A Mini Project Report on

Uber Data Analysis

Submitted in partial fulfilment of the requirements

for the award of the degree of

Bachelor of Engineering

in

Computer Science & Engineering (AI & ML)

Department

by

Smit Gawand
Gaurav Yadav
Aayush Vishwakarma
Vedanshu Nikam

Under the Guidance of

Prof. Ravindra Bhat

Vishwaniketan Institute of Management Entrepreneurship and Engineering Technology

Computer Science & Engineering (AI & ML) Department

UNIVERSITY OF MUMBAI

Academic Year 2022-2023

Computer Science & Engineering (AI & ML)

Department

CERTIFICATE

This is to certify that the project entitled **"Uber Data Analysis"** submitted by **"Smit Gawand, Gaurav Yadav, Aayush Vishwakarma, Vedanshu Nikam"** for the partial fulfillment of the requirement for award of a degree **Bachelor of Engineering** in **CSE(AIML).**,to the University of Mumbai, is a bonafide work carried out during academic year 2021-2022

Prof Ravindra Bhat Guide CSE (AIML) Department	Prof. S S Patil Head CSE (AIML) Department Vimeet Khalapur	Dr. B R Patil Principal Vimeet Khalapur

External Examiner

Place: ViMEET , Khalapur

Date

Acknowledgement

We have great pleasure in presenting the report on Mini Project **Uber Data Analysis .** We take this opportunity to express our sincere thanks towards my guide Prof. Ravindra Bhat **,**Department of CSE(AIML), ViMEET ,Khalapur for providing the technical guidelines and suggestions regarding line of work. We would like to express my gratitude towards his/her constant encouragement, support and guidance through the development of project.

We thank **Prof. Dr. S S Patil** Head CSE(AIML), ViMEET for his encouragement during progress meeting and providing guidelines to write this report.

We thank **Prof. Ravindra Bhat** MINI project coordinator, CSE(AIML), ViMEET for being encouraging throughout the course and for guidance.

We also thank the entire staff of ViMEET for their invaluable help rendered during the course of this work. We wish to express our deep gratitude towards all my colleagues of ViMEET fortheir encouragement.

Smit Gawand Gaurav Yadav Aayush Vishwakarma Vedanshu Nikam

Index

CHAPTER 1

This chapter contains the introduction of the project its history, application, advantages.

1.1 INTRODUCTION

Uber is a ride-hailing service that operates in numerous cities around the world. Uber's massive dataset is a treasure trove of information that can be analyzed to derive valuable insights and make data-driven decisions. Machine learning can play a crucial role in analyzing Uber's data and improving its services.

In this analysis, we will be focusing on the use of machine learning algorithms to analyze Uber's data and identify patterns in the data that can help improve its services. The data includes information on Uber's trips, drivers, passengers, and ratings. The goal is to use machine learning algorithms to gain insights into the data that can help Uber optimize its operations, increase revenue, and improve customer satisfaction.

1.1.1 History

Uber has been using data analysis and machine learning techniques since its inception in 2009. As the company grew rapidly, the need for data-driven decision-making became more apparent, and Uber started investing heavily in data science and machine learning.

One of the earliest use cases of machine learning at Uber was surge pricing. Surge pricing is a dynamic pricing system that adjusts prices based on supply and demand. To predict when and where demand would increase, Uber built machine learning models that analyzed historical data on trip requests, driver availability, and traffic patterns. These models helped Uber determine the optimal surge pricing rates and increase revenue.

1.1.2 Applications

1.Prediting Demand

2.Driver Allocation

3.Pricing Optimization

4.Route Planning

1.1.3 Advantages

- Improved customer experience
- Increased efficiency
- Fraud detection
- Predictive analytics

1.1.4 Outline

In this project further we will talk about , block diagram and so on.

Next chapter 2 includes the literature review of 8 papers, aim and , problem statement.

Chapter 3 includes the information about hardware and software

Chapter 4 includes Outcome, result.

Chapter 5 includes conclusion and future scope of the project further chapter 6 includes the reference is that we took.

CHAPTER 2

2. Literature Review

As we are researching on Uber and found what different researchers had done. So, they do research on the Uber dataset but on different factors. The rise of Uber as the global alternative has attracted a lot of interest recently. Our work on Uber's predicting pricing strategy is still relatively new. In this research, "Uber Data Analysis" we aim to shed light on Uber's Price. We are predicting the price of different types of Uber based on different factors. Some of the other factors that we found in other researches are:

Abel Brodeurand & Kerry Nield (2018) analyses the effect of rain on Uber rides in New York City after entering Uber rides in the market in May 2011, passengers and fare will decrease in all other rides such as taxi-ride. Also, dynamic pricing makes Uber drivers compete for rides when demand suddenly increases, i.e., during rainy hours. On increasing rain, the Uber rides are also increasing by 22% while the number of taxi rides per hour increases by only 5%. Taxis do not respond differently to increased demand in rainy hours than non-rainy hours since the entrance of Uber.

Surge Pricing is an algorithmic technique that Uber uses when there is a demand-supply imbalance. It occurs when there is a downward shift in both the rider's demand and driver's availability. During such a time of the rise in demand for rides, fares tend to usually high. Surge pricing is essential in a way that it helps in matching the driver's efforts with the demand from consumers. (Junfeng Jiao, 2018) did an investigation of Uber on surge multiplier in Austin, Texas founds that during times of high usage, Uber will enhance their prices to reflect this demand via a surge multiplier.

According to communications released by (Uber, 2015), this pricing is meant to attract more drivers into service at certain times, while also reducing demand on the part of riders. (Chen & Sheldon, 2016) While some research is mixed, in general, surge pricing does appear to control both supply and demand while keeping wait time consistently under 5 minutes.

Anna Baj-Rogowska (2017) analyses the user's feedback from social networking sites such as Facebook in the period between July 2016 and July 2017. Uber is one of the most dynamically growing companies representing the so-called sharing economy. It is also a basis for the ongoing evaluation of brand perception by the community and can be helpful in developing such a marketing strategy and activities, which will effectively improve the current rating and reduce possible losses. So, it can be concluded that feedback should be an important instrument to improve the market performance of Uber today.

Anderson (2014) concluded from surveying San Francisco drivers that driver behavior and characteristics are likely determining the overall vehicle miles traveled (VMT). Full-time drivers are likely to increase overall VMT, while occasional drivers are more likely to reduce overall VMT. We also analyze the research on the driving behavior of the driver while driving on the road. The driver has been categorized based on ages and genders that focus on their driving reactions from how they braking, speeding, and steer handling. For gender differences, male driver practice higher-risk of driving while female drivers are lacks of pre-caution over obstacles and dangerous spot. More or less, adult drivers which regularly drive vehicles can manage the vehicle quite well as compared with young drivers with less experience. In conclusion, the driver's driving behavior is related to their age, gender, and driving experiences.

Some papers take a comparison between the iconic yellow taxi and its modern competitor, Uber. (Vsevolod Salnikov, Renaud Lambiotte, Anastasios Noulas, and Cecilia Mascolo, 2014) identify situations when UberX, the cheapest version of the Uber taxi service, tends to be more expensive than yellow taxis for the same journey. Our observations show that it might be financially advantageous on average for travelers to choose either Yellow Cabs or Uber depending on the duration of their journey. However, the specific journey they are willing to take matters.

2.1 Issues & Challenges

a. **Overfitting in Regression Problem:-** Overfitting a model is a condition where a statistical model begins to describe the random error in the data rather than the relationships between variables. This problem occurs when the model is too complex. In regression analysis, overfitting can produce misleading R-squared values.

b. **Strip-plot and Scatter diagram:-** One problem with strip plots is how to display multiple points with the same value. If it uses the jitter option, a small amount of random noise is added to the vertical coordinate and if it goes with the stack option it increments the repeated values to the vertical coordinate which gives the strip plot a histogram-like appearance.
Scatter plot does not show the relationship for more than two variables.

c. **Label Encoding:-** It assigns a unique number(starting from 0) to each class of data which may lead to the generation of priority issues in the training of data sets. A label with high value may be considered to have high priority than a label having lower value but actually, there is no such priority relation between the attributes of the same classes.

d. **Computational Time:-** Algorithms like support vector machine(SVM) don't scale well for larger datasets especially when the number of features are more than the number of samples. Also, it sometimes runs endlessly and never completes execution.

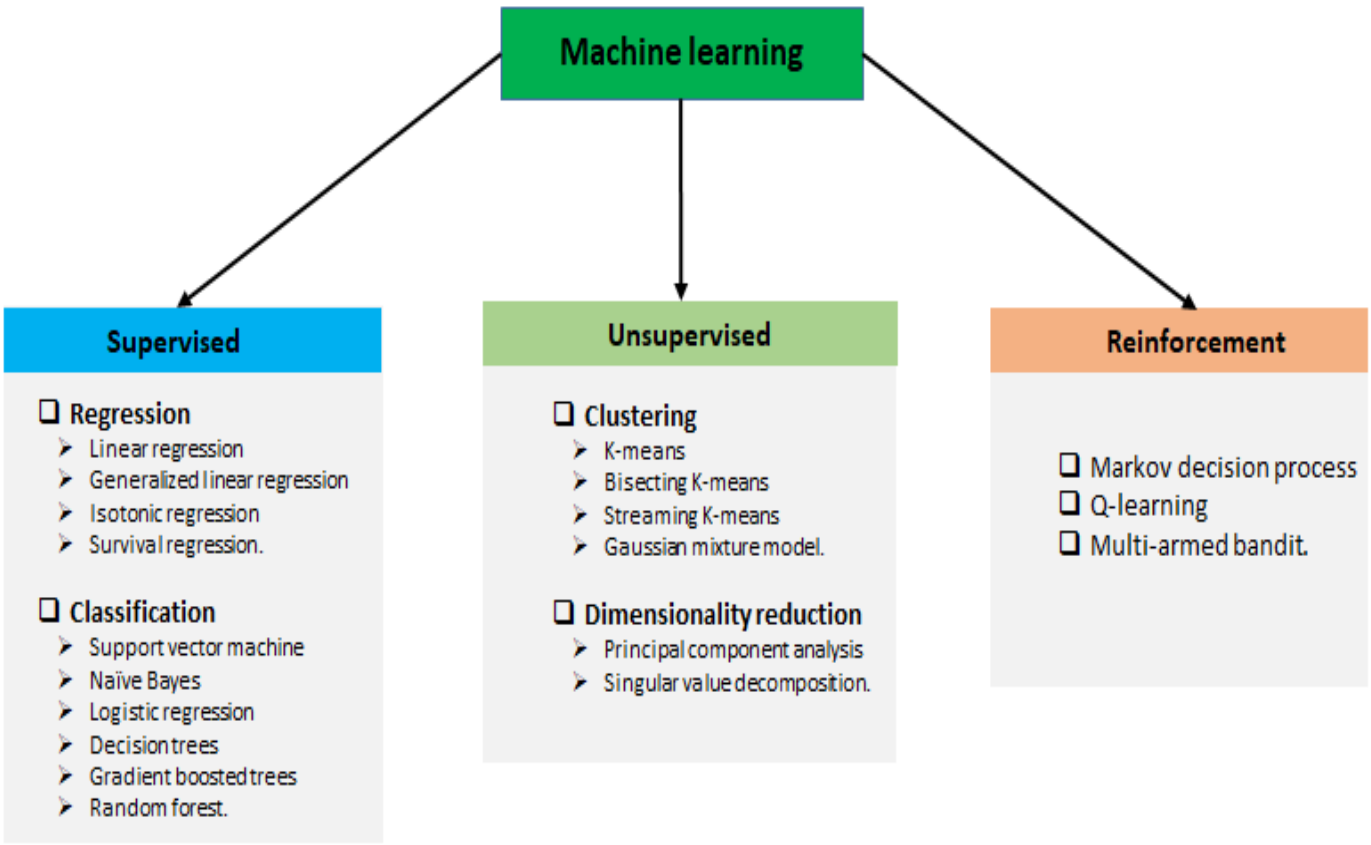
2.2 AIM

Complete Data Analysis and Exploration of Uber Dataset.

2.3 Objective

The objective is to first explore hidden or previously unknown information by applying exploratory data analytics on the dataset and to know the effect of each field on price with every other field of the dataset. Then we apply different machine learning models to complete the analysis. After this, the results of applied machine learning models were compared and analyzed on the basis of accuracy, and then the best performing model was suggested for further predictions of the label 'Price'.

2.4 FLOW CHART



Chapter 3

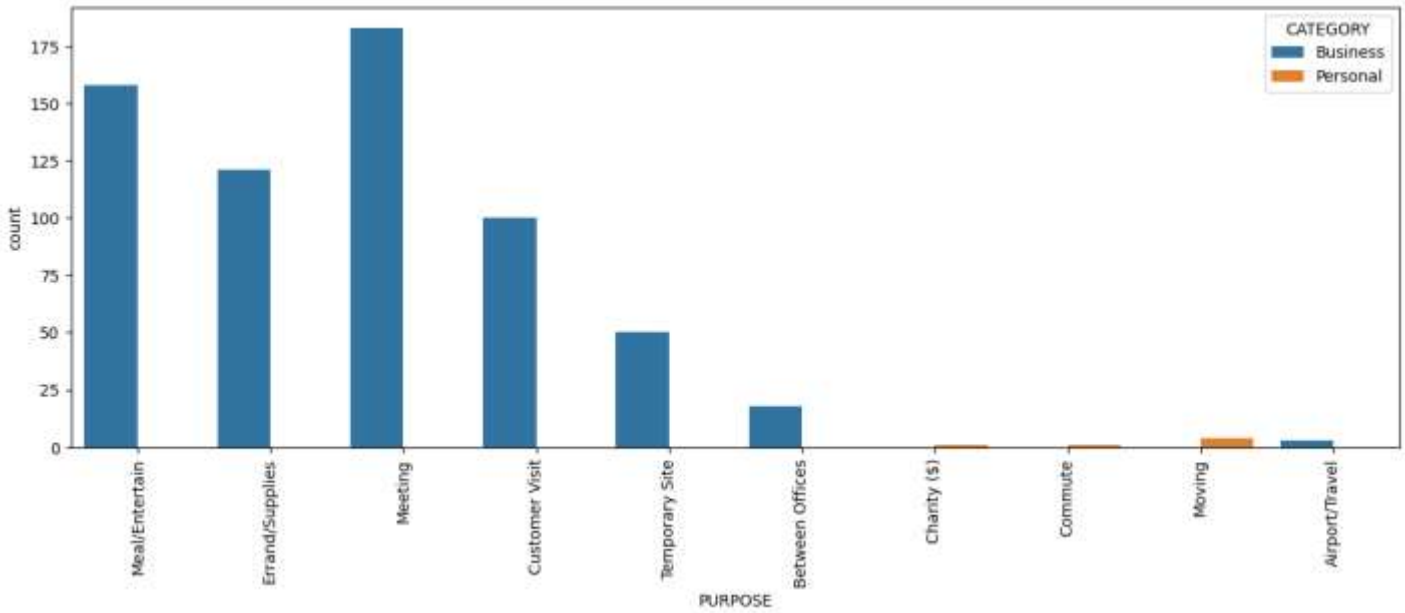
ALGORITHM

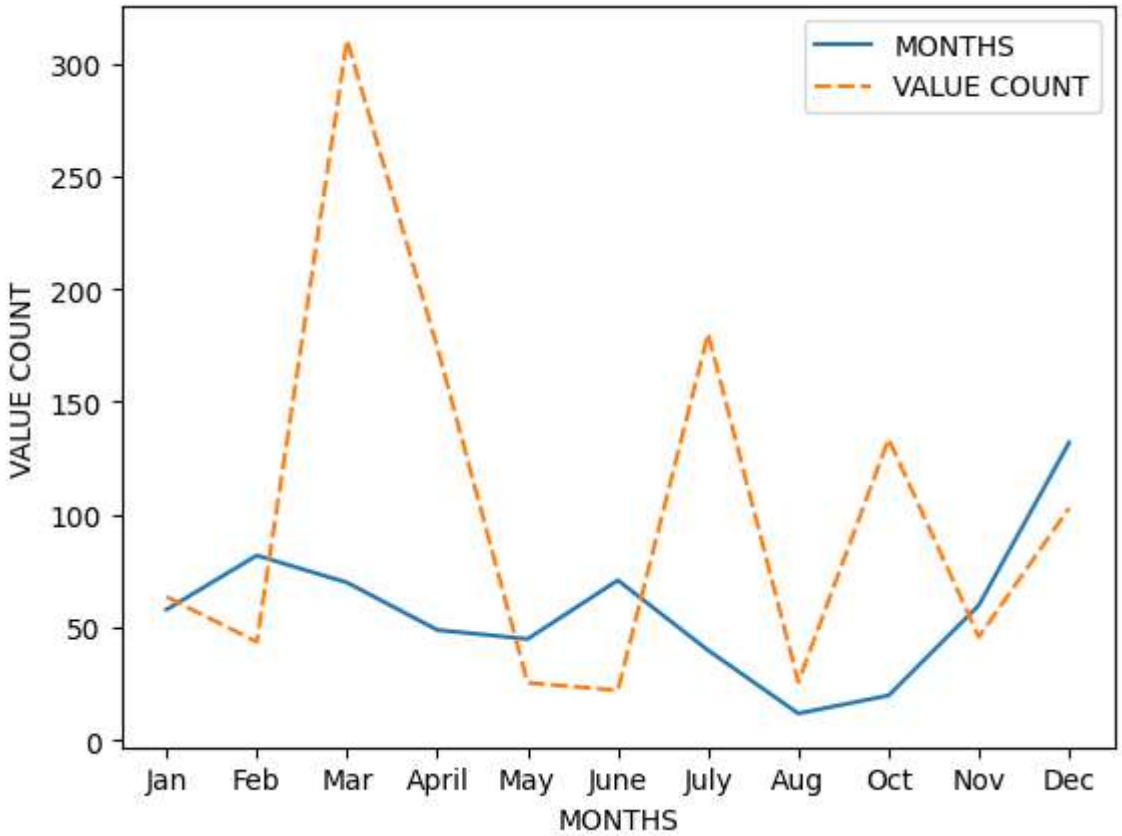
- Installing The Libraries
- Importing the dataset
- Filling the Missing Values
- Data Anaysis
- Implementing Machine Learning Models
- Predicting unseen Data
- Concluding the Report

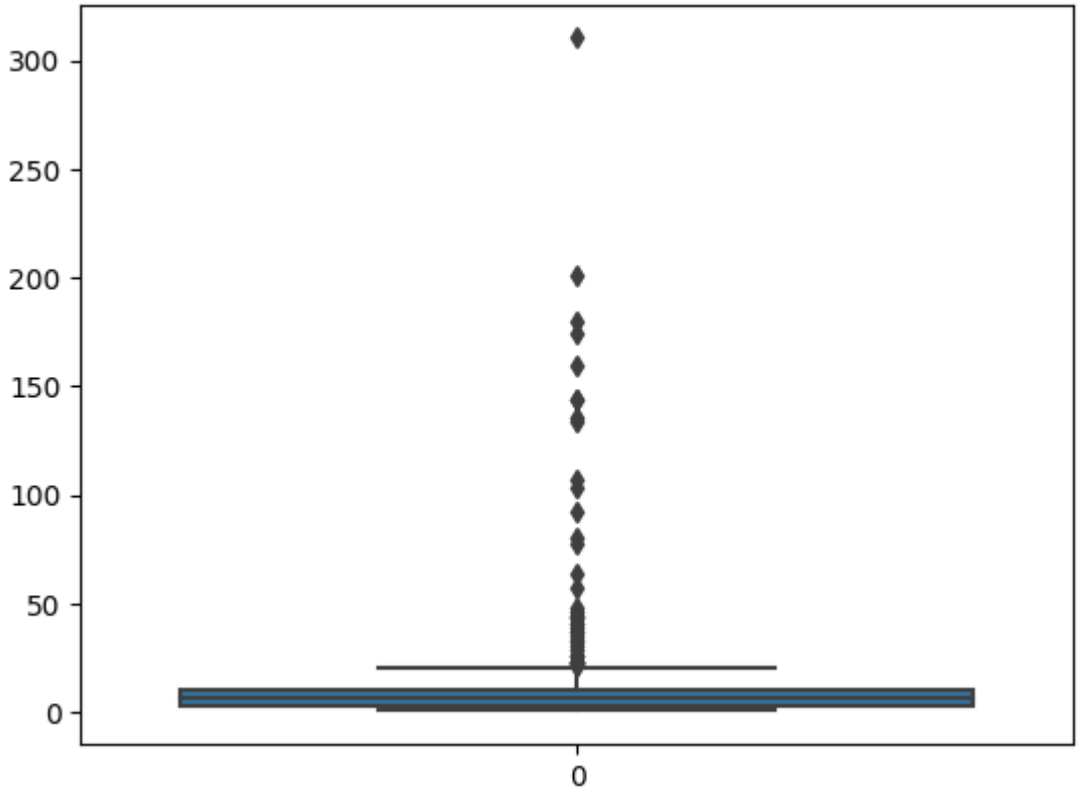
Chapter 4

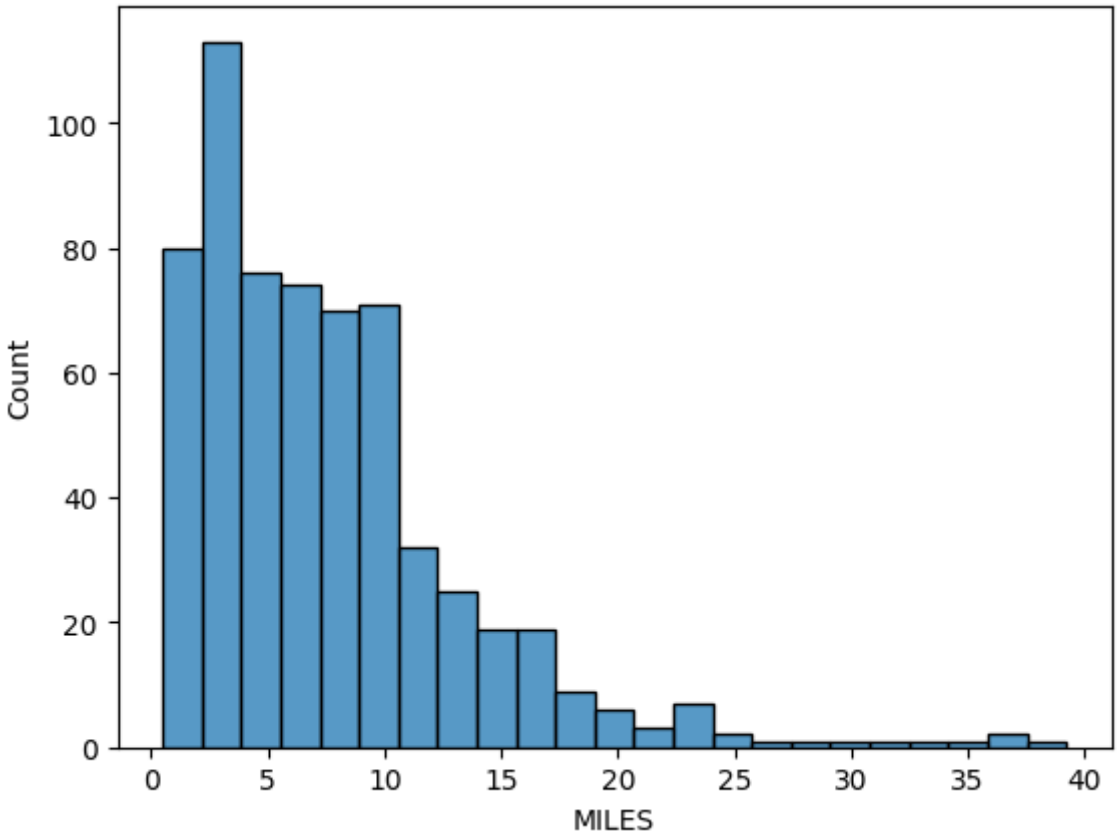
Result

Analysis of the given data :-






300
250
200
150
100
50
0
0


100
80
60
40
20
0
Count
0
5
10
15
20
25
30
35
40
MILES

Chapter 5

5.1 CONCLUSION

Before working on features first we need to know about the data insights which we get to know by EDA. Apart from that, we visualize the data by drawing various plots, due to which we understand that we don’t have any data for taxi’s price, also the price variations of other cabs and different types of weather. Other value count plots show the type and amount of data the dataset has. After this, we convert all categorical values into continuous data type and fill price Nan by the median of other values. Then the most important part of feature selection came which was done with the help of recursive feature elimination. With the help of RFE, the top 25 features were selected. Among those 25 features still, there are some features which we think are not that important to predict the price so we drop them and left with 8 important columns.

We apply four different models on our remaining dataset among which Decision Tree, Random Forest, and Gradient Boosting Regressor prove best with 96%+ accuracy on training for our model. This means the predictive power of all these three algorithms in this dataset with the chosen features is very high but in the end, we go with random forest because it does not prone to overfitting and design a function with the help of the same model to predict the price.

CHAPTER 6

Reference_link

Abel Brodeurand & Kerry Nield (2018) An empirical analysis of taxi, Lyft and Uber rides: Evidence from weather shocks in NYC

Junfeng Jiao (2018) Investigating Uber price surges during a special event in Austin, TX

Anna Baj-Rogowska (2017) Sentiment analysis of Facebook posts: The Uber Case Anastasios Noulas, Cecilia Mascolo, Renaud Lambiotte, and Vsevolod Salnikov (2014) OpenStreetCab: Exploiting Taxi Mobility Patterns in New York City to Reduce Commuter Costs

- https://github.com/ankita1112/House-Prices-Advanced-Regression/blob/master/Housing_Prediction_full.ipynb
- https://scikit-learn.org/stable/modules/multiclass.html
- https://www.codegrepper.com/code-examples/python/confusion+matrix+python
- https://topepo.github.io/caret/recursive-feature-elimination.html
- https://arxiv.org/pdf/1503.03021.pdf
- https://www.sciencedirect.com/science/article/abs/pii/S0167268118301598
- https://www.kaggle.com/punit0811/machine-learning-project-basic-linear-regression
- https://gdcoder.com/decision-tree-regressor-explained-in-depth/
- https://blog.paperspace.com/implementing-gradient-boosting-regression-python/