Daffodil International University

Introduction to mechine learning

project

Mahmuda akter lubna		ID: 181-35-2503	
Department:	:	Software Engineering	
Project tittle : Cab fair prediction report			

22th april , 2021

Table of Contents

Chapter 1: Abstract

Cab rides paint a vibrant picture of life in the city. The millions of rides taken each month can provide insight into traffic patterns, road blockage, or large-scale events that attract many New Yorkers. With ridesharing apps gaining popularity, it is increasingly important for taxi companies to provide visibility to their estimated fare and ride duration, since the competing apps provide these metrics upfront. Predicting fare and duration of a ride can help passengers decide when is the optimal time to start their commute, or help drivers decide which of two potential rides will be more profitable, for example. Furthermore, this visibility into fare will attract customers during times when ridesharing services are implementing surge pricing.

In order to predict duration and fare, only data which would be available at the beginning of a ride was used. This includes pickup and dropoff coordinates, trip distance, start time, number of passengers, and a rate code detailing whether the standard rate or the airport rate was applied. Linear regression with model selection, lasso, and random forest models were used to predict duration and fare amount.

Chapter 2: Introduction

Now a day's cab rental services are expanding with the multiplier rate. The ease of using the services and flexibility gives their customer a great experience with competitive prices.

2.1 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.

2.2 Data

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. Size of Dataset Provided: - 16067 rows, 7 Columns (including dependent variable) Missing Values: Yes
Outliers Presented: Yes 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 3: Literature review

The fare of a taxi ride is function of the mileage and the duration of the ride (sum of drop charge, distance charge and time charge). The drop charge is constant and the distance can easily be estimated but evaluating the duration is not a trivial task. It is the result of complex traffic processes that are nonlinear. One way to predict duration is by doing short term prediction with the help of real time data collection.

In [1] the authors tackle the problem by using data from from buses (GPS) and an algorithm based on Kalman filters. Using a similar approach, [2] uses real time data from smartphone placed inside vehicles. Estimating travel time for highways yields better results than in the cities. This allows for more accurate predictions.

In [3] the authors use a combination of traffic modelling, real time data analysis and traffic history to predict travel time in congested freeways. They try to overcome the assumption that real time analysis communication is instantaneous. A lot of other papers also work on freeways.

In [4] the prediction is done using Support Vector Regression (SVR) while in [5] Neural Networks (SSNN) are used. Predictive estimates of future transit times is a feature that was released in 2015 in the Google Maps API [6]. This shows the importance of being able to predict time travel without having real time data of traffic. We are trying to solve a similar problem: estimating ride duration without real time data, by analysing data collected from taxis. Being able to do such estimation would help making better future predictions. We are trying to solve a similar problem: estimating ride duration without real time data, by analysing data collected from taxis. Being able to do such estimation would help making better future predictions.

Chapter 4: Modeling

Pre-Processing

When we required to build a predictive model, we require to look and manipulate the data before we start modelling which includes multiple preprocessing steps such as exploring the data, cleaning the data as well as visualizing the data through graph and plots, all these steps is combined under one shed which is Exploratory Data Analysis, which includes following steps:

- Data exploration and Cleaning
- Missing values treament
- Outlier Analysis
- Feature Selection
- Features Scaling o Skewness and Log transformation • Visualization

➢ Modelling

Once all the Pre-Processing steps has been done on our data set, we will now further move to our next step which is modelling. Modelling plays an important role to find out the good inferences from the data. Choice of models depends upon the problem statement and data set. As per our problem statement and dataset, we will try some models on our preprocessed data and post comparing the output results we will select the best suitable model for our problem. 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:

1.randon search cv
2.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 5:Pre-Processing

5.1 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:

a. Separate the combined variables.
b. As we know we have some negative values in fare amount so we have to remove those values.
c. Passenger count would be max 6 if it is a SUV vehicle not more than that. We have to remove the rows having passengers counts more than 6 and less than 1.
d. There are some outlier figures in the fare (like top 3 values) so we need to remove those.
e. 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.

5.2 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 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

5.3 Selection of variables

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

- pickup_datetime
- pickup_longitude
- pickup_latitude
- dropoff_longitude
- dropoff_latitude
- Minute

Now only following variables we will use for further steps:

	fare_amount	passenger_count	year	Month	Date	Day of Week	Hour	distance
0	4.5	1.0	2009.0	6.0	15.0	0.0	17.0	1.030764
1	16.9	1.0	2010.0	1.0	5.0	1.0	16.0	8.450134
2	5.7	2.0	2011.0	8.0	18.0	3.0	0.0	1.389525
3	7.7	1.0	2012.0	4.0	21.0	5.0	4.0	2.799270
4	5.3	1.0	2010.0	3.0	9.0	1.0	7.0	1.999157
5	12.1	1.0	2011.0	1.0	6.0	3.0	9.0	3.787239
6	7.5	1.0	2012.0	11.0	20.0	1.0	20.0	1.555807
8	8.9	2.0	2009.0	9.0	2.0	2.0	1.0	2.849627
9	5.3	1.0	2012.0	4.0	8.0	6.0	7.0	1.374577
10	5.5	3.0	2012.0	12.0	24.0	0.0	11.0	0.000000

Variable Names	Variable Data Types
fare_amount	float64
passenger_count	object
year	object
Month	object
Date	object
Day of Week	object
Hour	object
distance	float64

5.4 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 8 variables including one dependent variable.

5.4.1 Below are the names of variables:

Independent variables is : passenger_count, year, Month, Date, Day of Week, Hour, distance

Our Dependent variable is: fare_amount

5.4.2 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

5.4.3 Dividing the variables into two categories basis their data types:

Continuous variables - 'fare_amount', 'distance'.

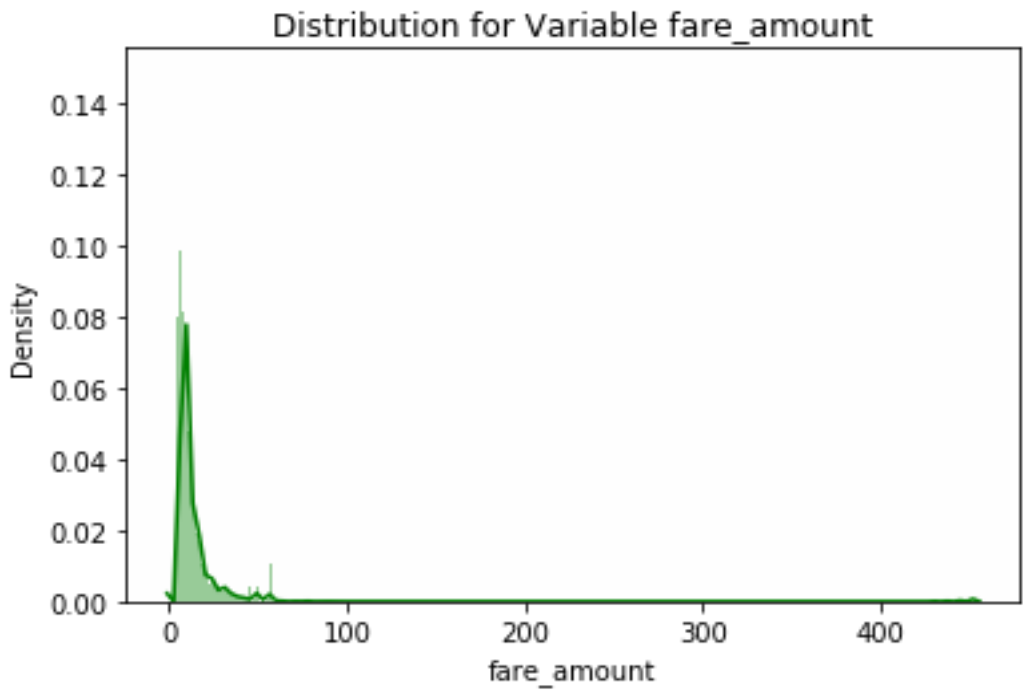
Categorical Variables - 'year', 'Month', 'Date', 'Day of Week', 'Hour',

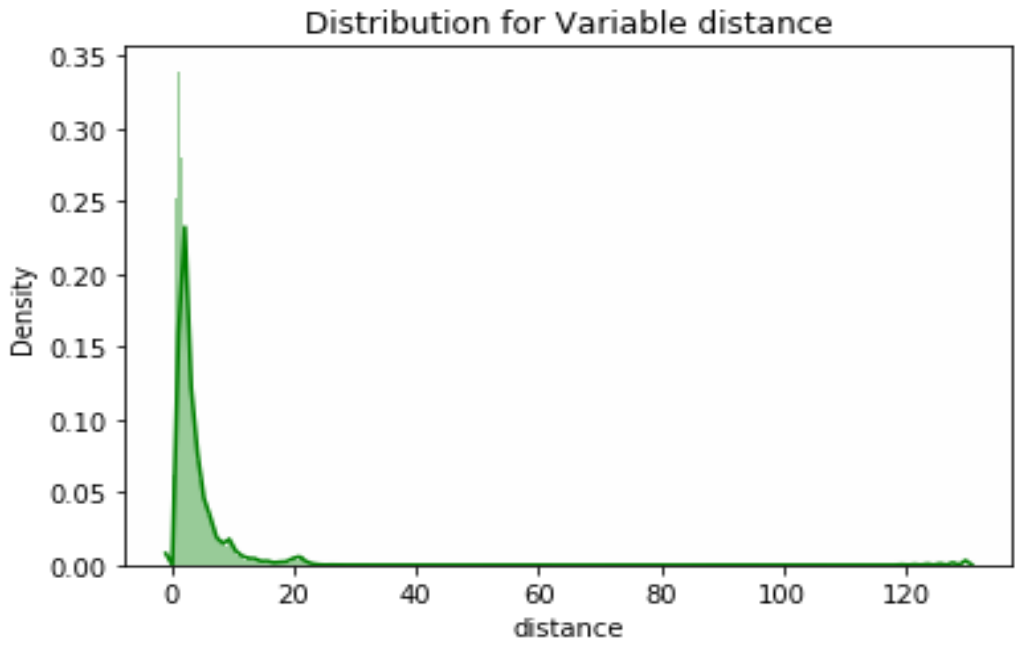
'passenger_count'

5.5 Feature Scaling

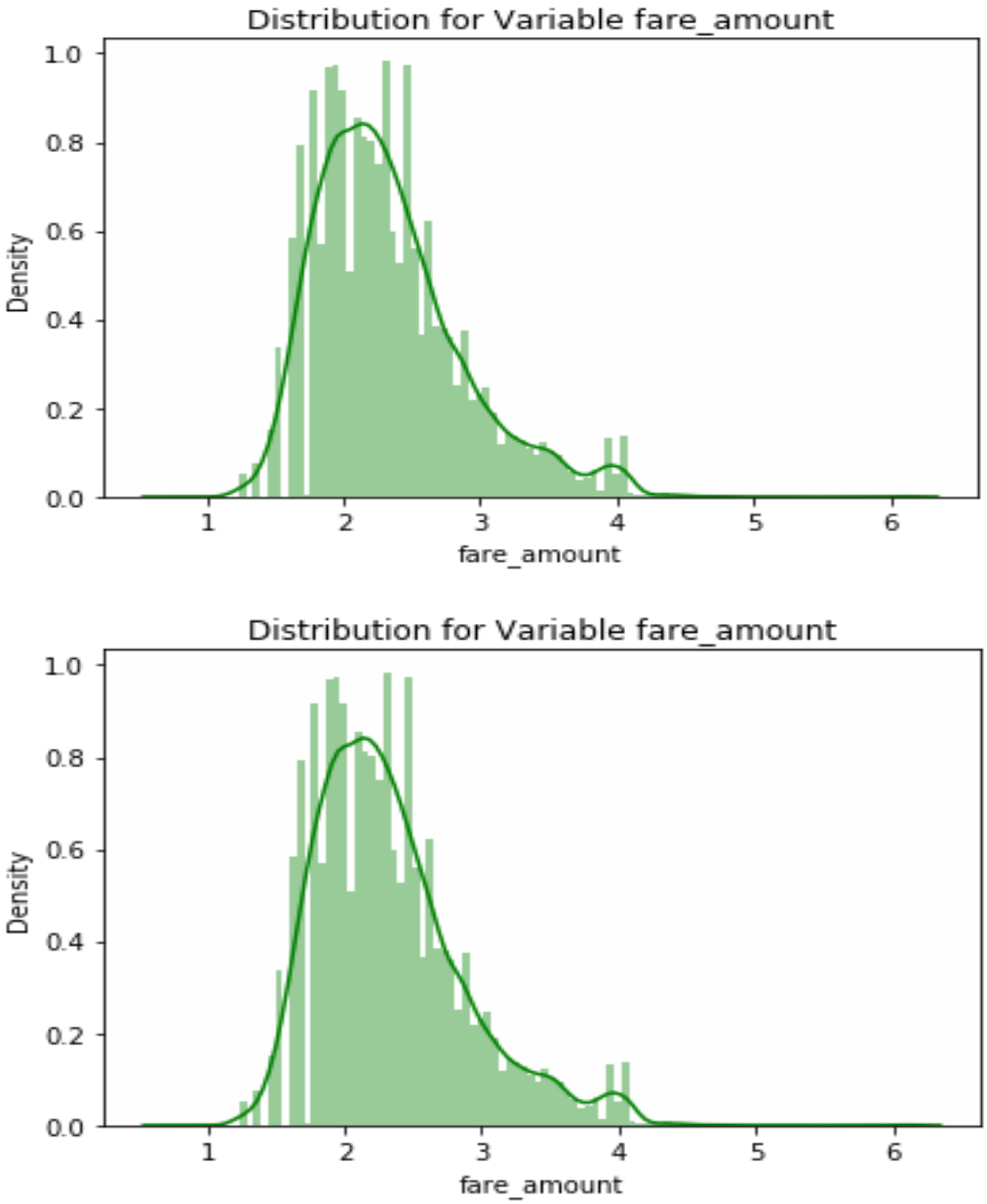
Skewness is asymmetry in a statistical distribution, in which the curve appears distorted or skewed either to the left or to the right. Skewness can be quantified to define the extent to which a distribution differs from a normal distribution. Here we tried to show the skewness of our variables and we 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:

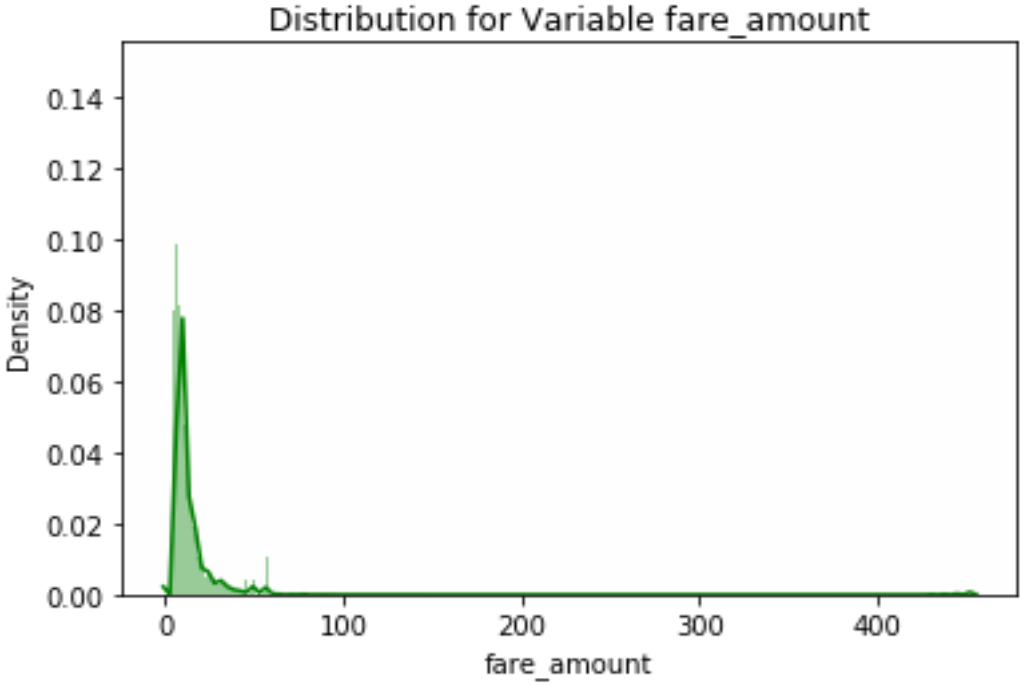

Distribution for Variable fare_amount
Density
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
100
200
300
400
fare_amount

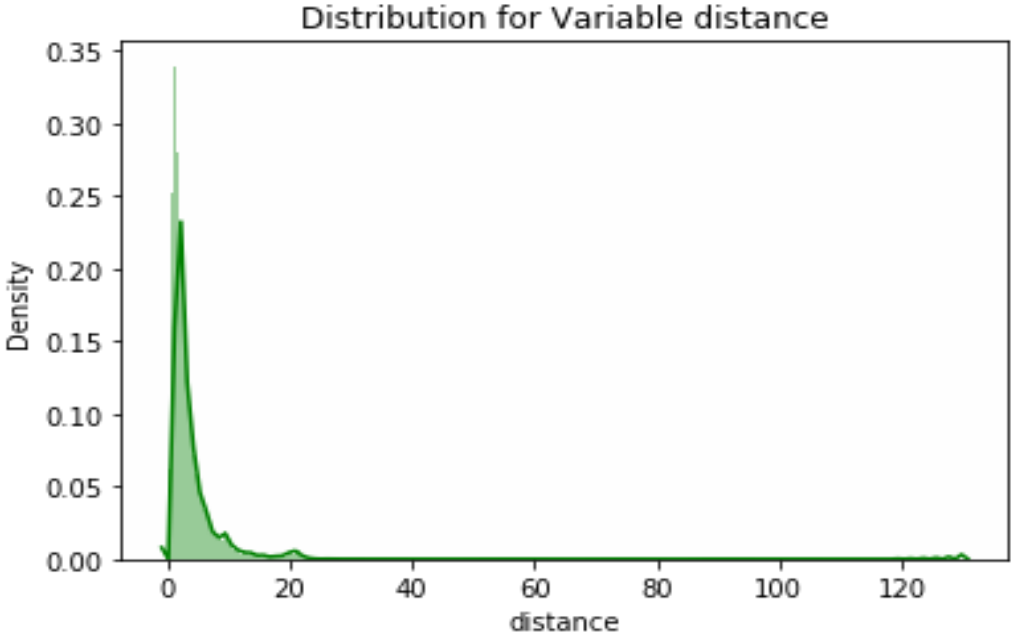

Distribution for Variable distance
Density
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
20
40
60
80
100
120
distance

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

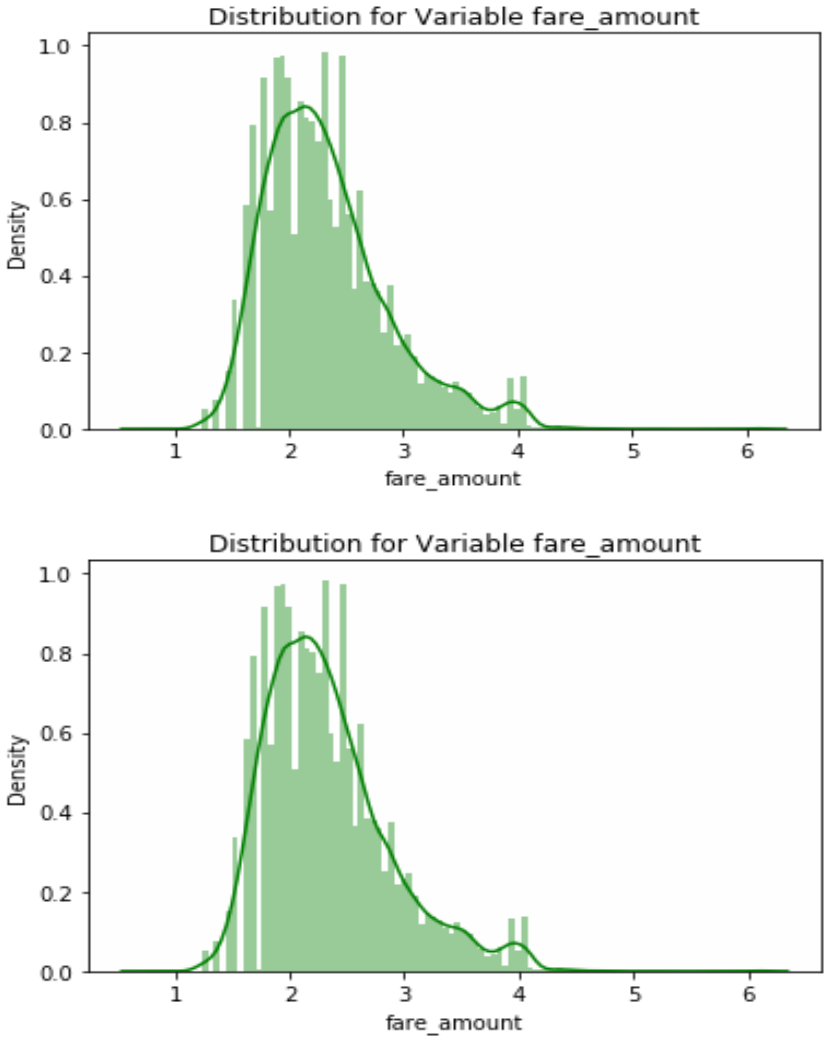


As our continuous variables appears to be normally distributed so we don't need to use feature scaling techniques like normalization and standardization for the same.


Distribution for Variable fare_amount
Density
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
100
200
300
400
fare_amount


Distribution for Variable distance
Density
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
20
40
60
80
100
120
distance

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



Chapter 6: Result discussion

6.1 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:

I. **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.

II. **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.

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	**RMSE**		**R Squared**	
	Train	**Test**	**Train**	**Test**
`Linear Regression	0.27	0.25	0.74	0.77
Decision Tree	0.30	0.28	0.70	0.70
Random Forest model	0.09	0.23	0.96	0.79
Gradient Boosting	0.22	0.22	0.82	0.81

Below table shows results post using hyper parameter tuning techniques:

Model Name	**Parameter**	**RMSE (Test)**	**R Squared (Test)**
Random Search CV	Random Forest	0.24	0.79
	Gradient Boosting	0.25	0.77
Grid Search CV	Random Forest	0.23	0.80
	Gradient Boosting	0.24	0.79

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 Random Search CV and Grid Search CV under which we have given the range of n_estimators, depth and CV folds.

6.2 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 Random Forest CV and Grid Search CV to apply cross validation technique and see changes brought about by that.
- After applying tunings Random forest model shows best results compared to gradient boosting.
- So finally, we can say that Random forest 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.