INSTITUTE FOR ADVANCED COMPUTING AND SOFTWARE DEVELOPMENT AKURDI, PUNE

Documentation On

“Predict the fare amount of future rides using regression analysis”

PG-DBDA SEP 2023

Submitted By:

Group No: 10

Omkar Nalawade	**230941225030**
Vivek Pawar	**230941225049**

Mrs. Priti Take
Project Guide

Mr. Rohit Puranik
Centre Coordinator

Acknowledgement

We would like to express our sincere gratitude to everyone who has contributed to the completion of our project.

First and foremost, we would like to thank our project guide **Mrs. Priti Take** mam for their constant guidance and support throughout the project. We extend our sincere thanks to our respected Centre Co- Ordinator, **Mr. Rohit Puranik,** for allowing us to use the facilities available

We would also like to express our appreciation to the faculty members of our department for their constructive feedback and encouragement. Their insights and suggestions have helped us to refine our ideas and enhance the quality of our work.

Furthermore, we would like to thank our families and friends for their unwavering support and encouragement throughout our academic journey. Their love and support have been a constant source of motivation and inspiration for us.

Thank you all for your valuable contributions to our project.

Omkar Nalawade	**230941225030**
Vivek Pawar	**230941225049**

ABSTRACT

For predicting the longer-term events predictive analysis use data which is archive. For capturing the trends which are important mathematical model is used from past data. The model then uses present data to predict the longer term or to derive actions to require optical outcomes tones of appreciation in recent time for predictive analytics thanks to development in support technology within areas of massive data in machine learning. Many industries use predictive analytics for making accurate forecast like giving the amount of fare for the ride within the city. These resource planning are enabled by the forecast as an example, cab fare can be predicted more accurately. A lot of factors are taken into consideration for a taxi start-up company . This project tries to know the patterns and use different methods for fare prediction. This project is developed for predicting the cab fare amount within a certain city. The project involves different steps like training, testing by using different variables like pickup, drop-off location for predicting cab fare.

TABLE OF CONTENTS

CHAPTER 1
INTRODUCTION

1.1 OVERVIEW

For prediction using systems machine learning is widely used all over the work. There are different types of app machines using machine learning for prediction some of them are supervised, unsupervised learning. In problems associated with business needs machine learning can also be called as predictive. Machine learning derived into different. Supervised learning is one of the most commonly used learning in machine learning , in order to train our data here we need some factors like the dataset and we have to use algorithms in order to predict the output of the program which we are doing. If we do not have a data set to train our model then the output which we get is less reliable and the fare will not be accurate. Unsupervised learning is like learning without the help of any dataset or external factor , it automatically selects a suitable algorithm and tries to predict the output which will be accurate most of the time. Because of computing technologiesthe machine learning which we are looking at today are not at all similar to the machine learning which was used before sometime. This specifically proves that systems can do the tasks which were assure to be done by humans onlyMachine learning can also be used in the domain of Artificial intelligence. It can ambe used in data mining. We can perform many things in affordable price and also can do many complex things easily with the help of the advance machine learning. It is mainly used to make accurate predictions by using different range of methods and algorithms it provides to the people. So it is widely used in many companies because we use the concept of machine learning daily in our day to day lives and also it will be used in future also as the scope of it is very high.

1.2 SCOPE OF THE PROJECT

The main contributions of this project therefore are:

- Data Analysis
- Dataset Preprocessing
- Training the Model
- Testing of Dataset

1.3 PROBLEM STATEMENT

The objective of this project is to predict Uber Fare amount. Need to design a system that predicts the fare amount for a Uber ride

CHAPTER 2

EXPERIMENTAL MATERIALS OR METHODS AND ALGORITHMSUSED

2.1 SOFTWARE REQUIREMENTS

Operating system	:	Windows 10
Languages used	:	Python
Python version	:	3.5 or 3.6
Notebooks	:	Google colab
Emulators	:	No emulators used
Software Libraries	:	Pandas, Matplotlib, Numpy, Seaborn,Math,Sklearn.

- OPERATING SYSTEM:

WINDOWS 10:

Windows 10 is a series of personal computer operating systems produced by Microsoft as part of its Windows NT family of operating systems. It is the successor to Windows 8.1, and was released to manufacturing on July 15, 2015, and broadly released for retail sale on July 29, 2015.Windows 10 receives new builds on an ongoing basis, which are available at no additional cost to users, in addition to additional test builds of Windows 10 which are available to Windows Insiders. Devices in enterprise environments can receive these updates at a slower pace, or use long-termsupport milestones that only receive critical updates, such as security patches, over their ten-year lifespan of extended support.

- JUPYTER NOTEOOK:

Google Colab, short for Google Colaboratory, is a cloud-based platform provided by Google that allows users to write, share, and execute Python code in a browser-based environment.

- PANDAS & NUMPY:

In computer programming, pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numericaltables and time series. It is free software released under the three-clause BSD license.

NumPy is a library for the Python programming language, adding supportfor large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

- MATPLOTLIB & SEABORN:

Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced byJohn Hunter in the year 2002.One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.

Seaborn is a Python data visualization library based on matplotlib. It providesa high- level interface for drawing attractive and informative statistical graphics.

2.2 HARDWARE REQUIREMENTS

There are big data that require big hardware.

Learning about big machine learning requires big data and big hardware.

- Processor – Intel Xeon E2630 v4 – 10 core processor, 2.4 GHz with Turboboost upto 3.1 GHz. 25 MB Cache
- Motherboard – ASRock EPC612D8A
- RAM – 8 GB DDR4 2133 MHz
- TB Hard Disk (7200 RPM) + 512 GB SSD
- GPU – NVidia TitanX Pascal (12 GB VRAM)
- Intel Heatsink to keep temperature under control
- Disk space: 2 GB
- Operating systems: Windows 10, macOS, and Linux
- Python* versions: 3.5.x, 3.6.x

2.3 ALGORITHMS USED

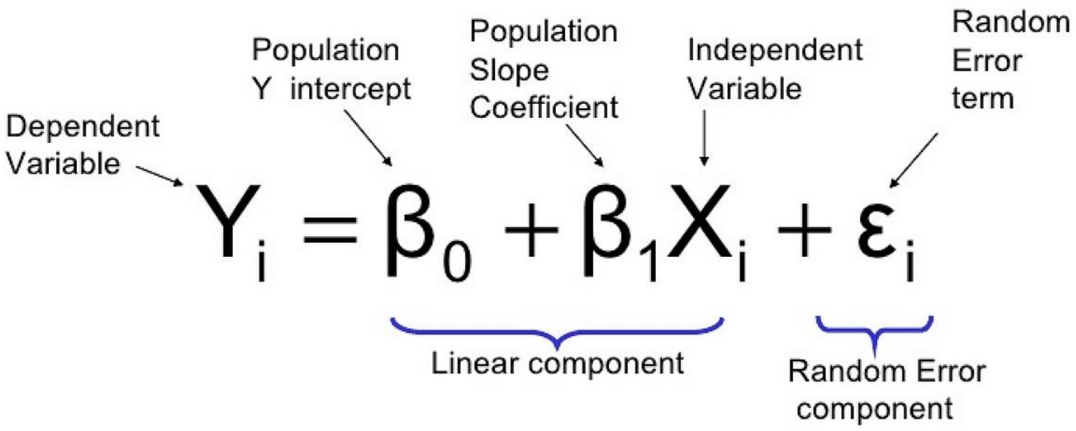
- **LINEAR REGRESSION**

It is a supervised machine learning algorithm . It is mostly used after the step of correlation . If we want to predict the value of an y variable by using the value of another variable then we can use Iinear algorithm. Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

When working with linear regression, our main goal is to find the best fit line that means the error between predicted values and actual values should be minimized.The best fit line will have the least error.

The different values for weights or the coefficient of lines (a0, a1) gives a different line of regression, so we need to calculate the best values for a0 and a1 to find the best fit line, so to calculate this we use cost function.



- Cost function-
    - The different values for weights or coefficient of lines (a0, a1) gives the different line of regression, and the cost function is used to estimate the values of the coefficient for the best fit line.
    - Cost function optimizes the regression coefficients or weights. It measures how a linear regression model is performing.
    - We can use the cost function to find the accuracy of the **mapping function**,which maps the input variable to the output variable. This mapping function is also known as **Hypothesis function**.

For Linear Regression, we use the **Mean Squared Error (MSE)** cost function, which is the average of squared error occurred between the predicted values and actual values.

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2$$

CHAPTER 3

SYSTEM DESIGN AND METHODOLOGY

3.1 METHODOLOGY

According to industry standards, the process of Data Analyzing mainly includes 6 main steps and this process is abbreviated as CRISP DM Process, which is Cross-Industry Process for Data Mining. And the Six main steps of CRISP DM Methodology for developing a model are:

1. Business understanding
2. Data understanding
3. Data preparation/Data Preprocessing
4. Modeling
5. Evaluation
6. Deployment

And in this project, all the above steps are followed to develop the model.

1. Business Understanding

It is important to understand the idea of business behind the data set. The given data set is asking us to predict fare amount. And it really becomes important for us to predict the fare amount accurately. Else, there might be great loss to the revenue of the firm. Thus, we have to concentrate on making the model most efficient.

2. Data Understanding

To get the best results, to get the most effective model it is really important toour data very well. Here, the given train data is a CSV file that consists 7 variables and 16067 Observation. A snapshot of the data provided.

fare_amount	pickup_datetime	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
4.5	2009-06-15 17:26:21 UTC	-73.844311	40.721319	-73.84161	40.712278	1
16.9	2010-01-05 16:52:16 UTC	-74.016048	40.711303	-73.979268	40.782004	1
5.7	2011-08-18 00:35:00 UTC	-73.982738	40.76127	-73.991242	40.750562	2
7.7	2012-04-21 04:30:42 UTC	-73.98713	40.733143	-73.991567	40.758092	1
5.3	2010-03-09 07:51:00 UTC	-73.968095	40.768008	-73.956655	40.783762	1
12.1	2011-01-06 09:50:45 UTC	-74.000964	40.73163	-73.972892	40.758233	1

The different variables of the data are:

- **fare_amount** : fare of the given cab ride.
- **pickup_datetime** : timestamp value explaining the time of ride start.
- **pickup_longitude** : a float value explaining longitude location of theride start.
- **pickup_latitude** : a float value explaining latitude location of the ride start.
- **dropoff_longitude** : a float value explaining longitude location of theride end**.**
- **dropoff_latitude** : a float value explaining latitude location of the ride end.
- **Passenger_cont** : an integer indicating number of passengers

Following table explains how the variables are categorized.

Independent Variables
pickup_datetime
pickup_longitude
pickup_latitude
dropoff_longitude
dropoff_latitude
passenger_count
dependent Variables
Fare_amount

Fig Independent variables Fig Dependent variables

From the given train data it is understood that, we have to predict fare amount, and other variables will help me achieve that, here pickup_latitude/longitude, dropoff_latitude/longitude this data are signifying the location of pick up and dropoff. It is explaining starting point and end point of the ride. So, these variablesare crucial for us. Passenger_count is another variable, that explains about how many people or passenger boarded the ride, between the pickup and drop off locations. And pick up date time gives information about the time the passenger is picked up and ride has started. But unlike pick up and drop off locations has start and end details both in given data. The time data has only start details and no time value or time related information of end of ride. So, during pre- processing of data we will drop this variable. As it seems the information of time is incomplete.

Also, there is a separate test data given, in the format of CSV file containing 9514 observations and 6 variables. All of them are the Independent variables. An in these data at the end we have to predict the fare or the target variable.

Following is a snap of the test data provided.

pickup_datetime	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
2015-01-27 13:08:24 UTC	-73.97332	40.7638054	-73.9814301	40.7438355	1
2015-01-27 13:08:24 UTC	-73.9868622	40.7193832	-73.9988861	40.7392006	1
2011-10-08 11:53:44 UTC	-73.982524	40.75126	-73.979654	40.746139	1

Fig 4.5 Test data

3 Data Preparation

The next step in the CRISP DM Process is, Data preprocessing. It is a data mining process that involves transformation of raw data into a format that helps us execute our model well. As, the data often we get are incomplete, inconsistent and also may contain many errors. Thus, Data preprocessing is a generic method to deal with such issues and get a data format that is easily understood by machine and that helps developing our model in best way. In this project also we have followed data pre-processing methods to rectify errors and issues in our data. And this is done by popular data preprocessing techniques, this are following below.

Note: I have removed the variable “Pick up date Time” as it is a timestamp valueand it shows only the start time of pick up time , whereas there is no drop off time, so in this data set it seems, it will have no impact in the target variable, andalso it lead to redundancy and model accuracy issues, so I preferred to drop it.

- ### Missing Value Analysis

Missing value is availability of incomplete observations in the dataset. This is found because of reasons like, incomplete submission, wrong input, manual error etc.These Missing values affect the accuracy of model. So, it becomes important to check missing values in our given data.

Missing Value Analysis in Given Data:

In the given dataset it is found that there are lot of values which are missing. It isfound in the following types:

1. Blank spaces : Which are converted to NA and NaN in R and Python respectively for further operations
2. Zero Values : This is also converted to NA and Nan in R and python respectively prior

further operations

3. Repeating Values : there are lots of repeating values in pickup_longitude, pickup_latitude, dropoff_longitude and dropoff_latitude. This will hamper our model, so such data is also removed to improve the performance.

Following the standards of percentage of missing values we now have to decide to accept a variable or drop it for further operations. Industry standards ask to follow following standards:

1. Missing value percentage < 30% : Accept the variable
2. Missing value percentage > 30 % : Drop the variable

It is found from the above graph plot that the there is no variable exceeding the30% range so we not need to exclude any of our variable.

i. Outlier Analysis

Outlier is an abnormal observation that stands or deviates away from other observations. These happens because of manual error, poor quality of data and it is correct but exceptional data. But, It can cause an error in predicting the target variables. So we have to check for outliers in our data set and also remove or replace the outliers wherever required.

- Outliers in this project:

In this dataset, I have found some irregular data, those are considered as outliers. These are explained below.

1. Fare_Amount :

I have always seen fare of a cab ride as positive, I have never seen any cab driver, giving me money to take a ride in his cab. But in this dataset, there are many instances where fare amount is negative. Given below are such instances:

fare_amount	pickup_datetime	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
-2.9	2010-03-09 23:37:10 UTC	-73.7895	40.6435	-73.7887	40.64195	1
-2.5	2015-03-22 05:14:27 UTC	-74	40.72063	-73.9998	40.72054	1
-3	2013-08-30 08:57:10 UTC	-73.9951	40.74076	-73.9959	40.74136	4

Fig Fare outliers

Passenger_count:

I have always found a cab with 4 seats to maximum of 8 seats. But in this dataset I have found passenger count more than this, and in some cases a large number of values. This seems irregular data, or a manual error. Thus, these are outliers and needs to be removed. Few instances are following.

Fare_amount	pickup_datetime	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
4.9	2010-07-12 09:44:33 UTC	-73.9832	40.73466	-73.9913	40.73892	456
6.1	2011-01-18 23:48:00 UTC	-74.0066	40.73893	-74.0108	40.71791	5334
8.5	2013-06-18 10:27:05 UTC	-73.9921	40.7642	-73.973	40.7627	535
8.1	2009-08-21 19:35:05 UTC	-73.9609	40.76156	-73.9763	40.74836	354

Fig 4.9 Passenger_count outlier

Location points:

When I checked the data it is found that most of the longitude points are within the 70 degree and most of the latitude points are within the 40 degree. This symbolizes all the data belongs to a specific location and a specific range. But I also found some data which consists location points too far from the average location point’s range of 70 Degree Longitude and 40 Degree latitude. It seems these far point locations are irregular data. And I consider this as outlier. I have collected the maximum and minimum values of location point as a reference to identify the outliers.

Variable	Minimum Value	Maximum Value
pickup_longitude	-74.43823	40.76613
pickup_latitude	-74.00689	401.08333
dropoff_longitude	-74.22705	40.80244
dropoff_latitude	-74.00638	41.36614

Following are the instances, where the values exceeds far from average locationpoints.

And I consider this as Outliers.

Fare_amount	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
15	40.72913	-74.0069	40.76337	-73.9616	1
52	40.73688	-74.0062	40.73689	-74.0064	6
15.5	40.76442	-73.9929	40.80244	-73.9507	1
6.5	40.74826	-73.9918	40.74037	-73.979	1
3.3	-73.9472	401.0833	-73.9514	40.77893	1

Fig Location Outliers

All this outliers mentioned above happened because of manual error, or interchange of data, or may be correct data but exceptional. But all these outliers can hamper our data model. So there is a requirement to eliminate or replace such outliers. And impute with proper methods to get better accuracy of the model. In this project, I used mode method to impute the outliers in passenger count and mean for location Points and fare amount.

Haversine formula

The Haversine formula is a mathematical formula used to calculate the shortest distance between two points on the surface of a sphere, given their latitudes and longitudes. It's commonly used in geospatial applications, such as navigation or determining distances between locations on Earth. The formula is based on spherical trigonometry and is particularly useful when working with small distances over the Earth's surface.

Here's the formula:

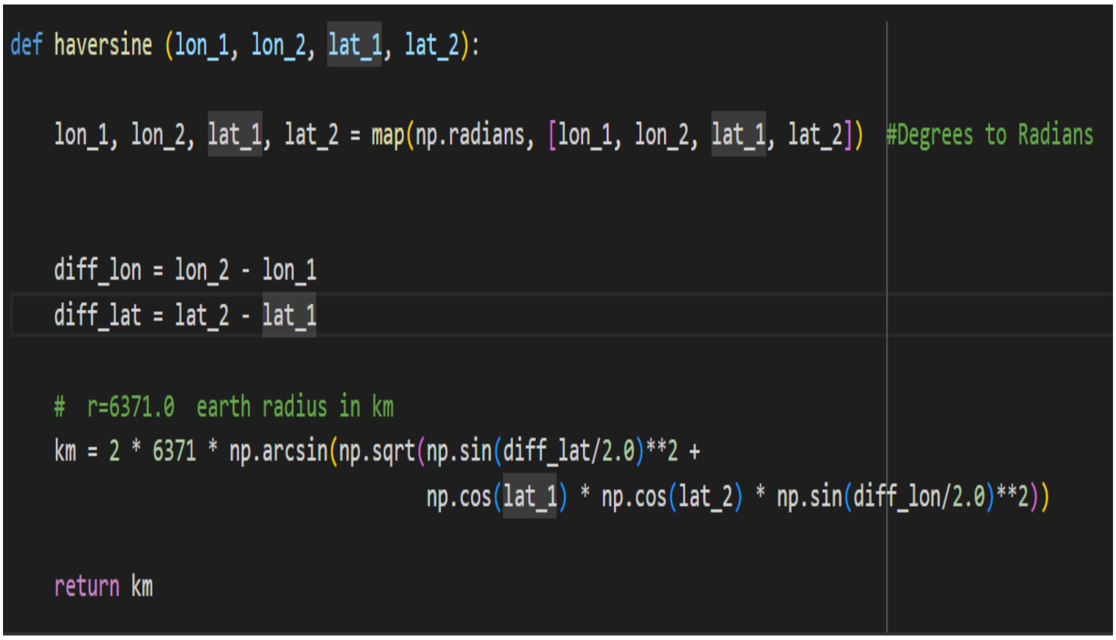
$$a = \sin^2\left(\frac{\Delta\text{lat}}{2}\right) + \cos(\text{lat}_1) \cdot \cos(\text{lat}_2) \cdot \sin^2\left(\frac{\Delta\text{long}}{2}\right)$$

$$c = 2 \cdot \text{atan2}\left(\sqrt{a}, \sqrt{1-a}\right)$$

$$d = R \cdot c$$

Where:

- Δlat is the difference in latitude between the two points (in radians),
- Δlong is the difference in longitude between the two points (in radians),
- lat_1 and lat_2 are the latitudes of the two points (in radians),
- R is the radius of the Earth (mean radius = 6,371 km), and
- d is the distance between the two points along the surface of the sphere (also in kilometers if R is in kilometers).



```
def haversine (lon_1, lon_2, lat_1, lat_2):

    lon_1, lon_2, lat_1, lat_2 = map(np.radians, [lon_1, lon_2, lat_1, lat_2])  #Degrees to Radians


    diff_lon = lon_2 - lon_1
    diff_lat = lat_2 - lat_1

    #  r=6371.0  earth radius in km
    km = 2 * 6371 * np.arcsin(np.sqrt(np.sin(diff_lat/2.0)**2 +
                                      np.cos(lat_1) * np.cos(lat_2) * np.sin(diff_lon/2.0)**2))

    return km
```

Feature Selection

Sometimes it happens that, all the variables in our data may not be accurate enough to predict the target variable, in such cases we need to analyze our data, understand our data and select the dataset variables that can be most useful for our model. In such cases we follow feature selection. Feature selection helps by reducing time for computation of model and also reduces the complexity of the model.

After understanding the data, preprocessing and selecting specific features,there is a process to engineer new variables if required to improve the accuracy of the model.

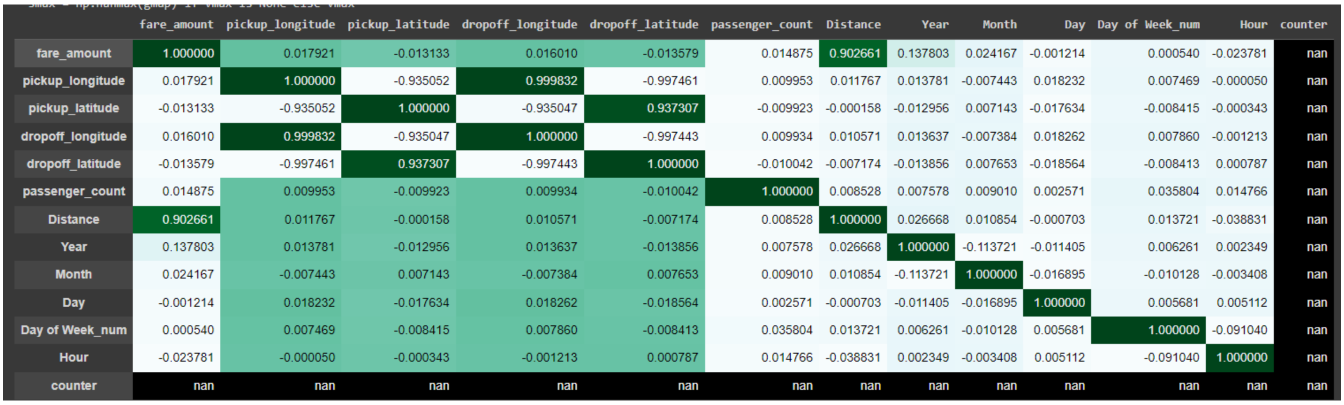
In this project the data contains only the pick up and drop points in longitude andlatitude. The fare_amount will mailnly depend on the distance covered between these two points. Thus, we have to create a new variable prior further processing the data. And in this project the variable I have created is Distance variable (dist), which is a numeric value and explains the distance covered between the pick up and drop of points. After researching I found a formulacalled The haversine formula, that determines the distance between two points on a sphere based on their given longitudes and latitudes. These formula calculates the shortest distance between two points in a sphere.

The function of haversine function is described, which helped me to engineer our new variable, Distance.

Correlation Analysis:

In some cases it is asked that models require independent variables free from collinearity issues. This can be checked by correlation analysis for the categorical variables and continuous variables. Correlation analysis is a process that is defined to identify the level of relation between two variables.

In this project, our Predictor variable is continuous, so we will plot a correlation table that will predict the correlation strength between independent variables andthe ‘fare_amount’ variable



	fare_amount	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count	Distance	Year	Month	Day	Day of Week_num	Hour	counter
fare_amount	1.000000	0.017921	-0.013133	0.016010	-0.013579	0.014875	0.902661	0.137803	0.024167	-0.001214	0.000540	-0.023781	nan
pickup_longitude	0.017921	1.000000	-0.935052	0.999832	-0.997461	0.009953	0.011767	0.013781	-0.007443	0.018232	0.007469	-0.000050	nan
pickup_latitude	-0.013133	-0.935052	1.000000	-0.935047	0.937307	-0.009923	-0.000158	-0.012956	0.007143	-0.017634	-0.008415	-0.000343	nan
dropoff_longitude	0.016010	0.999832	-0.935047	1.000000	-0.997443	0.009934	0.010571	0.013637	-0.007384	0.018262	0.007860	-0.001213	nan
dropoff_latitude	-0.013579	-0.997461	0.937307	-0.997443	1.000000	-0.010042	-0.007174	-0.013856	0.007653	-0.018564	-0.008413	0.000787	nan
passenger_count	0.014875	0.009953	-0.009923	0.009934	-0.010042	1.000000	0.008528	0.007578	0.009010	0.002571	0.035804	0.014766	nan
Distance	0.902661	0.011767	-0.000158	0.010571	-0.007174	0.008528	1.000000	0.026668	0.010854	-0.000703	0.013721	-0.038831	nan
Year	0.137803	0.013781	-0.012956	0.013637	-0.013856	0.007578	0.026668	1.000000	-0.113721	-0.011405	0.006261	0.002349	nan
Month	0.024167	-0.007443	0.007143	-0.007384	0.007653	0.009010	0.010854	-0.113721	1.000000	-0.016895	-0.010128	-0.003408	nan
Day	-0.001214	0.018232	-0.017634	0.018262	-0.018564	0.002571	-0.000703	-0.011405	-0.016895	1.000000	0.005681	0.005112	nan
Day of Week_num	0.000540	0.007469	-0.008415	0.007860	-0.008413	0.035804	0.013721	0.006261	-0.010128	0.005681	1.000000	-0.091040	nan
Hour	-0.023781	-0.000050	-0.000343	-0.001213	0.000787	0.014766	-0.038831	0.002349	-0.003408	0.005112	-0.091040	1.000000	nan
counter	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan

4.13 Correlation matrix

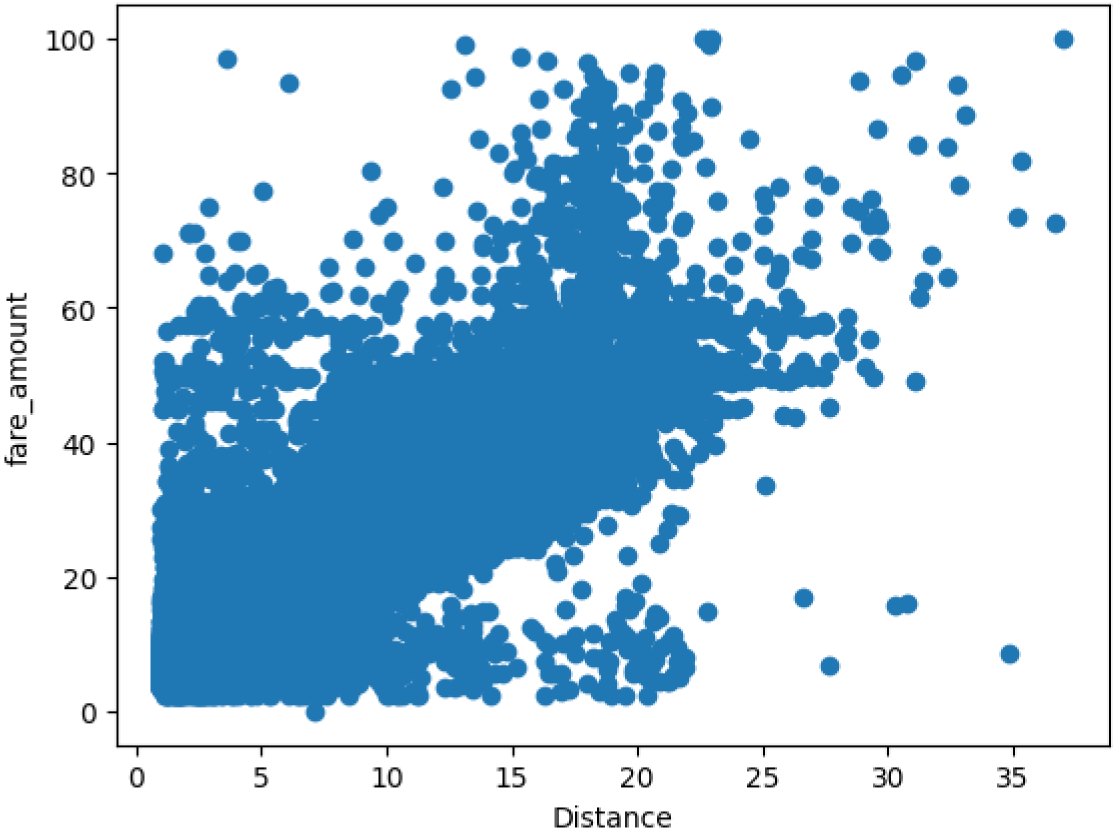
- From the above plot it is found that most of the variables are highly correlated with each other, like fare amount is highly correlated with distance variable.
- All the red charts represents that variables are highly correlated. And as there is some not red and blue charts, which represents negative correlation, it can be summarized that our dataset has strong or highly positive correlation between the variables.
- Because all the variables are numeric the important features are extracted using the correlation matrix. All the variables are important for predicting thefare_amount since none of the variables have a high correlation factor (considering the threshold as 0.9), so all the variables for model building arekept.

MODEL DEVELOPMENT

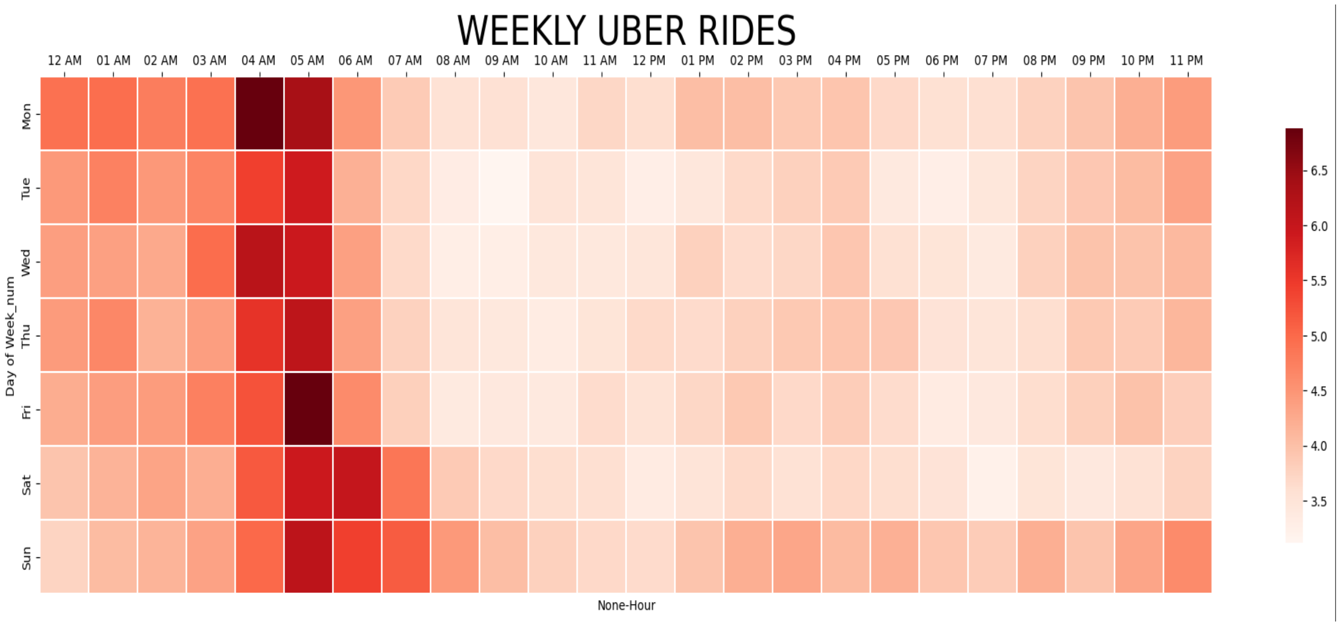
After all the above processes the next step is developing the model based onour prepared data.

In this project we got our target variable as “fare_amount”. The model has to predict a numeric value. Thus, it is identified that this is a Regression problem statement. And to develop a regression model, the various models that can beused are Random Forest and Linear Regression.

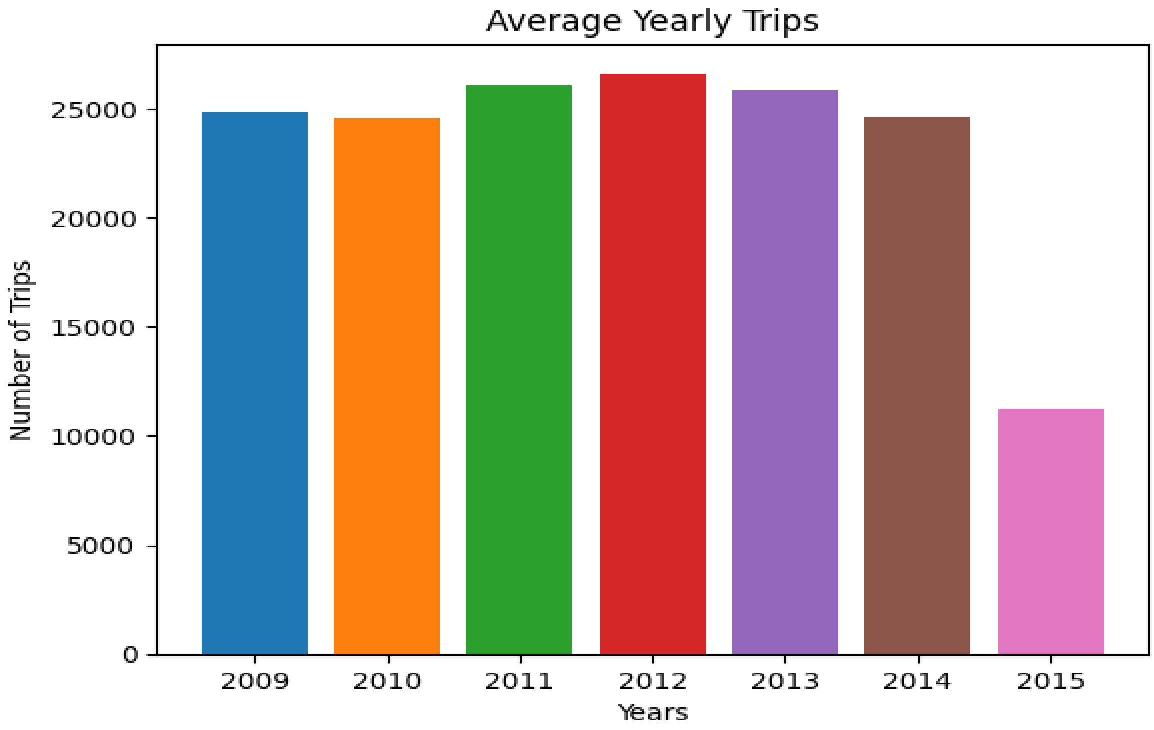
Scatter plot



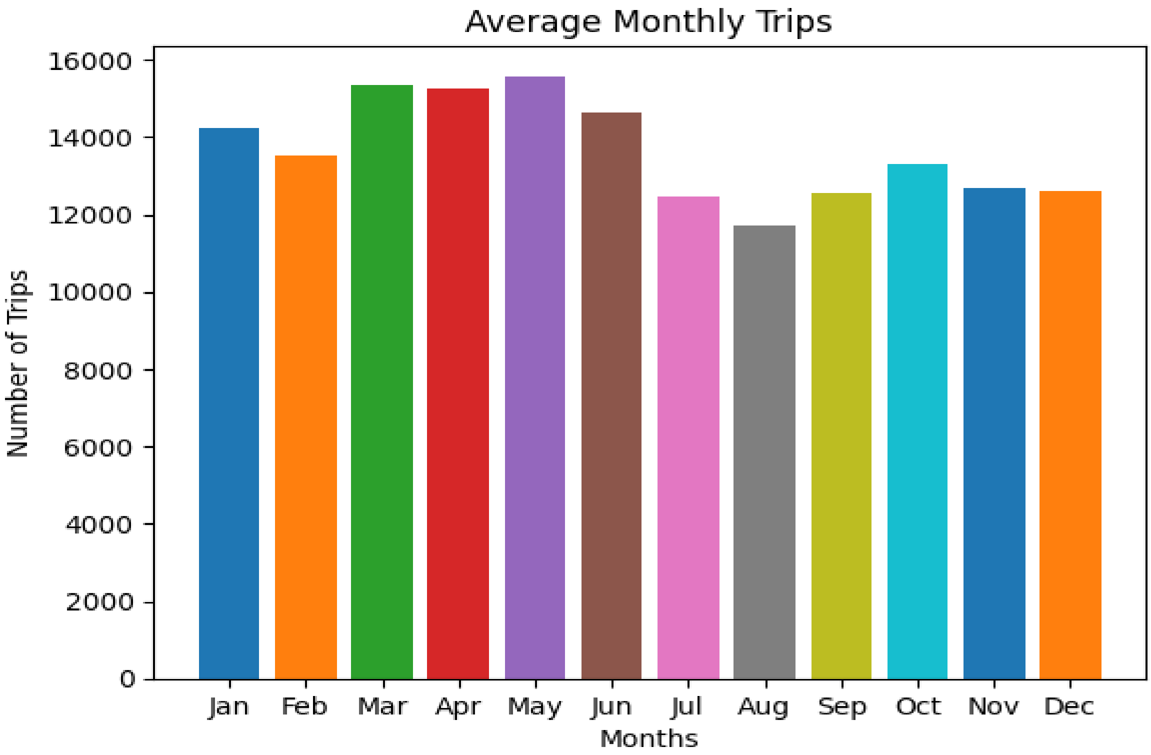
Heat map (weekly uber rides)



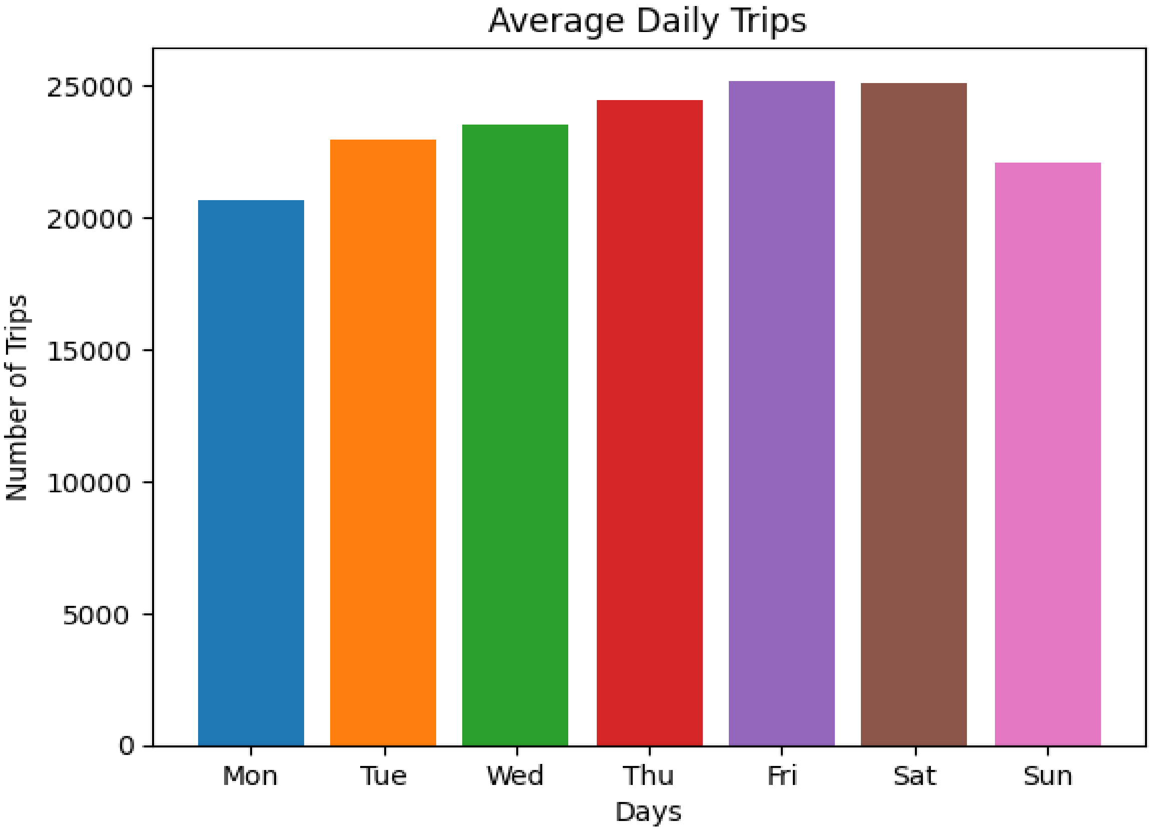
AVERAGE YEARLY TRIPS



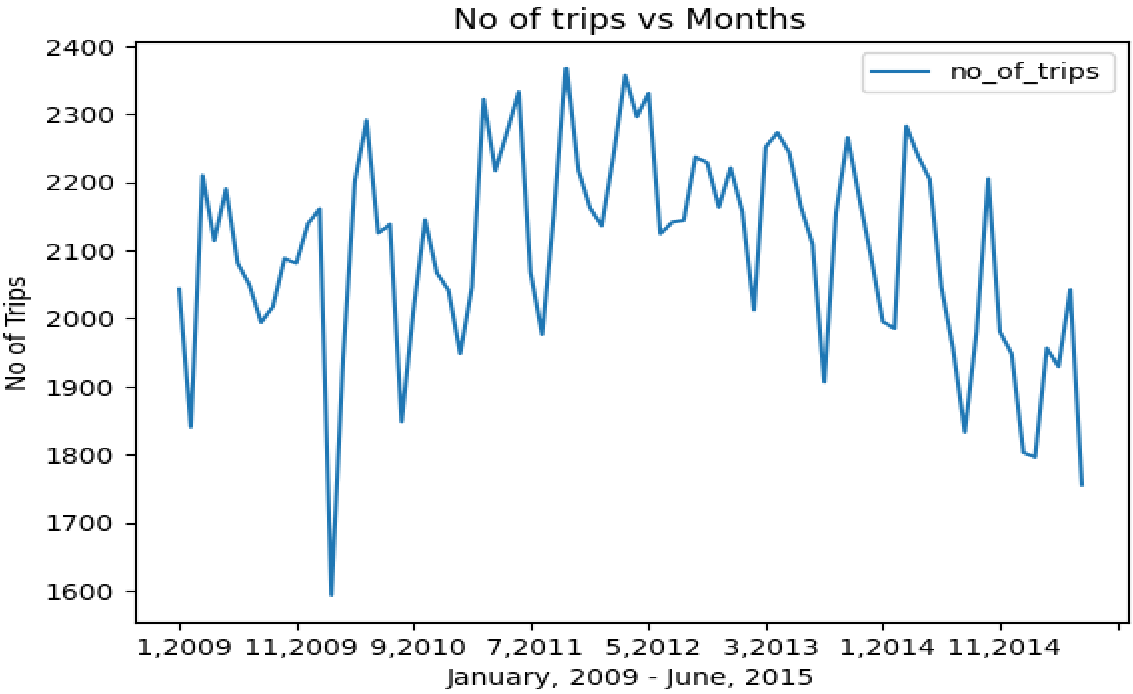
AVERAGE MONTHLY TRIPS



AVERAGE DAILY TRIP



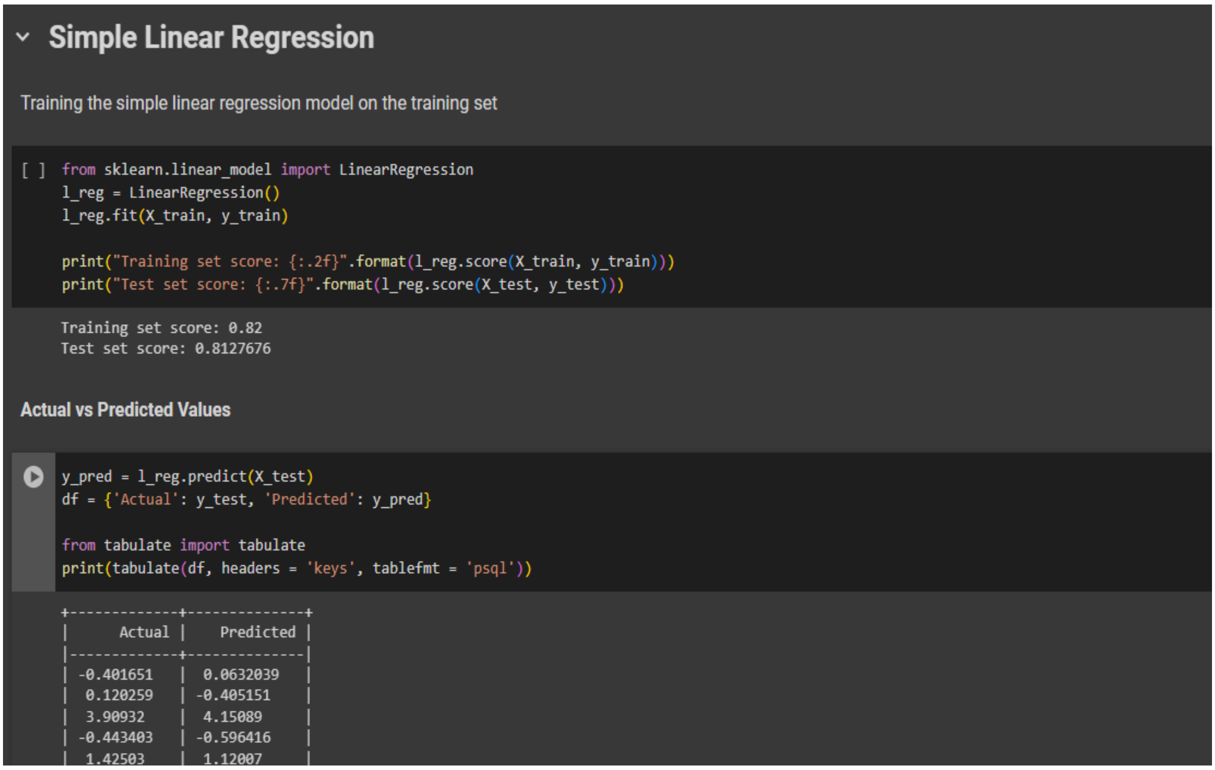
NO OF TRIPS VS MONTH



Linear Regression

It is a supervised machine learning algorithm . It is mostly used after the step of correlation . If we want to predict the value of an y variable by using the value of another variable then we can use Iinear algorithm. Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.



Simple Linear Regression

Training the simple linear regression model on the training set

```
from sklearn.linear_model import LinearRegression
l_reg = LinearRegression()
l_reg.fit(X_train, y_train)

print("Training set score: {:.2f}".format(l_reg.score(X_train, y_train)))
print("Test set score: {:.7f}".format(l_reg.score(X_test, y_test)))
```

```
Training set score: 0.82
Test set score: 0.8127676
```

Actual vs Predicted Values

```
y_pred = l_reg.predict(X_test)
df = {'Actual': y_test, 'Predicted': y_pred}

from tabulate import tabulate
print(tabulate(df, headers = 'keys', tablefmt = 'psql'))
```

```
+------------+-------------+
|     Actual |   Predicted |
|------------+-------------|
| -0.401651  |  0.0632039  |
|  0.120259  | -0.405151   |
|  3.90932   |  4.15089    |
| -0.443403  | -0.596416   |
|  1.42503   |  1.12007    |
```

CHAPTER 4

RESULT AND DISCUSSION

So, now we have developed few models for predicting the target variable, now the next step is to identify which one to choose for deployment. To decide these according to industry standards, we follow several criteria. Few among this are, calculating the error rate, and the accuracy. MAE and MAPE is used in our project.MAE or Mean Absolute Error, it is one of the error measures that is used to calculate the predictive performance of the model. In this project we will apply this measure to our models

Method	Mae Error(in Percentage)
Linear Regression	24.95

Fig Mean absolute error

The second matrix to identify or compare for better model is Accuracy. It is theratio of number of correct predictions to the total number of predictions made. Accuracy= number of correct predictions / Total predictions made.
It can also be calculated from MAE as Accuracy = 1- MAPE.

Method	Accuracy (in Percentage)
Linear Regression	82.00

Fig Accuracy

CHAPTER 6
CONCLUSION AND FUTURE WORK

CONCLUSION

Prediction of cab fare is one of the essential usage in automobile industry.The purpose study is to predict the cab fare using random forest and linear regression algorithms. Thus, the predicting the cab fare we have achieved our objective of our project. The accuracy for linear regression is 82.00%

The quality of a regression model depends on the matchup of predictions against actual values. In regression problems, the dependent variable is continuous. In classification problems, the dependent variable is categorical. Random Forest can be used to solve both regression and classification problems. Decision trees are nonlinear; unlike linear regression, there is no equation to express the relationship between independent and dependent variables. Out of the three models left, Random Forest is the best model as it has the lowest RMSE score and highest R- Squared score, which explains the highest variability and tells us how well the model fits in this data.

FUTURE WORK

As is known, with an increase in the number of features; underlying equations become a higher-order polynomial equation, and it leads to overfitting of the data. Generally, it is seen that an overfitted model performs worse on the testing data set, and it is also observed that the overfitted model performs worse on additional new test data set as well. A kind of normalized regression type -Ridge Regression may be further considered.

REFERENCES

[1] Gunjan panda,Supriya p.panda.”Machine learning using exploratory analysis to predict cab fare”. International Journal for Research in Applied Science & Engineering Technology (IJRASET) Aug 2019.

[2] Kelareva, Elena. "Predicting the Future with Google Maps APIs." Web blog post. Geo Developers Blog,https://maps-apis.googleblog.com/2015/11/predicting- future-with-google-maps-apis.html Accessed 15 Dec. 2016.

[3] Wu, Chun-Hsin, Jan-Ming Ho, and Der-Tsai Lee. "Travel-time prediction with support vector regression." IEEE transactions on intelligent transportation systems 5.4 (2004): 276-281.

[4] Van Lint, J. W. C., S. P. Hoogendoorn, and Henk J. van Zuylen. "Accurate freeway travel time prediction with state-space neural networks under missing data." Transportation Research Part C: Emerging Technologies 13.5 (2005): 347- 369.

[5] Vanajakshi, L., S. C. Subramanian, and R. Sivanandan. "Travel time predictionunder heterogeneous traffic conditions using global positioning system data from buses." IET intelligent transport systems 3.1 (2009): 1-9.

[6] Weijie Wang and Yanmin Lu, Analysis of the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) in Assessing Rounding Model, ICMEMSCE, IOP Publishing, 324 (2018), doi:10.1088/1757-899X/324/1/012049

[7] X. Qian, S. V. Ukkusuri.: Time-of-Day Pricing in Taxi Markets. IEEE Transactions on Intelligent Transportation Systems, Vol. 18 June 2017.

[8] Yildirimoglu, Mehmet, and Nikolas Geroliminis. "Experienced travel time prediction for congested freeways."Transportation Research Part B:Methodological 53 (2013): 45-63.