Find The Top-Rated Restaurant using Classification Techniques

SUBMITTED BY: GURCHARAN KAUR

STUDENT ID: 0775710

Table of Contents

Abstract

For capstone project, we choose the dataset based on the restaurants in Bengaluru. This dataset contains 51717 records and 17 attributes. This dataset basically contains the information regarding the restaurants dine in, takeout and online order options, reviews, and type of restaurants like casual dining, pubs, bars and café, type of cuisine and all. Bengaluru is the best place for foodies. The number of restaurants is increasing day by day.

New eateries are opening consistently, rivalry is getting to increment. This project focuses on demography and its food culture of the area. Predominantly it will help eateries with choosing their plan, food, and costs of the particular area. With the proper analysis of the project, it will be useful for individuals for picking eateries in light of many variables. It will likewise attempt to address the inquiries in view of the cafés and interest of the foodies. All information was scratched from Zomato and having different sort of data like 6 to 7 classes of eateries Buffet, Pubs, Bars, Cafes, Deliveries, Dine out, Drinks and night life. In the Bengaluru city, there are in excess of 12000 cafés, and it is serving dishes from everywhere the world. The majority of the eateries are serving same food. Consequently, it is difficult for new eateries to compete with the well-established restaurants.

The main goal of this project to find best top-rated restaurants in Bengaluru city and does the cost of food affect the rating of restaurants. As we mentioned in the research questions, we will try to figure out which cuisines are famous, about dine-in and takeout option in Bengaluru. These days, mostly people do not have time to cook food at home, so they are preferring restaurant food. With such an overwhelming demand of restaurants, it has become important to study the demography of a location.

Keywords: Classification techniques such as logistic regression, KNN, Gaussian Naïve Bayes, Random Forest classifier and Decision Tree regression.

Research Questions:

1. How many restaurants have online order options?
2. How many restaurants have table-booking option?
3. How many types of service are present in the restaurants?
4. Find the top 5 restaurants name in Bengaluru.
5. What kind of cuisine is most popular in the locality?
6. How many numbers of restaurants in each neighborhood**?**
7. Does cost effects the ratings of the restaurants in cities of Bengaluru?

We have these questions in our mind and with the help of these questions; we will try to find out the factors that would affect opening of a new restaurant in a locality. As this dataset also contains the reviews for each of the restaurant, from which we will find out the overall rating for the area. Moreover, we can also find which cuisine is popular in the area.

Tools:

All formulation and data visualization is done by using Python Programming Language.

GitHub Source:

https://github.com/Gurcharankaur710

Introduction:

Bangalore (officially known as Bengaluru) is the capital and biggest city of the Indian territory of Karnataka. With a populace of more than 15 million, Bangalore is the third-biggest city in India and 27th biggest city in the world.

Bangalore is one of the most ethnically diverse city in the country, with more than 51% of the city's populace being travelers from different pieces of India.

Bangalore is now and again referred to as the "Silicon Valley of India"(or "IT capital of India") in light of its job as the country's driving data technology (IT) exporter.

Bangalore has an interesting food culture. Eateries from everywhere the world can be found here in Bengaluru, with different sorts of foods.

Overall, it is possible that Bangalore is the best spot for foodies.

The food business is always at a rise in Bangalore, with 12,000 or more eateries presently active in the city, the number is yet expanding.

The developing number of cafés and dishes in Bangalore draws in me to assess the information to get a few experiences, some interesting facts, and figures.

Literature Review:

In the advanced world, popularity of food application is expanding systematically a direct result of its usefulness about book and request for food in couple of snaps on telephone for their #1 spots by looking into their reviews and rating of different clients. Requests are developing. However, it has become hard to compete with the current eateries, in the in the midst of developing interest. Everybody is serving a similar food.

Zomato Bangalore is such sort of dataset, which contains detail data about the eateries in each corner in Bangalore. We start with cleaning our dataset to clear the null values. Prior to going to café, the significant thing, which everybody does, is to actually look at review and rating of the eatery. Subsequently, we have numerous perspectives in our dataset that are dependent to many elements to rate our dataset. Classification algorithm is the most applicable data mining strategies used to apply in analysis. This algorithm is most common in few data analysis. Out of them many gives better classification accuracy.

We analyze various components of the dataset with our target attribute and concoct different visual guide to find which all components are highly co-related with our target variable. Then in light of those co-relation components, we can construct predictive models to predict the rate of the particular restaurant in view of the given arrangement of components. It is a real time dataset, so we can begin from Data Exploratory process like dealing with Nan values, Null values, and erase duplicacy. Our target variable is "Rates" attribute.

We examine the relationship of various features in the dataset with respect to Rates. We will visualize the relation of any remaining dependant features with respect to the target variable, and along this, the most related components that influences the target variable. From that point onward, we will execute different modeling structures like linear regression, data visualization on our dataset. These modeling will provide us with the accuracy of our prediction and afterward we will get to be aware of which model give the most optimised and right readings.

The survey exhibited that the web based food movement strategy is extraordinarily demandable, potential and money capable. This space is rapidly growing an immediate aftereffect of the size of market. Every human requirements to eat on various times and assortment in a day .So it ensures rehash all together and creating business. In view of repeat clients, Profit edges are high.

Mentioning on the web is these days is plan or a way of life. Mentioning on the web is much pleasant and more reasonable than eat out.

As per review, Zomato and Swiggy both have controlled over 68% part of the general business of online food transport. They have gained the present circumstance by applying different inventive methods, which attract the larger part more. Then again, Food panda and Ubereats is taking benefit from huge present base of their parent applications - Ola and Uber taxis. These associations made separate applications related with food industry for driving up the arrangements.

Zomato has produced for food conveyance similarly with respect to cafe revelation also. Finding cafes, recognizing notable dishes, glancing through courses of action and mentioning food in all cases application made it in lead.

Taha Yasin Demir performed numerous estimations on this equivalent dataset as he performed Exploratory Data Analysis, Correlation Analysis, Hyper parameter optimization, Restaurant clustering with Pca and K-Means, PCA (Principal Component Analysis) - Dimension Reduction.

Purnasai Gudikandula additionally performed Feature engineering, utilized response coded feature, but random forest regressor is dominating the race. Then different calculations like NLP features, NN models, linear regression, Ridge Regression, Lasso Regression.

Serhat Murat Alagoz and Haluk Hekinoglu (2012) believed that internet business is developing so quick around the world, the food business is additionally demonstrating and increment development. Both proposed the TAM-Technology Acceptance Model to concentrate on the internet based food-requesting applications. Their investigation expressed that mentality of individuals toward online food requesting is a result of simplicity and handiness of online food

handling process and above all their confidence in web based business sites and not many outside impacts.

H.S. Sethu and Bhavya Saini (2016), their thought was to investigate the client's insight, conduct and fulfillment of online food requesting and conveyance applications. It demonstrates that internet based food requesting applications save their time because of simple accessibility. They likewise tracked down that visibility of their number one food whenever and consistently admittance to web, and free information are the principle explanations behind utilizing the applications.

As demonstrated by Varsha Chavan, et al, (2015), the usage of innovative cell portable point of collaboration so that buyers could see demand and follow has helped the cafes in conveying orders from purchasers immediately. The extension in employments of innovative cell phones and PCs are giving stage for organization industry. Their Analysis gathered that this cycle is useful, suitable and easy to use, which is depended upon to better systematically in coming times.

According to Leong Wai Hong (2016), the inventive progress in various endeavors has changed the strategy to create. Effective frameworks can help with working at the value and usefulness of an eatery. The use of online food movement system is acknowledged that it can lead the cafes business grow at times and will help the restaurants with working with critical business on the web.

As per Attreysam examination on Bangalore cafes and scenes across different regions, which will be important for foodies who are living or recently moving to Bangalore and will really need to finish up the locale they can research premise their food tendencies. The pieces of information got from this examination will moreover be relevant to existing cafes owners and to people expecting to open another eatery.

Then, He made groups of area that have comparable kind of venues. Created map for all areas and made clusters utilizing K-means clustering to track down comparable areas and produce bits of knowledge.

According to Amit Verma to predict the rating for new eateries in Bangalore by assessing following elements: Analyzing demography of the area, Effect of rating on the kind of the eatery, assisting new cafés with choosing their subject, menus, food, cost and so forth, Additional offices given by the cafés like web-based conveyance and table reservation.

He learned about the client reviews using Natural toolkit libraries to find about the client different preferences. He performed linear regression, Random Forest, Decision Tree.

As indicated by Payal Bhandari Linear Regression is a direct way to deal with demonstrating the association between a scalar response (and ward variable) and something like one explanatory variables (or independent variables).

• Decision Tree Regression - regression decision tree creates relapse or arrange models as a tree structure. It isolates a dataset into progressively little subsets while at the same time a connected decision tree is continuously developed.

• Random Forest Generator or random decision tree are learning strategy for classification, regression and different undertakings working by building decision tree and yielding the class that is the classification or regression.

She looked for - What kind of a food is a better known in an area, Do the entire region loves veggie sweetheart food. If without a doubt, is that region populated by a particular group of people for e.g., Jain, Marwari's, Guajarati's who are by and large veggie lover. This kind of assessment ought to be conceivable using the data, by focusing on the components.

For example:

• Area of the café.

• Approx. Cost of food Theme based eatery or not, which region of that city serves that cooking styles with generally outrageous number of cafes.

• The necessities of people who are attempting to get the best food of the area.

• Is a particular region notable for its own kind of food?

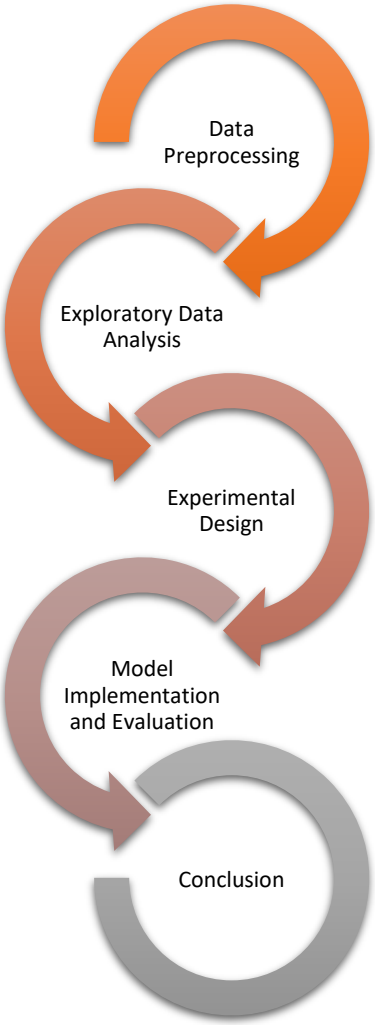
As indicated by Chirag Samal-Sentiment Analysis of Reviews of the dataset to recognize the sensations of the clients towards Restaurants. Sentiment Analysis is the computational task of thus sorting out what feelings a writer is imparting in message. Opinion is routinely illustrated as a matched capability (good versus skeptical), yet it can in like manner be an all the more fine-grained, for example, recognizing the inclination a maker is conveying (like dread, satisfaction or outrage).

Data is being used to make structures that are more successful, and here Recommendation Systems become a vital variable. Suggestion Systems are a kind of information filtering structures as they work on the idea of recorded records and gives things that are more appropriate to the chase thing or are connected with the pursuit history of the client. They are dynamic information isolating systems, which modify the information coming to a client considering his tendencies, significance of the information, etc. Recommender systems are used comprehensively for proposing films, articles, restaurants, spots to visit, things to buy, etc. He utilized Content Based Filtering. This methodology uses only information about the portrayal and attributes of the things clients has as of late consumed to show clients' inclinations. Accordingly, these computations endeavor to recommend things that resemble those that a client appreciated beforehand (or is investigating in

the present). In particular, unique candidate things are differentiated, and things as of late evaluated by the client and the best-it are proposed to match things. He utilized Exploratory Data Analysis-Visualization, Rate forecast, Sentiment Analysis of reviews, Recommendation System.

Above all are the information about the people who performed different calculations, algorithms according to their convenient way.

Methodology:



Data Details:

This dataset contains 51717 records and 17 attributes from which we have 16 object attributes and 1 numeric attribute. In this dataset, we have all attributes of object datatype; we should assign the fitting datatype to the attributes. After assigning appropriate data type, now we are having 2 objects, 5 categorical attributes, and 5 numeric datatypes

Further, we dropped 5 attributes, since that attributes was not giving useful information. The dropped attributes are URL, Phone, Address, Dish liked and Menu item.

Table 1: Information of Object data type.

Attributes	Description	Total Values
Name	It addresses the name of the restaurant in Bengaluru.	8792
Reviews	It has the list of tuples, which are containing the surveys for the restaurants, and each tuple conveys two values, rating and review by the client who visited eatery.	22513

Table 2: Information about Categorical attribute.

Attributes	Description	No. of Levels	Counts
Restaurant Type	It contains the data on type of eatery.	93	Quick bites: 19328 Casual dining: 10316
Location	It gives information of the location of the restaurants. We have 3 best locations.	93	BTM: 5130 HSR: 2522 Koramangala 5th Block: 2503

Cuisines	It depicts the type of cuisine in the restaurants.	2723	North Indian: 2952 Chinese: 2381 South Indian: 1826
Service Type	It shows the type of service in the restaurant.	7	Delivery: 25888 Dine-Out: 17763
City	It contains the information of the area restaurant is located. We have top 3 cities.	30	BTM: 3268 Koramangala 7th Block: 2935 Koramangala 5th Block: 2834

Table 3: Detail about Numerical attributes.

Attribute	Description	mean	std	min	25%	50%	75%	max
Online order	It means that restaurant is allowing online order or not.	0.58	0.49	0	0	1	1	1
Book Table	It contains that table booking options available or not.	0.12	0.33	0	0	0	0	1
Votes	It shows the votes of the restaurant.	283.96	804.31	0	7	41	198	16832
Rate	It is having the rating between 0 to 5.	3.70	0.39	1.8	3.5	3.7	3.9	4.9
Cost	It is having the information of cost of food for 2 person.	555.55	437.49	40	300	400	650	6000

Data Preprocessing

Detailed Data Dictionary:

In this, we dealt with each attribute in our dataset. To begin with, we converted the attribute to suitable data type. From that point forward, we really look at five number summary. Then, we checked the levels of the all-categorical attributes. We additionally look at the missing values in our dataset. Further, I fill the missing values, as I referenced previously. Then I performed EDA, One-hot encoding, Splitting train test, Sampling strategy, Classification Models.

Missing value:

Attribute	No. Of Missing Values
Rate	7757
Location	21
Restaurant Type	227
Cuisines	45
Cost	345

Criteria for cleaning the missing values:

As in the above table total 5 attributes (rate, location, restaurant type, cuisines, cost) are having missing data which we have to remove, because of this our dataset is imbalanced. We replaced the Na values of these attributes with the help of mean and mode.

Rate, Cost: We replaced missing value with mean for these attributes.

Location, Restaurant type, Cuisines: In these attributes, we fill the missing value with mode.

Now, our data is cleaned and there is no missing or null values in the dataset.

Exploratory Data Analysis:

Exploratory Data Analysis assists with giving knowledge of a dataset to understand the structure. It extract the significant parameters and connections between various factors.

1. Graphical representation of online order option or not.

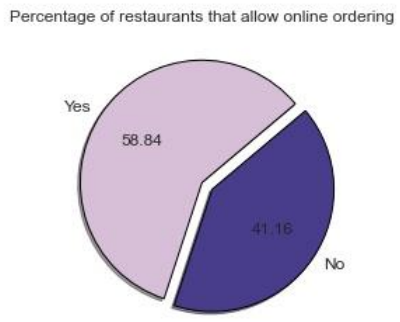


Figure 1: Pie chart of Online Order

Explanation: This pie charts shows the information about the restaurants having option for online order or not. The greater part (58.84%) of the restaurants in the Bangalore city having choice for online order. However, 41.16% restaurants are not giving web-based request choice.

2. Graphical representation of Table booking option or not.

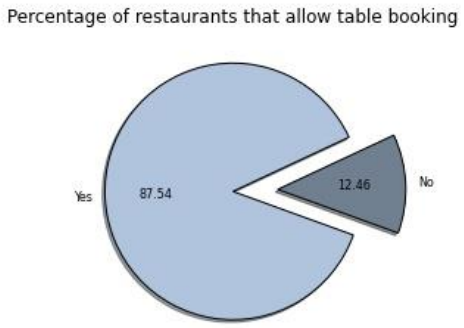


Figure 2: Pie chart of Table Booking option

Explanation: The pie chart demonstrate that how much restaurants are having option of pre booking of the table or not. The greater part (87.54%) of the eateries giving table booking choice and just 12.46% of the eateries are not having table-booking choice.

3. Type of services in the restaurants.

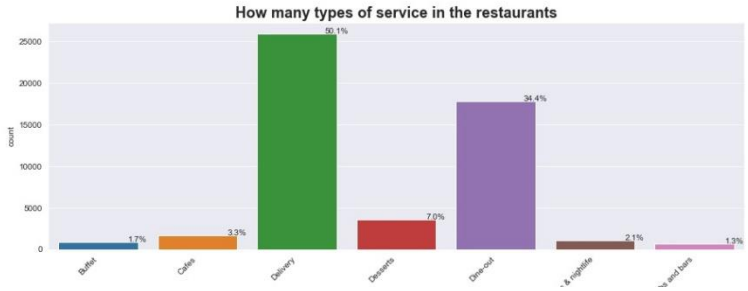


Figure 3: Bar chart of Type of Services

Explanation: There are all out 7 kind of services in the eateries. Top-notch services is the delivery option. From these services, clients like to Dine-out and do delivery service, as contrast with different sorts like buffet, cafes and bars and so on.

4. Graphical representation of Top 5 restaurants types in Bengaluru.

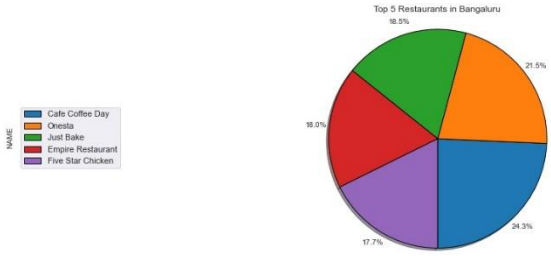


Figure 4: Pie chart of Top 5 restaurants

Explanation: This pie graph gives the data about the best five restaurants in Bangalore. Top first eatery is Café Coffee Day (24.3%); second one Onesta (21.5%), third one is Just Bake (18.5%)

5. Graphical representation of Top 10 cuisines.

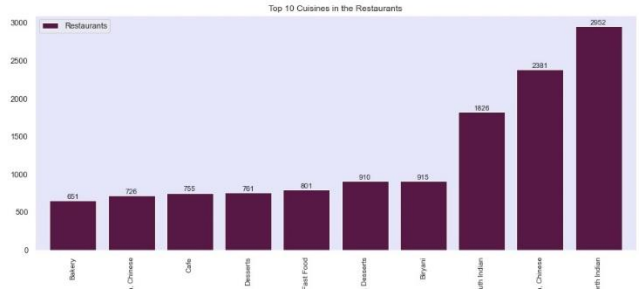


Figure 5: Bar chart of Top 10 Cuisines

Explanation: The bar graph shows the main 10 cuisine which are mostly preferred by customer. The absolute first food, which is famous, is North Indian cuisine. Second one is Chinese. Notwithstanding, among top 10 Bakery is toward the end.

6. Number of restaurants in each neighbourhood.

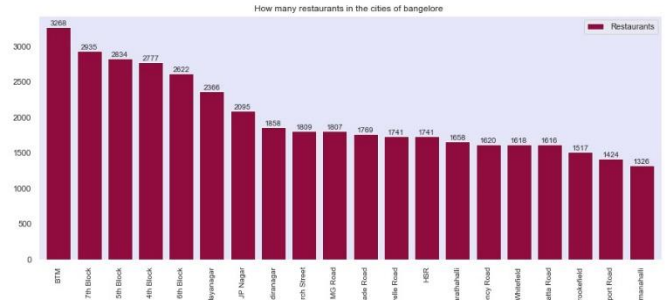


Figure 6: Bar chart of restaurants in each neighbourhood

Explanation: This bar diagram shows the count of eateries in every neighborhood of the Bangalore city. BTM (3268) is having biggest number of restaurants. In contrast, Old airport road and Kammanahalli areas are having most minimal number of restaurants.

7. Restaurant's rating distribution.

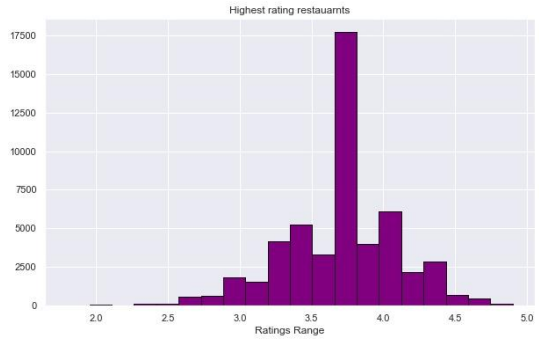


Figure 7: Bar chart of Rating distribution

Explanation: The bar outline exhibits the data about the most noteworthy rating distribution. In the entire Bangalore city, 17500 eateries are having most noteworthy rating, which is vary between 3.5 to 4.0.

8. Cost Effecting rating of the restaurants or not.

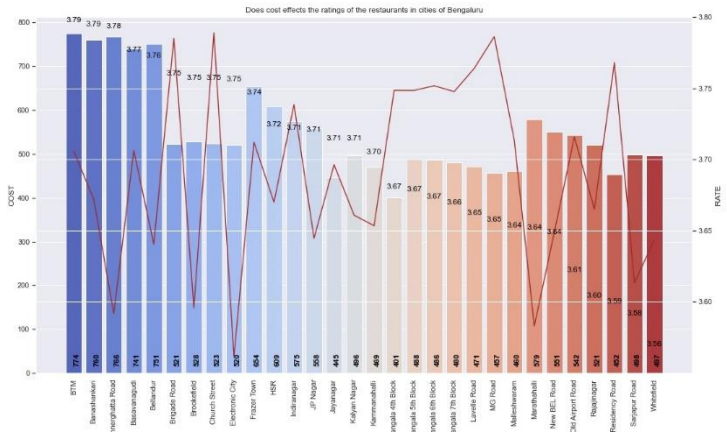


Figure 8: Cost Effect rating or not

Explanation: Above are the best 30 urban communities wherein clients likes to eat in and takeout. In the referenced urban communities, the expense of 2 man lies in the middle 367 to 448 and rating is change between 3.64 to 3.52. By examining the above diagram, we can say that cost doesn't impact the rating of the eateries in Bengaluru.

9. Graphical representation of numerical attributes.

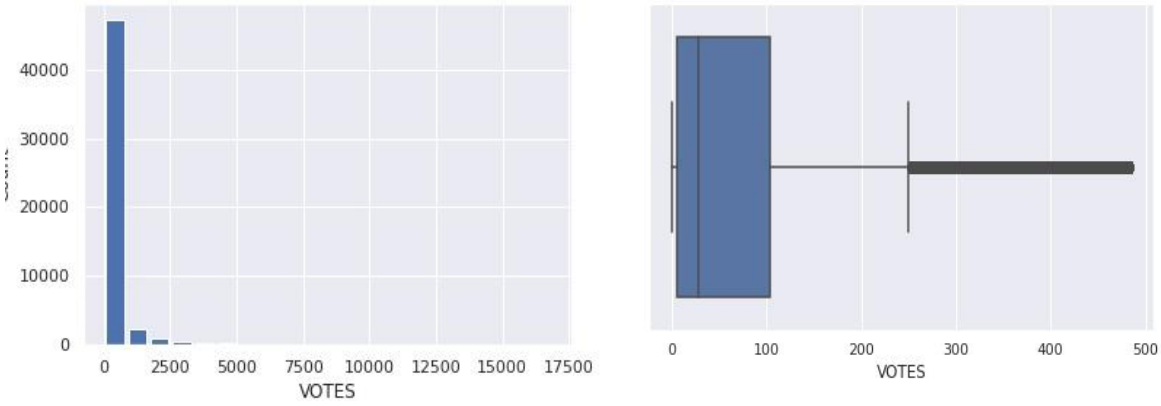


Figure 9: Histogram and Boxplot of Votes attribute

Explanation: Above is histogram of Votes attribute, which shows that the greater part of the information falls into right side, and that implies its right skewed histogram.

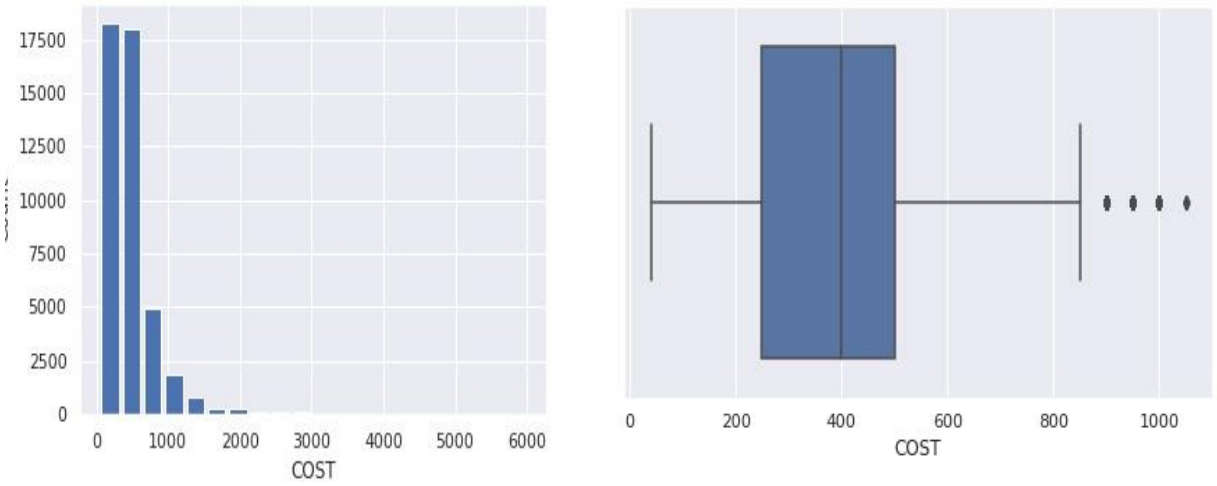


Figure 10: Histogram and Boxplot of Cost attribute

Explanation: The above histogram is of Cost attribute. It is right skewed histogram as it showing data falls into the right side (i.e. lies between 0 to 2000).

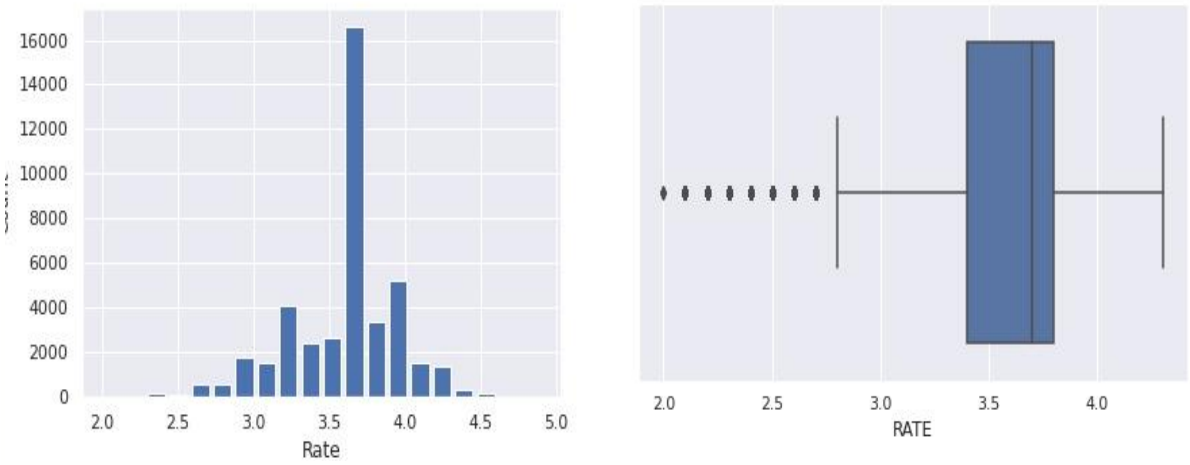


Figure 11: Histogram and Boxplot of Rate attribute

Explanation: This histogram shows the value of the Rate attribute. The value of Rate attribute vary between 2.5 to 4.5.

10. Scatter plot of showing relationship between cost and rating attributes

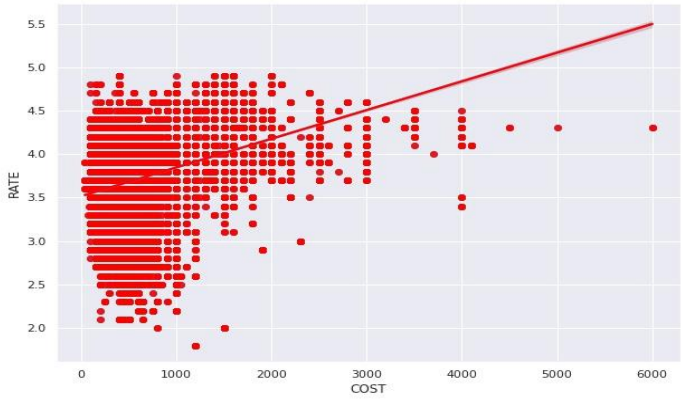


Figure 12: Scatter plot of Cost and Rate attributes

Explanation- After analyzing the scatter plot, we can express that there is solid relationship among cost and rate, which makes sense because with higher cost is ensure that both of food and dining experience is good.

11. Scatter plot-showing relation between cost and number of votes attribute.

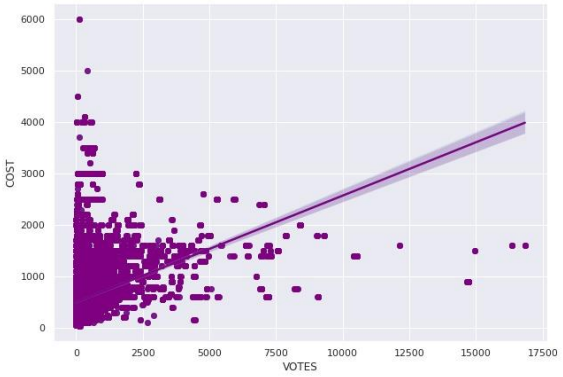


Figure 13: Scatter plot of Cost and Votes attribute

Explanation- After analysing the scatter plot, we can see that there is a high thickness of focuses that are both under cost 1000 and under 2500 votes. A justification for this is that there are more eateries that over around that cost range. There is yet a connection among cost and votes yet not exactly as solid and we might actually see that the variety is higher as found in the more extensive blue region around the line. This intends that there could be certain outliers present in the data.

12. Correlation:

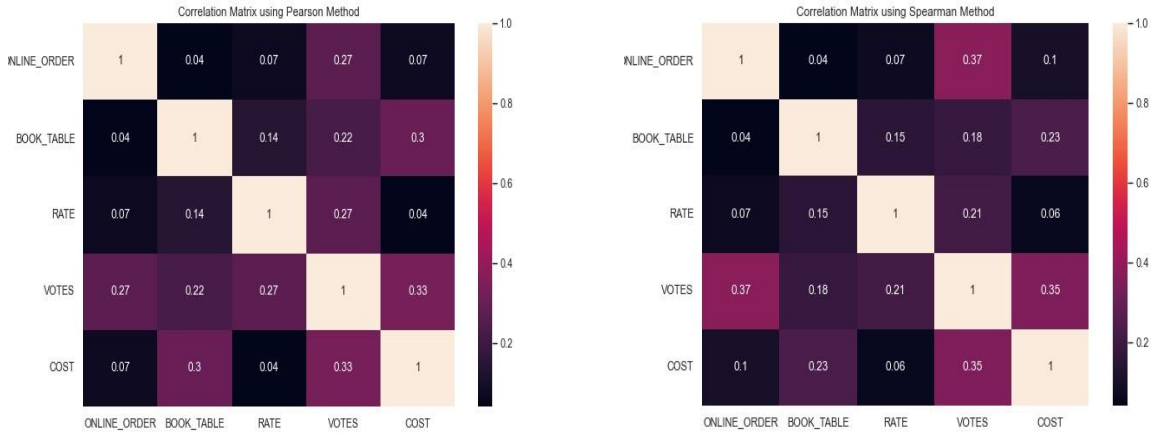


Figure 14: Correlation matrix using Pearson and Spearman method

Explanation: The above chart illustrates the connection between two factors utilizing two unique strategies for correlation matrix, which shows what the adjustment of one variable means for other variable. The worth changes between - 1 to 1. There is no solid connection has been seen between these the variables as a whole.

Variance of the Attributes:

This is the variance table of various attributes. Each attribute is having some change. In this manner, I will keep all attributes, I will not drop any of these attribute.

Attribute	Description
Online Order	0.241918
Book Table	0.028185
Rate	0.115013
Votes	10362.66
Cost	41866.97

One-Hot Encoding Operation:

One hot encoding technique represents the categorical data into binary vectors. It is a common process prior to performing classification techniques. I performed one hot encoding procedure on categorical attributes- Online request, Book table, Location, Restaurant type, Cuisines and Service type.

Min Max Scaling:

It essentially shrinks the range, now somewhere in the range between 0 and 1 (or - 1 to 1 assuming there are negative qualities). It is method to normalise the data utilizing Python's min-max functions.

Min Max Scaler does not reduce the significance of outliers. The default range for the feature returned by Min Max Scaler is 0 to 1.

Experimental Design:

Train and Test split approach: The train_test_split work is for parting a single dataset for two distinct purposes: training and testing. The testing subset is for building your model. The testing

subset is for utilizing the model on unknown information to assess the performance of the model. In this 70% of the information is used in training set and 30% by testing.

Under sampling strategy:

This system refers to gathering of procedures to adjust our dataset. It eliminates the example from the training dataset, which has a place with the majority class. I performed under sampling method on my dataset, in light of the fact that before that my information was not stable. After performing this technique, I made my information stable.

Models Implementation and Evaluation:

In the modeling part, I used five models- Logistic Regression, K Nearest Neighbors, Gaussian Naïve Bayes, Random Forest Classifier and Decision Tree Classifier. All these models I have applied on my dataset.

Logistic Regression: It is use to predict the dependent variable value based on the given independent variable. Therefore, this technique shows the linear relationship between input and output variable.

K Nearest Neighbors: It is a supervised machine-learning algorithm. This algorithm represents the k nearest neighbor. It is used for classification and regression both.

Gaussian Naive Bayes: It is a special type of NB algorithm. It is used when the features having continuous values

Random Forest: A supervised Machine Learning Algorithm is used widely in Classification and Regression problems. It makes decision trees on different samples.

Decision Tree Classifier: It is a non-parametric learning algorithm used for classification and regression. The mail goal is to create a model that predicts the value of a target variable.

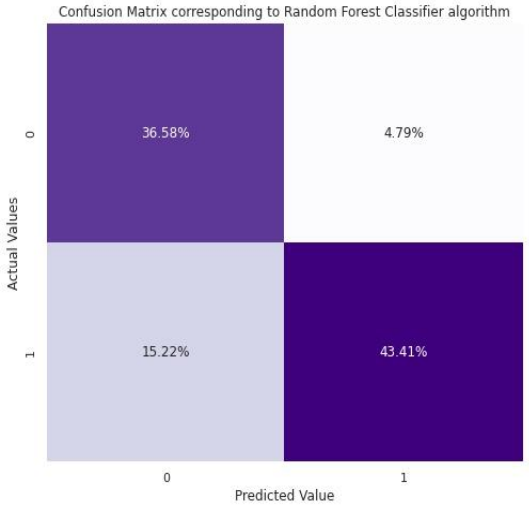
Table: Accuracy Matrix:

The table below shows accuracies of all the models. Among all models, Random Forest Classifier is the best-fit model on our dataset.

Model	Train-Test-Split
Naive Bayes	55.012322
Logistic Regression	60.847444
KNN	72.986724
Decision Tree	79.569123
Random Forest	80.411797

13. Confusion matrix and Receiver Operating Characteristic curve of the optimal model Random Forest Classifier Algorithm.

With the optimal model Random Forest Classifier, I made Confusion Matrix and ROC curve.



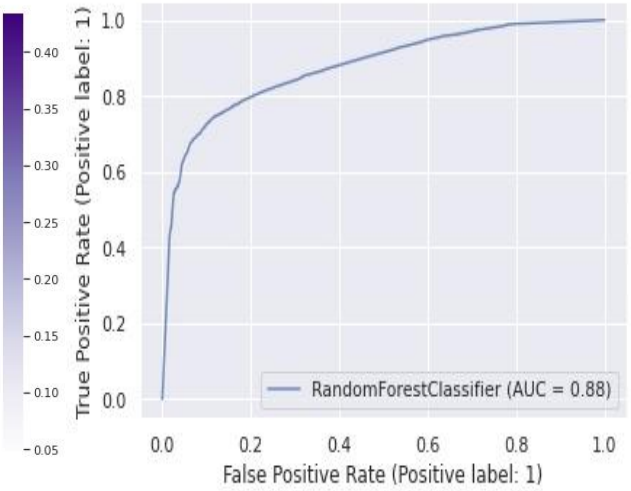


Figure 15: Confusion Matrix **Figure 16: ROC curve**

Explanation: The confusion matrix shows the accuracy of the optimal model. It shows that correspond to 0, only 4.79% values are wrong and 36.58% is right. However, corresponds to 1 is 15.22% are wrong values and 43.41% is right. So, it means almost 20% is predicted wrong and 80% is right correspondingly to 1.

Explanation: The AUC value vary between 0.5 to 1, where 0.5-0.6 auc denotes a bad classifier, 0.7 to 0.8 is considered acceptable, and 1 denotes an excellent classifier and our model gives the auc 0.8, which is much closed to 1. That means our roc curve gives the fair auc score.

Classification report:

	Precision	Recall	F1-score	Support
0	0.71	0.89	0.79	5204
1	0.91	0.74	0.82	7375

Conclusion:

We extracted data from CSV file, in this dataset many values were missing, and we did not throw up all values. Therefore, instead of removing NULL values, we tried to fill these. We have performed exploratory data analysis to answer the research questions. We used one-hot encoded features and performed different models.

Random Forest Regressor is the best-fit model, so we made correlation matrix and ROC curve with our best-fit model.

References:

https://medium.com/@attreysam/capstone-project-the-battle-of-neighbourhoods e6838bd09ddf

https://amitverma1305.wordpress.com/capstone-project/

https://www.kaggle.com/chirag9073/zomato-restaurants-analysis-and-prediction