Project Report

1. Zomato Customer Review Prediction

1.1 Introduction:

Machine learning gives computers the ability to learn without being explicitly programmed (Arthur Samuel, 1959). It is a subfield of computer science.

Machine learning (**ML**) is the scientific study of algorithms and statistical models that computer systems use to effectively perform a specific task without using explicit instructions, relying on patterns and inference instead. It is seen as a subset of artificial intelligence. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task. Machine learning algorithms are used in a wide variety of applications, such as email filtering, and computer vision, where it is infeasible to develop an algorithm of specific instructions for performing the task. Machine learning is closely related to computational statistics, which focuses on making predictions using computers. The study of mathematical optimization delivers methods, theory and application domains to the field of machine learning.

In today's digitized modern world, popularity of food apps is increasing due to its functionality to view, book and order for food by a few clicks on the phone for their favourite restaurant or cafes, by surveying the user ratings and reviews of the previously visited customers. Food app like Zomato provides a secular part where user can rate their experience of the visited restaurant or café. Zomato also provides columns for writing classified user reviews.

1.2 Objective Of Research:

The main objective behind this research is to improve in certain areas where we are lagging.By analyzing the reviews and ratings we can understand the problems of the customers. Customer votes are predicted with the help of these reviews and ratings, based on these votes we will identify the customer problems.

1.3 Problem Statement:

Here the problem statement is predicting the votes based on the customers reviews and rating.

1.4 Industry Profile:

Founded in 2008, Zomato is a leading platform for restaurant search & discovery, online food ordering, and restaurant table reservations. The company was founded by Deepinder Goyal and Pankaj Chaddah and is headquartered in Gurgaon (officially Gurugram). Zomato has been a pioneer in food ordering and restaurant discovery in India, which has benefitted both restaurants and customers.

Featuring a robust review system, Zomato allows foodies to find the best meals and restaurants in their neighborhood. A notable aspect about Zomato is that it is among the few companies that have gone global after starting operations in India. Zomato currently features more than 1 million restaurants globally on its platform

History:

The story of Zomato started when the founders noticed that people did not even knew the restaurants that were functional in their neighborhood. The founders thought that it would be a great idea to list all the restaurants on the web and provide their menus as well. This idea eventually led to the launch of FoodieBay in 2008. The startup initially catered to the Delhi-NCR region and after the service gained popularity, the founders decided to implement the idea across the country.

The founders decided to go for a rebranding exercise, which led to the transformation of FoodieBay into Zomato in 2010. Since then, Zomato has expanded operations to several new locations in the country. It has also launched international operations and now covers more than 10,000 locations across 24 countries globally. Millions of people across the globe use Zomato every day to find the best places to dine in their neighborhood.

Funding:

Zomato has received investments worth $443.8 million through 10 rounds of funding. Top investors include Ant Financial, Sequoia Capital, Temasek Holdings, Info Edge, and Vy Capital.

Acquisitions:

Zomato has acquired several companies over the years; with the most notable being the acquisition of US based Urbanspoon in 2015. Other acquisitions made by Zomato include Obedovat, Menu Mania, Lunchtime, MapleGraph, Sparse Labs, Gastronauci, NexTable, Cibando, Mekanist, and Runnr.

Competition:

Zomato competes with other restaurant discovery and food delivery platforms such as Swiggy, Dineout, Grubhub, Yelp, DoorDash, JustDial, etc.

About the Founders:

Zomato was founded by Deepinder Goyal and Pankaj Chaddah, both of whom are from IIT, Delhi. Deepinder Goyal currently serves as the Chief Executive Officer (CEO) at Zomato. Prior to launching Zomato, he used to work at Bain & Company as a Senior Associate Consultant. Pankaj Chaddah is the co-founder and prior to launching Zomato, he had worked at Bain & Company as a Senior Analyst and Associate Consultant.

2.Review of Literature:

Applying statistical techniques and machine learning algorithms on available data may guide zomato in identifying hidden problem in door to door delivery. Implementing data mining techniques to predict customer votes may give companies a zomato restaurents edge in improving the relationship with customers. Customers provide review and ratings based on their experience. Based on their reviews and ratings votes are predicted. Generally for prediction we use regression models. Here also we used regression models. Among those regression models we used multilinear regression model.

3.Data Collection:

It is hard to know in advance, what kind of data will be helpful in future. We considered dataset which consists of previous years data about zomato ratings. Using previous data we can easily predict present situation.

We need a training data set. It is the actual data set used to train the model for performing various actions. Zomato data set is downloaded from kaggle data repository. The dataset contains 9 attributes and 9551 instances. The attributes present in the dataset are:Average_cost_for_two,Table_booking,Online_delivery,delivering_now,Price_range,Aggregate_rating,Rating_color,Rating_text,Votes.Based on the data correlation function, It is identified that Price_range, Aggregate_rating are the independent variables and votes is the dependent variable.Based on the Price_range and Aggregate_rating we will predict the votes.

4.Methodology:

4.1 Exploratory Data Analysis:

Exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modelling or hypothesis testing task. Exploratory data analysis was promoted by John Tukey to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments. EDA is different from initial data analysis (IDA),[1] which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed. EDA encompasses IDA.

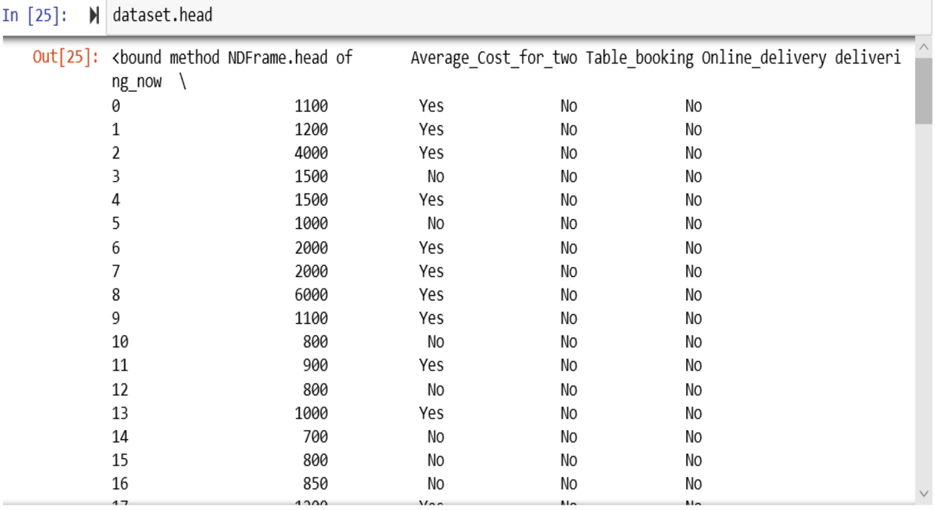
The objectives of EDA are to:

• Suggest hypotheses about the causes of observed phenomena • Assess assumptions on which statistical inference will be based

• Support the selection of appropriate statistical tools and techniques

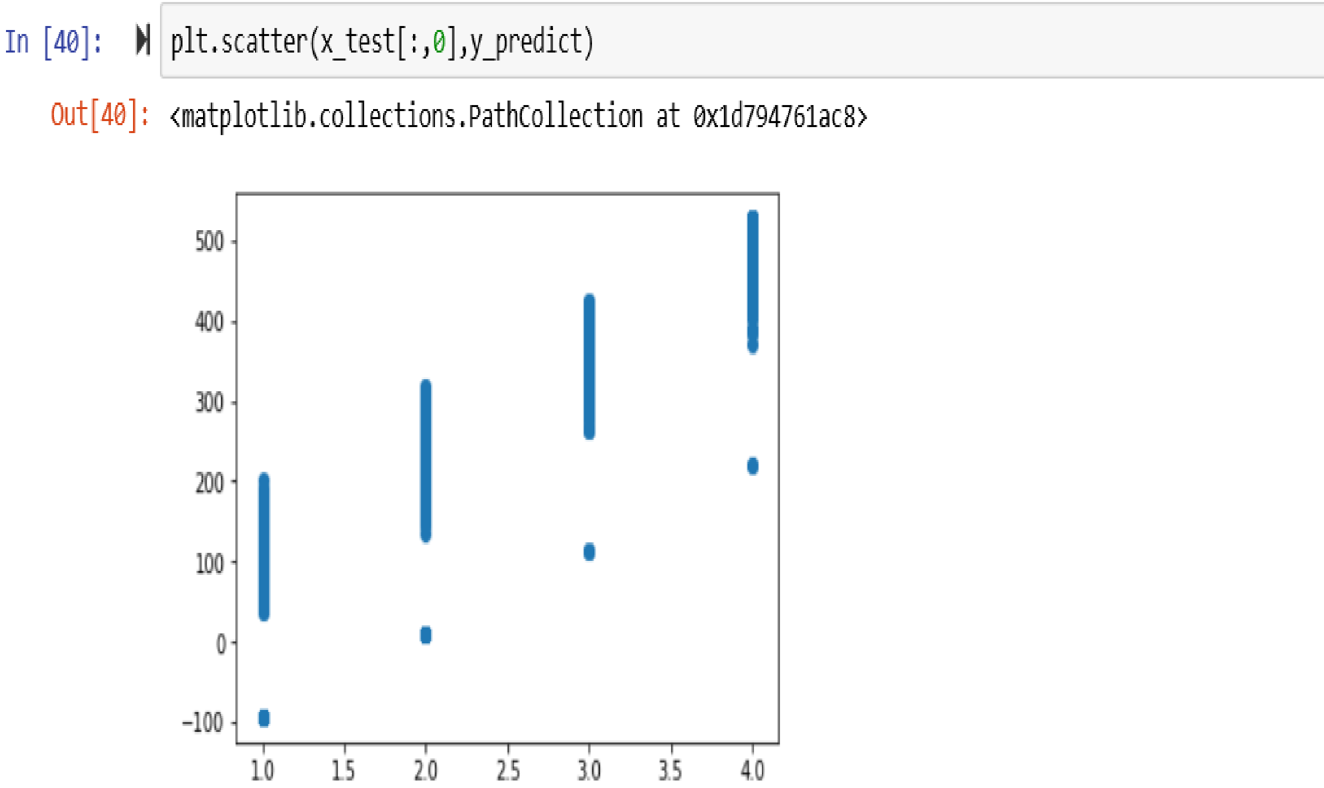
• Provide a basis for further data collection through surveys or experiments

4.1.1 Figures and table:



```
In [25]:  dataset.head

Out[25]: <bound method NDFrame.head of       Average_Cost_for_two Table_booking Online_delivery deliveri
         ng_now  \
         0                     1100           Yes              No           No
         1                     1200           Yes              No           No
         2                     4000           Yes              No           No
         3                     1500            No              No           No
         4                     1500           Yes              No           No
         5                     1000            No              No           No
         6                     2000           Yes              No           No
         7                     2000           Yes              No           No
         8                     6000           Yes              No           No
         9                     1100           Yes              No           No
         10                     800            No              No           No
         11                     900           Yes              No           No
         12                     800            No              No           No
         13                    1000           Yes              No           No
         14                     700            No              No           No
         15                     800            No              No           No
         16                     850            No              No           No
         17                    1200           Yes              No           No
```

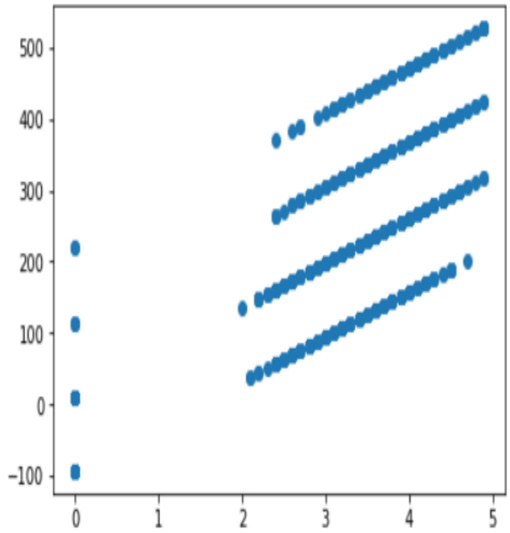


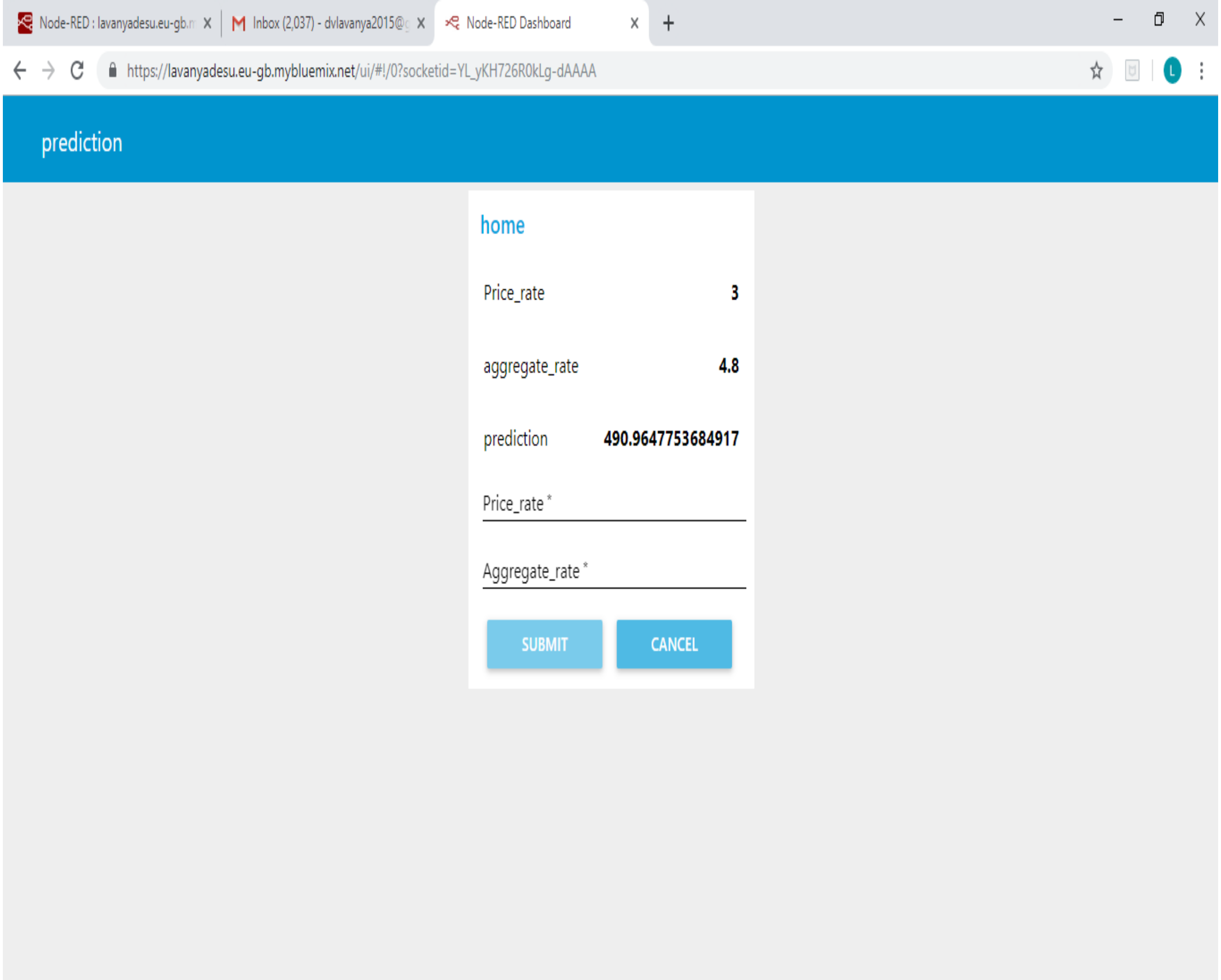
```
In [40]:  plt.scatter(x_test[:,0],y_predict)

Out[40]: <matplotlib.collections.PathCollection at 0x1d794761ac8>
```

```
In [41]: plt.scatter(x_test[:,1],y_predict)

Out[41]: <matplotlib.collections.PathCollection at 0x1d7947c12b0>
```




Node-RED : lavanyadesu.eu-gb.n
Inbox (2,037) - dvlavanya2015@
Node-RED Dashboard
https://lavanyadesu.eu-gb.mybluemix.net/ui/#!/0?socketid=YL_yKH726R0kLg-dAAAA
prediction
home

Price_rate 3

aggregate_rate 4.8

prediction 490.9647753684917

Price_rate *

Aggregate_rate *

4.2 Statistical techniques and visualization

Numpy:

NumPy stands for ‘Numerical Python’ or ‘Numeric Python’. It is an open source module of Python which provides fast mathematical computation on arrays and matrices. Since, arrays and matrices are an essential part of the Machine Learning ecosystem, NumPy along with Machine Learning modules like Scikit-learn, Pandas, Matplotlib, TensorFlow, etc. complete the Python Machine Learning Ecosystem.

NumPy provides the essential multi-dimensional array-oriented computing functionalities designed for high-level mathematical functions and scientific computation. Numpy can be imported into the notebook using

NumPy’s main object is the homogeneous multidimensional array. It is a table with same type elements, i.e, integers or string or characters (homogeneous), usually integers. In NumPy, dimensions are called axes. The number of axes is called the rank.

There are several ways to create an array in NumPy like np.array, np.zeros, no.ones, etc. Each of them provides some flexibility.

Some of the important attributes of a NumPy object are:

- Ndim: displays the dimension of the array
- Shape: returns a tuple of integers indicating the size of the array
- Size: returns the total number of elements in the NumPy array
- Dtype: returns the type of elements in the array, i.e., int64, character
- Itemsize: returns the size in bytes of each item
- Reshape: Reshapes the NumPy array

NumPy array elements can be accessed using indexing. Below are some of the useful examples:

- A[2:5] will print items 2 to 4. Index in NumPy arrays starts from 0

- A[2::2] will print items 2 to end skipping 2 items
- A[::-1] will print the array in the reverse order
- A[1:] will print from row 1 to end

Pandas:

Similar to NumPy, Pandas is one of the most widely used python libraries in data science. It provides high-performance, easy to use structures and data analysis tools. Unlike NumPy library which provides objects for multi-dimensional arrays, Pandas provides in-memory 2d table object called Dataframe. It is like a spreadsheet with column names and row labels.

Hence, with 2d tables, pandas is capable of providing many additional functionalities like creating pivot tables, computing columns based on other columns and plotting graphs. Pandas can be imported into Python using:

```
>>> import pandas as pd
```

Some commonly used data structures in pandas are:

Series objects: 1D array, similar to a column in a spreadsheet

DataFrame objects: 2D table, similar to a spreadsheet

Panel objects: Dictionary of DataFrames, similar to sheet in MS Excel

Pandas Series object is created using pd. Series function. Each row is provided with an index and by defaults is assigned numerical values starting from 0. Like NumPy, Pandas also provide the basic mathematical functionalities like addition, subtraction and conditional operations and broadcasting.

Pandas dataframe object represents a spreadsheet with cell values, column names, and row index labels. Dataframe can be visualized as dictionaries of Series. Dataframe rows and columns are simple and intuitive to access. Pandas also provide SQL-like functionality to filter, sort rows based on conditions.

New columns and rows can be easily added to the dataframe. In addition to the basic functionalities, pandas dataframe can be sorted by a particular column.

Dataframes can also be easily exported and imported from CSV, Excel, JSON, HTML and SQL database. Some other essential methods that are present in dataframes are:

head(): returns the top 5 rows in the dataframe object

tail(): returns the bottom 5 rows in the dataframe

info(): prints the summary of the dataframe

describe(): gives a nice overview of the main aggregated values over each column.

MatpltoLib:

Matplotlib is a 2d plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments. Matplotlib can be used in Python scripts, Python and IPython shell, Jupyter Notebook, web application servers and GUI toolkits. Matplotlib. Pyplot is a collection of functions that make matplotlib work like MATLAB. Majority of plotting commands in pyplot have MATLAB analogs with similar arguments.

4.3 Data Modeling and visualization:

We consider dataset from https://github.com/xoraus/ML-zomatoPrediction-/ t/BBC.csv and modified it into 700 rows and 4 columns.

Imported libraries are numpy, pandas, matplotlib. NumPy stands for ‘Numerical Python’ or ‘Numeric Python’. It is an open source module of Python which provides fast mathematical computation on arrays and matrices. Pandas is one of the most widely used python libraries in data science. It provides high-performance, easy to use structures and data analysis tools. Matplotlib is a 2d plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments. Matplotlib can be used in Python scripts, Python and IPython shell, Jupyter Notebook, web application servers and GUI toolkits.

A library is essentially a collection of modules that can be called and used. A lot of the things in the programming world do not need to be written explicitly ever time they are required. There are functions for them, which can simply be invoked. This is a list for most popular Python libraries for Data Science.

A lot of datasets come in CSV formats. We will need to locate the directory of the CSV file at first (it's more efficient to keep the dataset in the same directory as your program) and read it using a method called *read_csv* which can be found in the library.

Sometimes you may find some data are missing in the dataset. We need to be equipped to handle the problem when we come across them. Obviously you could remove the entire line of data but what if you are unknowingly removing crucial information? Of course we would not want to do that. One of the most common idea to handle the problem is to take a mean of all the values of the same column and have it to replace the missing data.

Sometimes our data is in qualitative form, that is we have texts as our data. We can find categories in text form. Now it gets complicated for machines to understand texts and process them, rather than numbers, since the models are based on mathematical equations and calculations. Therefore, we have to encode the categorical data.

Now we need to split our dataset into two sets—a Training set and a Test set. We will train our machine learning models on our training set, i.e our machine learning models will try to understand any correlations in our training set and then we will test the models on our test set to check how accurately it can predict. A general rule of the thumb is to allocate 80% of the dataset to training set and the remaining 20% to test set. For this task, we will import *test_train_split* from *model_selection* library of scikit.

The final step of data preprocessing is to apply the very important feature scaling. It is a method used to standardize the range of independent variables or features of data.

5.Findings and Suggestions:

Zomato is the largest food website in India & that I guess ensures that the readership is huge. A very common problem with food reviews on the net is that many eateries mark ‘this review is not useful’ for non-favourable reviews & ‘useful’ for favourable reviews. Zomato should be able to curb that by not allowing one email id/facebook login etc to post more than one comment.

6. Conclusion:

A look into multilinear regression which includes the investigation of the subject, information mining methods, information mining forms, information mining calculations and its usage bitterly to make it more intelligent to the clients. Presenting the crude information subsequent to preparing and executing the information mining procedures in intuitive way to the clients for better understanding. From that crude information using multilinear regression we predicted the votes. Based on those votes they analyze in which area they have to improve.

: