Project Title

Ranking of Restaurants

By:

Rajkumar S. Jagdale

Email Id:

rajkumarjagdale@gmail.com

Index

Ranking of Restaurants

1. Introduction:

Zomato is an Indian restaurant search and discovery service founded in 2008 by Deepinder Goyal and Pankaj Chaddah. It currently operates in 24 countries and provides information and reviews of restaurants, including images of menus where the restaurant does not have its own website and also online delivery. In this project, we have taken Zomato dataset of restaurants of Banglore city. This dataset is publically available on kaggle and we performed Exploratory Data Analysis and classification purpose. In Exploratory Data Analysis, we have achieved some important insights and also performed different Supervised Machine Learning Algorithms. In Machine Learning Algorithms, Regression, Naïve Bayes, Support Vector Machine, KNN, algorithms have been applied and calculated accuracy of each classifier. We got _____% Accuracy which is highest accuracy.

2. Objectives

The Project work focused on Exploratory Data Analysis and classification of ranking of Restaurants of Zomato dataset from Bangalore City. For experimentation purpose, Zomato dataset have been considered for analysis. Objectives of this project work as given below:

1. Study of Zomato Dataset and its attributes.
2. Pre-Processing the dataset and get ready for EDA and classification
3. Implementation of EDA and Machine Learning Algorithms
4. Visualization of EDA results and Accuracy of Classifiers.

3. Statements of problem:

1. The large quantity of various reviews and ratings on different on Zomato Restaurants has made it difficult to decide exact which place is better for us. These online reviews and ratings create difficulty for customer to properly select the Restaurant and take decision. Perfect and accurate analysis different attributes like reviews, Ratings Cost, Location etc. should help the customer to choose the Restaurants matching his/her needs with accuracy.

2. EDA Analysis and Classification helps this decisions with various methodology/

Algorithm for analyzing the data but without desired accuracy or perfection.

3. Tis Project's aim is achieving desired accuracy in decision making by exploring EDA and using ML classifiers considering the parameters of Zomato Restaurants dataset and find out some important insights and calculate the accuracy of Machine Learning Classifiers.

4. Dataset Preparation and Pre-processing:

For every Data Science Problem, Data is important and developer / Researchers spend time on Data Preparation and Pre-Processing. This project also need of Data Preparation and Pre-Processing. For this steps different techniques have been used and made dataset ready for EDA and classification.

4.1 Dataset Preparation

For Analysis of any data, Data Preparation is important step. While preparing the dataset following steps need to be follow.

4.1.1 Data collection

There are different data sources where we get different types of datasets like Text Data, Video Data, Audio data and many more. In This Project, we have been downloaded dataset from kaggle which is publically available. This Zomato dataset has 51717 rows (records) and 17 Columns (Attributes). It contains following attributes in Zomato Dataset:

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)			

4.1.2 Data Visualization

Data Visualization reduces time to understand dataset that why id important in Data Analytics. In Data Science, There are many tools and techniques for data visualization like Matplotlib, Seaborn etc.

4.1.3 Labelling

Before Pre-processing, Labelling is necessary in EDA. We have already labelled dataset.

4.1.4 Data Selection

If there is huge amount of data is present then there is to more confusion to select proper dataset for analysis purpose. In this project we have selected the dataset which is proper for our EDA and classification.

4.2 Data Pre-processing

In this step, we performed different techniques for pre-processing the dataset and made clean for further analysis.

4.2.1 Data Formatting

In the data formatting we performed operations like hanging the Columns Names to proper names, Remove the NaN values from the dataset etc. we formatted data properly and did further process of cleaning.

4.2.2 Data cleaning

Data cleaning step is very important in Data analytics. It helps to delete unwanted data and made useful data only for further analysis. We have deleted unnecessary Columns like 'url','dish_liked' and 'phone' etc which are not important for EDA and Classification too. Removed the Duplicates records which increases redundancy of the data which affects on EDA process. We also removed '/5' from Rates and replace proper rating and converted into Numerical form and also converted cost object type into numerical type.

4.3 Data Transformation

Data transformation is important in feature extraction. Features should be in numerical form in Machine Learning techniques. Scaling helps to convert all numerical numbers into one scale.

4.3.1 Scaling:

In this step, categorical data has been converted into numerical. Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data pre-processing step.In this all attributes like online_order, book_table, rate, votes, location, rest_type, cuisines, cost, menu_item are encoded using LabelEcoder.

4.3.2 Feature Extraction:

When the input data to an algorithm is too large to be processed and it is suspected to be redundant then it can be transformed into a reduced set of features. The selected features are expected to contain the relevant information from the input data, so that the desired task can be performed by using this reduced representation instead of the complete initial data.

In this project, we have selected features like online_order, book_table, rate, votes, location, rest_type, cuisines, cost and menu_item as input features and rate is target feature.

4.4 Dataset Splitting

Splitting the dataset into Train set and Test set in very important for train the model and calculate accuracy of model. Training data is used to train an algorithm, typically making up a certain percentage of an overall dataset along with a testing set. Test data is used to see how well the machine can predict new answers based on its training. In this project, from Zomato dataset, while applying Machine Learning Classifier 80 % data is used for training and 20 % data is used for testing purpose.

4.5 Proposed Methodology

In this project following approach has been used for EDA and Classification Analysis.

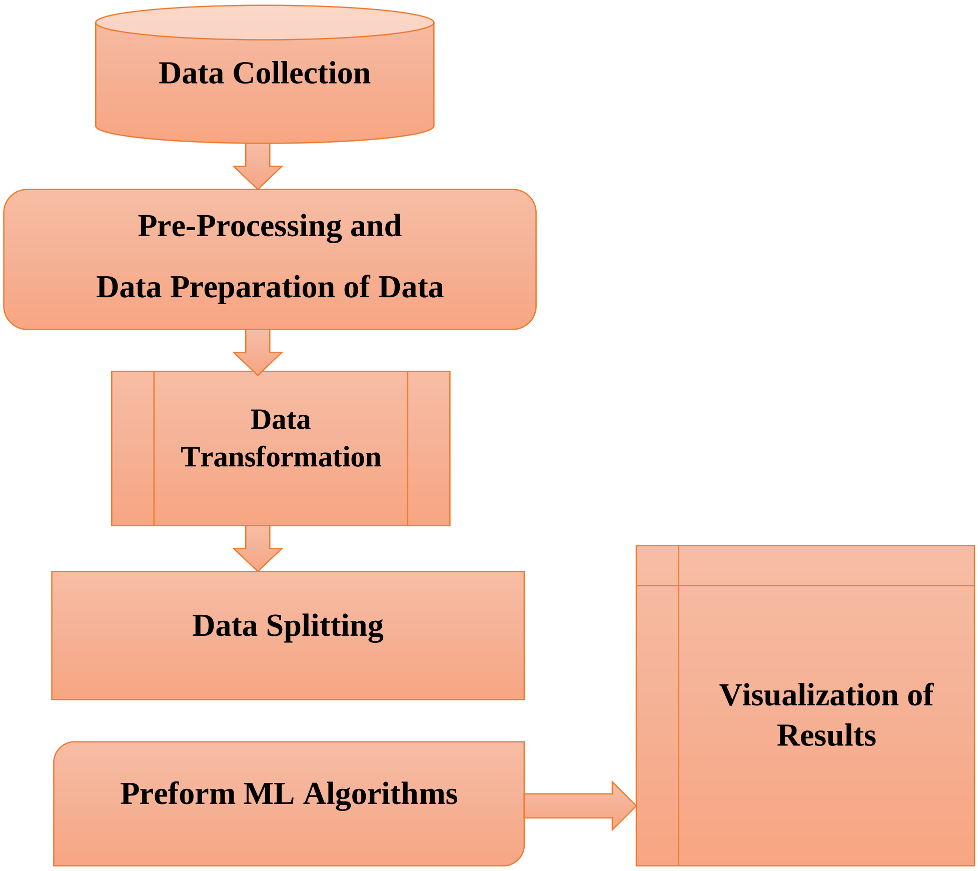

Fig. 1 Flowchart of Proposed Methodology

5. Model Training and Testing

Training a model simply means learning (determining) good values for all the weights and the bias from labeled examples. In supervised learning, a machine learning algorithm builds a model by examining many examples and attempting to find a model that minimizes loss; this process is called empirical risk minimization.

The usage of the word "testing " in relation to Machine Learning models is primarily used for testing the model performance in terms of accuracy/precision of the model. It can be noted that the word, "testing" means different for conventional software development and Machine Learning models development.

6. Supervised Machine Learning Algorithms

In Supervised learning, you train the machine using data which is well "labelled." You want to train a machine which helps you predict how long it will take you to drive home from your workplace is an example of supervised learning. Regression and Classification are two types of supervised machine learning techniques

6.1 Logistic Regression

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

6.2 Decision Tree

Decision tree builds *classification* or regression models in the form of a *tree* structure. It breaks down a data set into smaller and smaller subsets while at the same time an associated *decision tree* is incrementally developed. ... A *decision* node has two or more branches. Leaf node represents a *classification* or *decision*.

6.3 K-**Nearest Neighbors (**KNN**)**

K-Nearest Neighbors (KNN) is one of the simplest algorithms used in Machine Learning for regression and classification problem. KNN algorithms use data and classify new data points based on similarity measures (e.g. distance function). Classification is done by a majority vote to its neighbors.

6.4 Random Forest

The *random forest* is a classification algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated *forest* of trees whose prediction by committee is more accurate than that of any individual tree.

6.5 Support Vector Machine

A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an SVM model sets of labeled training data for each category, they're able to categorize new text. So you're working on a text classification problem.

6.6 Gradient Boosting

Gradient boosting is a type of machine learning boosting. It relies on the intuition that the best possible next model, when combined with previous models, minimizes the overall prediction error. The key idea is to set the target outcomes for this next model in order to minimize the error.

6.7 Naïve Bayes

It is a classification technique based on Bayes Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

7. Result Analysis and Discussion

After successfully applying Supervised Machine Learning Algorithms we got highest accuracy for Random Forest classifier i.e. 94.90 % and lowest accuracy for Naïve bayes 45 %.

Table 1. Accuracy Table for ML Classifiers

Sr. No.	ML Classifier Name	Accuracy (%)
1	Logistic Regression	59.29
2	Decision Tree	93.67
3	KNN	89.81

4	Random Forest	**94.90**
5	Support Vector Machine	56.05
6	Gradient Boosting	69.18
7	Naive bayes	**45**

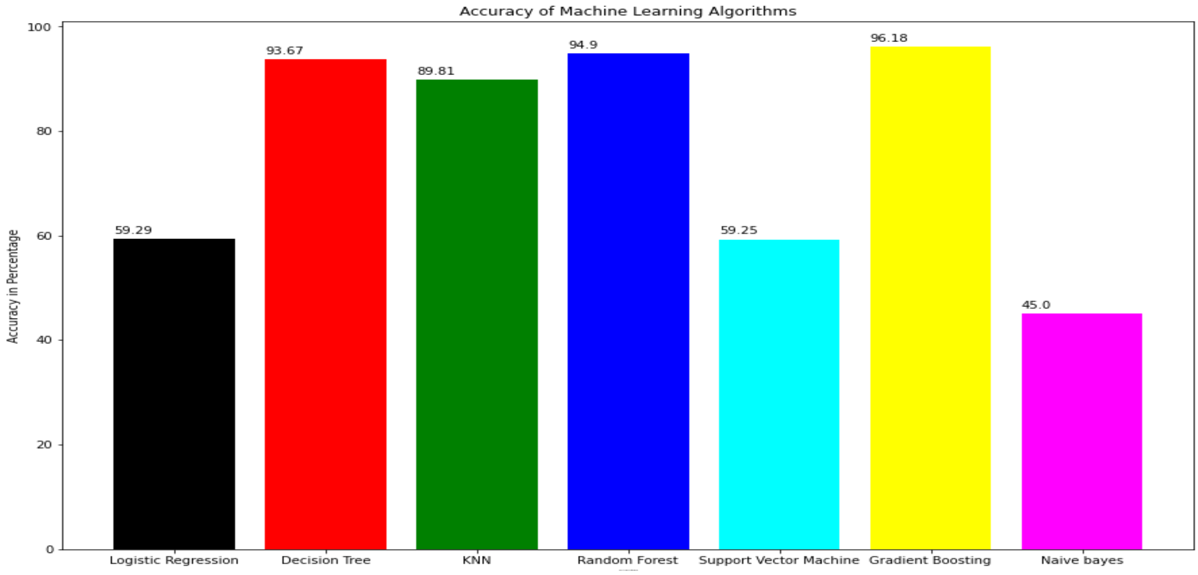

Fig 2. Bar Graph of Accuracy of Ml Classifiers

8. Exploratory Data Analysis

Exploratory Data Analysis as the name suggests is an approach (mostly graphical/visual) to discover insights from the data set, summarize the key relationships, identifying underlying parameters, etc. EDA helps organisations to study the data at hand and unlock patterns or trends to further define their KPIs (Key Performance Indicators)

8.1 Correlation between different variables

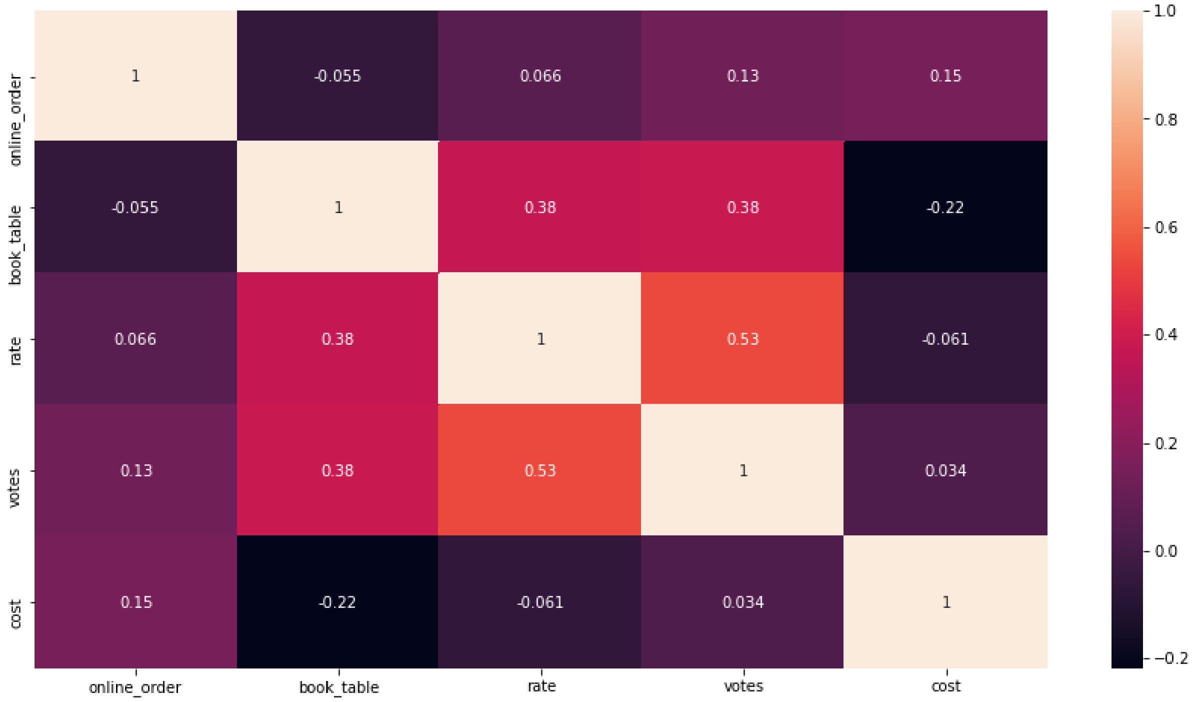

Fig 3. Correlation between different variables

8.2 Restaurants delivering Online or not

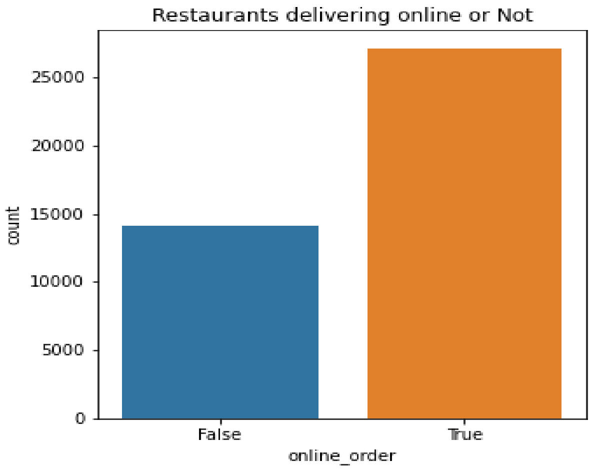

Fig 4 . Restaurants delivering online or not

8.3 Restaurants allowing table booking or not

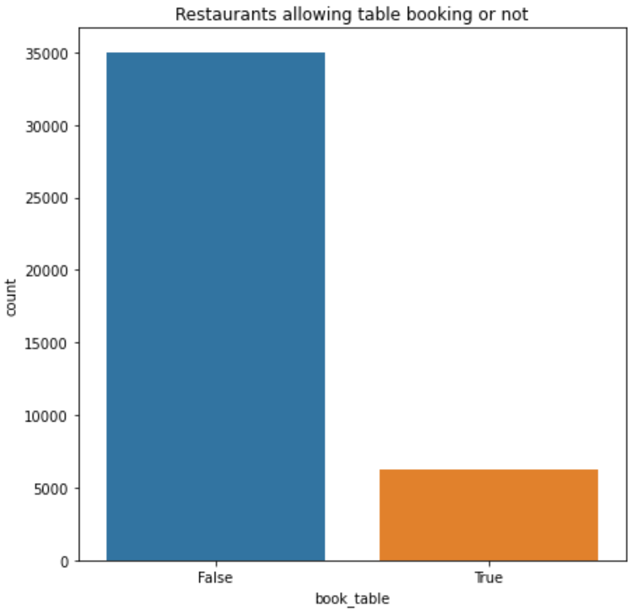

Fig 5. Restaurants allowing table booking or not

8.4 Types of Services

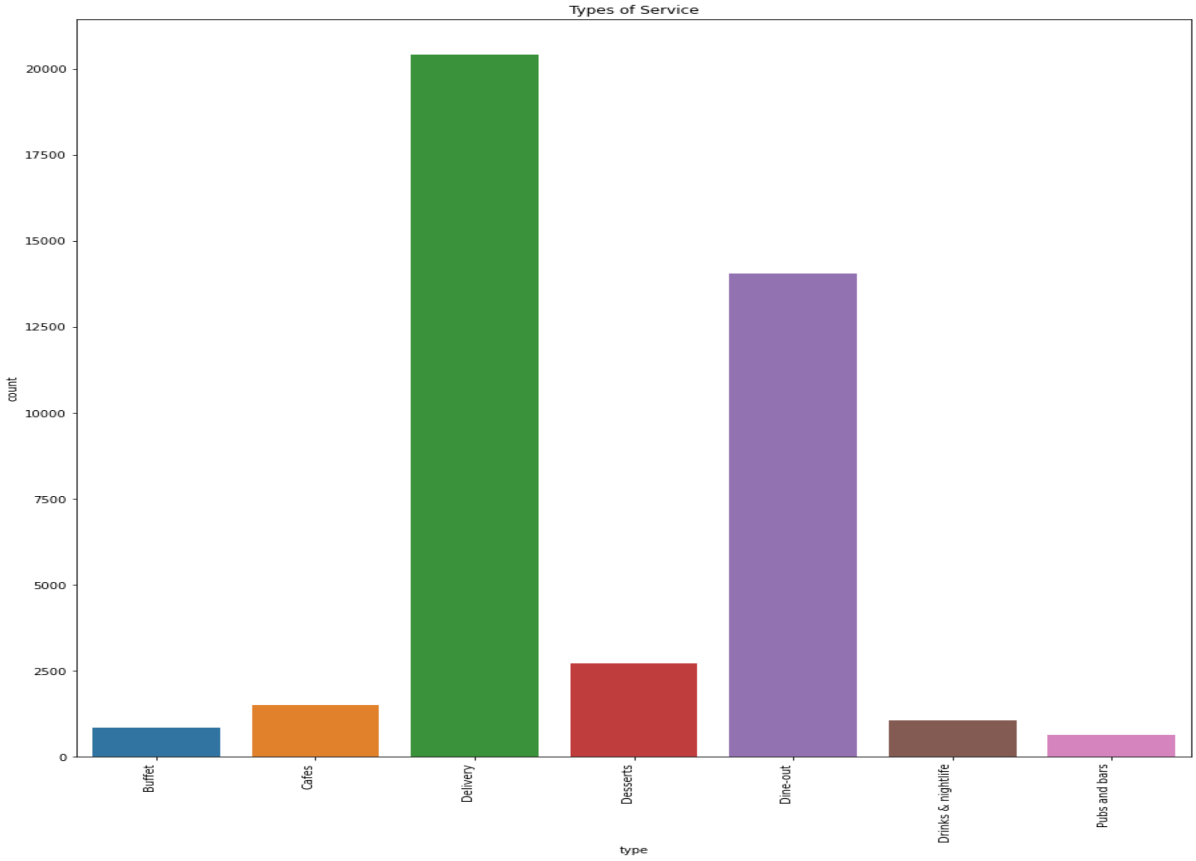

Fig 6 . Types of Services

8.5 Plot of Type Vs Rating

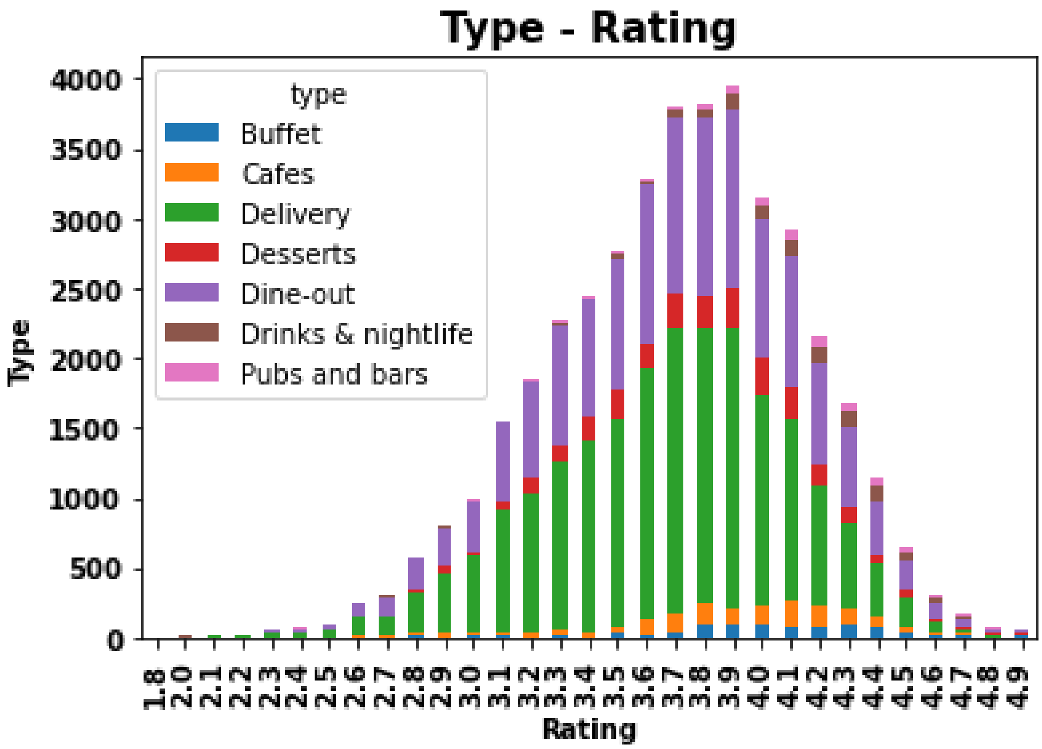

Fig 7. Plot of Type Vs Rating

8.6 Scatter Plot of Cost Vs Rating

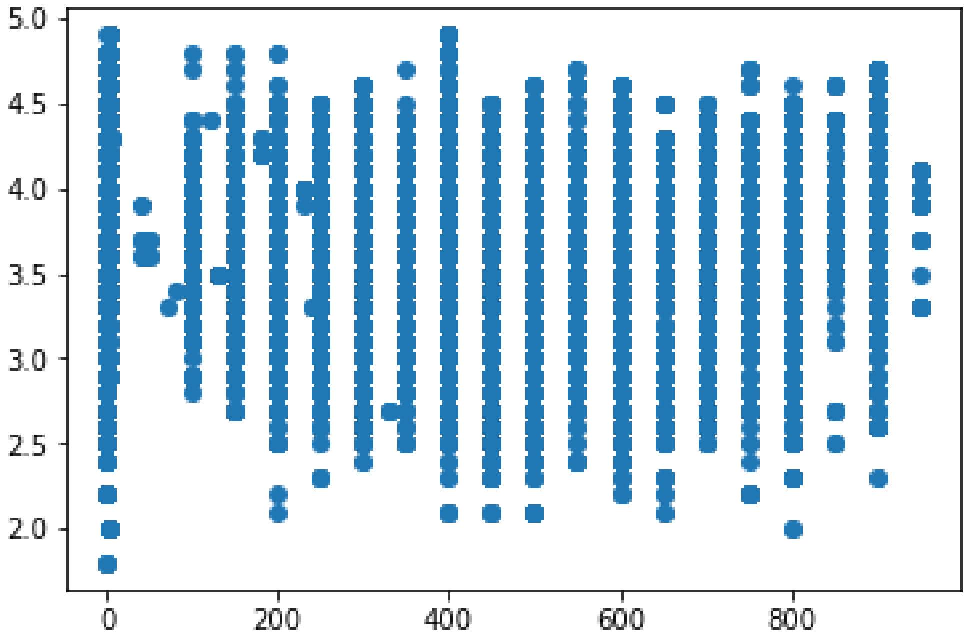

Fig 8. Scatter Plot of Cost Vs Rating

8.7 Histogram of distribution of restaurant rating

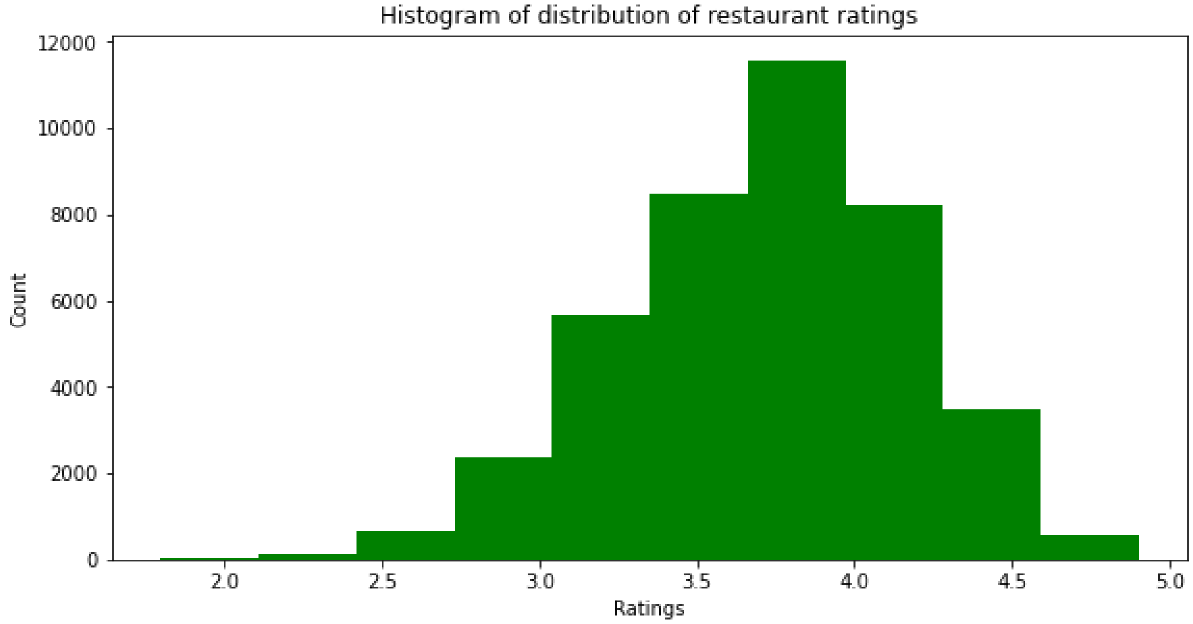

Fig 9 . Histogram of distribution of restaurant ratings

8.8 Top 15 Restro with maximum number of outlets

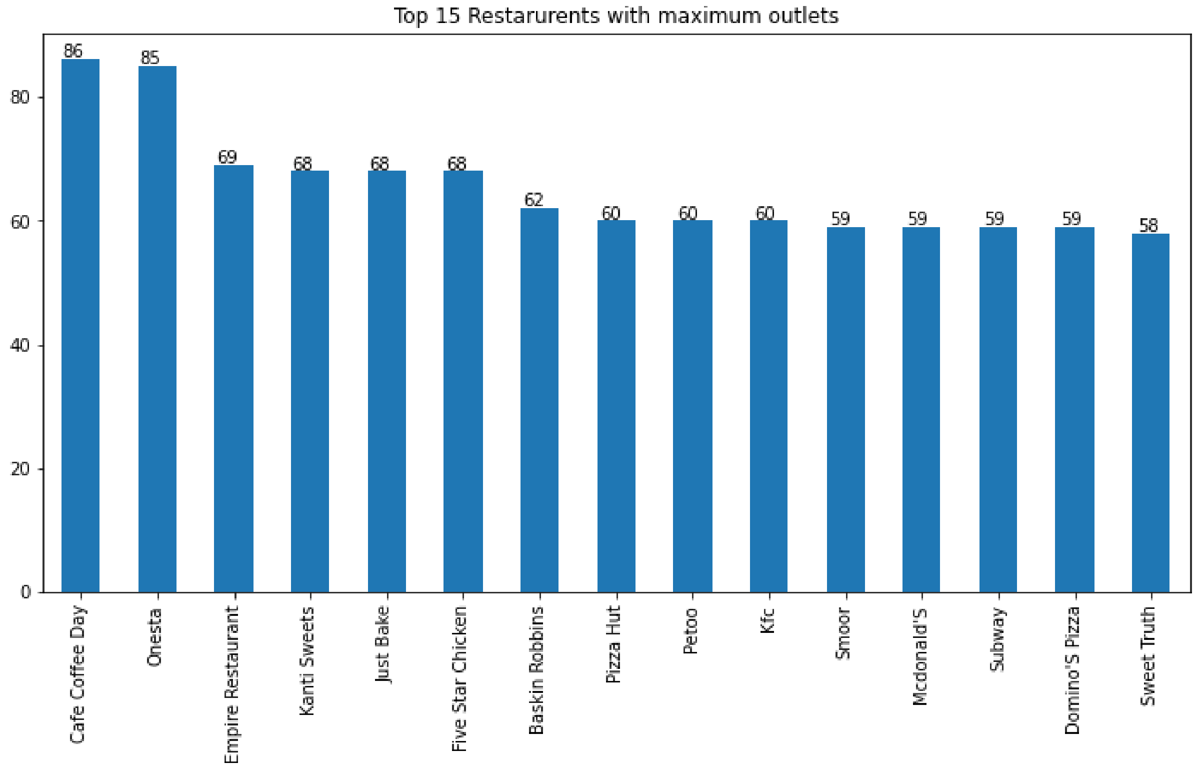

Fig 10 . Top 15 Restro with maximum number of outlets

8.9 Most famous restaurant chains in Bengaluru

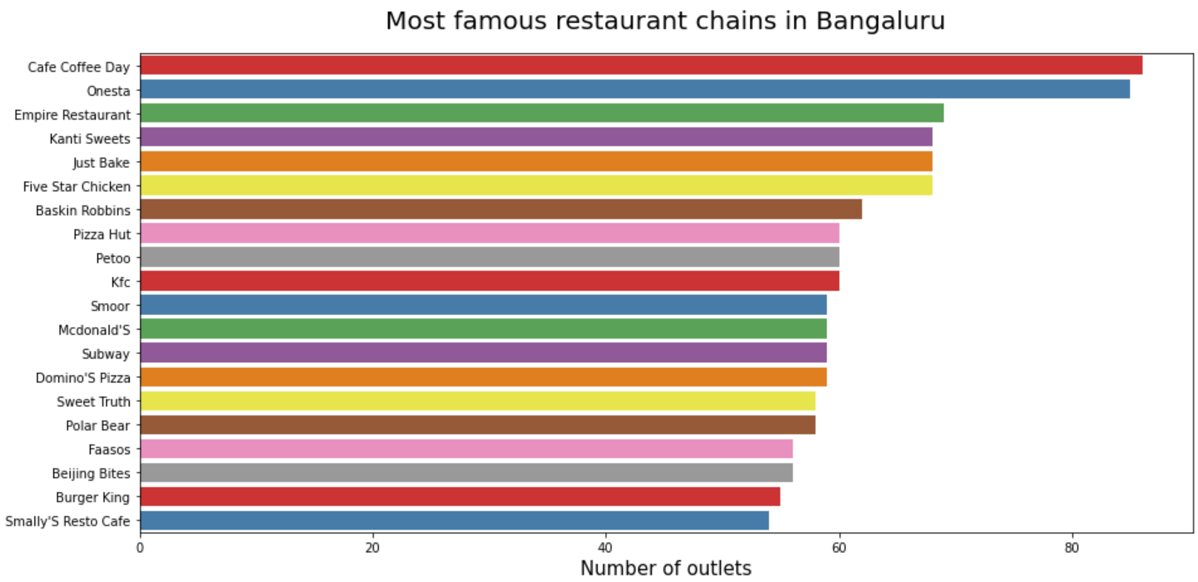

Fig 11. Most famous restaurant chains in Bengaluru

8.10 Restaurant Count by Location

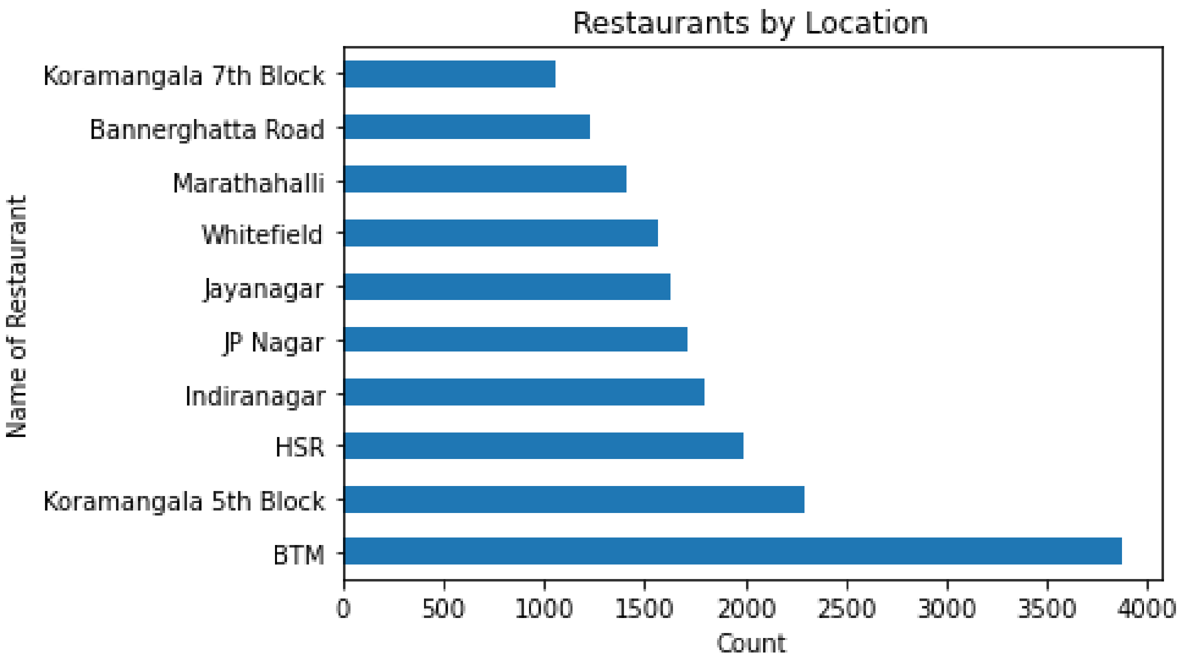

Fig 12. Restaurant Count by Location

9. Conclusions and Future Scope

After completion of this small project, we can conclude that we got highest accuracy for Random Forest classifier i.e. 94.90 % and lowest accuracy for Naïve bayes 45 %. And In future scope we can increase in dataset size and also apply Deep Learning algorithms. We also can do web based application on Restaurants Rating system online which is useful for common people to find out proper and popular Restaurants in any location.

References:

1. https://www.kaggle.com/himanshupoddar/zomato-bangalore-restaurants
2. https://dzone.com/articles/qa-blackbox-testing-for-machine-learning-models
3. https://medium.com/@chiragsehra42/decision-trees-explained-easily-28f23241248
4. https://towardsdatascience.com/understanding-random-forest-58381e0602d2
5. https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/
6. https://www.displayr.com/gradient-boosting-the-coolest-kid-on-the-machine-learning-block/
7. https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/