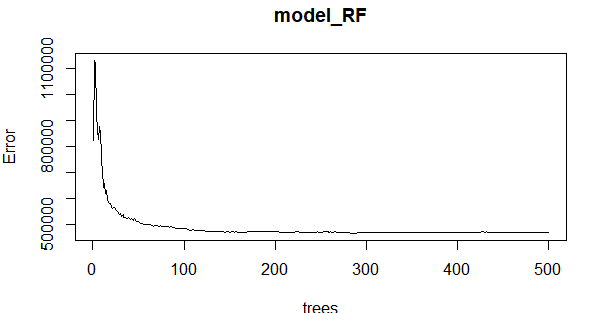
### ***2.2 Modeling***

#### ***2.2.1 Model Selection:*** For modelling, we are going to use some famous models to our data-set and will conclude the result according to it.

##### ***a) Decision Tree:*** Decision tree is a rule. Each branch connects nodes with “and” and multiple branches are connected by “or”. It can be used for classification and regression. It is a supervised machine learning algorithm. Accept continuous and categorical variables as independent variables. Extremely easy to understand by the business users. Split of decision tree is seen in the below tree. Here , is the tree for our mode

##### ***b) Random Forest:*** Random Forest or decision tree forests are an ensemble learning method for classification, regression and other tasks. It consists of an arbitrary number of simple trees, which are used to determine the final outcome. In the regression problem, their responses are averaged to obtain an estimate of the dependent variable. Using tree ensembles can lead to significant improvement in prediction accuracy (i.e., better ability to predict new data cases). The goal of using a large number of trees is to train enough that each feature has a chance to appear in several model--> As we increase the number of trees the error count decrease until a point (100 trees) and then becomes constant. Error vs number of trees to be used graph is as follows:

.



--> We can also call our model and get below details

Call:

randomForest(formula = fare\_amount ~ ., data = data\_train, importance = TRUE, ntree = 500)

Type of random forest: regression

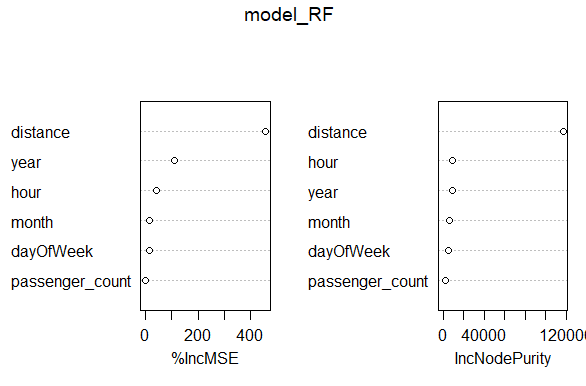
Number of trees: 500

No. of variables tried at each split: 2

Mean of squared residuals: 4.049305

% Var explained: 71.36

--> We can check the importance of our variables in Random Forest Model with (varImpPlot(MODELNAME))



The first graph shows that if a variable is assigned values by random permutation by how much will the MSE increase. Higher the value, higher the importance. On the other hand, node purity is measured by the Gini index which is the difference between before and after split on that variable.

##### ***c) Linear Regression: L***inear regression is the most basic type of regression and commonly used predictive analysis. Linear regression is an approach for modelling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression. --> Following is the summary of the Linear model:

Call:

lm(formula = fare\_amount ~ ., data = train\_cab\_final)

Residuals:

Min 1Q Median 3Q Max

-11.9659 -1.3063 -0.3959 0.8551 17.9954

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -7.197e+02 2.138e+01 -33.669 < 2e-16 \*\*\*

passenger\_count 1.013e-01 3.716e-02 2.727 0.0064 \*\*

month 5.585e-02 5.755e-03 9.705 < 2e-16 \*\*\*

year 3.593e-01 1.062e-02 33.825 < 2e-16 \*\*\*

dayOfWeek 4.729e-02 1.005e-02 4.706 2.56e-06 \*\*\*

hour 7.345e-03 3.349e-03 2.193 0.0283 \*

distance 2.018e+00 1.313e-02 153.711 < 2e-16 \*\*\*

---

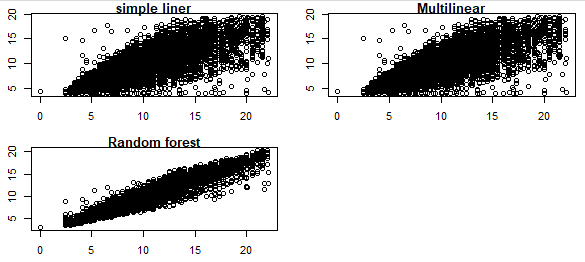
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.112 on 11436 degrees of freedom

Multiple R-squared: 0.6847, Adjusted R-squared: 0.6846

F-statistic: 4139 on 6 and 11436 DF, p-value: < 2.2e-16

#### **2.2.2 Visualizing models :** We can see the plots of our predicted model to understand it better



# *Chapter 3*

## *Conclusion*

1. ***Model Evaluation:*** Model evaluation is done on basis of evaluation metrics or error metrics. Evaluation metrics explain the performance of a model. An important aspect of evaluation metrics is their capability to discriminate among model results. Simply, building a predictive model is not our motive. But, creating and selecting a model which gives high accuracy on out of sample data. Hence, it is crucial to check accuracy or other metric of the model prior to computing predicted values. In our data as we applied regression models we have error metrics like Mean square error (MSE), MAPE, Root mean square error (RMSE), Mean absolute error (MAE). ***With all Variables***

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Language/Model |  | Python |  |  |  |  | R |  |
| MODELS | MSE | RMSE | MAPE | R-SQ | Score | MSE | RMSE | MAPE |
| Decision Tree | 20.24369617 | 4.49929952 | 24.06285692 | 0.783105317 | 0.82 |  | 1.32 | 13.63 |
| Random Forest | **17.78135176** | **4.216794014** | **22.88214811** | **0.809487327** | 0.97 |  | 0.95 | 10.52 |
| Linear regression | 26.61443151 | 5.158917668 | 26.94376004 | 0.714848087 | 0.66 |  |  |  |
| Linear Regression Model Using Ridge | 27.80949915 | 5.273471262 | 27.87948324 | 0.702043912 | 0.64 |  | 2.213257 | 23.71183 |

***With distance only***

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Language/Model |  | Python |  |  |  |  | R |  |
| MODELS | MSE | RMSE | MAPE | R-SQ | Score | MSE | RMSE | MAPE |
| Decision Tree | 21.07802666 | 4.591081209 | 25.94061573 | 0.774166146 | 0.77 |  |  |  |
| Random Forest | 28.35212599 | 5.324671444 | 28.80842963 | 0.696230108 | 0.96 |  | 0.95 | 10.52 |
| Linear regression | 31.3680148 | 5.600715562 | 28.72551651 | 0.663917321 | 0.59 |  | 2.213257 | 23.71183 |
| Ridge Regression | 31.3680357 | 5.600717427 | 28.72558271 | 0.663917097 | 0.59 |  |  |  |

### *Model Selection :* We can see that all models perform comparatively on average and therefore we select **random forest** classifier models for better prediction. From the above plots of Actual Vs Predicted values, we can infer that values of Random forest falls on straight line indicating random forest fits better than the other three models. Also amongst the three models, Random forest has best R-sq. (Coef. of determination). Hence we’ll fix Random Forest as our model.

### *Execution/Answer :* Applying the prediction model on test data, we get below prediction distribution.

------------------------------------------------------- END OF DOCUMENT -------------------------------------------------