VISVESVARAYA TECHNOLOGICAL UNIVERSITY

JNANA SANGAMA, BELAGAVI – 590 018

An Internship Project Report
on

Uber Fare Price Prediction

Submitted in partial fulfilment of the requirements for the VIII Semester of the degree of **Bachelor of Engineering in Information Science and Engineering** of Visvesvaraya Technological University, Belagavi

by

R Jaswanth
1RN19IS117

Reshma G
1RN19IS119

Under the Guidance of

Mr. T S Bhagavath Singh
Associate Professor
Department of ISE

ESTD:2001
An Institute with a Difference

Department of Information Science and Engineering

RNS Institute of Technology

Dr. Vishnuvaradhan Road, Rajarajeshwari Nagar post,
Channasandra, Bengaluru-560098

2022-2023

RNS INSTITUTE OF TECHNOLOGY

Dr. Vishnuvaradhan Road, Rajarajeshwari Nagar post,
Channasandra, Bengaluru - 560098

DEPARTMENT OF INFORMATION SCIENCE AND ENGINEERING

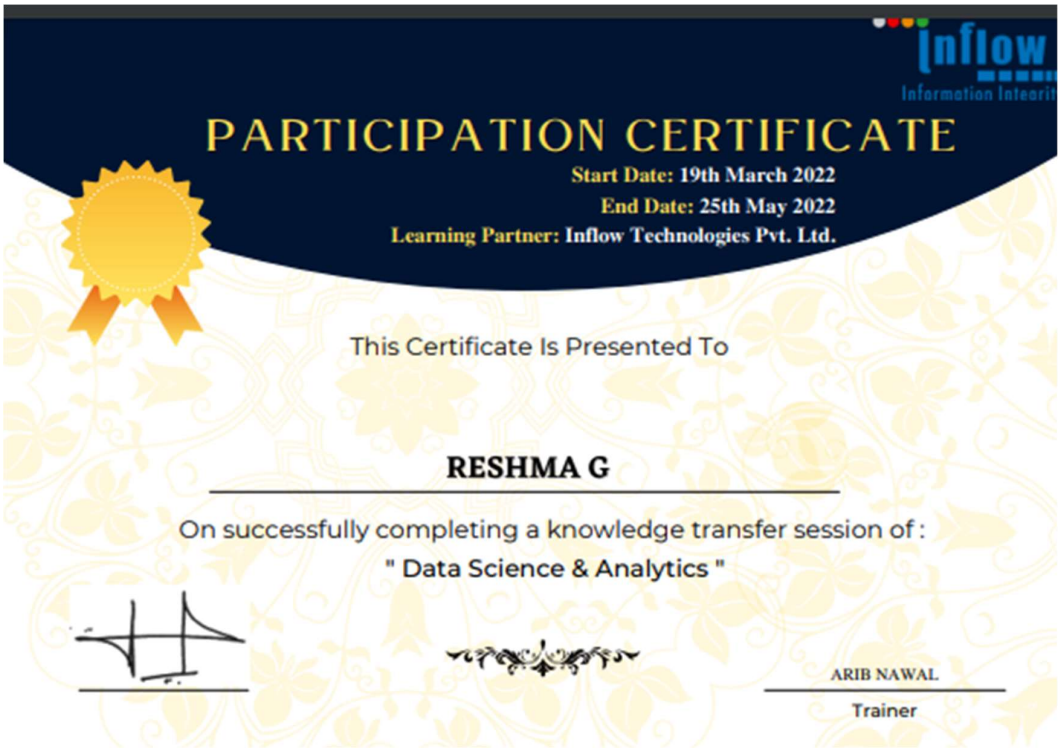
CERTIFICATE

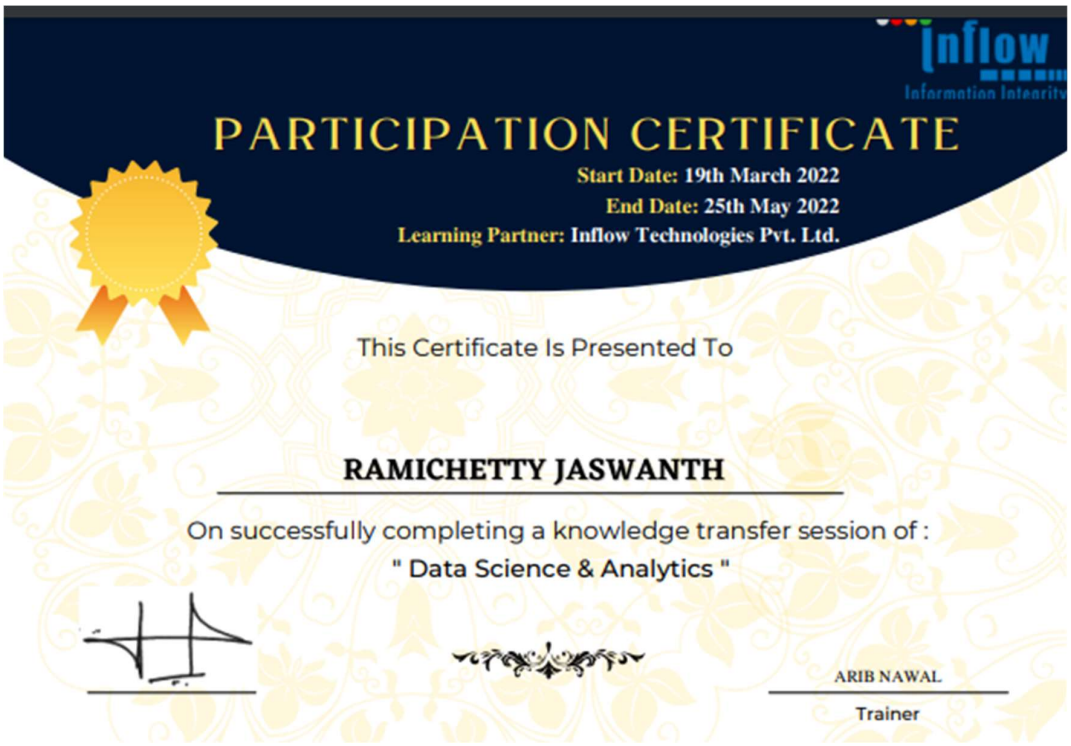
Certified that the Internship work entitled ***Uber Fare Price Prediction*** has been successfully completed by **R.Jaswanth (1RN19IS117)** and **Reshma.G (1RN19IS119),** bonafide students of **RNS Institute of Technology, Bengaluru** in partial fulfilment of the requirements of 8th semester for the award of degree in **Bachelor of Engineering in Information Science and Engineering** of **Visvesvaraya Technological University, Belagavi** during the academic year **2022-2023**. The internship report has been approved as it satisfies the academic requirements in respect of internship work for the said degree.

Mr. T S Bhagavath Singh	**Dr. R Rajkumar/Mr. Pramoda R**	**Dr. Suresh L**
Internship Guide	Internship Coordinator	Professor and HoD
Associate Professor	Associate Professor	Department of ISE
Department of ISE	Department of ISE	RNSIT

External Viva

Name of the Examiners	**Signature with Date**
1. _____	**1.** _____
2. _____	**2.** _____


Inflow
Information Integrity
PARTICIPATION CERTIFICATE
Start Date: 19th March 2022
End Date: 25th May 2022
Learning Partner: Inflow Technologies Pvt. Ltd.
This Certificate Is Presented To
RESHMA G
On successfully completing a knowledge transfer session of :
" Data Science & Analytics "
ARIB NAWAL
Trainer


Inflow
Information Integrity
PARTICIPATION CERTIFICATE
Start Date: 19th March 2022
End Date: 25th May 2022
Learning Partner: Inflow Technologies Pvt. Ltd.
This Certificate Is Presented To
RAMICHETTY JASWANTH
On successfully completing a knowledge transfer session of :
" Data Science & Analytics "
ARIB NAWAL
Trainer

DECLARATION

We, **R JASWANTH [USN:1RN19IS117] & RESHMA G [USN:1RN19IS119],** students of VII Semester BE, in Information Science and Engineering, RNS Institute of Technology hereby declare that the Internship work entitled ***Uber Fare Price Prediction*** has been carried out by us and submitted in partial fulfillment of the requirements for the *VII Semester degree of* ***Bachelor of Engineering in Information Science and Engineering*** *of Visvesvaraya Technological University, Belagavi* during academic year 2022-2023.

Place : Bengaluru

Date :

R JASWANTH 1RN19IS117
RESHMA G 1RN19IS119

ABSTRACT

This project is about the world's largest taxi company Uber. In this project, we will predict the fare amount in dollars for their future transactional process. We don't even remember some people refuses a ride in the dead of the night or on late evenings. All this – and much more, with the advent of online cab services, through which a user can book a ride in a matter of a few minutes. Since, enormous people are using this cab service, so it is important to manage the data, which result into benefitting them by implementing other business ideas. This Dataset contains following fields: 1)Key:- A unique id(Identifier) for each passenger 2)fare_amount:- The cost of the trip 3)pickup_datetime:- the time of pickup the passenger 4)passenger_count - The number of passengers in the vehicle (driver entered value) 5)pickup_longitude - the longitude where the meter was engaged 6)pickup_latitude - the latitude where the meter was engaged 7)dropoff_longitude – the longitude where the meter was disengaged 8)dropoff_latitude - the latitude where the meter was disengaged.

ACKNOWLEDGMENT

At the very onset we would like to place our gratefulness to all those people who helped us in making the Internship a successful one.

Coming up, this internship to be a success was not easy. Apart from the sheer effort, the enlightenment of the very experienced teachers also plays a paramount role because it is they who guided us in the right direction.

First of all, we would like to thank the **Management of RNS Institute of Technology** for providing such a healthy environment for the successful completion of internship work.

In this regard, we express sincere gratitude to our beloved Principal **Dr. M K Venkatesha,** for providing us all the facilities.

We are extremely grateful to our own and beloved Professor and Head of Department of Information science and Engineering, **Dr. Suresh L**, for having accepted to patronize us in the right direction with all his wisdom.

We place our heartfelt thanks to **Mr. T S Bhagavath Singh,** Associate Professor, Department of Information Science and Engineering for having guided internship and all the staff members of the department of Information Science and Engineering for helping at all times.

We thank **Mr. Arib Nawal, Inflow Technologies Pvt. Ltd.,** for providing the opportunity to be a part of the Internship program and having guided us to complete the same successfully.

We also thank our internship coordinators **Dr. R Rajkumar** Associate Professor,and **Mr. Pramoda R,** Assistant Professor, Department of Information Science and Engineering. We would thank our friends for having supported us with all their strength and might. Last but not the least, we thank our parents for supporting and encouraging us throughout. We have made an honest effort in this assignment.

R JASWANTH

RESHMA G

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

In recent years, a concept known as the “sharing economy” has taken the market by storm, giving rise to a number of truly revolutionary businesses. While a number of companies have cashed in on this trend, the sharing economy’s undisputed king is Uber as a ride-sharing company that empowers anyone to art earning money with their vehicle and enables those needing a lift to quickly and affordable to find a ride. The amount of success Uber has been able to achieve in their short history is remarkable.
Uber’s disruptive technology, explosive growth, and constant controversy make it one of the most fascinating companies to emerge over the past decade. The almost ten-year-old company soon grew to become the highest valued private start up company in the world.

Sometimes it’s easy to give up on someone else’s driving. This is less stress, more mental space and one uses that time to do other things. Yes, that’s one of the ideas that grew and later became the idea behind Uber

The company offer passenger boarding services that allow users to rent cars with drivers through websites or mobile apps. Whether traveling a short distance or traveling from one city to another, these services have helped people in many ways and have actually made their lives very easy. Hence taking this into consideration we have created an program to predict the uber fare price very precisely and accurately.

1.1 Background

Uber was first founded in 2009 by Garrett Camp and Travis Kalanick under the name Uber Cab. At the time, Camp had recently spent 800 dollars hiring a private car to transport him and his friends on New Year’s Eve, and he was trying to figure out a way he could make the service more affordable to the average person. Camp reasoned that allowing multiple people to share the cost of the service would drive it down, and Uber Cab was born came the first Uber Cab employee. In 2011, the company’s name was shortened to Uber, and in 2012, Uber rolled out Uber X - a service which allowed people to work for Uber driving their own car. Since then,

Uber has been on the cutting edge of a number of transportation services and technologies, from self-driving cars, to a carpooling service, and even a helicopter service. Today, Uber operates in 300 cities across 6 continents, and in 2016 Uber grossed 20 billion dollars. Interestingly enough, Uber actually 2.8 billion dollars on that 2016 gross, showing just how committed the company is to continuing to push the envelope and develop new services and technologies that will revolutionize the transportation industry. While a lot of factors played into launching Uber into the level of recognition and success the company enjoys, one of those factors was the ever-recognizable Uber logo.

1.2 Existing System

Uber is an on-demand car service that allows users to request a ride through their Android or iPhone app. Once a driver is signaled through the app, it usually takes less than 10 minutes for a car to arrive at your door. During this brief wait period, Uber allows you to track the cars location so you know exactly when to expect its arrival. The app provides a cashless payment process by charging all Uber rides to an on file credit card attached to your account. One thing to note though, sometimes Uber enacts surge pricing during peak travel times, meaning your fare could double, triple, or even cost you up to 7X the normal amount! On the plus side, when surge pricing is in effect you will be notified within the app before you request a ride. Uber is currently available in 67 countries!

Uber Fare Finder calculates the cost of your Uber ride. Enter your pickup and dropoff locations, and get the fare estimate and surge pricing information for Uber's primary services (UberX, Uber Black, Uber XL, etc.) available in your area.

1.3 Proposed System

The proposed techniques are used for Predicting the Fare amount to be paid in dollars. We have done prediction using the machine learning algorithm such as Linear Regression,by using a various input fields such as 1)Key:- A unique id(Identifier) for each passenger 2)fare_amount:- The cost of the trip 3)pickup_datetime:- the time of pickup the passenger 4)passenger_count - The number of passengers in the vehicle (driver entered value) 5)pickup_longitude - the longitude where the meter was engaged 6)pickup_latitude - the latitude where the meter was engaged 7)dropoff_longitude - the longitude where the meter was disengaged 8)dropoff_latitude - the latitude where the meter was disengaged.

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a “least squares” method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).

Chapter 2

LITERATURE REVIEW

[1] Data Analysis of Uber and Lyft Cab Services by Shashank H, The author tells us the popularity of uber in the recent years and about the urban citizens who are benefited by the uber. Later the author compares the difference between the competitive taxis and uber and defines new way of calling and also the new way of paying for cabs, the author also tells us about the importance of data produced by the cabs daily and also about the visualization and analysis of data. After that the author tells how the different time and different environments will have an effect on passengers to make different choices.

[2] A survey suggests that the Cab fares vary in step with different factors like place, time of the day, and so forth. Cabs as well, wherein the fare depends upon the wide variety of passengers, visitors, so on. The vendor has facts about all of the factors, but the buyers can get admission to the records that is constrained and we cannot expect the price lists. Uber and Ola use factors like traffic in a specific vicinity, and call for and supply elements motive of the paper is to investigate the factors that have an impact on the deviation in the tariffs and the way they're associated with the trade inside the prices.

[3] In the previous generation, the fare changed into handiest depending on distance, however with the enhancement, in technologies, the cab's fare is dependent on lots of things like time, area, number of passengers, traffic, quantity of hours, base fare, and so forth. The look at is based on Supervised mastering whose one application is prediction, in device getting to know. This research aims to observe predictive evaluation, which is a method of analysis in Machine Learning. Many corporations like Ola, Uber, and many others makes use of Artificial Intelligence and system learning technology to find the answer to correct fare prediction ha

Chapter 3

ANALYSIS

System analysis is a process of gathering and interpreting facts, and diagnosing problems and information about the Breast Cancer Prediction System to recommend improvements to the system. It is a problem-solving activity that requires intensive communication between the system users and system developers. System analysis or study is an important phase of any system development process.

3.1 Introduction

The system is studied to the minutest detail and analysed. The system analyst plays the role of the interrogator and dwells deep into the working of the present system. The system is viewed as a whole and the input to the system is identified. The outputs from the organizations are traced to the various processes.

System analysis is concerned with becoming aware of the problem, identifying the relevant and decisional variables, analysing and synthesizing the various factors and determining an optimal or at least a satisfactory solution or program of action. A detailed study of the process must be made by various techniques like interviews, questionnaires etc. The data collected by these sources must be scrutinized to arrive at a conclusion. The conclusion is an understanding of how the system functions. This system is called the existing system.

The designer now functions as a problem solver and tries to sort out the difficulties that the enterprise faces. The solutions are given as proposals. The proposal is then weighed with the existing system analytically and the best one is selected. The proposal is presented to the user for endorsement by the user. The proposal is reviewed on user request and suitable changes are made. This is a loop that ends as soon as the user is satisfied with the proposal.

A preliminary study is a process of gathering and interpreting facts, and using the information for further studies on the system. The preliminary study is problem-solving activity that requires intensive communication between the system users and system developers. It does various feasibility studies. In these studies, a rough figure of the system activities can be

obtained, from which the decision about the strategies to be followed for effective system study and analysis can be taken. This feasibility study can be implemented.

3.2 Hardware Requirements

The Hardware requirements are very minimal and the program can be run on most of the machines. Table 3.1 gives details of hardware requirements.

Table 3.1: Hardware Requirements

Processor	Intel Core i3 processor
Processor Speed	1.70 GHz
RAM	4 GB
Storage Space	40 GB
Monitor Resolution	1024*768 or 1336*768 or 1280*1024

3.3 Software Requirements

The software requirements are description of features and functionalities of the system. Table 3.2 gives details of software requirements.

Table 3.2: Software Requirements

Operating System	Windows 8.1
IDE	VS code
Tools	Power BI
Libraries	Pandas, Numpy, Streamlit, Matplotlib, Seaborn

3.3.1 VS code

Visual Studio Code is a source-code editor that can be used with a variety of programming languages,including C#, Java, JavaScript, Go, Node.js, Python, C++, C, Rust and Fortran It is based on the Electron framework,[21] which is used to develop Node.js web applications that run on the Blink layout engine. Visual Studio Code employs the same editor component (codenamed "Monaco") used in Azure DevOps (formerly called Visual Studio Online and Visual Studio Team Services).

Out of the box, Visual Studio Code includes basic support for most common programming languages. This basic support includes syntax highlighting, bracket matching, code folding,

and configurable snippets. Visual Studio Code also ships with IntelliSense for JavaScript, TypeScript, JSON, CSS, and HTML, as well as debugging support for Node.js. Support for additional languages can be provided by freely available extensions on the VS Code Marketplace.

3.3.2 Google Colaboratory

Google Colaboratory,is a cloud-based environment for writing documents with live code, visualizations, and narrative text. For those who are familiar with Jupyter notebooks, Colab notebooks are the same, including the .ipynb extension. Unlike Jupyter and Atom (our previous editor for code and reports), however, Colab requires no setup on your computer! It also provides a large amount of free computing power and easy document sharing.A Colab notebook consists of text cells, code cells, and outputs of code cells.

1. Text cells are written in Markdown, a markup language we'll learn about in the next section. This means they can contain formatted text, images, HTML, LaTeX, and more.

2. Code cells are written in Python. We can also insert a system/terminal command by prefixing a line with !.

3. Outputs of code cells appear below their corresponding cell. They can include text, graphics, and information about errors that occurred while executing the code.

3.3.3 Power BI

Power BI is an interactive data visualization software product developed by Microsoft with a primary focus on business intelligence. It is part of the Microsoft Power Platform. In March 2016, Microsoft released an additional service called Power BI Embedded on its Azure cloud platform. Power BI is a collection of software services, apps, and connectors that work together to turn unrelated sources of data into coherent, visually immersive, and interactive insights. Data may be input by reading directly from a database, webpage, or structured files such as spreadsheets, CSV, XML, and JSON. It was released on 11th July 2011.Power BI Features:

- Power BI supports powerful data discovery and exploration that enables users to answer important questions in seconds.

- No prior programming knowledge is needed; users without relevant experience can start immediately with creating visualizations using Power BI.
- It can connect to several data sources that other BI tools do not support. Power BI enables users to create reports by joining and blending different datasets.
- Power BI Server supports a centralized location to manage all published data sources within an organization.
- Power BI software's drag-and-drop features, intuitive drill-down capabilities, and natural language querying (that we'll share more about later), new users and data analysts alike can quickly create visualizations and dashboards to gain insight almost instantly.
- Power BI software's role-based permissions user can manage who has access to what data down to the row, and user can even define who can make changes to every data source or workbook.
- All views and dashboards in Power BI software are mobile and tablet compatible.
- Power BI software's Ask Data feature can be a time-saver when it comes to asking simple questions and needing to create quick visualizations.
- Additionally, sharing in Power BI software is especially powerful because, instead of sending a static report, user can share a dashboard that is interactive and provides more than just a single view.
- Power BI software's expansive and supportive community ignites learning across the organization and equips employees with extensive training and tutorial options, collaborative forums, and support. It is part of the Microsoft Power Platform.

Chapter 4

DESIGN

The System Design Document describes the system requirements, operating environment, system and subsystem architecture, files and database design, input formats, output layouts, human-machine interfaces, detailed design, processing logic, and external interfaces.

4.1 Data Analysis Life Cycle

The data analysis lifecycle describes the process of conducting a data analytics project, which consists of six key steps based on the CRISP-DM methodology. Data analysis is the process of examining data sets in order to find trends and draw conclusions about the information they contain. Increasingly, data analytics is done with the aid of specialized systems and software. When presented with a data project, user will be given a brief outline of the expectations. From that outline, user should identify the key objectives that the business is trying to uncover.

Fig 4.1: Data Analysis Life Cycle

When presented with a small dataset, user can use tools like Excel, R, Python, Power BI Prep or Power BI Desktop to help prepare user data for its cleaning. Once user has organized and identified all the variables in the dataset, user can begin cleaning. Using different statistical modelling methods, user can determine which is the best. Interactive visualization tools like

Power BI are tremendously useful in illustrating user conclusions to clients. Being able to tell a story with user data is essential and necessary.

4.2 Libraries

A Python library is a collection of related modules. It contains bundles of code that can be used repeatedly in different programs. It makes Python Programming simpler and convenient for the programmer. As we don't need to write the same code again and again for different programs. Python libraries play a very vital role in fields of Machine Learning, Data Science, Data Visualization, etc.

4.2.1 NumPy

NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices. NumPy was created in 2005 by Travis Oliphant. It is an open-source project and user can use it freely. NumPy stands for Numerical Python.

Main Features

- NumPy aims to provide an array object that is up to 50x faster than traditional Python lists. The array object in NumPy is called ndarray, it provides a lot of supporting functions that make working with ndarray very easy. It is optimized to work with latest CPU architectures.
- We can use the functions in NumPy to work with code written in other languages. We can hence integrate the functionalities available in various programming languages. This helps implement inter-platform functions.
- It has the capability to perform complex operations of the elements like linear algebra, Fourier transform, etc. We have separate modules for each of the complex functions. We have the linalg module for linear algebra functions.

4.2.2 Pandas

Pandas is a Python package that provides fast, flexible, and expressive data structures designed to make working with ”relational” or ”labeled” data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real world data analysis in Python. Additionally, it has the broader goal of becoming the most powerful and flexible open source data analysis / manipulation tool available in any language. It is already well on its way towards this goal.

Main Features

- Easy handling of missing data (represented as NaN, NA, or NaT) in floating point as well as non-floating point data
- Size mutability: columns can be inserted and deleted from DataFrame and higher dimensional objects.
- Automatic and explicit data alignment: objects can be explicitly aligned to a set of labels, or the user can simply ignore the labels and let Series, DataFrame, etc. automatically align the data for user in computations.
- Powerful, flexible group by functionality to perform split-apply-combine operations on data sets, for both aggregating and transforming data.
- Make it easy to convert ragged, differently-indexed data in other Python and NumPy data structures into DataFrame objects
- Intelligent label-based slicing, fancy indexing, and subsetting of large data sets.
- Intuitive merging and joining data sets.
- Flexible reshaping and pivoting of data sets.
- Hierarchical labeling of axes (possible to have multiple labels per tick).

Robust IO tools for loading data from flat files (CSV and delimited), Excel files, databases, and saving/loading data from the ultrafast HDF5 format.

4.2.3 Matplotlib

Matplotlib is a low level graph plotting library in python that serves as a visualization utility. It was created by John D. Hunter. It is open source and we can use it freely. Matplotlib is mostly written in python, a few segments are written in C, Objective-C and Javascript for Platform compatibility.

Main Features

- Creates publication quality plots.
- Makes interactive figures that can zoom, pan, update.
- Customizes visual style and lausert.
- Export to many file formats.
- Can be embedded in JupyterLab and Graphical User Interfaces.
- Uses a rich array of third-party packages built on Matplotlib.

4.2.4 Seaborn

Seaborn is an amazing visualization library for statistical graphics plotting in Python. It provides beautiful default styles and color palettes to make statistical plots more attractive. It is built on the top of matplotlib library and also closely integrated to the data structures from pandas. It provides dataset-oriented APIs, so that we can switch between different visual representations for same variables.

Main Features

- Built in themes for styling matplotlib graphics.
- Visualizing univariate and bivariate data.
- Fitting in and visualizing linear regression models.
- Seaborn works well with NumPy and Pandas data structures.
- It comes with built in themes for styling Matplotlib graphics.

4.2.5 Streamlit

Streamlit is an open source app framework in Python language. It helps us create web apps for data science and machine learning in a short time. It is compatible with major Python libraries such as scikit-learn, Keras, PyTorch, SymPy(latex), NumPy, pandas, Matplotlib etc.

Main Features

- Streamlit allows you to re-use any Python code you have already written. This can save considerable amounts of time compared to non-Python based tools where all code to create visualizations needs to be re-written.
- No hidden state
- No callbacks
- Build apps in a dozen lines of Python with a simple API

4.3 Dataset

A data set is a collection of related, discrete items of related data that may be accessed individually or in combination or managed as a whole entity. A data set is organized into some type of data structure. For Breast Cancer Prediction, we have used a dataset from Kaggle. After a suspicious lump is found, the doctor will conduct a diagnosis to determine whether it is cancerous and, if so, whether it has spread to other parts of the body.

This breast cancer dataset was obtained from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

It contains ID number and features computed for each cell nucleus that are radius (mean of distances from center to points on the perimeter), texture (standard deviation of gray-scale values), area, smoothness (local variation in radius lengths), concavity (severity of concave portions of the contour), symmetry feature 'Diagnosis'(M = malignant, B = benign) is response variable and it takes value 1 in case of Cancer and 0 otherwise.

	Unnamed: 0	key	fare_amount	pickup_datetime	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
0	24238194	2015-05-07 19:52:06.0000003	7.5	2015-05-07 19:52:06 UTC	-73.999817	40.738354	-73.999512	40.723217	1.0
1	27835199	2009-07-17 20:04:56.0000002	7.7	2009-07-17 20:04:56 UTC	-73.994355	40.728225	-73.994710	40.750325	1.0
2	44984355	2009-08-24 21:45:00.00000061	12.9	2009-08-24 21:45:00 UTC	-74.005043	40.740770	-73.962565	40.772647	1.0
3	25894730	2009-06-26 08:22:21.0000001	5.3	2009-06-26 08:22:21 UTC	-73.976124	40.790844	-73.965316	40.803349	3.0
4	17610152	2014-08-28 17:47:00.000000188	16.0	2014-08-28 17:47:00 UTC	-73.925023	40.744085	-73.973082	40.761247	5.0

Fig 4.2: Description of Uber Fare Price Prediction Dataset

Chapter 5

IMPLEMENTATION DETAILS

Implementation is the process of defining how the system should be built, and ensuring that it is operational and meets quality standards. It is a systematic and structured approach for effectively integrating a software-based service or component into the requirements of end users.

5.1 Training/Testing

```
y=uber_2['fare_amount']
x=uber_2.drop(['fare_amount','pickup_datetime','Day of Week','pickup','drop off'],axis=1)
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.25,random_state=10)
```

The main difference between training data and testing data is that training data is the subset of the original data that is used to train the machine learning model, whereas testing data is used to check the accuracy of the model.

The training dataset is generally larger in size compared to the testing dataset. The general ratios of splitting train and test datasets are 80:20, 70:30, or 75:25.

The code block here aids in separating the X and the Y values, dividing the data into inputs parameters and outputs value format, and finally training and testing the bifurcated data.

We will be dividing the dataset into two main groups. One for training the model and the other for Testing our trained model's performance.

The random state hyperparameter in the train_test_split() function controls the shuffling process. With random_state=None, we get the different train and test sets across different executions and the shuffling process is out of control. With random_state=0, we get the same train and test sets across different executions.

42 is just a random number that helps to reproduce the same result after reusing the train_test_split. If we want, we can choose other numbers as well.

5.2 Algorithms

An “algorithm” in machine learning is a procedure that is run on data to create a machine learning “model.” Machine learning algorithms perform “pattern recognition.” Algorithms “learn” from data, or are “fit” on a dataset. There are many machine learning algorithms.

5.2.1 Linear Regression

```
from sklearn.linear_model import LinearRegression  #linear regression model
LR=LinearRegression()
LR.fit(x_train,y_train)
y_pred=LR.predict(x_test)
y_pred
```

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a “least squares” method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).

The Results of this model in our use case show that it boasts the highest accuracy of 0.97202

5.3 Code Segment

Code segment of Web application

```
def user_report():
    pickup_longitude=st.sidebar.slider("Pickup longitude",-180.0,180.0)
    pickup_latitude=st.sidebar.slider("PickupLatitude",90.0,90.0)
   dropoff_longitude=st.sidebar.slider("Dropoff Longitude",-180.0,180.0)
```

```
    dropoff_latitude=st.sidebar.slider(" Dropoff Latitude",-90.0,90.0)

      passenger_count=st.sidebar.slider(" Passenger Count",1,8)
    Distance=st.sidebar.slider("Distance",0.010000,53.070000)
    Year=st.sidebar.slider("Year",2009,2015)
    Month=st.sidebar.slider("Month",1,12)
    Day=st.sidebar.slider("Day",1,31)
    DayofWeek_num=st.sidebar.slider("Day of Week num",0,6)
    Hour=st.sidebar.slider("Hour",0,23)
    counter=st.sidebar.slider("Counter",0,1)

user_report_data={
       'pickup_longitude':pickup_longitude,
       'pickup_latitude':  pickup_latitude,
       'dropoff_longitude': dropoff_longitude,
       'dropoff_latitude': dropoff_latitude,
       'passenger_count':passenger_count,
       'Distance':Distance,
       'Year': Year,
       'Month': Month,
       'Day': Day,
       'Day of Week_num':DayofWeek_num,
       'Hour':Hour,
       'counter': counter
        }
    report_data=pd.DataFrame(user_report_data,index=[0])
    return report_data
user_data=user_report()
st.header('Uber data')
st.write(user_data)
Fare=model.predict(user_data)
st.subheader('Uber fare')
st.subheader(np.round(Fare[0],2)
```

Chapter 6

TESTING

Testing is vital for the success of any software. no system design is ever perfect. Testing is also carried out in two phases. The first phase is during the software engineering that is during module creation. The second phase is after the completion of the software. this is system testing which verifies that the whole set of programs hanged together.

6.1 Types of Testing

There are numerous types of software testing techniques that can be used to ensure changes to the code work as expected. Not all testing is equal though, and also explore how some testing practices differ.

White Box Testing

In this technique, the close examination of the logical parts through the software is tested by cases that exercise species sets of conditions or loops. all logical parts of the software were checked once. errors that can be corrected using this technique are typographical errors, logical expressions which should be executed once may be getting executed more than once and errors resulting from using wrong controls and loops. When the box testing tests all the independent part within a module, the logical decisions on their true and the false side are exercised, all loops and bounds within their operational bounds were exercised and internal data structure to ensure their validity were exercised once.

Black Box Testing

This method enables the software engineer to devise sets of input techniques that fully exercise all functional requirements for a program. black box testing tests the input, the output and the external data. it checks whether the input data is correct and whether we are getting the desired output.

Alpha Testing

The alpha testing proceeds until the system developer and the Credit Card agrees that the provided system is an acceptable implementation of the system requirements. Acceptance testing is also sometimes called alpha testing. Bespoke systems are developed for a card.

Beta Testing

On the other hand, when a system is to be marked as a software product, another process called beta testing is often conducted. During beta testing, a system is delivered to a number of potential users who agree to use it. The Credit Cards then report problems to the developers. This provides the product for real use and detects errors which may not have been anticipated by the system developers.

Unit Testing

Each module is considered independently. it focuses on each unit of software as implemented in the source code. it is white-box testing.

Integration Testing:

Integration testing aims at constructing the program structure while at the same constructing tests to uncover errors associated with interfacing the modules. modules are integrated by using the top-down approach.

Validation Testing

Validation testing was performed to ensure that all the functional and performance requirements are met.

System Testing:

It is executing programs to check logical changes made in it with intention of finding errors. A system is tested for the online response, the volume of interactions, recovery from failure etc. System testing is done to ensure that the system satisfies all the user requirements.

6.2 Implementation and Software Specification Testing

This phase of the systems development life cycle refines hardware and software specifications, establishes programming plans, trains users and implements extensive testing procedures, to evaluate the design and operating specifications and/or provide the basis for further modification.

Technical Design

This activity builds upon specifications produced during the new system design, adding detailed technical specifications and documentation.

Test Specification Planning

This activity prepares detailed test specifications for individual modules and programs, job streams, subsystems, and for the system as a whole.

Programming and Testing

This activity encompasses the actual development, writing, and testing of program units or modules.

User Training

This activity encompasses writing user procedure manuals, preparation of user training materials, conducting training programs, and testing procedures.

Acceptance Test

A final procedural review to demonstrate a system and secure user approval before a system becomes operational.

Installation Phase

In this phase the new Computerized system is installed, the conversion to new procedures is fully implemented, and the potential of the new system is explored.

System Installation

The process of starting the actual use of a system and training user personnel in its operation.

Review Phase

This phase evaluates the successes and failures during a systems development project, and measures the results of a new Computerized Transport system in terms of benefits and savings projected at the start of the project.

Development Recap

A review of a project immediately after completion to find successes and potential problems in future work.

Post Implementation Review

A review, conducted after a new system has been in operation for some time, evaluates actual system performance against original expectations and projections for cost-benefit improvements. Also identifies maintenance projects to enhance or improve the system.

Chapter 7

DISCUSSION OF RESULTS

In this chapter the results of the study are presented and discussed with reference to the aim of the study, which was to determine the best model for breast cancer prediction.

7.1 Results

The Linear Regression model is used to give comparison between the actual value and the predicted values. So that the fare amount in dollars is predicted.

```
[39] df2=pd.DataFrame({'Actual values':y_test,'Predicted values':y_pred})
     df2
```

	Actual values	Predicted values
1414	5.7	8.853714
7180	14.5	8.266214
16699	6.5	9.758287
5404	5.7	5.539214
18681	17.0	17.147625
...	...	...
9620	7.7	8.767381
15592	8.5	7.439466
20607	6.5	7.691384
16992	22.9	27.380450
25301	5.5	8.053183

6485 rows × 2 columns

Fig 7.1.1 Linear regression

7.2 MODEL DEPLOYMENT RESULTS

Model deployment is the process of putting machine learning models into production. The homepage contains eight input fields upon which being entered will predict the fare amount to be paid in dollors. This model is deployed using Streamlit web Application Programming Interface.

For the following values, predicted output is the fare amount to be paid in dollars.

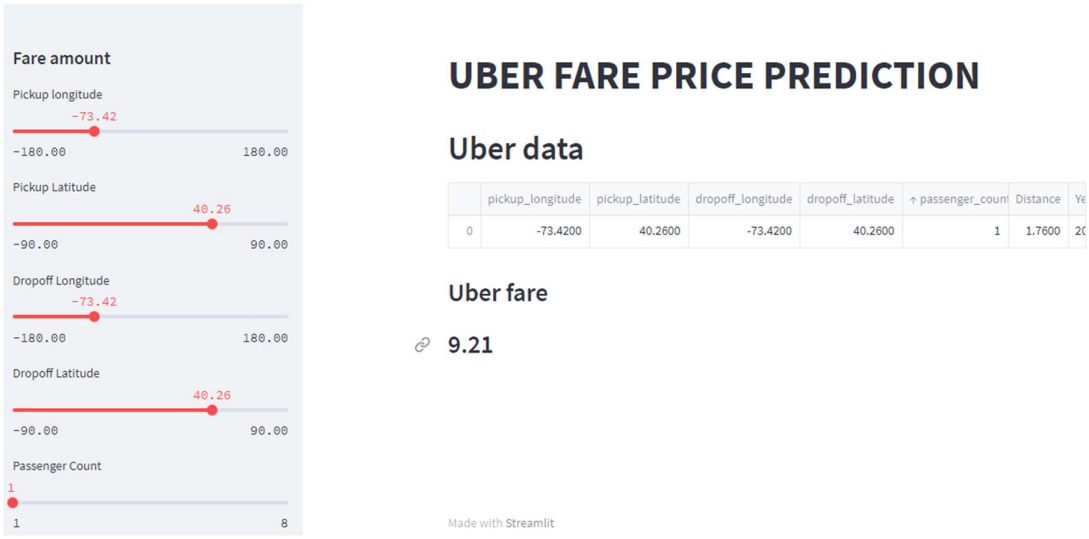


Fig 7.2.1: Prediction

7.3 Visualization of results in Power BI

A scatter plot (aka scatter chart, scatter graph) uses dots to represent values for two different numeric variables. The position of each dot on the horizontal and vertical axis indicates values for an individual data point. Scatter plots are used to observe relationships between variables. This is a scatter plot showing the distance vs fare amount. It is almost a linear plot with few outliers.

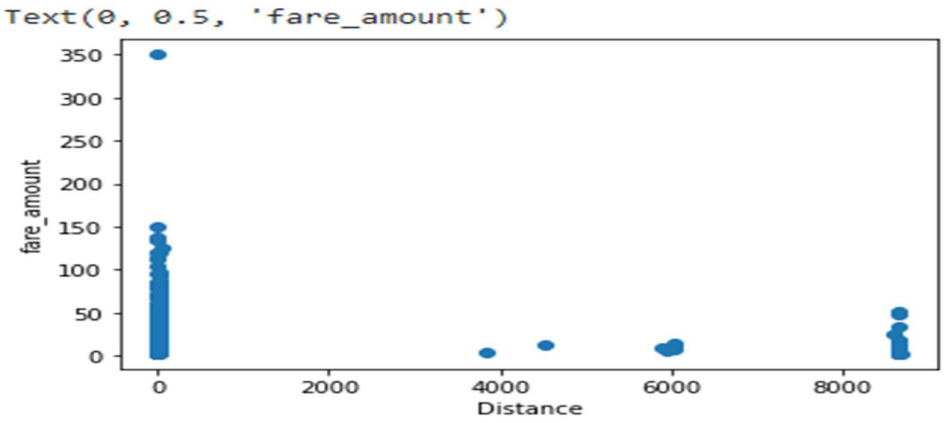


Fig 7.3.1: Distance vs fare amount

A bar plot or bar chart is a graph that represents the category of data with rectangular bars with lengths and heights that is proportional to the values which they represent. The bar plots can be plotted horizontally or vertically. A bar chart describes the comparisons between the discrete categories. One of the axis of the plot represents the specific categories being compared, while the other axis represents the measured values corresponding to those categories.This bar graph is plotted average number of trips in a particular year

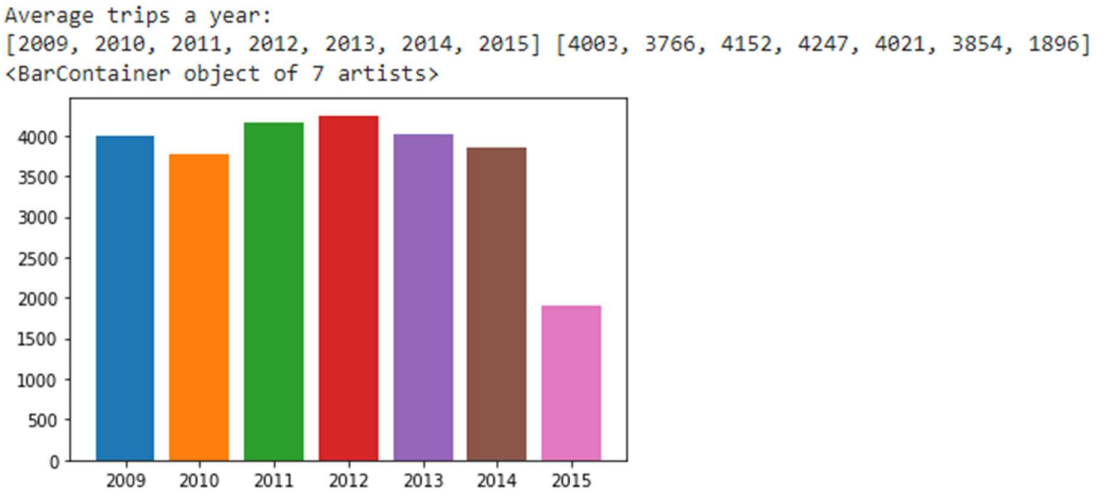


Fig 7.3.2: Number of trips in a year

A bar graph is a specific way of representing data using rectangular bars where the length of each bar is proportional to the value they represent. It is a graphical representation of data using bars of different heights. In real life, bar graphs are commonly used to represent business data. This bar graph is plotted average number of trips in a particular month.

```
Average trips a Month:
['Jan', 'Feb', 'Mar', 'Apr', 'May', 'Jun', 'Jul', 'Aug', 'Sep', 'Oct',
<BarContainer object of 12 artists>
```

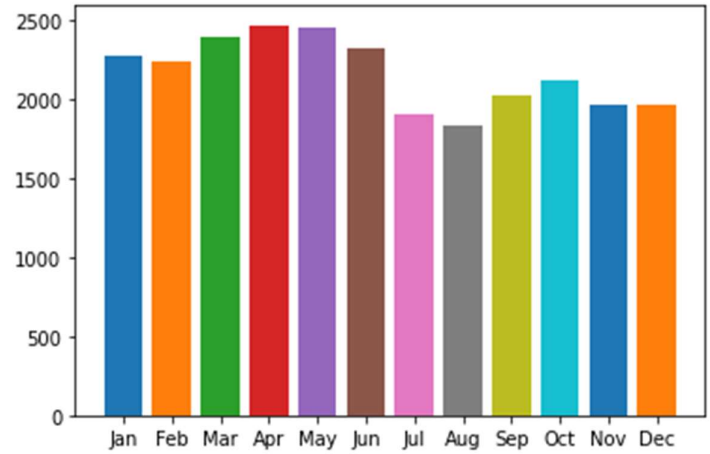


Fig 7.3.3: Number of trips in a month

A bar graph is a specific way of representing data using rectangular bars where the length of each bar is proportional to the value they represent. It is a graphical representation of data using bars of different heights. In real life, bar graphs are commonly used to represent business data. This bar graph is plotted average number of trips in a particular month.

```
Average trips by Days:
['Mon', 'Tue', 'Wed', 'Thu', 'Fri', 'Sat', 'Sun'] [3140, 3770, 3766, 3857, 3973, 4075, 3358]
<BarContainer object of 7 artists>
```

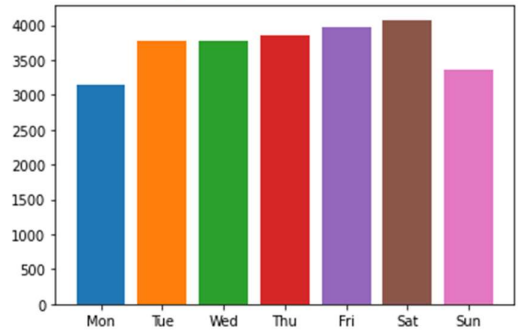


Fig 7.3.4: Number of trips in a day

A line chart is a graphical representation of an asset's historical price action that connects a series of data points with a continuous line. This is the most basic type of chart used in finance, and it typically only depicts a security's closing prices over time. This line chart is plotted between year vs trips.

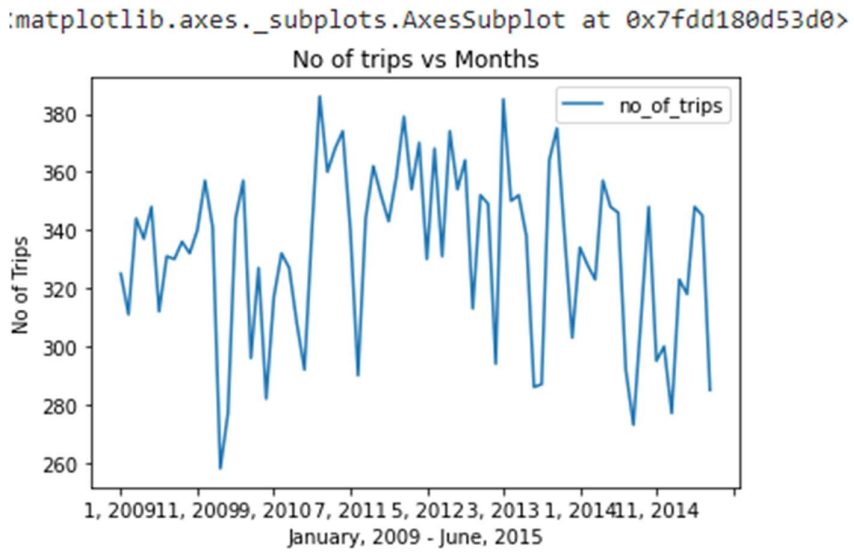


Fig 7.3.5: year vs trips

A heatmap is a two-dimensional graphical representation of data where the individual values that are contained in a matrix are represented as colours.The Seaborn package allows the creation of annotated heatmaps which can be tweaked using Matplotlib tools as per the creator's requirement.

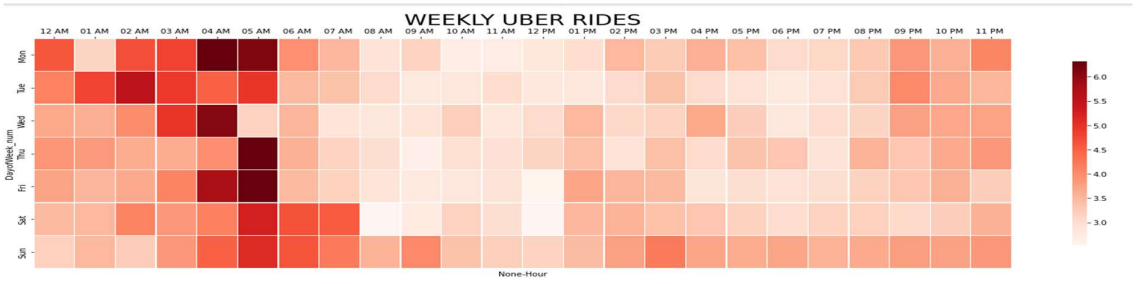


Fig 7.3.6: Weekly uber rides

Displot function provides access to several approaches for visualizing the univariate or bivariate distribution of data, including subsets of data defined by semantic mapping and faceting across multiple subplots.

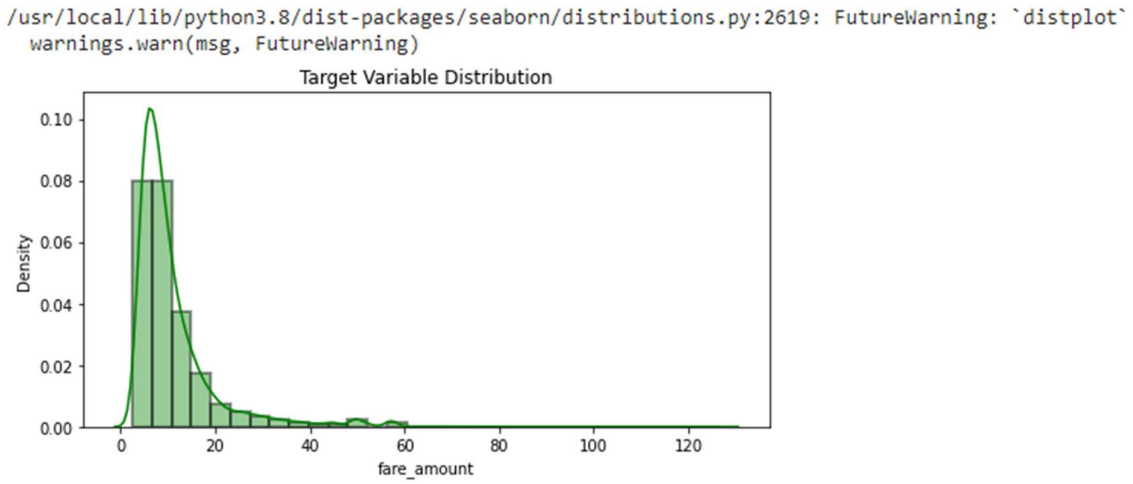


Fig 7.3.7: Density vs fare amount

Chapter 8

CONCLUSION AND FUTURE WORK

8.1 Conclusion

Here are some of the key outcomes of the project:

- The Dataset was large enough total of 2M samples & after preprocessing 18.4% of the data samples were dropped.
- Visualising the distribution of data & their relationships, helped us to get some insights on the feature-set.
- The features had high multicollinearity, hence in Feature Extraction step, we shortlisted the appropriate features with VIF Technique.
- Testing multiple algorithms with default hyperparamters gave us some understanding for various models performance on this specific dataset.
- While, Polynomial Regression (Order-5) was the best choise, yet it is safe to use multiple regression algorithm, as their scores were quiet comparable & also they're more generalisable.

8.2 Future work

- We can make the interface more appealing to everyone.
- We can introduce more accurate algorithms in the future to enhance the price prediction.
- We can use various types of regressions in the code.
- We can further make the interface more user friendly.
- We can also introduce chat bots to help people easily.

REFERENCES

[1] A retrospective on Taxi Hailing methodology Sanskar Shah Student, Department of Information Technology, K J Somaiya College of Engineering, Maharashtra, India. International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 07 Issue: 09 | Sep 2020 www.irjet.net p-ISSN: 2395-0072

[2] Banerjee, Pallab & Kumar, Biresh & Singh, Amarnath & Ranjan, Priyeta & Soni, Kunal. (2020). Predictive Analysis of Taxi Fare using Machine Learning. International Journal of Scientific Research in Computer Science, Engineering and Information Technology. 373-378. 10.32628/CSEIT2062108.

[3] Chao, Junzhi. (2019). Modeling and Analysis of Uber's Rider Pricing. 10.2991/aebmr.k.191217.127.

[4] Faghih, Sabiheh & Safikhani, Abolfazl & Moghimi, Bahman & Kamga, Camille. (2017). Predicting Short-Term Uber Demand Using Spatio-Temporal Modeling: A New York City Case Study.

[5] Khandelwal, K., Sawarkar, A. ., & Hira, S. (2021). A Novel Approach for Fare Prediction Using Machine Learning Techniques. International Journal of Next-Generation Computing, 12(5). https://doi.org/10.47164/ijngc.v12i5.451

[6] Kunal, Arora & Kaur, Sharanjit & Sharma, Vinod. (2021). Prediction of Dynamic Price of Ride-On-Demand Services Using Linear Regression. International Journal of Computer Applications & Information Technology. 13. 376-389.

[7] Predicting real-time surge pricing of ride-sourcing companies Matthew Battifaranoa , Zhen (Sean) Qiana,b, A Department of Civil and Environmental Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, United States b Heinz College, Carnegie Mellon University, Pittsburgh, PA 15213, United States

[8] Zhao, Kai & Khryashchev, Denis & Huy, Vo. (2019). Predicting Taxi and Uber Demand in Cities: Approaching the Limit of Predictability. IEEE Transactions on Knowledge and Data Engineering. PP. 1-1. 10.1109/TKDE.2019.2955686