

BSE AIML Internship

Students name and department

1.Shreya Kumari (computer science and engineering)

2.Vaishnavi Kharbade (computer science andengineering)

3.Rakshanda Mankar (computer science andengineering)

Name of the dataset: Zomato Bangalore Restaurants

**1.Introduction of the technology used (Python)**

The technology used for dataset is Python.Machine learning is nothing but using data to make a machine to make intelligent decisions. It is based on recognizing and learning through patterns in data. Intelligent algorithms are then built by extracting, processing, defining, cleaning, arranging and then understanding the data. To perform such tedious tasks, why should one use Python?

### Simple and consistent

### Libraries and Frameworks

### Platform Independence

### Great Community Base

### Python can do the same tasks as an R programming language does. The major features of python are data wrangling, engineering, web scraping and so on. Python is also having the tools that help in implementing machine learning at a large scale. Python is one of the simplest languages to maintain and it is more robust than R. Now a days Python is having the cutting edge API. This API is quite helpful in machine learning and AI.

### 2.Problem statement formulation

### Our dataset consist of “Zomato Bangalore Restaurant” which is completely based on various types of Restaurant. The problem statements are:

* By Rating we can analyze that restaurant, cafes, cuisines and types of Restaurants will shut down in future or not.
* According to Approximate Cost we will analyze will the cost increase in future or will decrease in future.
* Predicting the restaurants on the basis of the cuisine.
* Analysis of Restaurants based on their Online Delivery.
* Analysis of Restaurants based on their online order and how rating is related to it.

**3.Introduction of the Data Set selected (Different traits of the dataset)**

**Data preprocessing** is a **data** mining technique that involves transforming raw **data** into an understandable format. Real-world **data** is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. **Data preprocessing** is a proven method of resolving such issues.

The Dataset we selected is “Zomato Bangalore Restaurant” which have 17 columns namely (url, address, name, online\_order, book\_table, rate, votes, phone, location, rest\_type, dish\_liked, cuisines, approx.\_cost.for.two.people, reviews\_list, menu\_item, listed\_in.type, listed\_in.city) and total dataset completely comprises of 56251.

The dataset consist of URL column which provides us with the link of the restaurant from where people can book or order the dishes, address which provides the exact location where the café or the restaurant is situated, Name of the restaurant or café, online order which let the people know whether online order is available or not, book a table column also let us know if we can book a table or not, rate column helps us to know how good the restaurant or café is if the rating is low it means the place is not that good if the rating is high then the place is more recommended .The rate is given out of 5, the votes column provides us with the number of votes for that particular restaurant. More the number of votes more is the place good and recommended to other people. Location column is somewhat similar to address just that in location we only get to know about the area and no the complete address. The rest type column provides with the information about the type of restaurant it can be casual dining ,café, quick bites ,etc. The dish liked column letus know about the most loved dish of that particular place. The dish which is loved the most by the customers. Cuisines column let us know about which type of cuisine is available at that place, Cuisine can be of any type North Indian, South India, Chinese, Continental, Salsa, Maharashtrian, etc.The approximate cost for two people tells us about the approximate cost which will be required for two people at that particular place.

Reviews list gives us the reviews of all the people gave about that restaurant on Zomato. Reviews are basically the comments or the thoughts about that place. Menu items consist of the dishes in the menu. In the data set the menu items are not provided for all restaurants.They are provided for some of them. The listed in type column let us know about the type of facility available at restaurant like café,buffet,delivery,dine-out. The listed in city let the customer know about the city in which the café or the restaurant is situated.

**4.List of the operations like Data Preprocessing & ML Algorithm used for prediction**

We have done Data Preprocessing on inappropriate rows which have garbage valuesand null values included in dataset. There are so many garbage values in dataset which have no appropriate data so we have omit that garbage values from dataset and some rows have null values which are yet to describe in dataset so we have omit that particular data value from dataset so that after performing any predication the accuracy of dataset should be accurate. We deleted the columns URL, address and phone no. column. As we cannot predict or carry out the machine learning algorithms on the basis of these columns .Also replacing restaurants with their ratings given as New to NAN and dropping them finally.

The menu item column consist of the garbage values ([]) which be replaced with the NA values. We will be using the unsupervised data and will convert the data in supervised form to increase more accuracy.Machine learning algorithms which we will be using Decision Tree classification and Random Forest Algorithm.A random forest is a meta-estimator (i.e. it combines the result of multiple predictions) which **aggregates many decision trees.**

**1.Decision Tree Classification**

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy), each representing values for the attribute tested. Leaf node (e.g., Hours Played) represents a decision on the numerical target. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

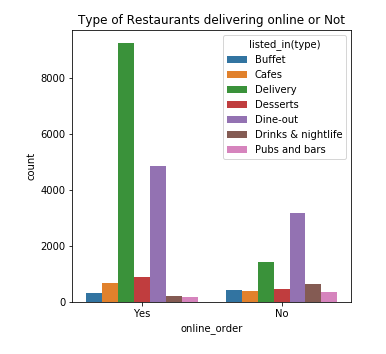
**2.Random Forest Regression**

Random forests or random decision forests are an[ensemblelearning](https://en.wikipedia.org/wiki/Ensemble_learning) method for [classification](https://en.wikipedia.org/wiki/Statistical_classification),regressionand other tasks that operate by constructing a multitude of [decisiontrees](https://en.wikipedia.org/wiki/Decision_tree_learning) at training time and outputting the class that is the [mode](https://en.wikipedia.org/wiki/Mode_(statistics)) of the classes (classification) or mean prediction (regression) of the individual trees.

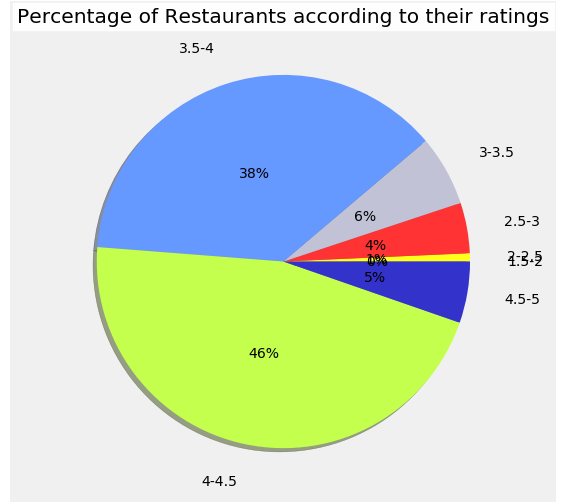
**4.Final outcome (Predictions with Accuracy percentage)**

****

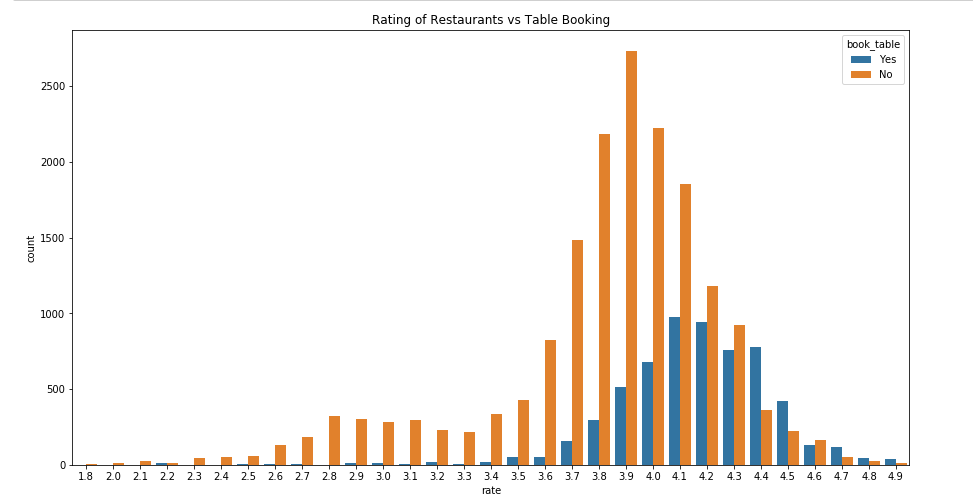
**Fig1 : Prediction with Accuracy based on rating**

****

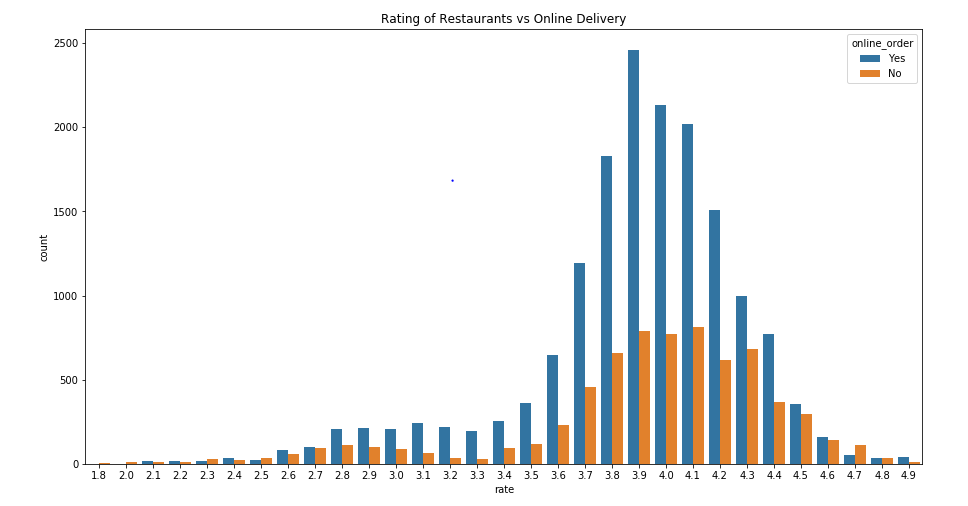
**Fig2 :Analysis of Restaurants based on their Online Delivery**



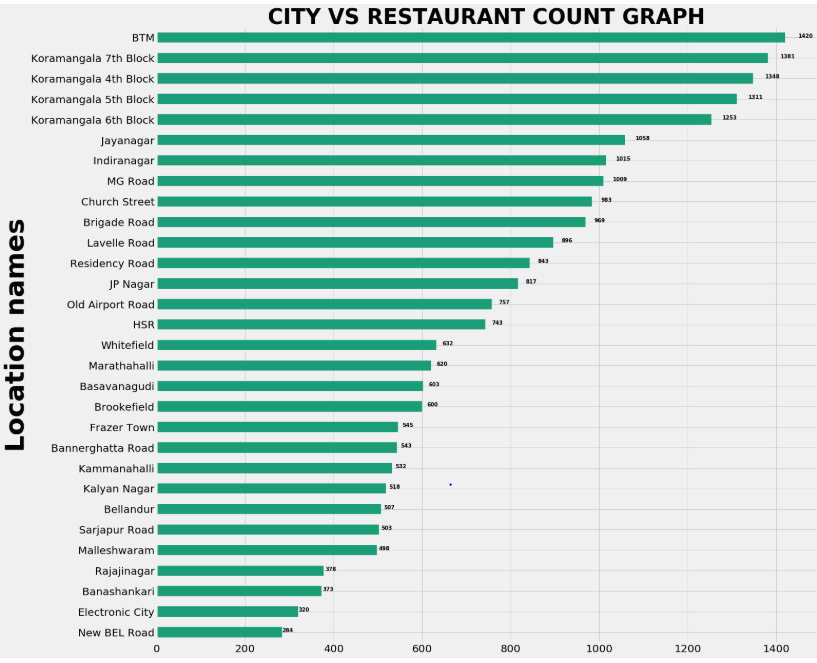
**Fig 3: Analysis based on ratings**

****

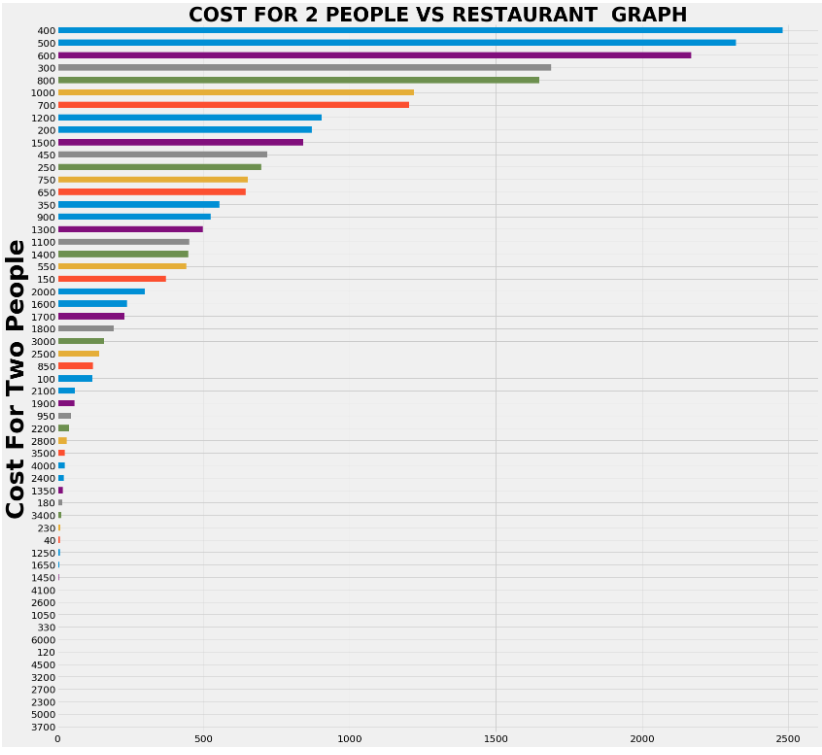
**Fig 4 : Rating of Restaurants vs Table Booking**

****

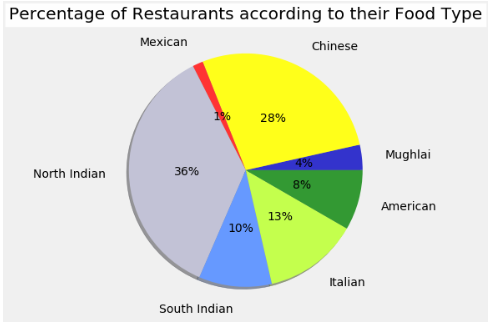
**Fig 5 : Rating of Restaurants vs Online Delivery**

****

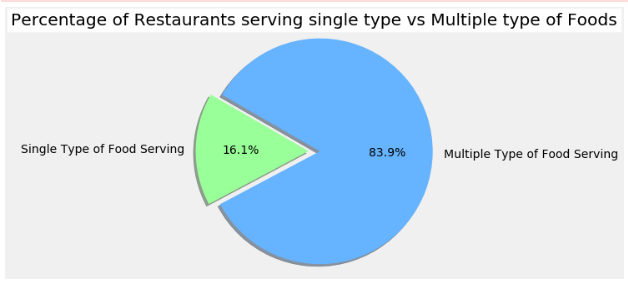
**Fig 6 : CITY VS RESTAURANT COUNT GRAPH**

****

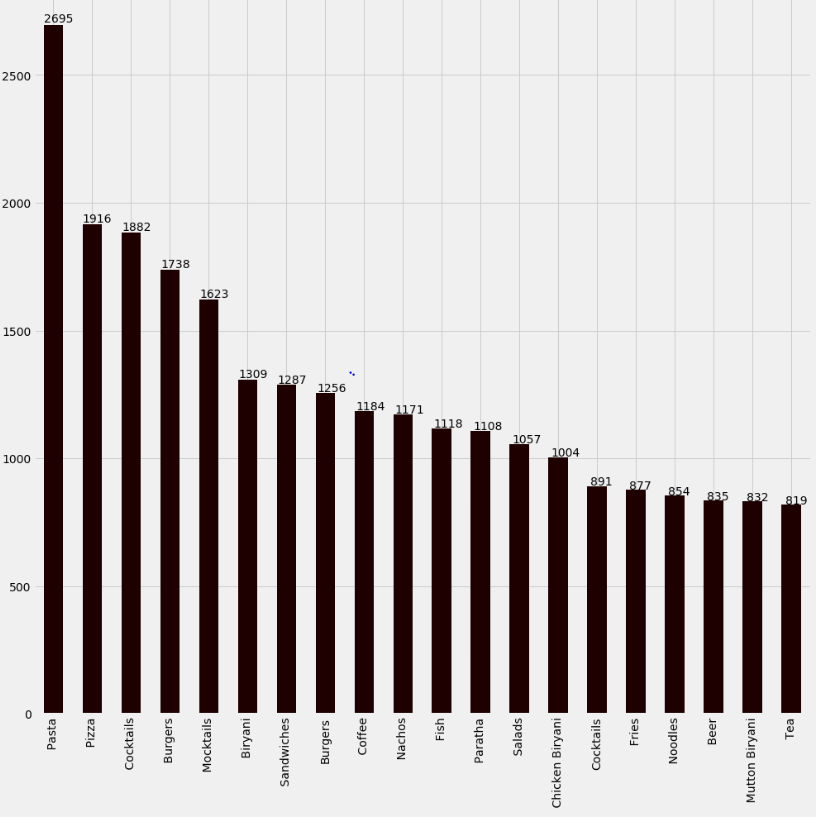
**Fig 7: COST FOR 2 PEOPLE VS RESTAURANT GRAPH**

****

**Fig 8 : Percentage of Restaurants according to their Food Type**

****

**Fig 9 : Percentage of Restaurant serving single type vs Multiple type of foods**

****

**Fig 10 : Top 20 Favorite Food counts**

**5. Conclusion of the project**

Hence we have concluded a number of features about existing restaurants of different areas in a city and analyses them to predict rating of the restaurant. This makes it an important aspect to be considered, before making a dining decision. Such analysis is essential part of planning before establishing a venture like that of a restaurant. Lot of researches have been made on factors which affect sales and market in restaurant industry. Various dine-scape factors have been analyzed to improve customer satisfaction levels. If the data for other cities is also collected, such predictions could be made for accurate.