Zomato Restaurants Data Analysis Using Machine Learning Algorithms

[1] Naved Choudhary, [2] Vijay Panwar, [3] Sonam Mittal, [4] Gaurav Sahu

[1] Student, B K Birla Institute of Engineering & Technology, Pilani

[2] Student, B K Birla Institute of Engineering & Technology, Pilani

[3] Assistant Professor, B K Birla Institute of Engineering & Technology, Pilani

[4] Assistant Professor, B K Birla Institute of Engineering & Technology, Pilani

[1]navedchoudhary821@gmail.com,[2]Vijaypanwar8445@gmail.com,[3]sonam.mittal2405@gmail.com,
[4]gauravsahu.87@gmail.com

ABSTRACT: Whenever we visit a new place, we want to go to the best restaurant or the cheapest restaurant, but a decent one. Or we can first look at the ratings or the reviews if we want to try food in some new restaurants. Zomato is one such app that provides users with ratings and reviews of restaurants across India. Ratings or reviews are considered to be one of the most significant/decisive variables that decide how good a restaurant is. We will therefore use the real time data set here in our project that has different factors/features that a user can look into about a restaurant. We restrict our data only to Bangalore City.

The Zomato dataset provides us with information on factors influencing the establishment of different types of restaurants at different locations in Bangalore, along with the overall ranking of each restaurant. It has 51717 rows and 17 columns in this dataset. We'd like to find the cheapest restaurant in Bengaluru here. Along with the same, we can discuss different other relationships such as the most expensive restaurant, the best location, the relationship between location and ranking, the number of restaurants in a location.

Since it is a real time data we would start with our Data Exploratory process like handling the Nan values, null values, dropping duplicates and other Transformations. Our target variable is the "Rates" column. We explore the relationship of the other features in the dataset with respect to Rates. we will the visualize the relation of all the other depend features with respect to our target variable, and hence find the most correlated features which effects our target variable. We would then implement the data in various modeling structures such as Random Forest, Linear Regression, and Decision Tree. These modeling will then give us the accuracy of prediction and then we could state which model gives us the most optimized and accurate readings [1].

Keywords: Pre-processing, Classification, Logistic Regression, Decision Tree, Random Forest

1. INTRODUCTION

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 favorite restaurant or cafes, by surveying the user ratings and reviews of the previously visited customers. Restaurant Rating also provides columns for writing classified user reviews[2].The fundamental concept of analyzing the Zomato dataset is to get a reasonable idea of the factors influencing each restaurant's aggregate ranking, setting up different types of restaurants at different locations, Bengaluru being one such city has more than 12,000 restaurants serving dishes from all over the world. The industry has not been saturated with new restaurants opening every day yet, and demand is growing day by day. However, it has become difficult for new restaurants to compete with existing restaurants, amid growing demand. They mostly serve the same food. Bengaluru is India's IT capital. They mostly serve the same food.

Bengaluru is India's IT capital. Most people here are mostly reliant on restaurant food because they don't have time to prepare for themselves. It has thus become important to research the demographics of a place with such an enormous demand for restaurants. What kind of food in a town is more prevalent. The whole town enjoys vegetarian food. If yes, then that place is inhabited, for example, by a specific sect of people. Jain, Marwaris, mainly vegetarian Gujaratis. Using the data, this form of analysis can be done by analyzing various factors .

Bangalore(officially known as Bengaluru) is the capital and largest city of the Indian state of Karnataka. With a population of over 15 million, Bangalore is the third largest city in India and 27th largest city in the world.Bangalore is one of the most ethnically diverse cities in the country, with over 51% of the city's population being migrants from other parts of India.
Bangalore is sometimes referred to as the "Silicon Valley of India" because of its role as the nation's leading information technology (IT) exporter.Bangalore has a unique food culture. Restaurants from all over the world can be found here in Bengaluru, with various kinds of cuisines.
Some might even say that Bangalore is the best place for foodies.The growing number of restaurants and dishes in Bangalore is what attracts me to inspect the data to get some insights, some interestingfacts and figures.So, in this article I will be analyzing the Zomato restaurant data for the city, Bangalore.

2. LITERATURE REVIEW

Zomato Bangalore is a data set such that, it gives us an in detail insight about what are the various restaurants that are located in various areas in Bangalore. It also gives us details about the various ratings of the restaurants based on the cuisines, cost, location and various other features. We start with cleaning the data to reduce or clear the null values and then proceed to transform a few columns with basic calculations. One of the most important aspects before going to any of the restaurants is that we look at the reviews and ratings of the restaurants. Hence we find various aspects in our dataset which are dependent on various features to predict the rate of our dataset. Classification algorithm is one of the greatest significant and applicable data mining techniques used to apply in analysis. Classification algorithm is the most common in several data analysis. Many of them show good classification accuracy.

We compare various feature of the data set with our target column and come up with various visual aid to find which all features are highly co-related to our target variable. Based on these co-relation features we then build predictive models to predict the rate of any particular restaurant based on the given set of features. Since it is a real time data we would start with our Data Exploratory process like handling the Nan values, null values, dropping duplicates and other Transformations. Our target variable is the "Rates" column.

We explore the relationship of the other features in the dataset with respect to Rates. we will the visualize the relation of all the other depend features with respect to our target variable, and hence find the most correlated features which effects our target variable. We would then implement the data in various modeling structures such as Random Forest, Linear Regression, Decision Tree. These modeling will then give us the accuracy of our prediction and then we could state which model gives us the most optimized and accurate readings.

3. DATASET DESCRIPTION

The dataset is taken from kaggle, you can find it [3].
Courtesy of Himanshu Poddar, the data is accurate to that available on the Zomato website until 15 March 2019.

The dataset contains the following features:

1. url : This feature contains the url of the restaurant on the Zomato website

2. address : This feature contains the address of the restaurant in Bangalore
3. name : This feature contains the name of the restaurant
4. online_order : whether online ordering is available in the restaurant or not
5. book_table : table book option available or not
6. rate : contains the overall rating of the restaurant out of 5
7. votes : contains total number of upvotes for the restaurant
8. phone : contains the phone number of the restaurant
9. location : contains the neighborhood in which the restaurant is located
10. rest_type : restaurant type
11. dish_liked : dishes people liked in the restaurant
12. cuisines : food styles, separated by comma
13. approx_cost(for two people) : contains the approximate cost of meal for two people
14. reviews_list : list of tuples containing reviews for the restaurant, each tuple consists of two values, rating and review by the customer
15. menu_item : contains list of menus available in the restaurant
16. listed_in(type) : type of meal
17. listed_in(city) : contains the neighborhood in which the restaurant is located.

3.1 DATA PRE-PROCESSING

The Dataset contained 17 Attributes.

- Records with null values were dropped from ratings columns and were replaced in the other columns with a numerical value.
- Values in the ‘Rating’ column were changed. The ‘/5’ string was deleted. For eg. If the rating of a restaurant was 3.5/5, it was changed to 3.5.
- Using Label Encoding from sklearn library, encoding was done on columns like book_table, online_order, rest_type, listed_in (city).

3.2 EXPLORATORY DATA ANALYSIS

A lot of effort went into the EDA as it gives us a detailed knowledge of our data.

Exploratory Data Analysis (EDA) is an approach/philosophy for data analysis that employs a variety of techniques (mostly graphical) to

- maximize insight into a data set;
- uncover underlying structure;
- extract important variables;
- detect outliers and anomalies;
- test underlying assumptions;
- develop parsimonious models;
- Determine optimal factor settings [4].

3.3 RANDOMIZATION AND SPLITTING OF DATASET

The features selected in the preceding step were approved to develop classification models. Initially the dataset was randomized to obtain an arbitrary permutated sample. It was followed by splitting of the dataset into training (70% of the dataset) and test (30%) sets.

3.4. CLASSIFICATION ALGORITHMS

3.4.1. Linear Regression:

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable[5]. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model. Before attempting to fit a linear model to observed data, a modeler should first determine whether or not there is a relationship between the variables of interest. This does not necessarily imply that one variable causes the other (for example, higher SAT scores do not cause higher college grades), but that there is some significant association strength of the relationship between two variables. If there appears to be no association between the proposed explanatory and dependent variables (i.e., the scatter plot does not indicate any increasing or decreasing trends), then fitting a linear regression model to the data probably will not provide a useful model. A valuable numerical measure of association between two variables is the correlation coefficient, which is a value between -1 and 1 indicating the strength of the association of the observed data for the two variables. A linear regression line has an equation of the form Y = a + bX, where X is the explanatory variable and Y is the dependent variable. The slope of the line is b, and a is the intercept (the value of y when x = 0)

3.4.2. Random Forest:

Random forests or random decision forests are an ensemble learning technique for classification [6].Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of over fitting. The below diagram explains the working of the Random Forest algorithm.

Working Of Random Forest Algorithm

We can understand the working of Random Forest algorithm with the help of following steps –

- **Step 1** – First, start with the selection of random samples from a given dataset.
- **Step 2** − Next, this algorithm will construct a decision tree for every sample. Then it will get the prediction result from every decision tree.
- **Step 3** − In this step, voting will be performed for every predicted result.
- **Step 4** − At last, select the most voted prediction result as the final prediction result [7].

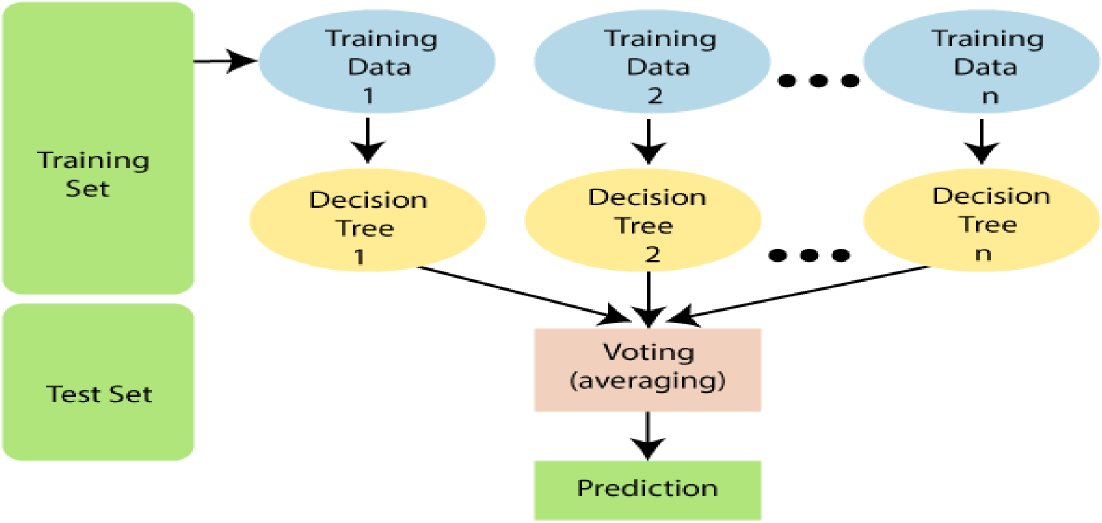

Fig 1. Example of Random Forest

The pseudo code for random forest algorithm can split into two stages. First, in which ‘n' random trees are created, this forms the random forest. In the second stage, the outcome for the same test feature from all decision trees is combined. Then the final prediction is derived by assessing the results of each decision tree or just by going with a prediction that appears the most times in the decision trees[8].

3.4.3. Decision Tree:

- Decision Tree calculation has a place with the supervised learning algorithms. Decision Tree is a supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.
- In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
- The decisions or the test are performed on the basis of features of the given dataset.
- It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.
- It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.

- In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm.
- A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into sub trees. Below diagram explains the general structure of a decision tree[9]:

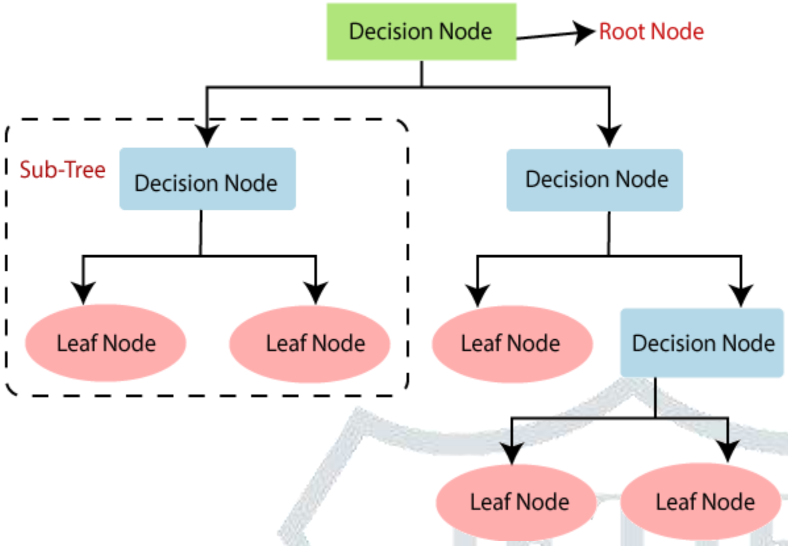

Fig 2. Example of Decision Tree

Types of Decision Trees

Types of decision trees are based on the type of target variable we have. It can be of two types:

1. **Categorical Variable Decision Tree:** Decision Tree which has a categorical target variable then it called a Categorical variable decision tree.
2. **Continuous Variable Decision Tree:** Decision Tree has a continuous target variable then it is called Continuous Variable Decision Tree [10].

4. RESULTS

Algorithms	Accuracy
Linear Regression	24%
Random Forest	87%
Decision Tree	85%

In this model, we have considered various restaurant records with characteristics such as name, average cost, place, whether online order is accepted, we can book a table, restaurant type. This model would help business owners forecast their ranking on the criteria taken into account in our model and enhance the experience of the customer. Different algorithms have been used, but on Random Forest the final model is eventually chosen, which provides the highest precision compared to others.

5. CONCLUSIONS

This paper analyses a variety of characteristics of current restaurants in different areas of a city and analyzes them to predict restaurant ratings. This makes it an important thing to take into consideration before making a dining decision. Before creating a venture like that of a restaurant, such research is an important part of planning. There has been a lot of research into variables impacting revenue and the competition in the restaurant industry. To enhance customer satisfaction rates, numerous dine-scape variables have been analyzed. If data is also collected for other citrics, such predictions could be made for accuracy

6. REFERENCES

[1] https://colab.research.google.com
[2] Chirath Kumarasiri, Cassim Faroo,"User Centric Mobile Based Decision-Making System Using Natural Language Processing (NLP) and Aspect Based Opinion Mining (ABOM) Techniques for Restaurant Selection". Springer 2018. DOI: 10.1007/978-3-030- 01174-1_4
[3] https://www.kaggle.com/himanshupoddar/zomato-bangalore-restaurants
[4] Atharva Kulkarni,Divya Bhandari,Sachin Bhoite.A study of Restaurants Rating Prediction using Machine Learning Algorithms. International Journal of Computer Applications Technology and Research, 2019, p.377-378.DOI: 10.7753/IJCATR0809.1008
[5] Linear Regression, Retrieve from: HTTPS://WWW.toworddatascience.com/Linear_ Regression.
HTM, LAST Accessed: 5 Octobor, 2019
[6] L. Breiman, Random Forests. Machine Learning, 45(1), (2001); 5–32. https://doi.org/10.1023/A:1010933404324
[7]https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_classification_algorithms_random_forest.htm
[8] https://www.techleer.com/articles/107-random-forest-supervised-classification-machine-learning-algorithm
[9] Decision Trees, Retrieve from: https://dataaspirant.com/2017/01/30/how-decision-treealgorithm-works/, Last Accessed: 5 Octobor,2019
[10] https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html