Find The Top-Rated Restaurant using Classification Techniques

SUBMITTED TO: SAVITA SEHARAWAT

SUBMITTED BY: AMANDEEP KAUR

STUDENT ID: 0773293

Table of Contents

Abstract

For capstone project, I 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 area. With the proper analysis of the project, it will be useful for individuals for picking eateries considering 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. Most 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, I 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?

I have these questions in our mind and with the help of these questions; I 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 I will find out the overall rating for the area. Moreover, I can also find which cuisine is popular in the area.

Tools:

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

GitHub Source:

https://github.com/Amandeepkaur293

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 cities 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") considering 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, during 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 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 requirement 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 principal 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 maps 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?

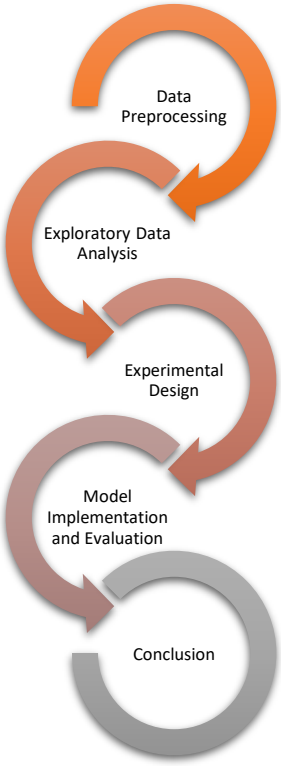
As indicated by Chirag Samal-Sentiment Analysis of Reviews of the dataset to recognize the sensations of the clients towards Restaurants. Sentiment Analysis is the computational task of thus sorting out what feelings a writer is imparting in message. Opinion is routinely illustrated as a matched capability (good versus skeptical), yet it can in like manner be a 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 relate to 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). 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:

I have contained 51717 records and 17 attributes from which I have 16 object attributes and 1 numeric attribute. In this dataset I have all the objects so, I must assign the appropriate datatype to the attributes. After assigning the appropriate data type to the attribute, I have 2 objects, 5 categorical attributes, and 5 numeric data types. Further, I dropped 5 attributes, because that attribute was not giving useful information. The dropped attributes are URL, Phone, Address, Dish liked and Menu item.

Table 1: Information of Object Attributes

Attributes	Description	Total Values
Name	It represents the name of the restaurant in Bengaluru.	8792
Reviews	It has the list of tuples, which are containing the reviews for the restaurant, and each tuple carries two values, rating and review by the customer who visited restaurant.	22513

Table 2: Information about Categorical Attributes

Attributes	Description	No. of Levels	Counts
Restaurant Type	It contains the information on type of restaurant.	93	Quick bites: 19328 Casual dining: 10316
Location	It carries information area in which a restaurant is situated. I have 3 best locations.	93	BTM: 5130 HSR: 2522 Koramangala 5th Block: 2503
Cuisines	It represents the type of cuisine in the restaurants.	2723	North Indian: 2952 Chinese: 2381 South Indian: 1826

Attributes	Description	No. of Levels	Counts
Service Type	It represents the type of meal in the restaurant.	7	Delivery: 25888 Dine-Out: 17763
City	It carries the information of the area restaurant is listed. 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 accepting online order or not.	0.58	0.49	0	0	1	1	1
Book Table	It carries the table booking options available in the restaurant or not.	0.12	0.33	0	0	0	0	1
Votes	It contains the total number of ratings for the restaurant with actual date and time.	283.96	804.31	0	7	41	198	16832
Rate	It contains the rating between 0 to 5.	3.70	0.39	1.8	3.5	3.7	3.9	4.9
Cost	It carries the information of cost of food for 2 persons.	555.55	437.49	40	300	400	650	6000

Data Preprocessing

Detailed Data Dictionary:

In this, I worked on each attribute in our dataset. First, I converted the data type to appropriate data type. After that, I checked five number summaries. Then, I checked the levels of the categorical attributes. I also checked the missing values in our dataset. Further, I fill the missing values, as I mentioned above. Then I performed EDA, One-hot encoding, Splitting train test, Sampling strategy, Classification Models.

Table 4: Missing Values

Attribute	Missing Values
Rate	7757
Location	21
Restaurant Type	227
Cuisines	45
Cost	345

Criteria for cleaning the missing values:

You can see that total 5 attributes (rate, location, restaurant type, cuisines, cost) have missing data which I have to remove because it imbalanced our dataset. I replaced the Na values for these attributes with the help of mean and mode.

Rate, Cost: In these attributes, I replaced missing value with mean.

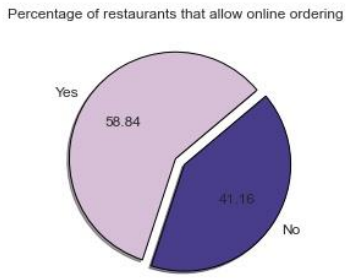
Location, Restaurant type, Cuisines: For these attributes, I fill the missing value with mode.

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

Exploratory Data Analysis:

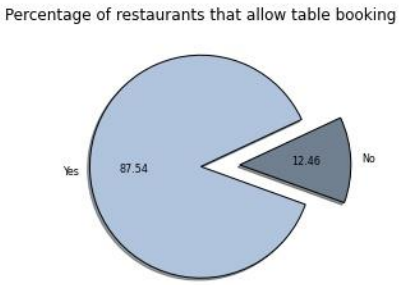
Exploratory Data Analysis helps to give insight of a dataset to understand the structure. It extracts the important parameters and relationships between different variables.

Figure 1: Graphical representation of online order option or not.



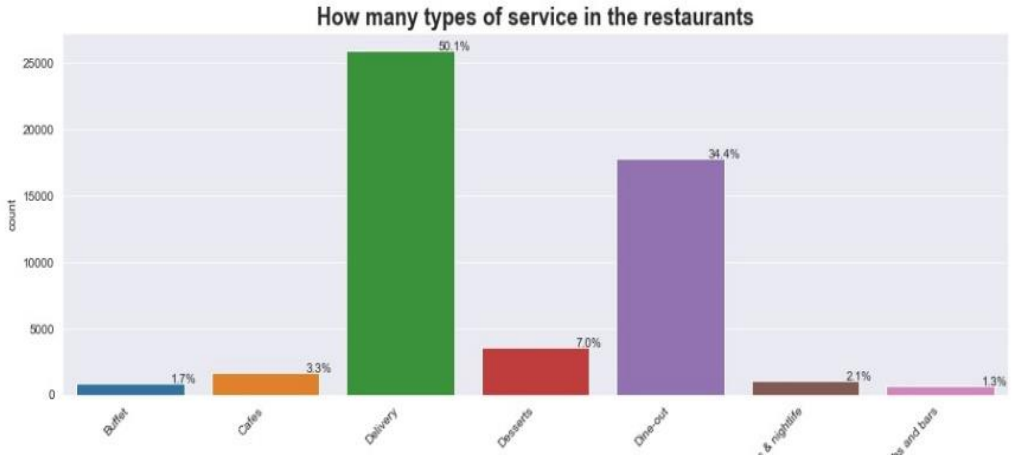
Explanation: This pie chart gives the information about the restaurants doing online order or not. More than half (58.84%) of the restaurants in the Bangalore city are providing option for online order. However, 41.16% restaurants are not having option of online order.

Figure 2: Graphical representation of online order option or not.



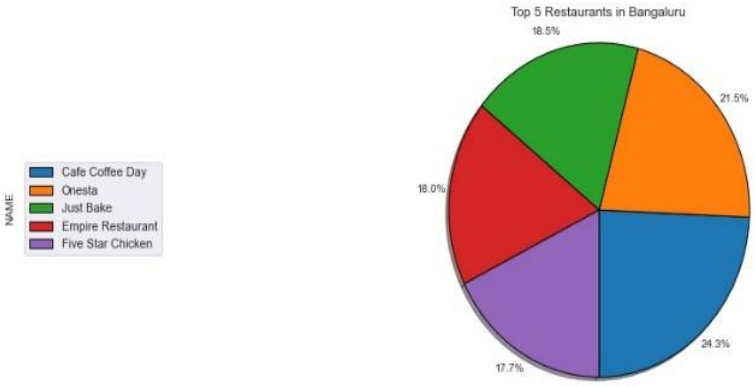
Explanation: It shows how much restaurants are allowing to pre booking of the table or not. More than half (87.54%) of the restaurants providing table booking option and only 12.46% of the restaurants are not having table-booking option.

Figure 3:Type of services in the restaurants.



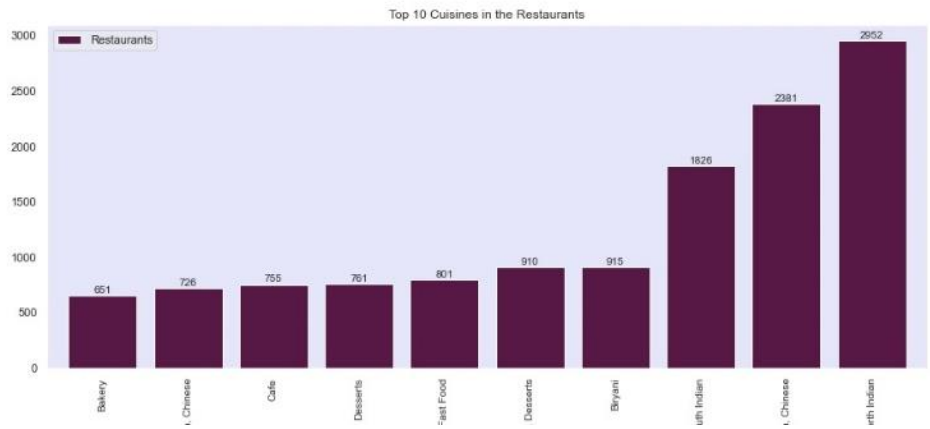
Explanation: There are total 7 type of services in the restaurants such as buffet, cafes, delivery, desserts, dine-out, clubs & nightlife, pubs, and bars. Top-notch service is the delivery option. From these services, customers like to Dine-out and delivery type of service, as compared to the other types like buffet, cafes, and bars etcetera.

Figure 4: Graphical representation of Top 5 restaurant types in Bengaluru



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

Figure 5: Graphical representation of Top 10 cuisines

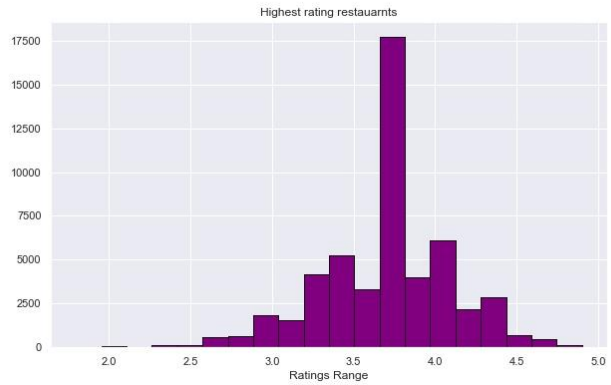


Explanation: The bar chart shows the top 10 cuisines which are mostly liked by the customer. The very first cuisine, which is popular, is the North Indian cuisine. Second one is Chinese and the third one is south Indian. However, among top 10 cuisines, Bakery is least famous cuisines in people of Bangalore.

Figure 6: Number of restaurants in each neighbourhood.

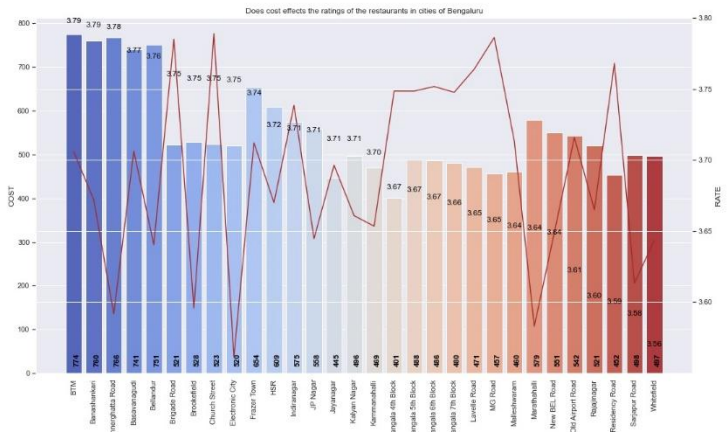
Explanation: This bar chart represents the count of restaurants in each neighbourhood of the Bangalore city. BTM (3268) is having highest number of restaurants. On the contrary, old airport road and Kammanahalli locations are having lowest number of restaurants.

Figure 7: Restaurant's rating distribution.



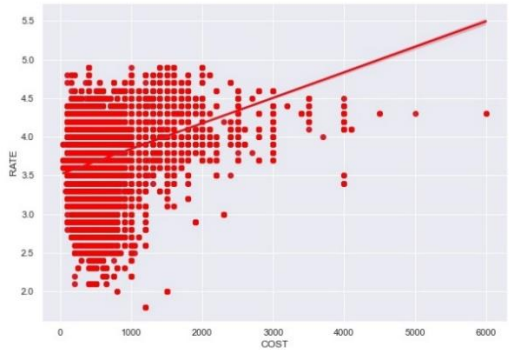
Explanation: The bar chart demonstrates the information about the highest rating distribution. In the whole Bangalore city, 17500 restaurants are having highest rating, which is vary between 3.5 to 4.0.

Figure 8: Does cost effect the rating of the restaurants or not.



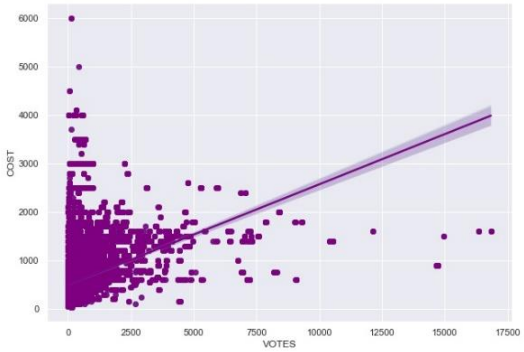
Explanation: Above are the top 30 cities in which customers likes to dine in and takeout. In the mentioned cities, the cost of 2 person lies in between 774 to 448 and rating is varied between 3.79 to 3.52. By analysing the above graph, we can say that cost does not affect the rating of the restaurants in Bengaluru.

Figure 9: Scatter plot of showing relationship between cost and rating attributes



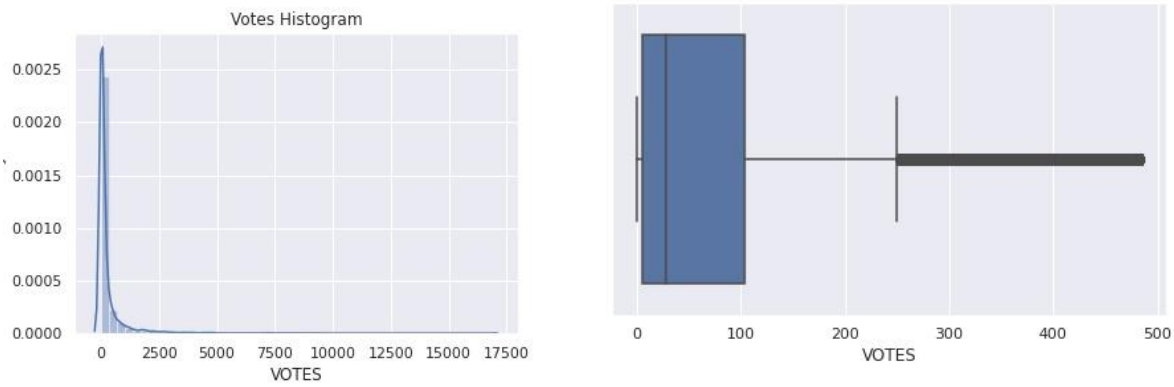
Explanation: By analysing the scatter plot, we can say that there is strong correlation between cost and rate, which makes sense because those with higher cost will ensure that both of food and dining experience is good.

Figure 10: Scatter plot-showing relation between cost and number of votes attribute



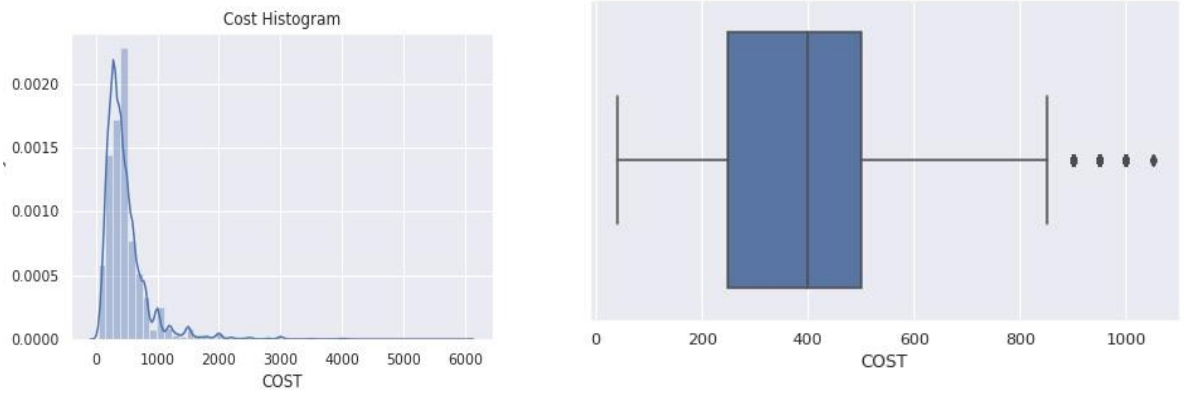
Explanation: By analysing the scatter plot, we can see that there is a high density of points that are both under cost 1000 and under 2500 votes. A reason for this is that there are more restaurants that over around that price range. There is still a correlation between cost and votes but not quite as strong and we can even see that the variation is higher as seen in the wider blue area around the line. This means that there could be certain outliers present in the data.

Figure 11: Histogram and Boxplot of Numerical attribute (Votes)



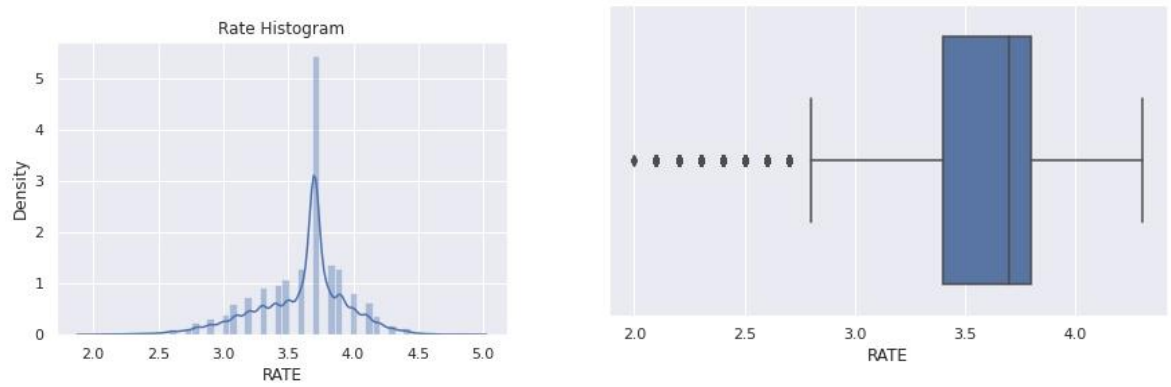
Explanation: Above is histogram of Votes attribute, which shows that most of the data falls into right side, which means its right skewed histogram. Variation lies between 0 to 2500. Basically, Box plot use the IQR method to display data and outliers(shape of the data) but in orde r to be get a list of identified outlier, we will need to use the mathematical formula and retrieve th e outlier data. That's why we use IQR score method to remove the outliers from our dataset.

Figure 12: Histogram and Boxplot of Cost attribute



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

Figure 13: Histogram and Boxplot of Rate attribute

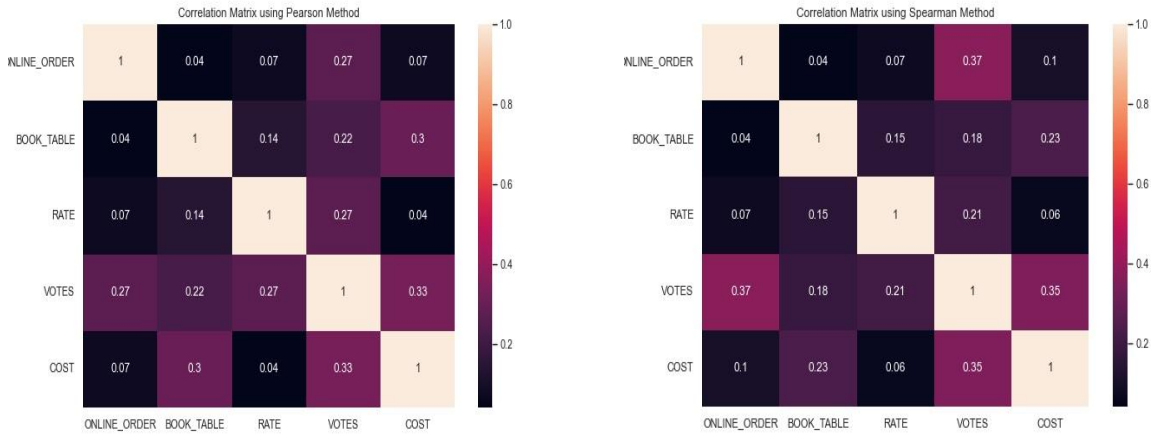


Explanation: This histogram shows the value of the Rate attribute. The value of Rate attribute lies between 2.5 to 4.5.

Correlation Matrix:

I made correlation matrix with two methods such as Pearson and spearman. These two methods are very popular. Generally, we can use Pearson's correlation when we think the variables relationship is linear and Spearman's correlation when we think the relationship is monotonic.

Figure 14: Correlation matrix using Pearson and Spearman method



Explanation: The above chart illustrates the relation between two variables using two different methods of correlation matrix, which shows how the change in one variable affects another variable. The value varies between -1 to 1. There is no strong correlation has been observed between these all the variables.

Variance of the Attributes:

Basically, Variance helps in measuring how far a number or value of a dataset is from the mean or average value. This is the variance table of different attributes.

Table 5: Variance Table

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 before performing classification techniques. I performed one hot encoding technique on our categorical attributes- Online order, Book table, Location, Restaurant type, Cuisines and Service type.

Min Max Scaling:

Min-Max scaling is a normalization technique that enables us to scale data in a dataset to a specific range using each feature's minimum and maximum value. In this we subtract the Minimum from

all values, thereby marking a scale from Min to Max. Then divide it by the difference between Min and Max. The result is that our values will go from zero to 1.

Experimental Design:

Train and Test split approach:

The train-test-split function is for splitting a single dataset for two different purposes: training and testing. The testing subset is for building your model. The testing subset is for using the model on unknown data to evaluate the performance of the model. In this 70% of the data is used by training set and 30% by testing.

Under sampling strategy:

This strategy refers to a group of techniques to balance our dataset. It removes the example from the training dataset, which belongs to the majority class. I performed the under-sampling technique on my dataset, because before that my data was not stable. After performing this technique, I made my data stable.

Model Implementation and Evaluation

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

- Logistic Regression: It perform 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 supervised learning algorithm used for classification and regression. The mail goal is to create a model that predicts the value of a target variable.

Accuracy Matrix:

This is comparison of accuracies of all the models. I use linear regression, naïve bayes, knn, decision tree and random forest classifier for our dataset. Out of all models, Random Forest Classifier is the best-fit model on our dataset which gives the 80% accuracy for our dataset.

I also created the confusion matrix according to optimal model and made the roc curve correspond to Random Forest classifier.

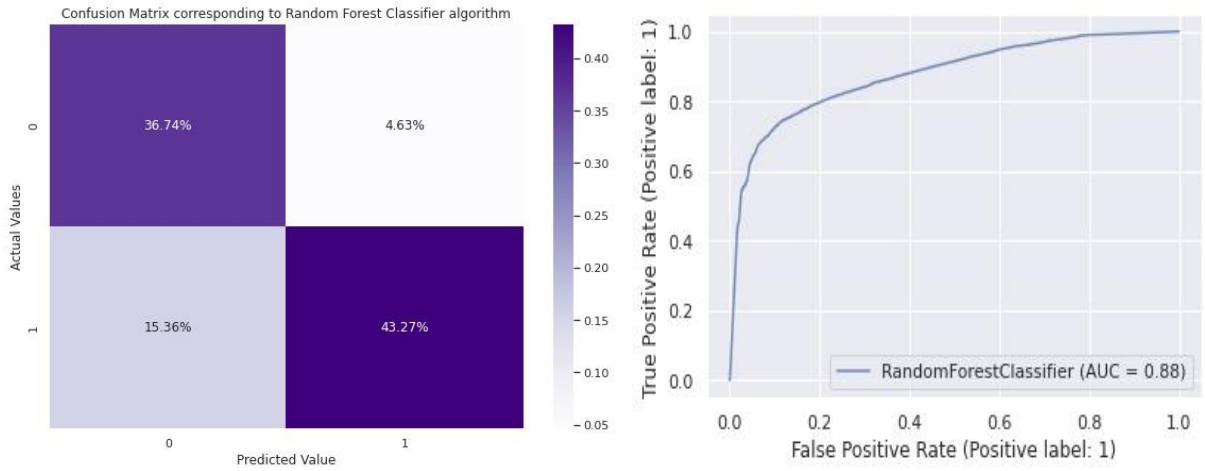
Table 6: Accuracy Matrix

Model	Train-Test-Split
Naive Bayes	55.012322
Logistic Regression	60.847444
KNN	72.986724
Decision Tree	79.569123
Random Forest	80.340250

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 and ROC curve



Explanation: The confusion matrix shows the accuracy of the optimal model. It predicted that correspond to 0, only 4.63% is wrong and 36.74% is right. Whereas corresponds to 1 is 15.36% are wrong values and 43.27% is right. So almost 20% is predicted wrong and 80% is right.

Explanation for ROC curve: Generally, if the AUC value lies 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. It means our roc curve gives the fair auc score.

Table 7: Classification Report

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

Explanation of Classification Report:

- **Precision:** Accuracy of positive predictions.
- **Recall:** Fraction of positives that were correctly identified.
- **F1-score:** The F1 score is a weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0.
- **Support:** Support is the number of actual occurrences of the class in the specified dataset.

Conclusion:

My dataset is all about the restaurants in the cities of Bangalore and my dataset is available on the Kaggle. I collected data from CSV file from the Kaggle, from which half of values were missing, and I did not throw up all values, instead of removing NULL values, I tried to fill the values with mean and mode. I have done exploratory data analysis to answer all the research questions. I used one-hot encoded features; I also normalized our data with the help of min-max scaling, I also do train-test-split and do under sampling strategy to stable my dataset and tried different classification models. Random Forest classification is the best fit model of our dataset, so we made confusion matrix corresponding to RFC and made the ROC curve with our optimal model. Moreover, we find out top rated restaurants are present in the BTM city and the minimum rating of all these restaurants are lies under between 3.79 to 3.52. Also, we can say that the average rating of new restaurants in Bangalore would be lies in between the 3.

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