**A**

**Minor Project Report on**

**“UBER & LYFT DATA ANALYSIS TO PREDICT CAB FARE”**

In partial fulfillment of requirements for the degree of

**Bachelor of Technology (B. Tech.)**

in

**Computer Science and Engineering**

**Submitted by**

Ms. Tapisha Purohit (170351)

**Under the Guidance of**

Ms. Aditi Kajala

*CSE*

**SCHOOL OF ENGINEERING AND TECHNOLOGY**

**Mody University and Science and Technology Lakshmangarh**

**Distt. Sikar-332311**

December 2020

**ACKNOWLEDGEMENT**

I sincerely express my gratitude to my guide for her benevolent guidance in completing the report on “**UBER & LYFT DATA ANALYSIS”**. Her kindness and help have been the source of encouragement for me.

I am grateful to her for the guidance, inspiration and constructive suggestions that helpful us in the preparation of this project.

Tapisha Purohit (170351)

**CERTIFICATE**

This is to certify that the minor project report entitled **“UBER & LYFT DATA ANALYSIS TO PREDICT CAB FARE”** submitted by Ms. Tapisha Purohit, as a partial fulfillment for the requirement of B. Tech. VII Semester examination of the School of Engineering and Technology, Mody University of Science and Technology, Lakshmangarh for the academic session 2019-2020 is an original project work carried out under the supervision and guidance of Ms. Aditi Kajala has undergone the requisite duration as prescribed by the institution for the project work. CSE-f2-002

**PROJECT GUIDE: HEAD OF DEPARTMENT**

**Approval Code:** AUT\_20\_CSE\_F23\_07 **Signature:**

**Name:** Mrs. Aditi Kajala  **Name**: Dr. A Senthil

**Date: 1**8/10/2020 **Date:**

**EXAMINER-I: EXAMINER-II**

**Name:** Dr. Anand Sharma  **Name:** Mr. Rajesh Yadav

**Dept.:** CSE  **Dept.:** CSE

**ABSTRACT**

*This project is about a cab company who want to predict the fare for future transactional. As, nowadays there are no. of cab companies like Uber, Ola, lyft, etc. and these deliver services to lakhs of customers daily. It’s becomes really important to manage data properly. So I decided to work on Uber & Lyft data analysis to predict fare price and to estimate it.*

*The key value propositions of cab is offering scheduling flexibility to their driver-partners. According to a report, 83% of drivers prefers having a job that allow them to choose their own schedule and they can use this to maximize their expected revenue. It also provides heat maps of the customer demand that allows drivers to target the high-demand regions.*

*In this project, I’m going to work with the CAB & WEATHER dataset where cab consists of the data related to distance, cab type, time-stamp, destination, source, price, product id and name. Where weather data set consist of data related to temperate, locations, clouds, pressure, rain, humidity and wind.*

**TABLE OF CONTENT**

|  |  |  |
| --- | --- | --- |
| **SR. NO.** | **TOPIC** | **PAGE NO.** |
| **1.** | **Introduction** | 1-3 |
| 1.1 | Present System | 2 |
| 1.2 | Proposed System | 2-3 |
| **2.** | **System Design** | 4-6 |
| 2.1 | System Flowchart | 4 |
| 2.2 | Data tables (database if taken)/ Decision Tree/ Data Dictionary. | 5-6 |
| **3.** | **Hardware & Software details** | 7 |
| **4.** | **Implementation work details** | 8-9 |
| 4.1 | Real life application’s | 8 |
| 4.2 | Data implementation and program execution | 9 |
| **5.** | **Source code** | 10-12 |
| **6.** | **Input/output Screens/ Model’s Photograph** | 13-15 |
| **7.** | **System Testing** | 16 |
| **8.** | **Conclusion** | 17 |
| 8.1 | Limitations | 17 |
| 8.2 | Scope for future | 17 |
| **9.** | **Bibliography** | 18 |
| **10.** | **Annexures** | 19 |

**TABLE OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **SR NO.** | **TOPIC** | **PAGE NO.** |
| 2.1 | System Design | 5 |
| 2.2 | Flowchart of cab fare prediction | 5 |
| 2.3 | Data table of Weather dataset | 6 |
| 2.4 | Data table of cab\_rides | 6 |
| 4.1 | Graph of linear regression | 8 |

**Chapter 1: Introduction**

**1. INTRODUCTION**

Uber & Lyft is an American based company that offers vehicles for hire, package delivery, food delivery, transportation, The Company is based in San Francisco and has many worldwide operations in metropolitan areas. It’s one of the largest providers in economy and is also in the development of self-driving cars. In US Uber has a 30% market share for food delivery & 70% market share for ride-sharing.

Now a day’s cab rental services are expanding with multiplier rate. The ease of using these services & flexibility gives customer a great experience with ruthless price. Basically the service is generally accessed via mobile app. The users have to set up a personal profile with a Name, phone number, other info & Payment preference (i.e. credit card, cash, or some other way). When the service is complete, the customer may be given option to provide a boon to driver, which is also billed to customer's payment method.

The objective of this project is to predict the CAB FARE for the future transactional cases. Nowadays cab delivers services to lakhs of customers daily and it becomes really important to manage their data properly and estimate the fare price accurately to earn most revenues.

In this project, we are going to work with the CAB\_RIDES & WEATHER dataset where cab\_rides consists of the data related to distance, cab type, time-stamp, destination, source, price, product id and name. Where weather data set consist of data related to temperate, locations, clouds, pressure, rain, humidity and wind.

We are going to prepare this dataset to answer these questions:

* How many data we have for Uber & lyft?
* Data visualization for cab type i.e. Uber & Lyft.
* Show a bar plot for fluctuation of the price?
* Which hour is peak hour for the cab?
* How price & hour varies with each other?
* What is the connection between rain and price?
* What is the actual & predicted value?
* Show the graph between actual and predicted?

We are going to use jupyter notebook for the execution of this project and it will be coded in python programming language. The python libraries which we are going to use is Numpy for working with arrays, Pandas for Data cleaning, Seaborn & Plotly module, which is great visualization tool in python will be used in order to plot some insightful and intuitive graphs to answer the above given questions.

**1.1 PRESENT SYSTEM**

We have same project with a different datasets which exist in the present and the objective is to predict the cab fare amount. Suppose you are starting a new startup of a cab rental company and you have successfully collected all the historical data which you used in your project and successfully run the project. Now you have requirement to apply analytics for fare prediction and need to design a system that predicts the amount for a cab ride in the city. So the given below attributes which used to shows the result:

Pickup\_datetime- It indicates the timestamp from where the cab ride started.

Pickup\_longitude- It indicates the longitude coordinate of where the cab ride started.

pickup\_latitude - It indicates the latitude coordinate of where the cab ride started.

dropoff\_longitude - It indicates the longitude coordinate of where the cab ride ended.

dropoff\_latitude - It indicates the latitude coordinate of where the cab ride ended.

Passenger\_count – It indicates the No. of passengers in the cab ride.

By using above data we can see that:-

* Number of passengers & fare - By bar graph we can see that single passengers are most frequent traveler, & the highest fare also seems to come from cabs which carry just 1 passenger.
* Date of months & fares - Fare throughout the month mostly seem uniform.
* Hours & fares- 6pm to 11pm frequency of cab boarding is very due to peak hours and fare price during 2pm to 8 pm is bit high as we compared to all other time.
* Week day & Fare – Basically the fare is high on Friday, Saturday or may be during weekends.

**1.2 PROPOSED SYSTEM**

By using the same concept I came with a new data set and it consist of two different datasets i.e. CAB\_RIDES & WEATHER. Here in this project I’m not going to use the few attributes.

CAB\_RIDES data set consists of 8 attributes i.e.:

distance- It shows the total distance covered.

cab-type- It shows that the cab is Uber or Lyft.

time\_stamp - It shows the time.

destination – It shows the direction in which cab is going.

source – It tell us the source.

price- It tell the price in which area the passenger wants to go.

Surge\_id – It is 1 throughout all.

id – It tell the id of particular cab.

product\_id- It tell the particular cab id.

name – It tell us the name of particular cab.

WEATHER data set consists of 10 attributes:

temp - It tell the temperature of particular area.

location – It tell location.

clouds – It tell that how much clouds are there.

pressure – It tell us the pressure.

rain – It tell about the rain.

time\_stamp – It tell about time of particular area.

humidity – It tell the humidity.

wind – It tell the wind flowing in particular area.

By using above dataset we can get these answers:

* Data for Uber & Lyft is
* There is a fluctuation in price first it will goes up after that it will slowly goes down.

**Chapter 2: System Design**

**2. SYSTEM DEISGN**

In designing my project there are total 6 steps involved and they are as follows.

**1 DEFINE THE GOAL:**

Basically here we have to design a system that predict the fare of cab which give the passenger estimates.

**2 GET YOUE DATA SET:**

It’s time to start looking for your data, the second phase of a data analytics project. In my project I have two datasets one is Weather and the other is cab\_ride data set.

In weather dataset there are total of 10 attributes whereas in cab\_rides data set there is 8 attributes.

**3 Explore and Clean Your Data:**

The next data science step is the dreaded data preparation process that typically takes up to 80% of the time dedicated to a data project.

* Once you’ve gotten your data, it’s time to get to work on it in the third data analytics project phase.
* Start digging to see what you’ve got and how you can link everything together to achieve your original goal.
* The next step (and by far the most dreaded one) is cleaning your data.
* There is even missing data. It’s time to look at every one of your columns to make sure your data is homogeneous and clean.

**4 Build Helpful Visualizations:**

Now I have a nice dataset (or maybe several), so this is a good time to start exploring it by building graphs.

While dealing with large volumes of data, visualization is the best way to explore and communicate your findings and is the next phase of your data analytics project.

**5 MODEL TRAINING:**

For Training my model I am going to use Decision tree Algorithm because decision tree works great when there is overlapping of a data where linear regression fails to work in that case.

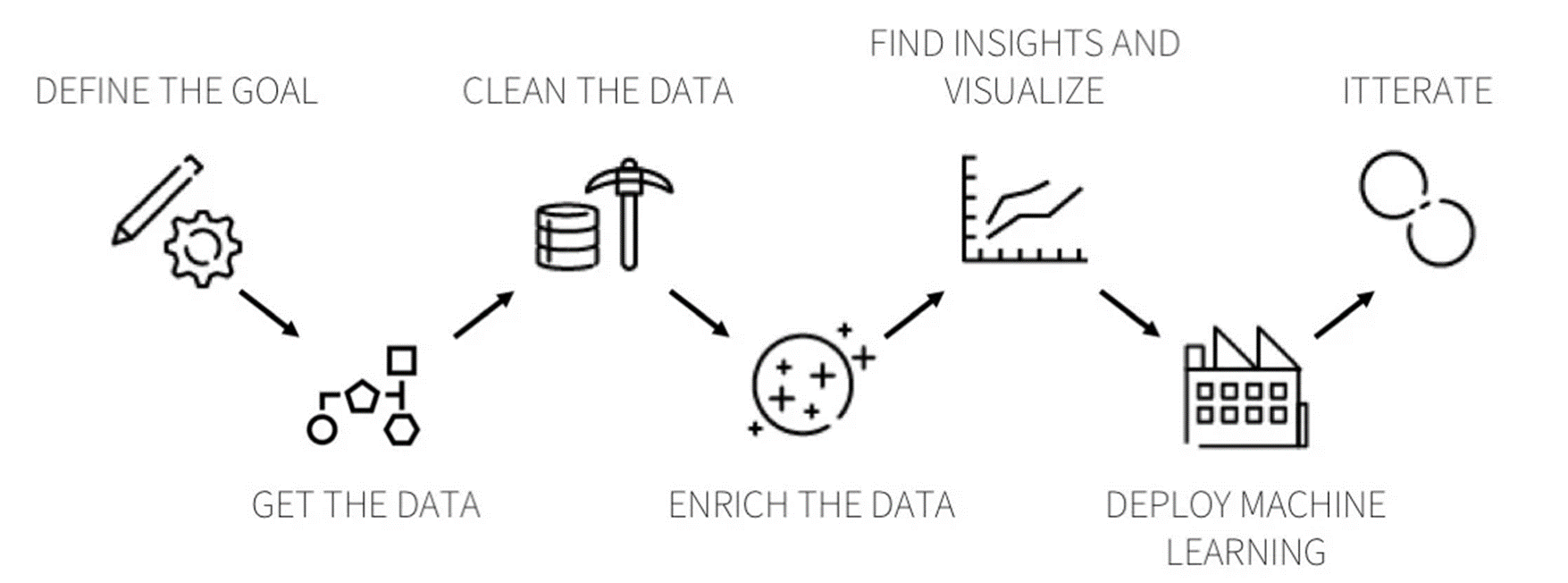
****

Fig 2.1 – System design

**2.1 SYSTEM FLOWCHART**

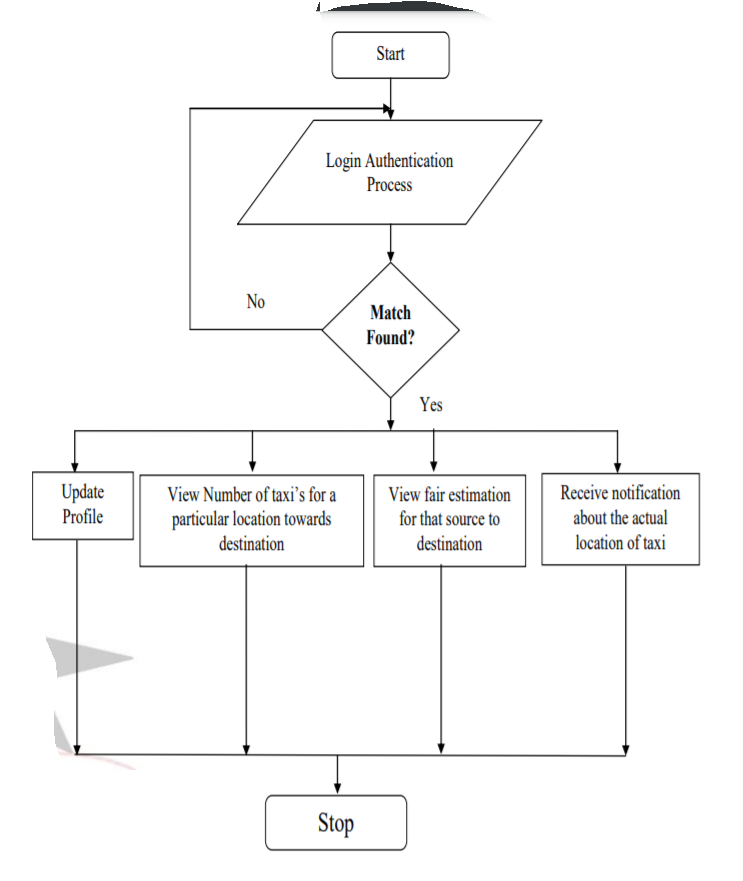
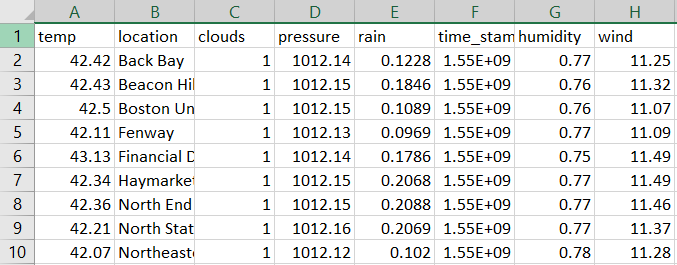


Fig 2.2- Flowchart of Cab fare prediction

**2.2 DATA TABLES**

In my project there are two dataset 1 is Weather and another is cab\_rides. So weather consists of (6277, 8) & cab\_rides consist of (693072, 10).



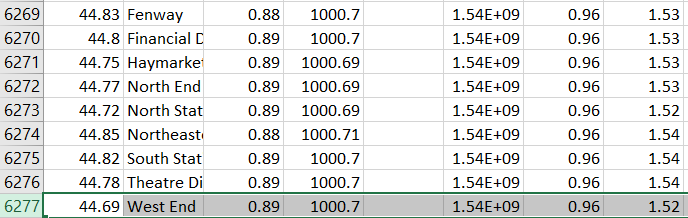
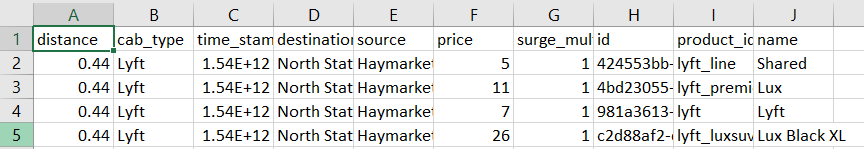


Fig 2.3- Data table of weather dataset



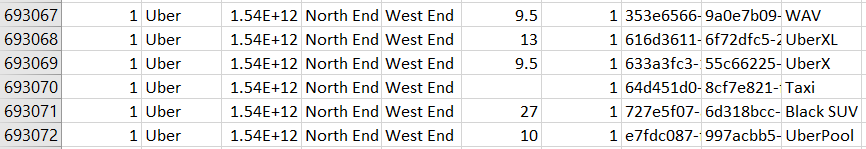


Fig 2.4 – Data table of cab\_rides

**Chapter 3: Hardware & Software Details**

**HARDWARE DETAILS –**

Device Name – Asus VivoBook 14

Model Number - X412FA-EK512T

Processor – Intel core i5 10th Generation

RAM – 8GB

Operating system- Windows 10 HOME

OS Architecture- 64 bit operating system

**SOFTWARE DETAILS-**

Here I’m going to use jupyter notebook for the execution of this project and it will be coded in python programming language. The python libraries which are used below:

* Numpy- For the computation and processing of the multidimensional and single dimensional array elements.
* Pandas - For data cleaning, manipulations and analysis.
* Seaborn & Plotly- for data visualization.
* Matplotlib – It’s a low level graph plotting library that serves as a visualization utility.
* Itertool – It’s a module that provides a various functions that work on iterators to produce complex iterators.
* gc – It provides the automatic garbage collector underlying memory management mechanism.
* os – It provides functions for interacting with operating system.
* Sys - It provides variables & functions which are used to manipulate diff parts of python runtime environment.
* %matplotlib inline – It set’s the backend of Matplotlib to the inline backend.

**Chapter 4: Implementation Work Details**

The implementation details of the Uber & Lyft cab fare prediction is explained in above topics.

**4.1 REALS LIFE APPLICATION**

While coming to Uber & Lyft cab fare prediction the real life applications of this software would be:

* While going to anywhere we can easily estimate the price.
* If there is any bad weather in the area in which passenger wants to go than they can charge according that and they can find another way.
* If there is any heavy traffic than go for alternate route.

**4.2 DATA IMPLEMENTATION & PROGRAM EXECUTION**

In my project I have used Linear Regression Algorithms for training.

**4.2.1 Linear Regression:**

Linear regression is one of the most known and easily understood algorithms in statistics and machine learning. It always attempts to model the relationship between any two variables by fitting a linear equation i.e., a straight line to the observing data. One variable is named to be an explanatory variable example -it can be your income, and the other is named to be a dependent variable example-can be your expenses. From a machine learning point of view, it is considered to be the simplest model that one can try out on your data. If you have a look that the data always follow a straight-line trend, linear regression is responsible to give you quick and exceptionally accurate results.

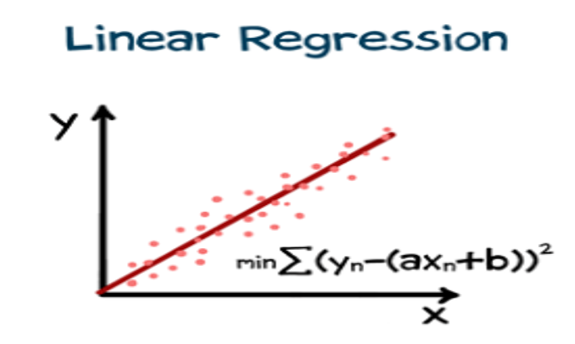


Fig 4.1 – Graph of linear regression

Here’s how we can perform linear regression

1. We have to plot or dependent variable Y-axis against the independent variable i.e., X-axis.
2. We have to plot a straight line and then measure correlation.
3. We have to keep changing the direction of the straight line until we can get the best correlation possible.
4. We extend from the given line to find possibly new values on y-axis.

**Chapter 5: Stimulation Code**

from sklearn.linear\_model import LogisticRegression,LinearRegression

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

import itertools

import gc

import os

import sys

%matplotlib inline

data=pd.read\_csv("Weather.csv")

data=pd.read\_csv("cab\_rides.csv")

cab\_data = pd.read\_csv("cab\_rides.csv")

weather\_data = pd.read\_csv("Weather.csv")

weather\_data

import datetime

cab\_data['datetime']= pd.to\_datetime(cab\_data['time\_stamp'])

cab\_data

weather\_data['date\_time'] = pd.to\_datetime(weather\_data['time\_stamp'])

cab\_data.columns

weather\_data.columns

cab\_data.shape

weather\_data.shape

cab\_data.describe()

weather\_data.describe()

a=pd.concat([cab\_data,weather\_data])

a['day']=a.date\_time.dt.day

a['hour']=a.date\_time.dt.hour

a.fillna(0,inplace=True)

a.columns

a.groupby('cab\_type').count()

a.groupby('cab\_type').count().plot.bar()

a['price'].value\_counts().plot(kind='bar',figsize=(100,50),color='blue')

a['hour'].value\_counts().plot(kind='bar',figsize=(10,5),color='blue')

import matplotlib.pyplot as plt

x=a['hour']

y=a['price']

plt.plot(x,y)

plt.show()

x=a['rain']

y=a['price']

plt.plot(x,y)

plt.show()

a.columns

x1=a[['distance', 'temp','clouds', 'pressure', 'humidity','wind','rain','day','hour','surge\_multiplier','clouds']]

y1=a['price']

# Using Skicit-learn to split data into training and testing sets

from sklearn.model\_selection import train\_test\_split

# Split the data into training and testing sets

x\_train, y\_train, x\_test, y\_test = train\_test\_split(x1, y1, test\_size = 0.25, random\_state = 42)

linear=LinearRegression()

linear.fit(x\_train,x\_test)

predictions=linear.predict(y\_train)

df = pd.DataFrame({'Actual': y\_test, 'Predicted': predictions})

df

df1 = df.head(25)

df1.plot(kind='bar',figsize=(26,10))

plt.grid(which='major', linestyle='-', linewidth='0.5', color='green')

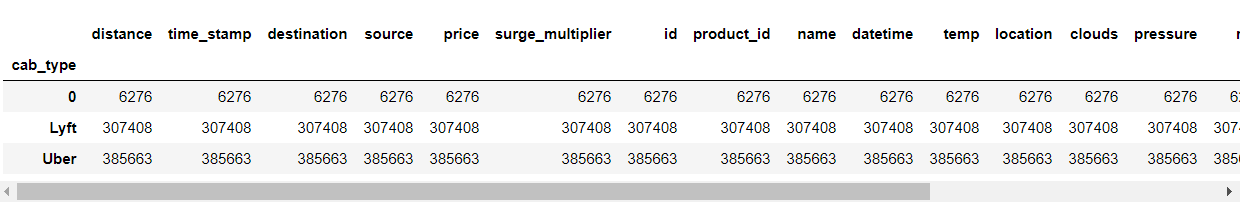
plt.grid(which='minor', linestyle=':', linewidth='0.5', color='black')

plt.show()

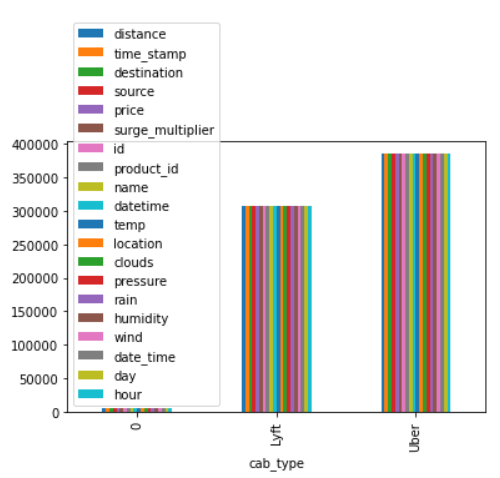
**Chapter 6: Input/output Screens/ Model’s Photograph**

**Final Result-**

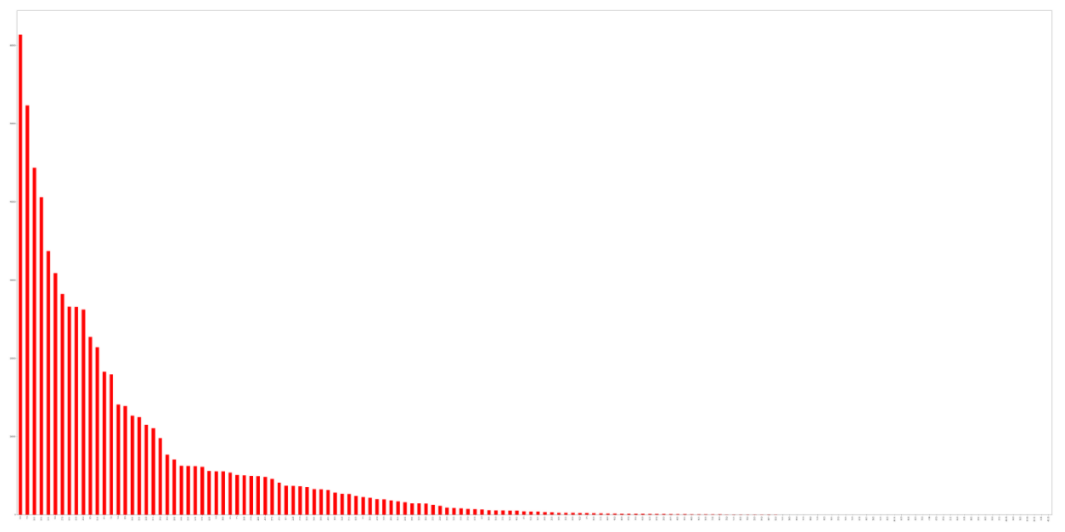
**Data for Uber & Lyft:**



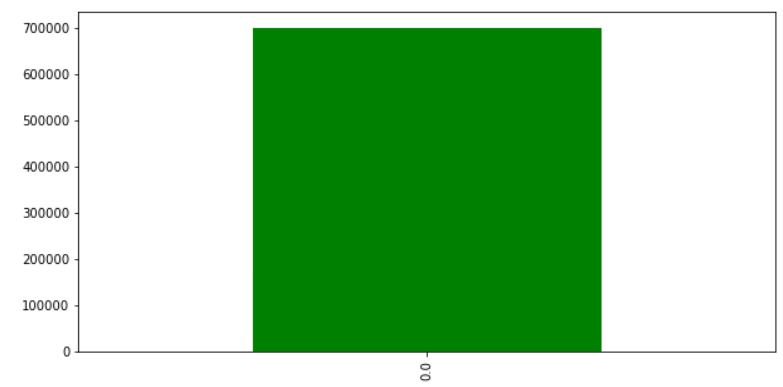
**Data visualization for Uber & Lyft:**



**Bar plot for fluctuation of the price?**



**Peak Hour for the cab**



## Price & hour variation

## 

## Connection between Rain & Price

## 

## Actual & Predicted values

## 

## Bar graph between actual and predicted

## 

**Chapter 7: System Testing**

I have trained model with an accuracy of a nearly 50% and then we have issued the data set for which we will attract the model performance and the accuracy which gives the model which gives the results for the prediction. We can test the model for giving all these a data input items and we can make the data items go through the model and check the accuracy whether the model give predicting the right amount and check with an mathematical matrix operations called as confusion Matrix where we predict the model performance by knowing the outcome of the model before itself so that we can recognize the results from our actual output and the prediction output from the model which will define the model performance.

**Chapter 8: Conclusion**

The project of Uber & Lyft data analysis is finally completed and for this, the developer should know about the basics of Python language. Data visualization makes it easier to understand the core values of the datasets and we can say Data science is really a very interesting field and this is one of the projects which prove it. So by

* The data for Uber & Lyft is
* There is a fluctuation in price first it will goes up after that it will slowly goes down.

To ease the dynamic price surge, we need to manage the ‘Demand & Supply’ of cabs through the high demand situations. Based on these results, we can expect that the demand will be high.

* As a Customer: We can plan our trips in advance so that we can avoid paying extra money.
* As an Uber& Lyft Driver: We can maximize profits by choosing to go on trips.
* As an Uber & Lyft Company: We can incentivize the cab Driver’s to spread across different locations and not to be clustered in a specific location of their interest, in order to manage the ‘supply and demand’ more efficiently, more profitably, and in a customer-friendly way, during these situations.

**8.1 LIMITATIONS**

* Basically machine learning needs enough time to let the algorithms learn and develop enough to fulfill their purpose with a considerable amount of accuracy and relevancy. It also needs massive resources to function. This can mean additional requirements of computer power for you**.**
* If there is a chance of rainfall but there is no rainfall so in that case it gives wrong information.
* Machine Learning is autonomous but highly susceptible to errors.

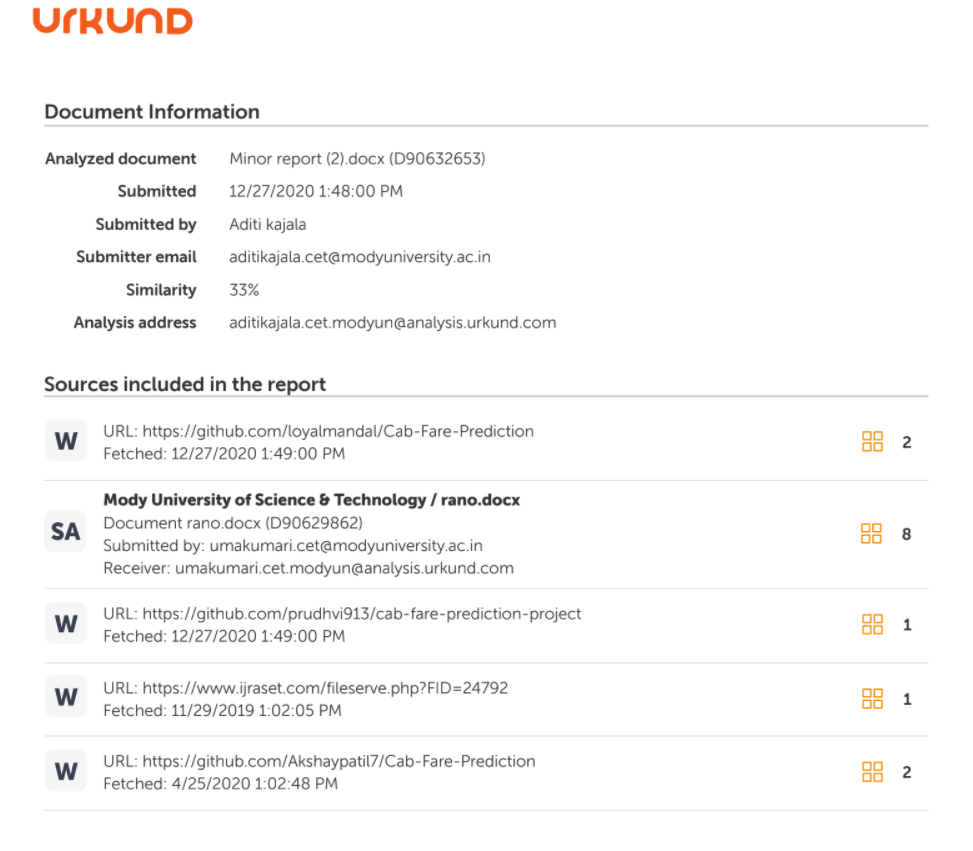
**8.2 SCOPE FOR FUTURE WORK**

In future, I plan to work on extraction of additional features and test other parameters. I believe that working on more datasets and attributes will produce good results if more features and data are extracted.

**Chapter 9: Bibliography**

1. X. Qian, S. V. Ukkusuri.: Time-of-Day Pricing in Taxi Markets. IEEE Transactions on Intelligent Transportation Systems, Vol. 18 June 2017.
2. Y. Bengio. Learning deep architectures for AI. Foundations and trends in Machine Learning, 2(1):1–127, 2009.
3. Sikora R. and Al-laymoun O.H., “A Modified Stacking Ensemble Machine Learning Algorithm Using Genetic Algorithms”, Journal of International Technology and Information Management. vol.23, No.1, 2014, pp.1-12.
4. L. Deng and D. Yu. Deep learning: Methods and applications. Foundations and Trends in Signal Processing, 7(3–4):197–387, 2014.
5. . J. Schmidhuber. Deep learning in neural networks: An overview. Neural Networks, 61:85– 117, 2015.
6. A. De Brébisson, É. Simon, A. Auvolat, P. Vincent, and Y. Bengio. (2015). “Artificial neural networks applied to taxi destination prediction.”

**Chapter 10: Annexures**



6/7

7/ 7