***Project Report***

***On***

***Cab fare Prediction***

AUTHOR: Saksham Sanjay Shivhare



INDEX



* 1. Problem Statement
  2. Data



* + - Pre-Processing
    - Modelling
    - Model Selection



* 1. Data exploration and Cleaning (Missing Values and Outliers)
  2. Creating some new variables from the given variables
  3. Selection of variables
  4. Some more data exploration
     + Dependent and Independent Variables
     + Uniqueness of Variables
     + Dividing the variables categories
  5. Feature Scaling



* 1. Linear Regression
  2. Decision Tree
  3. Random Forest
  4. Gradient Boosting
  5. Hyper Parameters Tunings for optimizing the results



* 1. Model Evaluation
  2. Model Selection
  3. Some Visualization facts



**Chapter 1**

# Introduction

* 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.

* 1. **Data**

1. Understanding of data is the very first and important step in the process of finding solution of any business problem. Here in our case our company has provided a data set with following features, we need to go through each and every variable of it to understand and for better functioning.
2. Size of Dataset Provided: - 16067 rows, 7 Columns (including dependent variable) Missing Values: Yes
3. Outliers Presented: Yes
4. Below mentioned is a list of all the variable names with their meanings:

|  |  |
| --- | --- |
| **Variables** | **Description** |
| **fare\_amount** | Fare amount |
| **pickup\_datetime** | Cab pickup date with time |
| **pickup\_longitude** | Pickup location longitude |
| **pickup\_latitude** | Pickup location latitude |
| **dropoff\_longitude** | Drop location longitude |
| **dropoff\_latitude** | Drop location latitude |
| **passenger\_count** | Number of passengers sitting in the cab |

**Chapter 2**

# Methodology

* **Pre-Processing**

In this type of predictive model we need to build a model which can predict the fare ( in this data set) or continuous variable. The data that we have received is a raw data which as missing values, noise, outliers and some NAN values. We need to carry to multiple data preprocessing steps to make date usable for predictive modeling. Below are the preprocessing steps used in this model.

* + Data exploration and Cleaning
  + Missing values treament
  + Outlier Analysis
  + Feature Selection
  + Features Scaling
  + Skewness and Log transformation
* **Modelling**

Once all the Pre-Processing steps has been done on our data set, we are ready to move to the next step which is modelling. Modelling plays an important role in deciding which algorithm is to be used to create final model. Choice of the model depends upon the nature of data and the target variable. We need to use regression model if the data in continuous and we need to use classification model is the date demands classification. . As per our data set following models need to be tested:

* + Linear regression
  + Decision Tree
  + Random forest,
  + Gradient Boosting
* We have also used hyper parameter tunings to check the parameters on which our model runs best. Following are two techniques of hyper parameter tuning we have used:
  + Random Search CV
  + Grid Search CV
* **Model Selection**

The final step of our methodology will be the selection of the model based on the different output and results shown by different models. We have multiple parameters which we will study further in our report to test whether the model is suitable for our problem statement or not.

**Chapter 3**

# Pre-Processing

**3.1 Data exploration and Cleaning (Missing Values and Outliers)**

The very first step which comes with any data science project is data exploration and cleaning which includes following points as per this project:

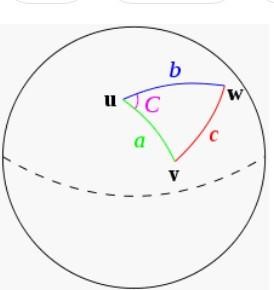
* separate the combined variable
* The fare amount value has some negative value which is not possible and should be removed
* The No. of passengers columns has some outlier no. of passengers cannot be more than 6 and some rows have 100 passengers which need to be treated
* Latitudes range from -90 to 90. Longitudes range from -180 to 180. We need to remove the rows if any latitude and longitude lies beyond the ranges**.**

**3.2 Creating some new variables from the given variables.**

Here in our data set our variable name pickup\_datetime contains date and time for pickup. So we tried to extract some important variables from pickup\_datetime:

* Year
* Month
* Date
* Day of Week
* Hour
* Minute

Also, we tried to Also, we tried to find out the distance using the haversine formula which says:

The **haversine formula** determines the great-circle distance between two points on a sphere given their longitudes and latitudes. Important in navigation, it is a special case of a more general formula in spherical trigonometry, the law of haversines, that relates the sides and angles of spherical triangles.

So our new extracted variables are:

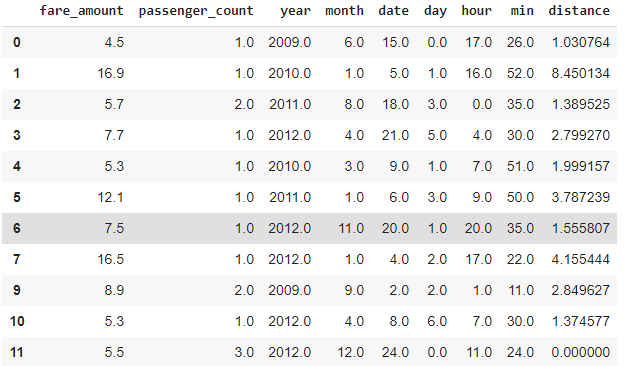
* + fare\_amount
  + pickup\_datetime
  + pickup\_longitude
  + pickup\_latitude
  + dropoff\_longitude
  + dropoff\_latitude
  + passenger\_count
  + year
  + Month
  + Date
  + Day of Week
  + Hour
  + Minute
  + Distance

3.3 **Selection of variables**

Now as we know that all above variables are of now use so we will drop the redundant variables:

* pickup\_datetime
* pickup\_longitude
* pickup\_latitude
* dropoff\_longitude
* dropoff\_latitude

Now only following variables we will use for further steps:



## Some more data exploration

In this report we are trying to predict the fare prices of a cab rental company. So here we have a data set of 16067 observations with 9 variables including one dependent variable.

* + 1. **Below are the names of Independent variables:**

**passenger\_count, year, Month, Date, Day of Week, Hour, distance, minute**

Our Dependent variable is: **fare\_amount**

* + 1. **Uniqueness in Variable**

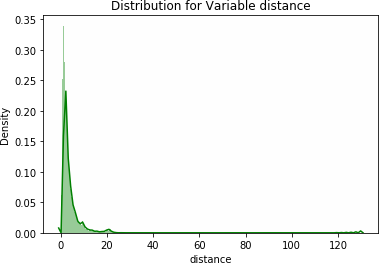
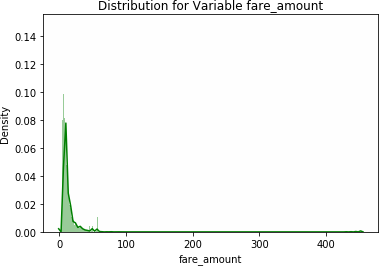
We need to look at the unique number in the variables which help us to decide whether the variable is categorical or numeric. So, by using python script ‘nunique’ we tried to find out the unique values in each variable. We have also added the table below:

|  |  |
| --- | --- |
| **Variable Name** | **Unique Counts** |
| fare\_amount | 450 |
| passenger\_count | 7 |
| year | 7 |
| Month | 12 |
| Date | 31 |
| Day of Week | 7 |
| Hour | 24 |
| distance | 15424 |

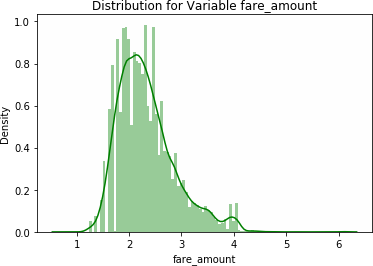
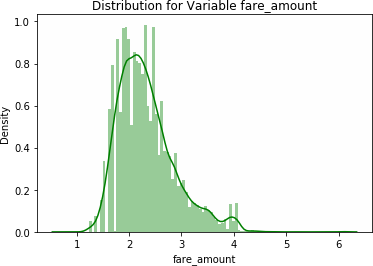
## Feature Scaling

**Skewness** is asymmetry in a statistical distribution, in which the curve appears distorted or skewed either to the left or to the right. Skewness can be quantified to define the extent to which a distribution differs from a normal distribution. Here we tried to show the skewness of our variables and we find that our target variable absenteeism in hours having is one sided skewed so by using **log transform** technique we tried to reduce the skewness of the same.

Below mentioned graphs shows the probability distribution plot to check distribution before log transformation:



Below mentioned graphs shows the probability distribution plot to check distribution after log transformation:



As our continuous variables appears to be normally distributed so we don’t need to use feature scaling techniques like normalization and standardization for the same.

**Chapter 4**

# Modelling

After a thorough preprocessing, we will use some regression models on our processed data to predict the target variable. Following are the models which we have built –

* Linear Regression
* Decision Tree
* Random Forest
* Gradient Boosting
* Lasso Regression
* Ridge Regression

Before running any model, we will split our data into two parts which is train and test data. Here in our case we have taken 80% of the data as our train data. Below is the snipped image of the split of train test.

* 1. **Linear Regression**

[Multiple linear regression](http://www.statisticssolutions.com/academic-solutions/membership-resources/member-profile/data-analysis-plan-templates/data-analysis-plan-multiple-linear-regression/) is the most common form of linear regression analysis. Multiple regression is an extension of simple linear regression. It is used as a predictive analysis, when we want to predict the value of a variable based on the value of two or more other variables. The variable we want to predict is called the dependent variable (or sometimes, the outcome, target or criterion variable).

* 1. **Decision Tree**

A tree has many analogies in real life, and turns out that it has influenced a wide area of machine learning, covering both classification and regression. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. As the name goes, it uses a tree-like model of decisions.

Below is the screenshot of the query we executed and the result shown, we will compare the results of each model in a combined table later on.

* 1. **Random Forest**

Random forests or random decision forests are an ensemble learning method for classification, regression and other task, 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 overfitting to their training set.

To say it in simple words: Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

* 1. **Gradient Boosting**

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

* 1. **Lasso and Ridge regression**

Ridge and Lasso might appear to work towards a common goal, the inherent properties and practical use cases differ substantially. If you’ve heard of them before, you must know that they work by penalizing the magnitude of coefficients of features along with minimizing the error between predicted and actual observations. These are called ‘regularization’ techniques. The key difference is in how they assign penalty to the coefficients:

Ridge Regression:

Performs L2 regularization, i.e. adds penalty equivalent to square of the magnitude of coefficients

Minimization objective = LS Obj + α \* (sum of square of coefficients)

Lasso Regression:

Performs L1 regularization, i.e. adds penalty equivalent to absolute value of the magnitude of coefficients

Minimization objective = LS Obj + α \* (sum of absolute value of coefficients)

* 1. **Hyper Parameters Tunings for optimizing the results**

Model hyper parameters are set by the data scientist ahead of training and control implementation aspects of the model. The weights learned during training of a linear regression model are parameters while the number of trees in a random forest is a model hyper parameter because this is set by the data scientist.

Hyper parameters can be thought of as model settings. These settings need to be tuned for each problem because the best model hyper parameters for one particular dataset will not be the best across all datasets. The process of hyper parameter tuning (also called hyper parameter optimization) means finding the combination of hyper parameter values for a machine learning model that performs the best - as measured on a validation dataset - for a problem.

We have two hyper parameters tuning techniques

* + - Random Search CV
    - Grid Search CV

1. **Random Search CV**: This algorithm set up a grid of hyper parameter values and select random combinations to train the model and score. The number of search iterations is set based on time/resources.
2. **Grid Search CV**: This algorithm set up a grid of hyper parameter values and for each combination, train a model and score on the validation data. In this approach, every single combination of hyper parameters values is tried which can be very inefficient.

**Chapter 5**

# Conclusion

* 1. **Model Evaluation**

The main concept of looking at what is called residuals or difference between our predictions f(x[I,]) and actual outcomes y[i].

In general, most data scientists use two methods to evaluate the performance of the model:

* + 1. **RMSE** (Root Mean Square Error): is a frequently used measure of the difference between values predicted by a model and the values actually observed from the environment that is being modelled.



*n*

*i*1

( *X*  *X*

*obs*,*i*

*n*

*mo del*,*i*

)2

*RMSE* 

* + 1. **R Squared(R^2):** is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression. In other words, we can say it explains as to how much of the variance of the target variable is explained.
    2. We have shown both train and test data results, the main reason behind showing both the results is to check whether our data is overfitted or not.

Below table shows the model results before applying hyper tuning:

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | **Model Name** | **RSME** | **R2\_Score** | | Linear Regression | 0.36054 | 0.5624 | | Decision Tree Regressor | 0.37582 | 0.52461 | | Random Forest Regressor | 0.2771 | 0.7415 | | Ridge | 0.3605 | 0.5624 | | Lasso | 0.5451 | -0.00014 | | Gradient boosting Algorithm | 0.2686 | 0.7571 | |

Below table shows results post using hyper parameter tuning techniques:

|  |  |  |  |
| --- | --- | --- | --- |
| **Model Name** | **Parameter** | **RMSE (Test)** | **R Squared (Test)** |
| **Grid Search CV** | Random Forest | 0.2764 | 0.74270 |
| Gradient Boosting | 0.27166 | 0.75161 |

Above table shows the results after tuning the parameters of our two best suited models i.e. Random Forest and Gradient Boosting. For tuning the parameters, we have used Grid Search CV under which we have given the range of n\_estimators, depth and CV folds.

* 1. **Model Selection**

On the basis RMSE and R Squared results a good model should have least RMSE and max R Squared value. So, from above tables we can see:

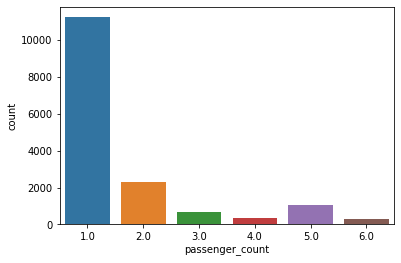
* From the observation of all RMSE Value and R-Squared Value we have concluded that,
* Both the models- Gradient Boosting Default and Random Forest perform comparatively well while comparing their RMSE and R-Squared value.
* After this, I chose Gradient boosting CV and Grid Search CV to apply cross validation technique and see changes brought about by that.
* After applying tunings Gradient boosting model shows best results compared to gradient boosting.
* So finally, we can say that Gradient boosting model is the best method to make prediction for this project with highest explained variance of the target variables and lowest error chances with parameter tuning technique Grid Search CV.

**Finally, I used this method to predict the target variable for the test data file shared in the problem statement. Results that I found are attached with my submissions.**

* 1. **Some more visualization facts:**

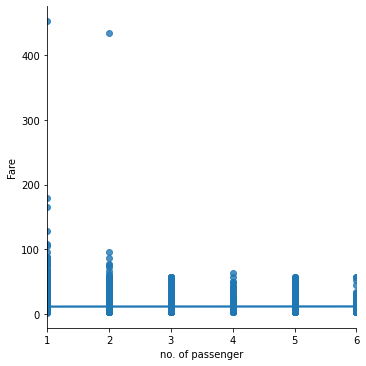
1. **Number of passengers and fare**

We can see in below graph that single passengers are the most frequent travelers, and the highest fare also seems to come from cabs which carry just 1 passenger.



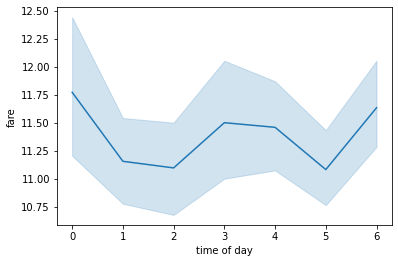
1. No. of passengers and the fare

Here we can see that fare was maximum for the single passenger travelling



1. Day and the fare

Here we can see that the fare has increased to day the weekends



1. Hour and the fare

Here we can see that fare was highest during the afternoon hours

