**PROJECT REPORT**

**ON**

**Zomato Restaurant Rating Prediction**

Logo

Description automatically generated

Submitted by-

Team 04

Gowtham Raghuraman - 001529954

Zarana Bhadricha - 001518953

**Abstract**

This report consists of the final findings of the Supervised Machine Learning algorithms used for the prediction of the rating of restaurants from Zomato in Banglore location in India.

The dataset used for the project has been taken from Kaggle. The data is then cleaned as per the requirement and then various models were built in Jupyter Notebook (Anaconda 3) to examine the models’ performance on certain parameters.

After a preliminary study of the available algorithms and data review, it became apparent that the problem fell under non-linear Regressor category. The study focuses on various algorithms by using classifiers like- Linear Regression, Decision Tree Regressor, Random Forest Regressor ,K-Nearest Neighbor Regression, Support Vector Machine and Gradient Boosting Regression.

The major finding is that the machine learning approach should be suitable for this problem due to many aspects, like a total of 17 features in the dataset.

**Table of Contents**

1. Introduction
   * Background
   * Motivation
   * Goal
2. Methodologies and Algorithms
   * Data Review
   * Software and Libraries used
   * Data Cleaning
   * Feature Selection
   * Exploratory Data Analysis
   * Methods used
     + Linear Regression
     + Decision Tree Regressor
     + Random Forest Regressor
     + Linear Support Vector Regressor
     + Knn Neighbor Regressor
     + Gradient Boosting
   * Feature Engineering
   * Feature Selection
3. Description of Dataset
4. Results and Analysis
5. Conclusion
6. Reference

**Introduction**

**Background**

Food culture of Bengaluru in India is vibrant and fascinating. Various cuisines right from United States to Japan, Russia to Antarctica, is available here. Delivery, Dine-out, Pubs, Bars, Drinks, Buffet, Desserts Bengaluru has it all. There has been a constant rise in the number of restaurants for two decades, currently which stands at approximately 8000 restaurants. Despite of this high number of restaurants the industry is far from reaching the saturation point while new restaurants are opening every day. However, it has become difficult for them to compete with already established restaurants. The key issues that continue to pose a challenge to them include high real estate costs, rising food costs, shortage of quality manpower, fragmented supply chain and over-licensing. With such an overwhelming demand of restaurants it has therefore become important to study the demography of a location. This project can not only shed light on what customers value the most about a restaurant, but also provide suggestions on what feature combinations one should choose when opening a new restaurant, and how likely this restaurant can succeed.

**Motivation**

Every business wants to know whether it can succeed in the future. For restaurants, rating on Zomato is one of the most important indicators. It not only reveals restaurants’ quality and services, but also helps to attract more customers. Also, what factors should be kept in mind if someone wants to open new restaurant. Does the demography of an area matters? Does location of a particular type of restaurant also depends on the people living in that area? Does the theme of the restaurant matters?. This project focuses on answering these questions and predicting ratings and popularity change of restaurants. With data from Zomato, we propose to use several machine learning methods including Linear Regression ,Decision Tree Regresison,Random Forest Regression, K-Nearest Neighbor Regressor ,Support Vector Regression and Gradient Bossting regression to make relevant prediction.

**Goal**

The primary objective of this project is to build a restaurant success predictor. We aim to visualize various factors that contribute to the popularity of a restaurant while studying the trend across the city of Bengaluru location wise. Our goal is to get helpful results like some mentioned below-

* To open a new restaurant what are the cusines that the customers like.
* The location of the restaurant and the most liked dish from various reviews.
* Features to include like online Booking for delivery and table booking.

**Methodology & Algorithm**

**Data Review**

The dataset shape is 51717 x 17 and based on the rating of each online order/dine-in experience of the customers. The size of data is 547.48 MB.

**Software and Libraries used**

The dataset is downloaded from Kaggle. And Jupyter notebook is used with several Libraries

That are-

* Numpy
* Pandas
* Matplotlib
* Seaborn
* Metrics
* Mean Squared Error
* Scikitlearn Tree
* DecisionTreeRegressor
* train\_test\_split
* Standard Scaler
* r2\_score
* accuracy\_score
* sklearn.ensemble
* RandomForestRegressor
* plotly.graph\_objects
* sklearn.preprocessing
* LabelEncoder
* OneHotEncoder
* Profile Reporting

And several others for Predicting models.

**Data Cleaning**

Data provided was very sparse with many missing values for the one of the feature column and almost 20% of data was missing for target column. Therefore, many rows had NAN values which can compromise the models, so to sort it, all the rows with NAN values are removed as it is a huge dataset with many records.

**Feature Selection**

#### The data originally had 16 columns with only one column has more null values”dish\_liked” and the rest of the columns had less than 1% null datas.

**Target Selection**

#### The target variable here is “rate” column where we’re trying to predict the column from the other selected features.

**Exploratory Data Analysis**

After Cleaning the data and sorting it, we have done feature selection on it using various pandas and matplotlib commands. To Perform Exploratory data Analysis (EDA) on the raw dataset to analyze and visualize the fundamental attributes of the provided data set. The data analysis is an integral part to arrive our conclusion of our prediction. We did check the null values for the columns and created a heatmap for the correlated values.we dropped the highly correleated values. The correlations that will interest us at this stage are between our target rate and the other features. The more a feature is correlated to the target (positively or negatively) the more it could help predicting the target.

#### **Models Used**

The Dependent Target Variable is rate which contains float digits from 1 to 5, which means, **1** being the least review from a customer on their and **5** being the best based on their experience in the restaurant.

As the Target Variable has many classe, the prediction model used is regression as the analysis is a form of predictive modelling technique which investigates the relationship between a dependent (target) and independent variables (predictor).

The dataset with its matrix of features (independent variables) is trained on various Regression models to predict the class value of the dependent target variable.

Different types of Regresison Models used in the project are-

* Linear Regression
* Decision Tree Regresison
* Random Forest Regression
* K-Nearest Neighbor Classification
* Support Vector Regression
* Gradient Boosting Regression

1. **Linear Regression**

Linear Regression attempts to model the relationship between two variables by fitting  linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. Linear Regression was not much success with the accuracy of the prediction of the target value of our dataset. The accuracy dropped less than 80%.

1. **Decision Tree Regression**

Decision Tree is a supervised machine learning model used to predict a target by learning decision rules from features. As the name suggests, we can think of this model as breaking down our data by making a decision based on asking a series of questions. A decision tree is represented as upside down where its root is at the top of the tree then it splits into branches and when it cannot further split then the end branch is called as decision.

1. **Random Forest Regressor**

A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. It worked the best and retured the best results out of the other results. In decision tree only one tree is made but in random forest, our algorithm randomly creates a specified number of decision trees. And chooses tree which is best for our model.

1. **K-Nearest Neighbor Regression**

KNN regression is a non-parametric method that, in an intuitive manner, approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighbourhood. Applying this dataset gave us a accuracy close to 90%.

1. **Support Vector Regression**

Support Vector Machine for regression implemented using libsvm using a parameter to control the number of support vectors.  The Linear SVR algorithm applies linear kernel method and it works well with large datasets. L1 or L2 method can be specified as a loss function in this model. The accuracy of this not not satisfactory.

1. **Gradient Boosting Regression**

GB Regression is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. The accuracy score of gradient Bossting was 51% was the least among all the regression methods.

**Feature Selection**

We selected features that have the maximum contribution to our output by this way we can reduce the compute time of our model. The more details on Feature selection are in the section of Data Preprocessing in this report.

**Visualization**

Visualization is done at classification models as well as EDA. Where bar graph, line graphs, pie charts etc. are used to represent data as well as data frames have been used Matplotlib,Seaborn and profile reporting are used.

**Data Source**

The dataset has been taken from Kaggle:

Zomato Restaurant restaurant aggregator and food delivery company and provides information, menus and user-reviews of restaurants from bangalore location in India.

<https://www.kaggle.com/himanshupoddar/zomato-bangalore-restaurants>

**Graphical user interface, application

Description automatically generated** A picture containing text, transport, lawn mower

Description automatically generated

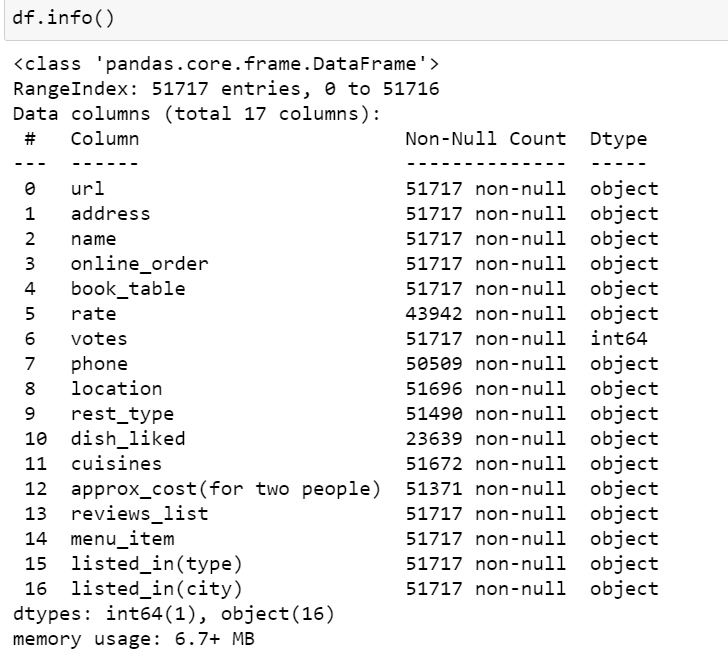
**Dataset**

The dataset shape is 51717 x 17 and based on the rating of each online order/dine-in experience of the customers. The size of data is 547.48 MB. The dataset is in the form of .csv file which was loaded.

List of Attributes:

|  |  |  |
| --- | --- | --- |
| **S.No** | **Attributes** | **Description** |
| 1 | url | The URL has the restaurant name and location concatenated to it of the website. |
| 2 | address | The address of the restaurant. |
| 3 | name | Name of the restaurant |
| 4 | online\_order | Whether the review is from online order experience based on the flag Yes/No. |
| 5 | book\_table | Whether the review is from dine in experience based on the flag Yes/No. |
| 6 | rate | Rating for the order/dine-in experience left by the customer post their experience. |
| 7 | votes | The no of votes provided from the customer. |
| 8 | phone | The contact number of the restaurant |
| 9 | location | The location of the restaurant in bangalore location |
| 10 | rest\_type | The type of the restaurant that it offers, whether casual dining,quick bites,café & take away etc. |
| 11 | dish\_liked | The dishes the customer liked from their experience. |
| 12 | cuisines | The cuisine the restaurant offers for their customers. |
| 13 | approx\_cost(for two people) | The average cost that will incur for people two for each experience. |
| 14 | reviews\_list | The rating and comments provided by the customer on their experience. |
| 15 | menu\_item | The menu of the restaurant available in the Zomato website. |
| 16 | listed\_in(type) | The type of the customer experience offered by restaurant like Buffet,Café & Takeaways to under if it can accommodate in person customer. |
| 17 | listed\_in(city) | The city that the restaurant is available. |

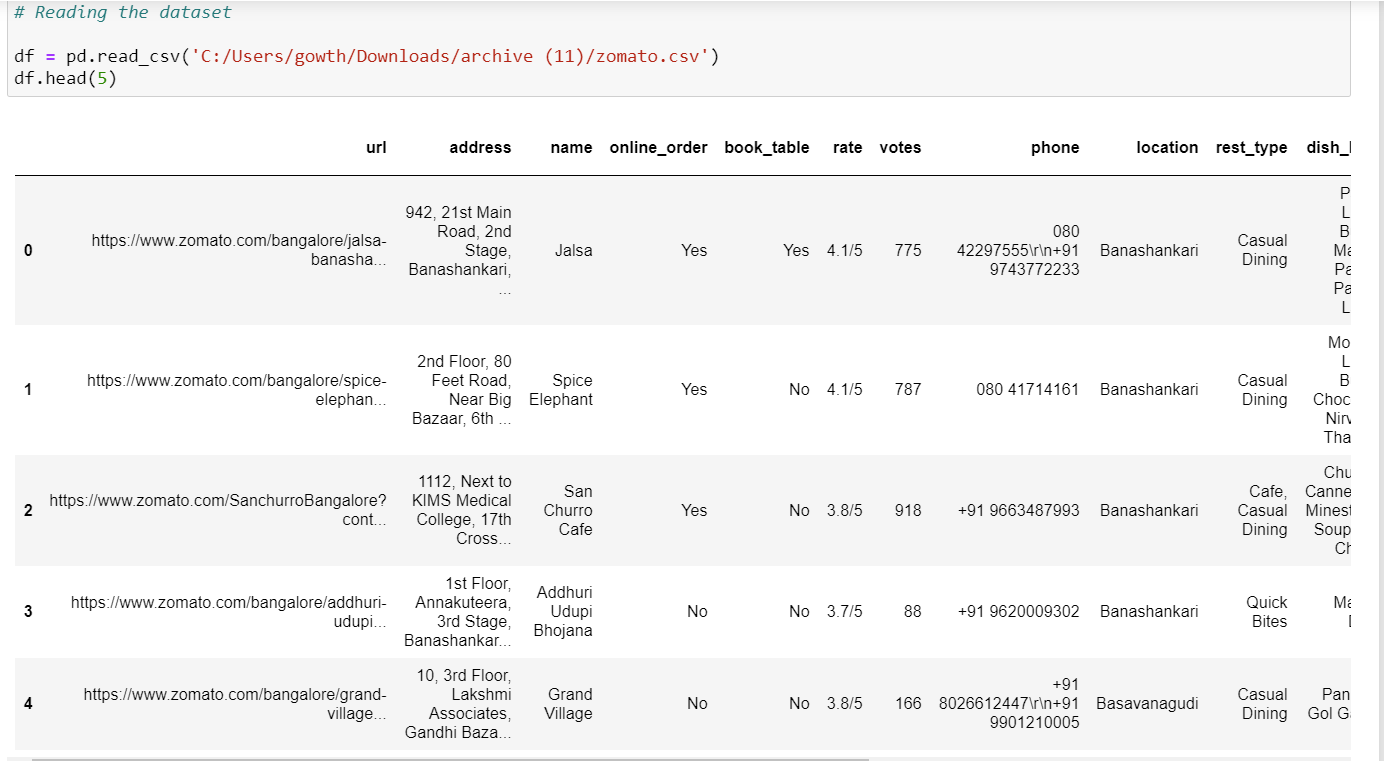
**DataFrame Information**



**Analysis**

**Data loading**

After loading the necessary libraries we loaded the dataset from local storage. Since the dataset contains the attribute names, we do not need to explicitly specify those during loading the dataset.



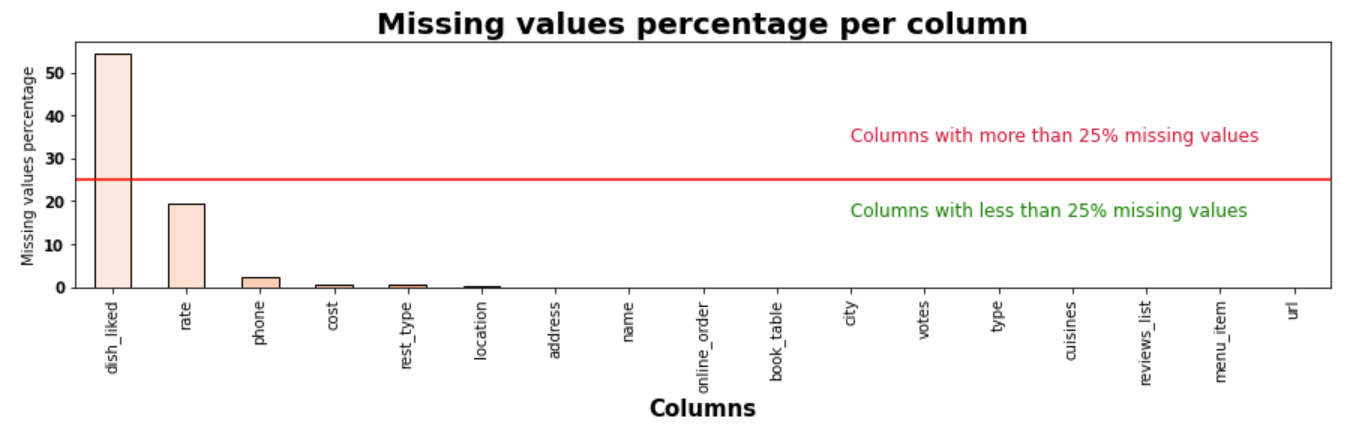
**Data Profile Reporting**

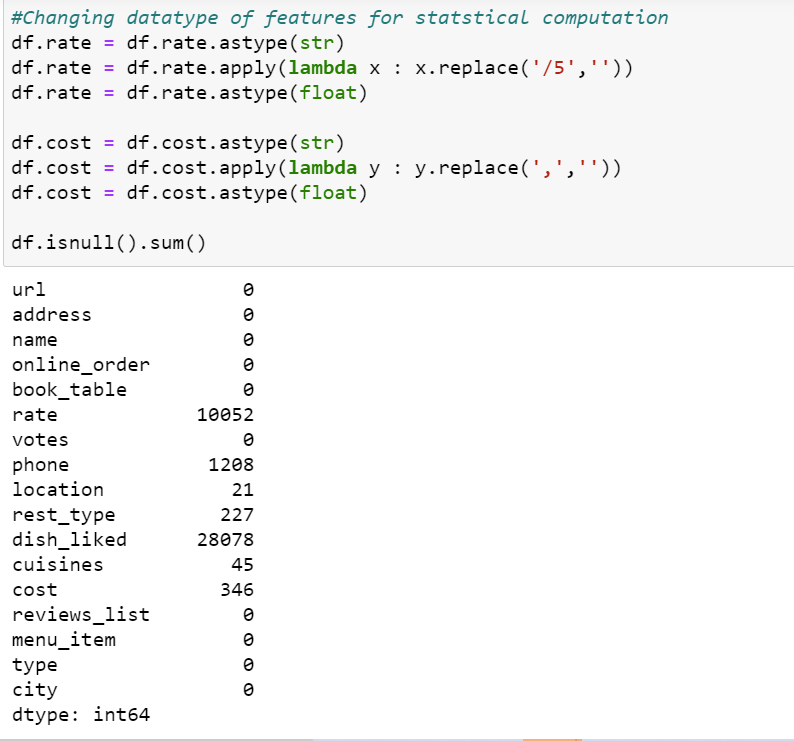
**We ran the profile reporting to view the general report of the data to view general.**



**Data Exploration & Data Cleaning**

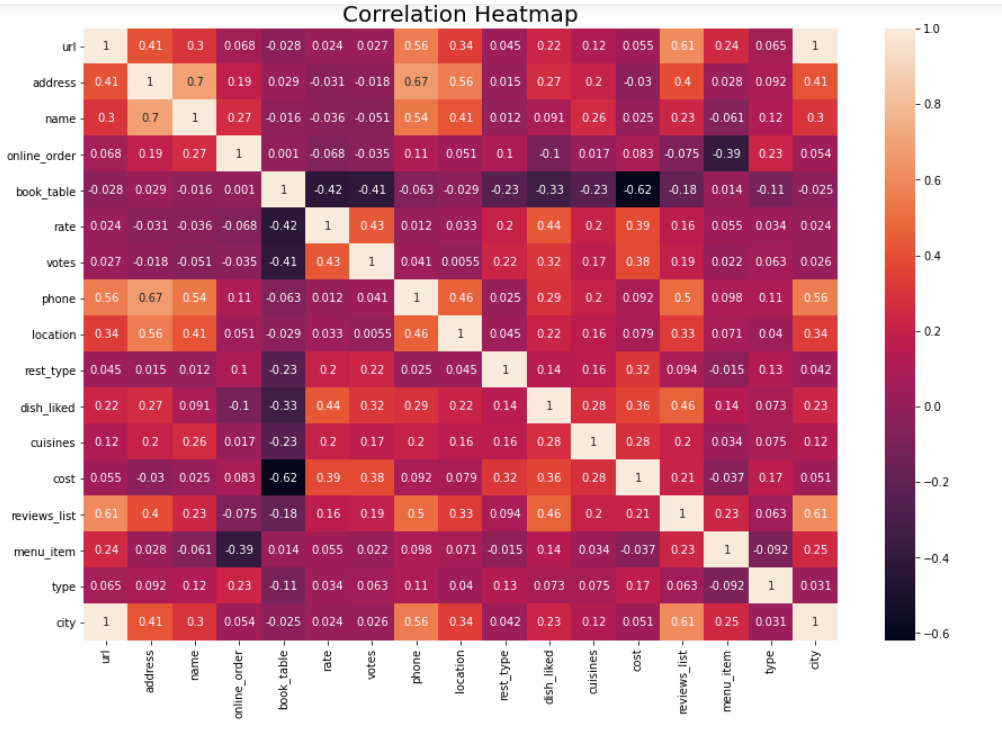
The given dataset has a volume of 51K rows and started to figure out the null values percentage for each feature in the dataset and to analyze to remove or replace the NaN values with relevant values to proceed with data preprocessing.



We analysed that dish\_liked column has the more number of null values among the chosen dataset which is almost 50% of the columns.

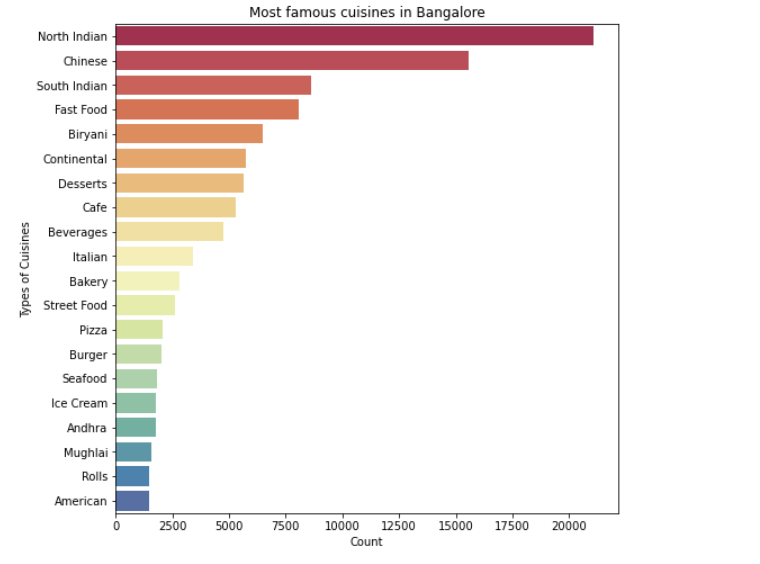
Then correlation matrix is formed with data



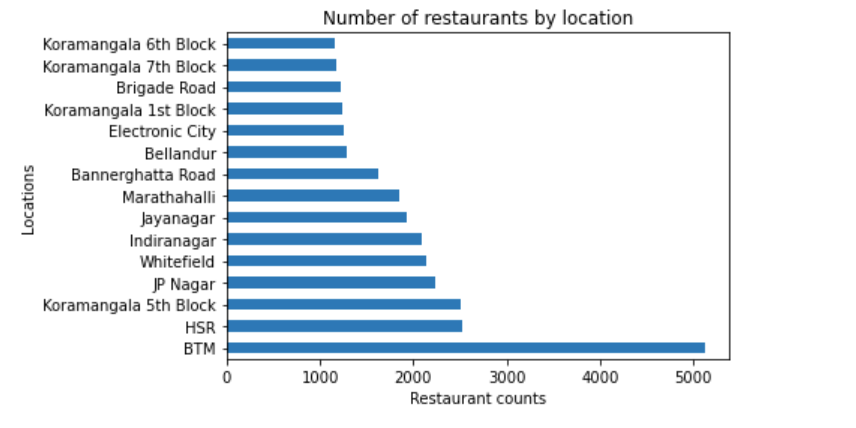
From the heatmap, we can see that some variables are positively correlated and some other negatively. The highest correlation 0.63 is between name and address and the lowest -0.44 between book\_table and approx\_cost\_2\_people. Highly correlated features maybe imply redondant information. So we might drop one the them. The correlations that will interest us at this stage are between our target rate and the other features. The more a feature is correlated to the target (positively or negatively) the more it could help predicting the target. And from the plot, one can see that the 2 top positively correlated features to the target are dish\_liked, vote. Which means the more a dish is liked and a restaurant receive higher votes the more the rate increase. One can also note from the plot that only one feature book\_table is negatively correlated to the target,which means that restaurant in which booking table is needed have low rate.

**Exploratory Data Analysis**

1. What are the top cuisines are in Bangalore?



### Does number of restaurants depends on area?

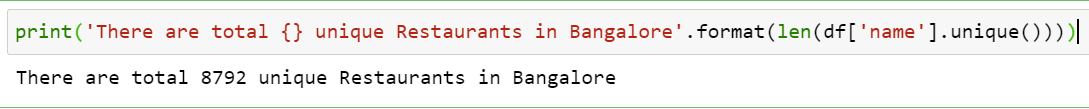


We can see from above barplot that BTM, HSR and Koramangala are top locations with maximum number of restaurants. We can infer below points.

1. The competition amomg these locations can be tough to break
2. Most of the foodie are in this area or prefer to go to these locations

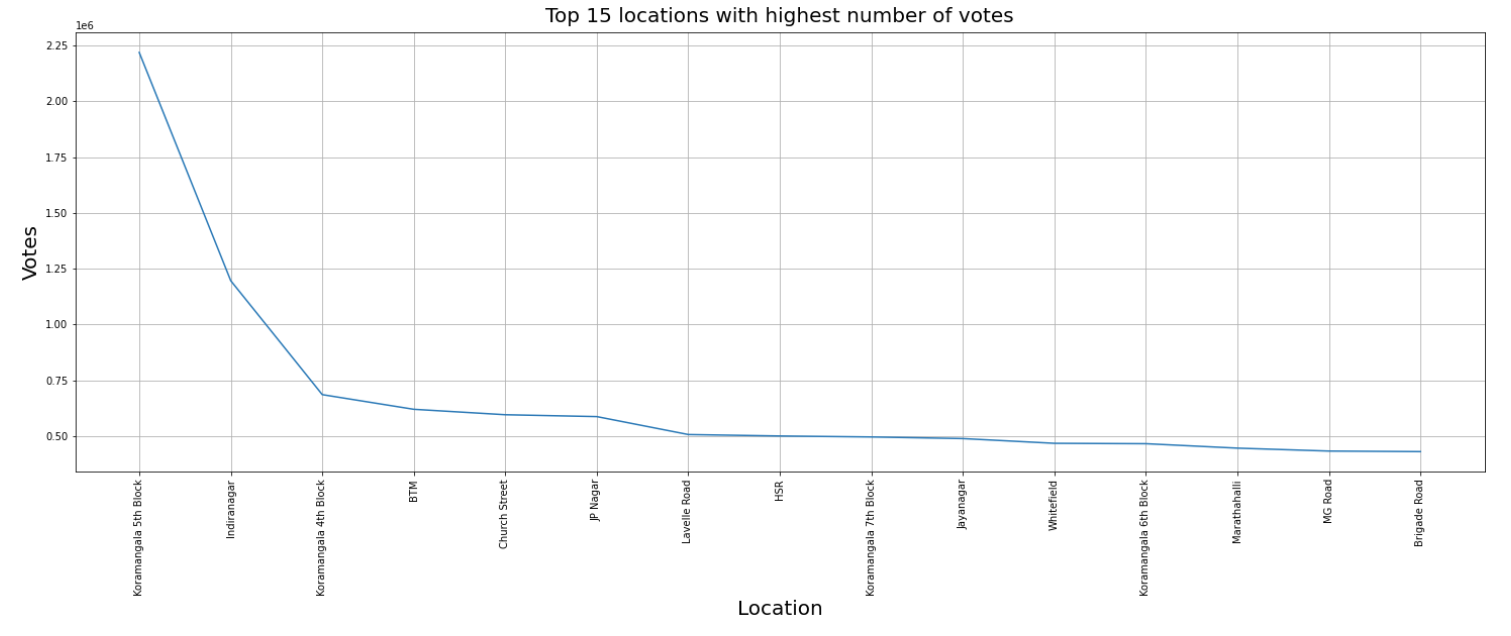
Let's refine the data further for better insights.

### How many unique restaurants does Bangalore have?



Since the Resturant names are duplicated in the dataset so we tried to find the unique restuarants in the location to analyse the data further.

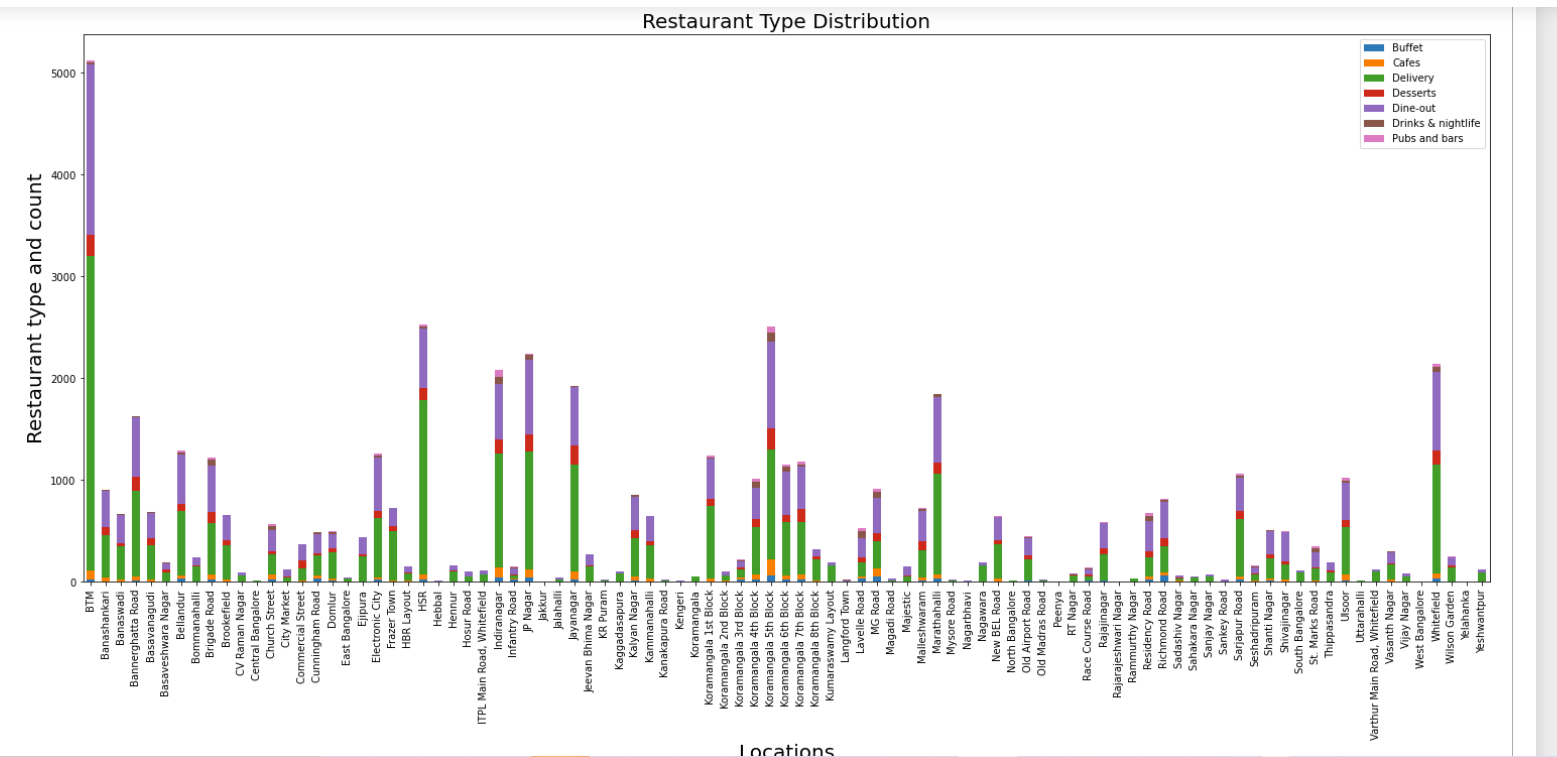
### Where are the good restaurants located?



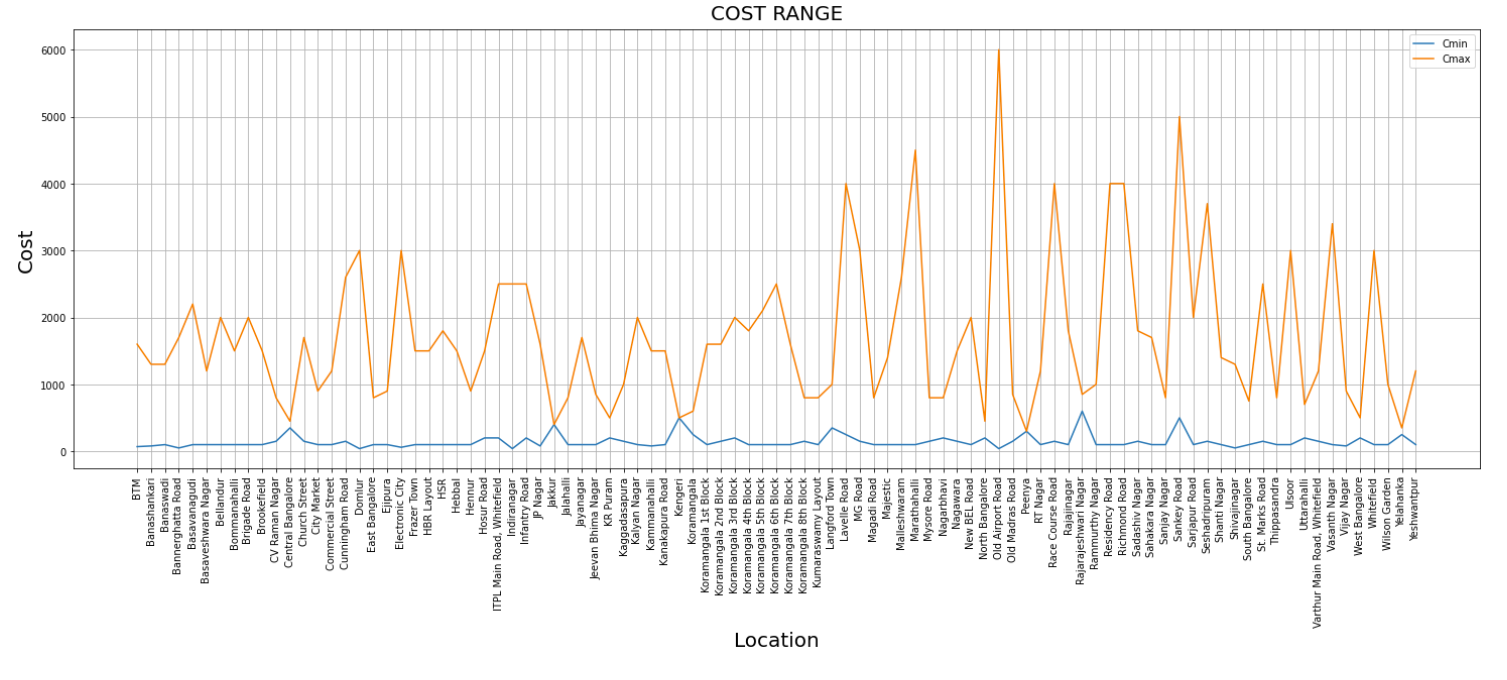
While opening a new restaurant, if initial investment is not much of a concern for you and you are looking to increase your chances to maximum, then why don't go with one of the best locations loved by zomato customers.

From the barchart, we can clearly see Koramangala dominates this scenario. Koramangala is the heart of Bangalore and attracts nearly all foodies. It might take heavy initial investment based upon it's popularity but it seems like a sure shot if you deliver well.

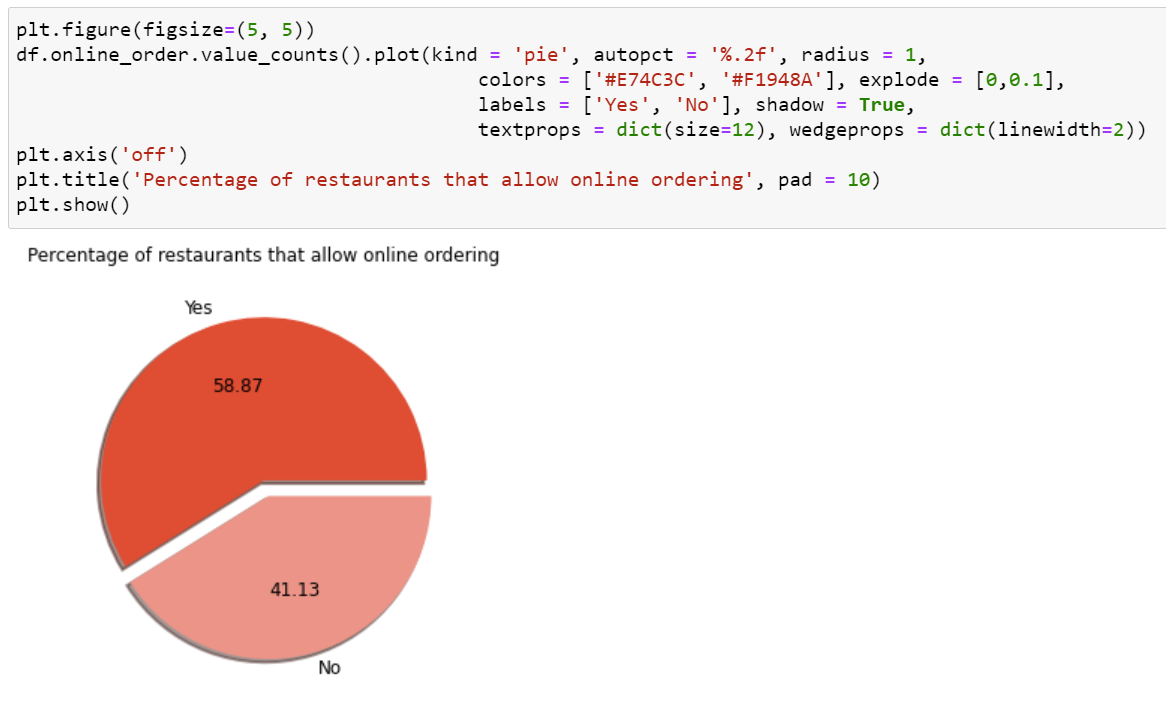
### What is the distribution of types of restaurants across different locations?



### What is the approximate spending in a particular area for two people?



### How do table booking and order online facilities impact the rating?





From the above Pie chart we can see that 87.5% of the restaurants have table booking facility available and 12.5% dont have table booking available.

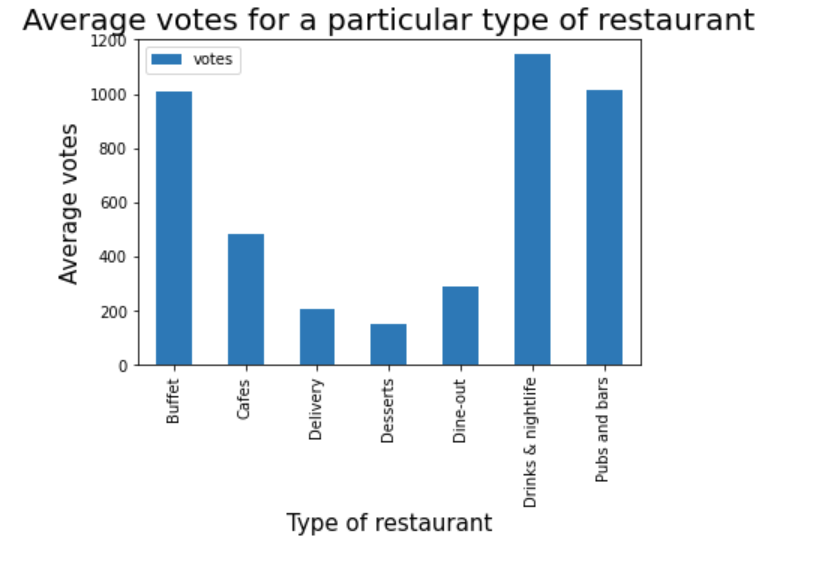
Now lets check how to rating affects the restaurant's overall rating based on these two services



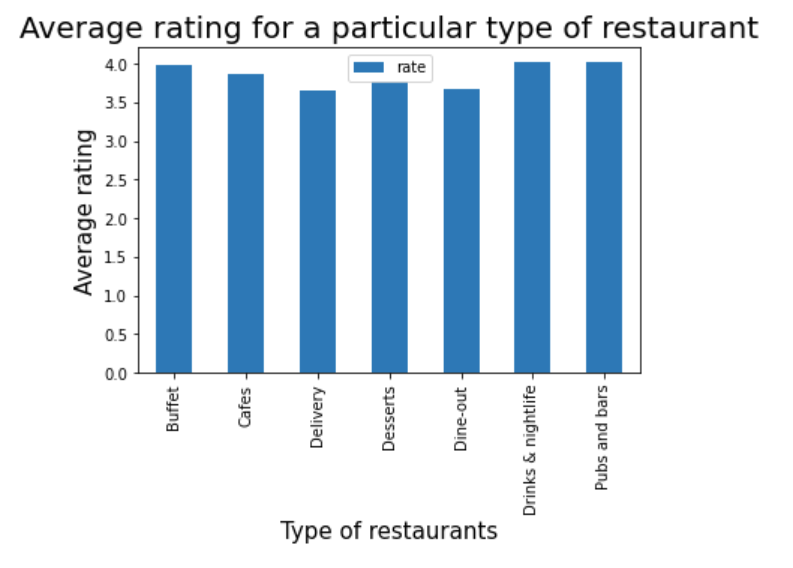
1 Refers to **Yes** and 0 refers to **No**

It can be observed that having the possibility to order online doesn't make a huge difference in the ratings. This might occur since there are only 10% more restaurants that have this option so the impact is not that relevant. On the other hand, restaurants that permit booking tables are much less, so they generally have a better rating since this option is much more appreciated.

### Does particular type of restaurants get more votes than other type?

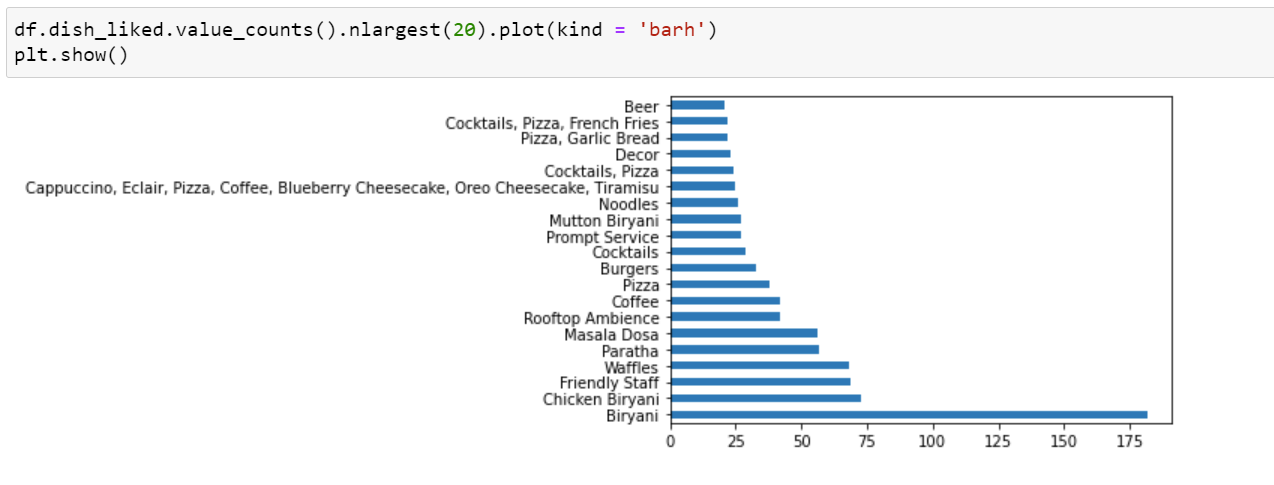


### Does particular type of restaurants get more rating than other type?



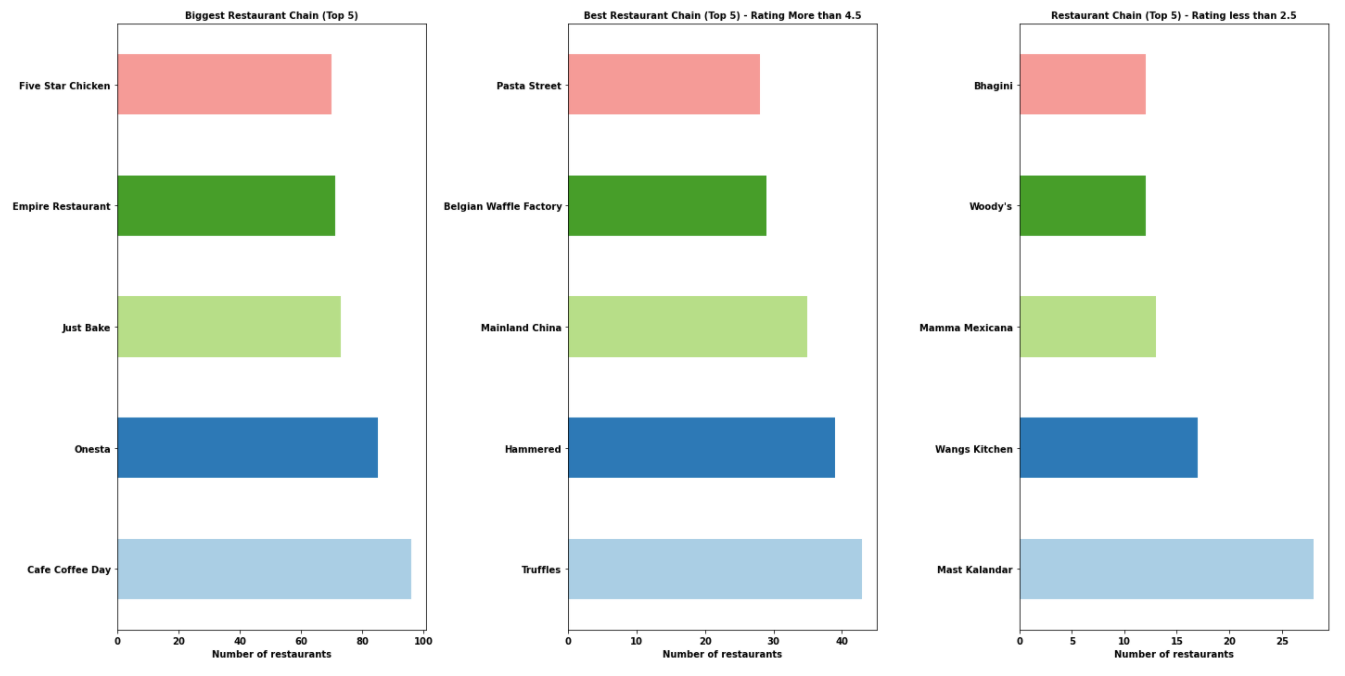
The above graph helps us to understand the customer who rates who dines or orders from a specific kind of restaurant to ensure its accuracy.

### What are the dishes enjoyed by people the most?



The graph above displays the top 20 dishes liked by people. Clearly, Biryani gains the top most position as compared to the rest of the dishes

### **Comparing Biggest Restaurant Chain and Best Restaurant Chain**

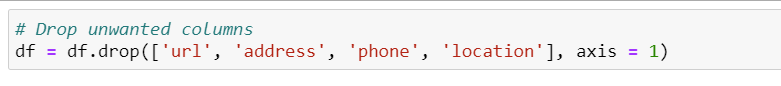


Cafe Coffee Day chain has over 90 cafes across the city that are listed in Zomato. On the other hand, Truffles - a burger chain has the best fast food restaurant (rating more than 4.5 out of 100), quality over quantity.

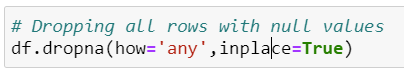
If you visit Banglore next time and if you want to check out a good restaurant over a weekend dont forget to try the food at Truffles, Hammered and Mainland China.

**Data Preprocessing**

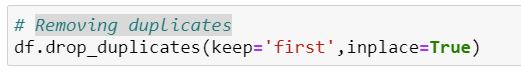
1. Dropping the unwanted columns like ‘url’,’address’,’phone’,’location’ which doesn’t have any null values and are highly correlated.



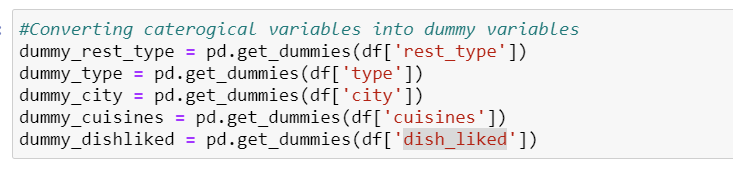
1. Dropping all rows with null values



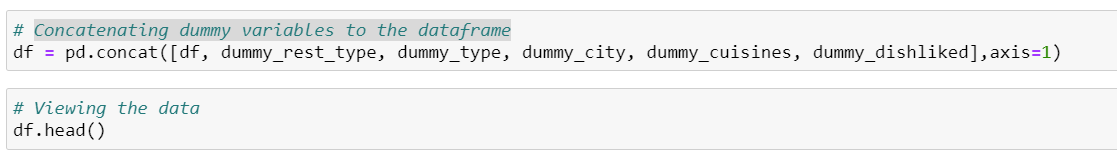
1. Removing duplicates



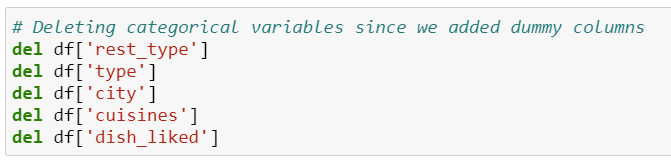
1. Converting caterogical variables(rest\_type, type, city, cuisines, dish\_liked) into dummy variables as one hot encoding.



1. Concatenating dummy variables to the dataframe



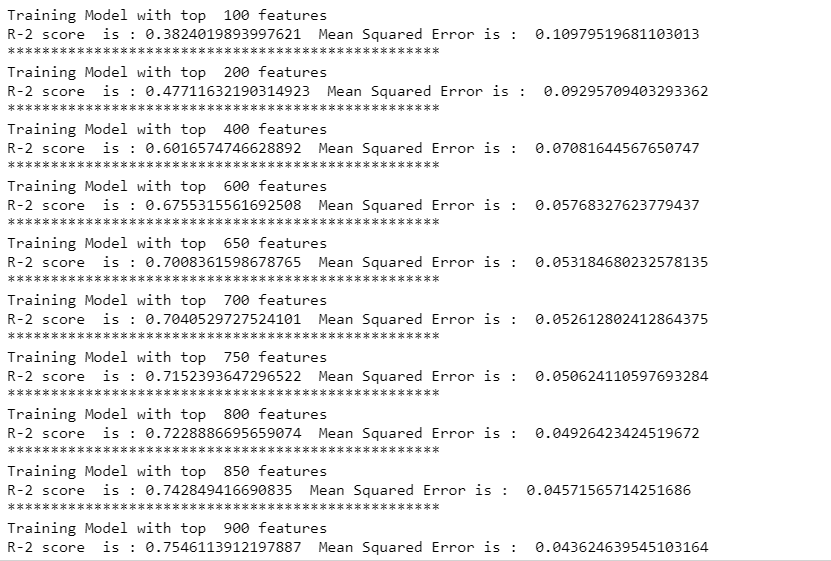
1. Deleting categorical variables since we added dummy columns



The highest correlation 0.63 is between name and address and the lowest -0.44 between book\_table and approx\_cost\_2\_people. Highly correlated features maybe imply redondant information. So we might drop one the them. The correlations that will interest us at this stage are between our target rate and the other features. The more a feature is correlated to the target (positively or negatively) the more it could help predicting the target. And from the plot, one can see that the 2 top positively correlated features to the target are dish\_liked, vote. Which means the more a dish is liked and a restaurant receive higher votes the more the rate increase. One can also note from the plot that only one feature book\_table is negatively correlated to the target. Which means that restaurant in which booking table is needed have low rate.

Initialising Standardscalar to Centering and scaling happen independently on each feature by computing the relevant statistics on the samples in the training set. Mean and standard deviation are then stored to be used on later data using [transform](https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html#sklearn.preprocessing.StandardScaler.transform). NaNs are treated as missing values: disregarded in fit, and maintained in transform.

**To find out the best features to apply in the models for prediction**

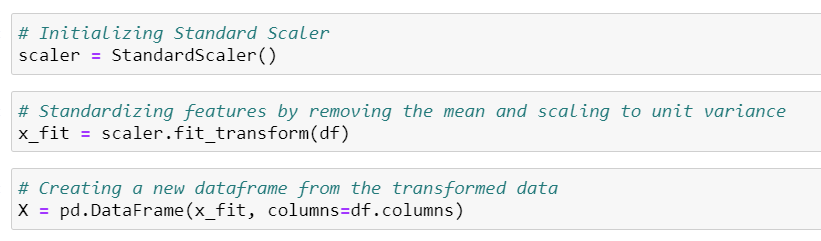


Choosing 800 features based on the above mean squared error and r2 value saturating at that point.

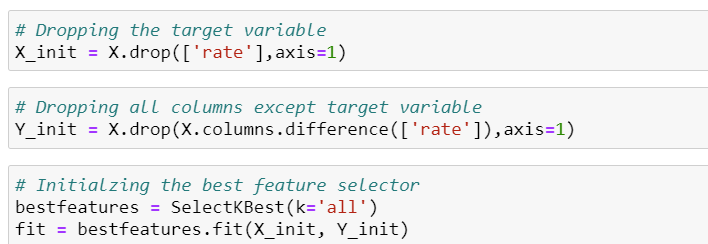
**plotting features v MSE loss**

**A picture containing text, shoji

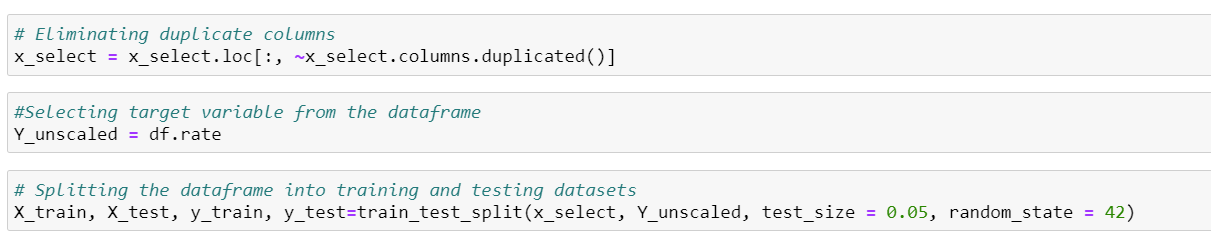
Description automatically generated**



Selecting all columns except target value in X\_init to train the model. Selecting the target column ‘rate’ to Y\_init.



* Deleting the duplicate columns after fitting the dish\_liked column in one hot encoding to x\_select .
* seleting the target variable to Y\_unscaled.
* Later, spliting the dataframe to test data and train data.

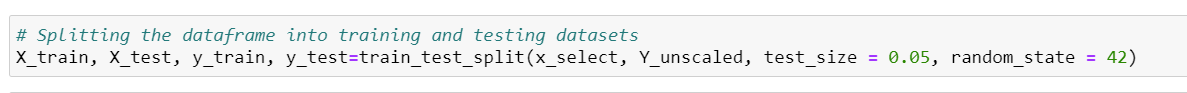


**Predict the Restaurant rating**

**Training and Testing Data**

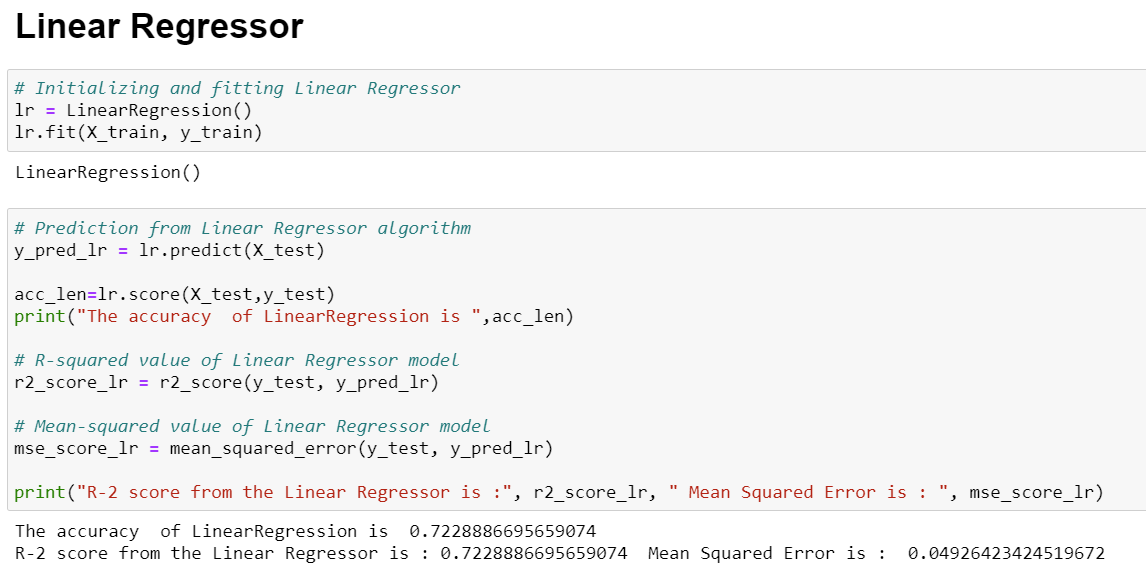
Data is separated in Training set as well as Testing set in ratio 10:2 to 10:3

And then classification models are applied on it

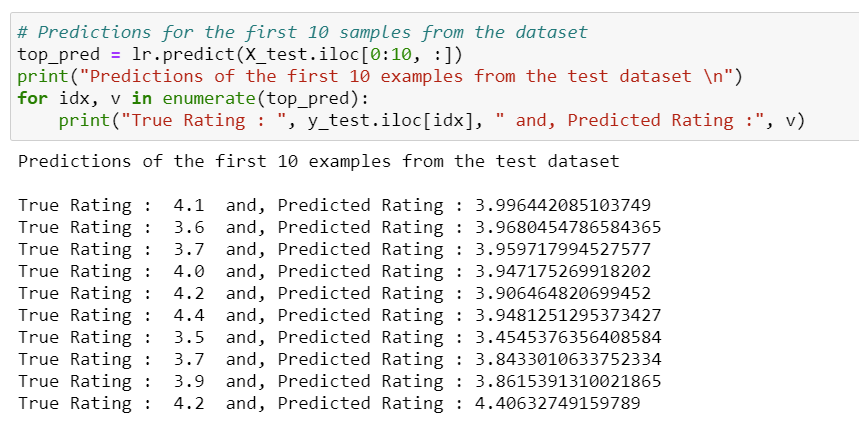


Applying and Analyzing all models

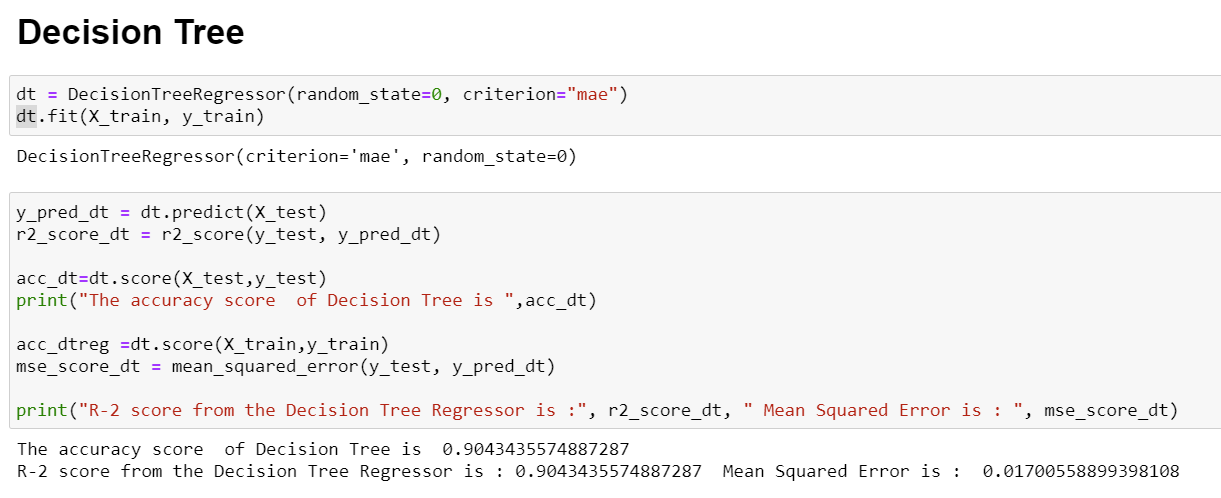
1. **Linear Regression:**

****

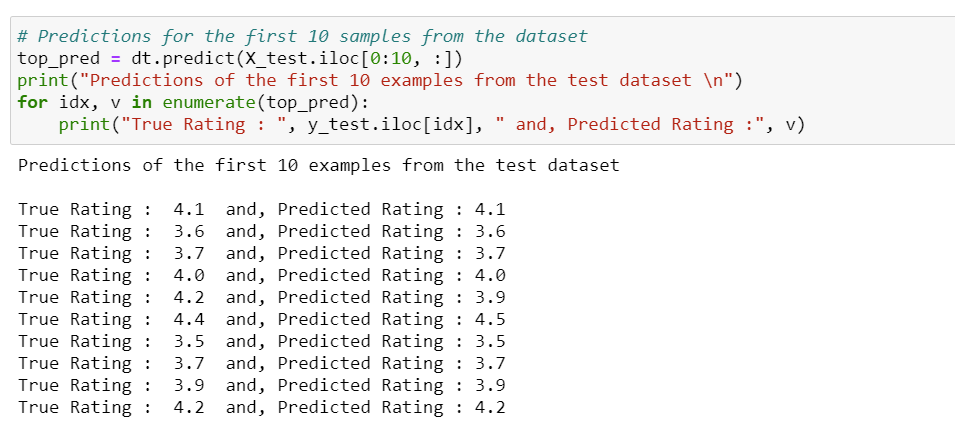
**Result :**The accuracy score of the linear Regression is 72% and not the accurate model for the dataset.



1. **Decision Tree:**

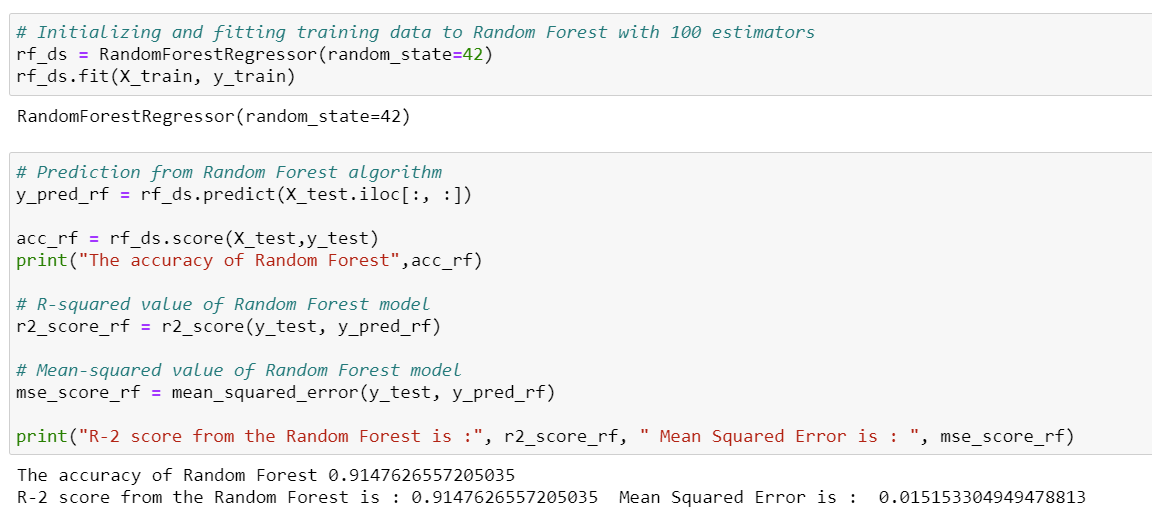


**Result:** We achived a Accuracy score for Decision tree is 90% which is considered as a good score.

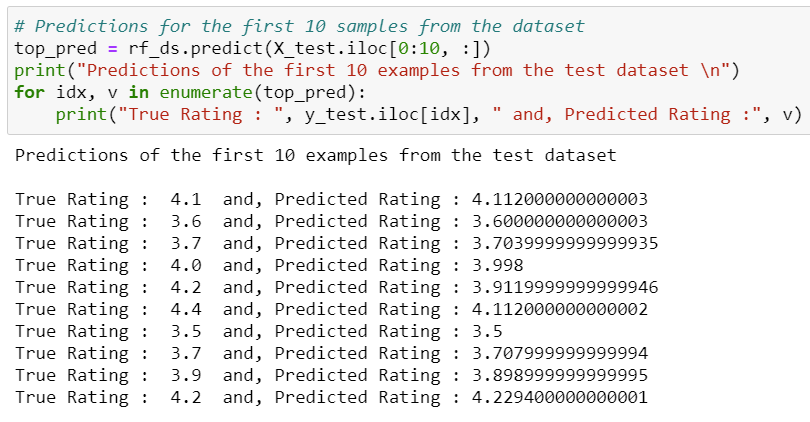


Displying the first 10 dataset with the predicted value and true dataset to figure out the prediction.

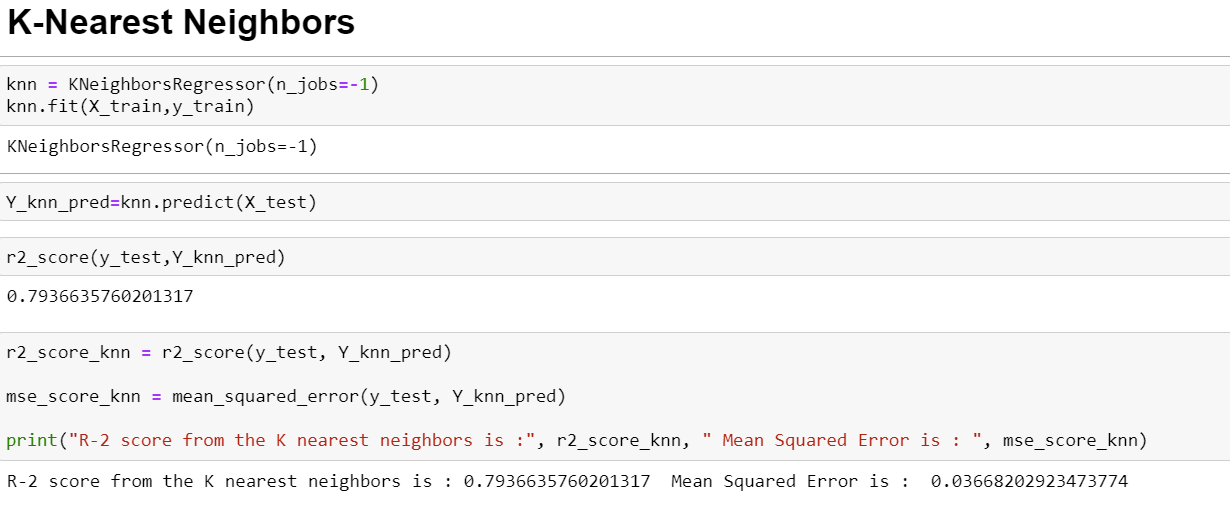
1. **Random Forest Regression**



**Result:**  The model provided the accuracy of 91% which is good score so far and the below are the 10 sample predicted values and true values.



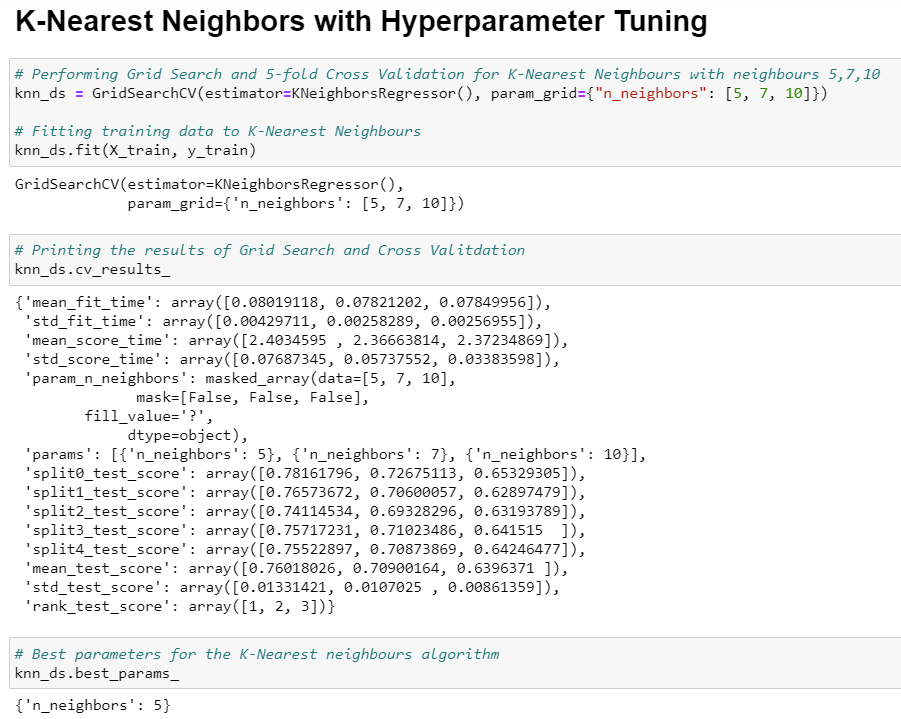
1. K-Nearest Neighbour Regression:



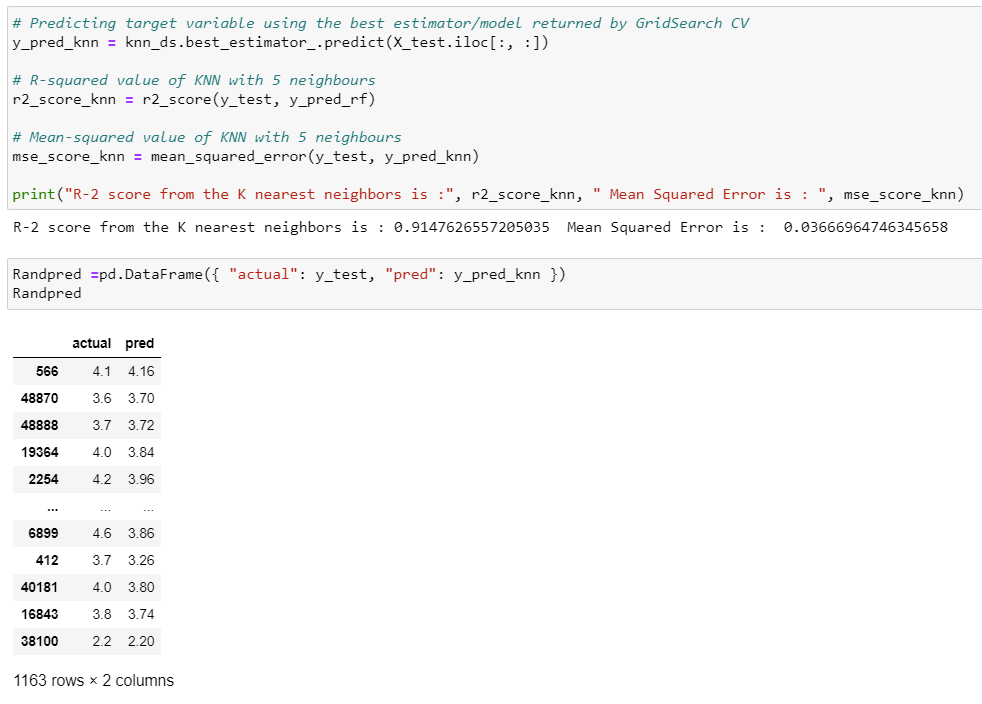
**Results :** The above model resulted in accuracy score of 79% which is low so and we applied hyperpamater to figureout if it could increase the accuracy score of the model.

**KNN Model with Hyperparameter tuning**

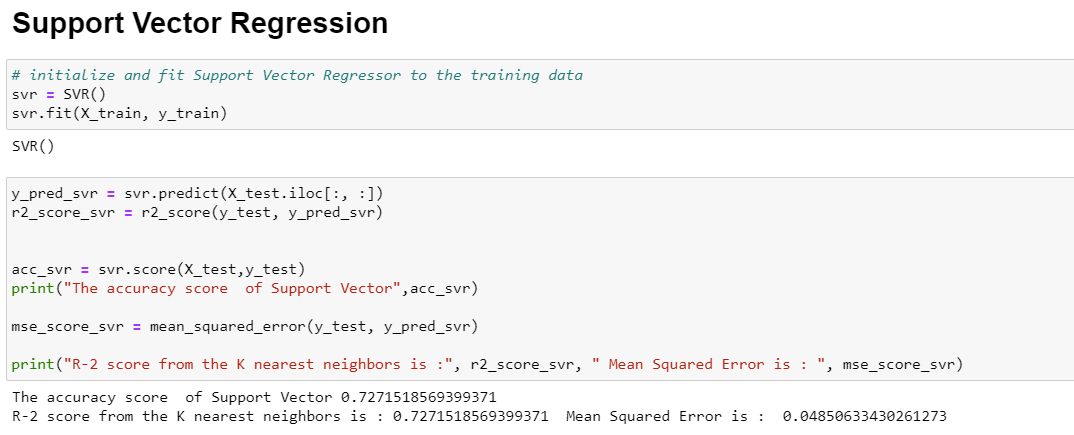
The hyperparameter is applied on n\_neighbours with values 5,7,10. After the GridsearchCV is applied it resulted in picking 5 as the best nearest neighbour.



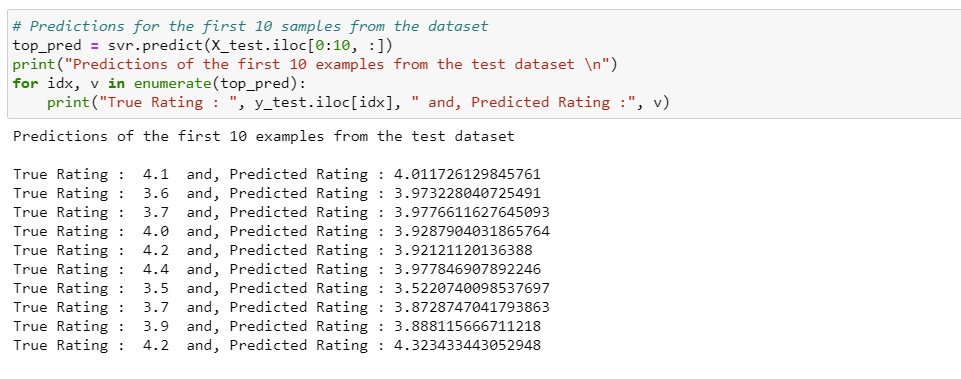
**Result:** The accuracy is predicted after applying the hyperparameter turning using GridsearchCV and the accuracy increased to 91% which is really good after the tuning method is applied.



1. **Support Vector Regression**

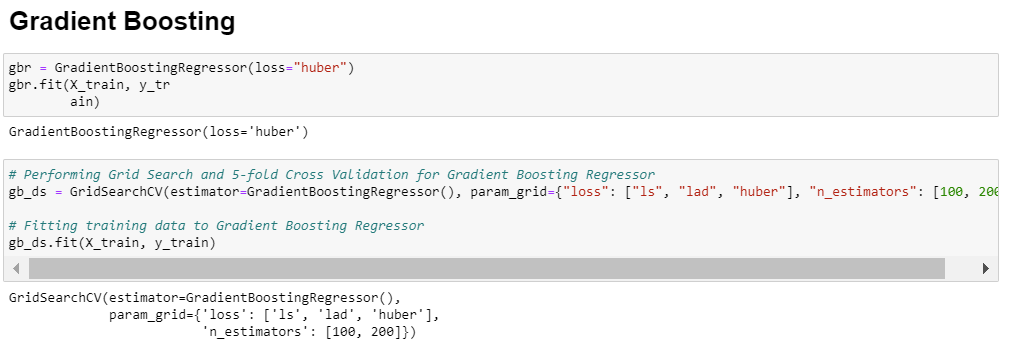


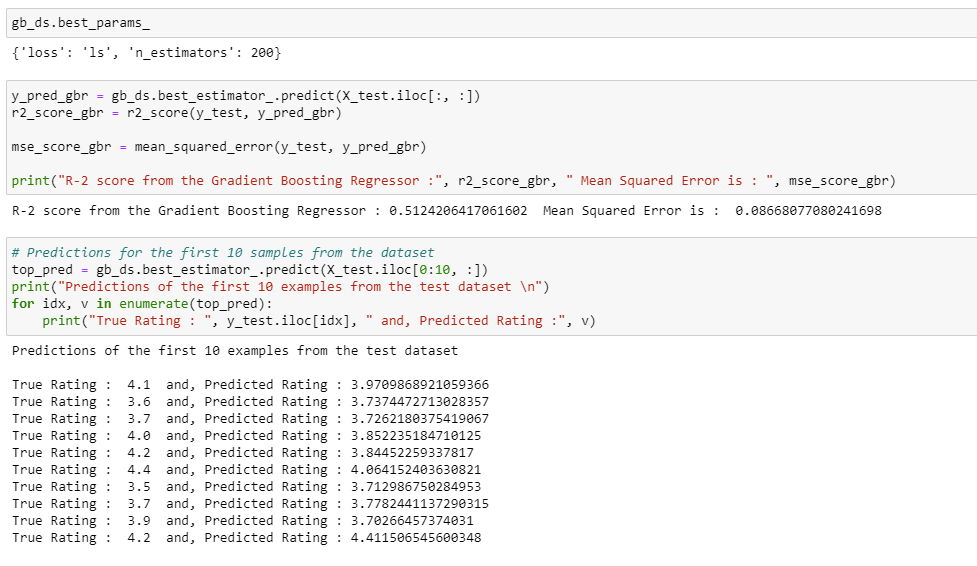
**Results:** The accuracy score of support vector regression is 72% which is least among the other models.



The above results are the ten predicted and true values of the 10 samples in the dataset.

1. **Gradient Boosting Regressor**

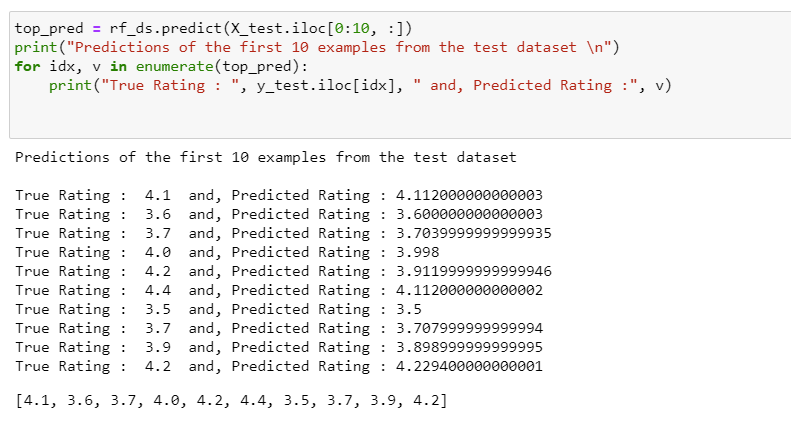




**Result** : The accuracy of the Gradient Boosting is 51% and its not a conventional method for the dataset and displayed the samples of datset with predicted and true values.

**A Comparision of the actual rating versus the rating predicted by the Random Forest model**

**True Values:**



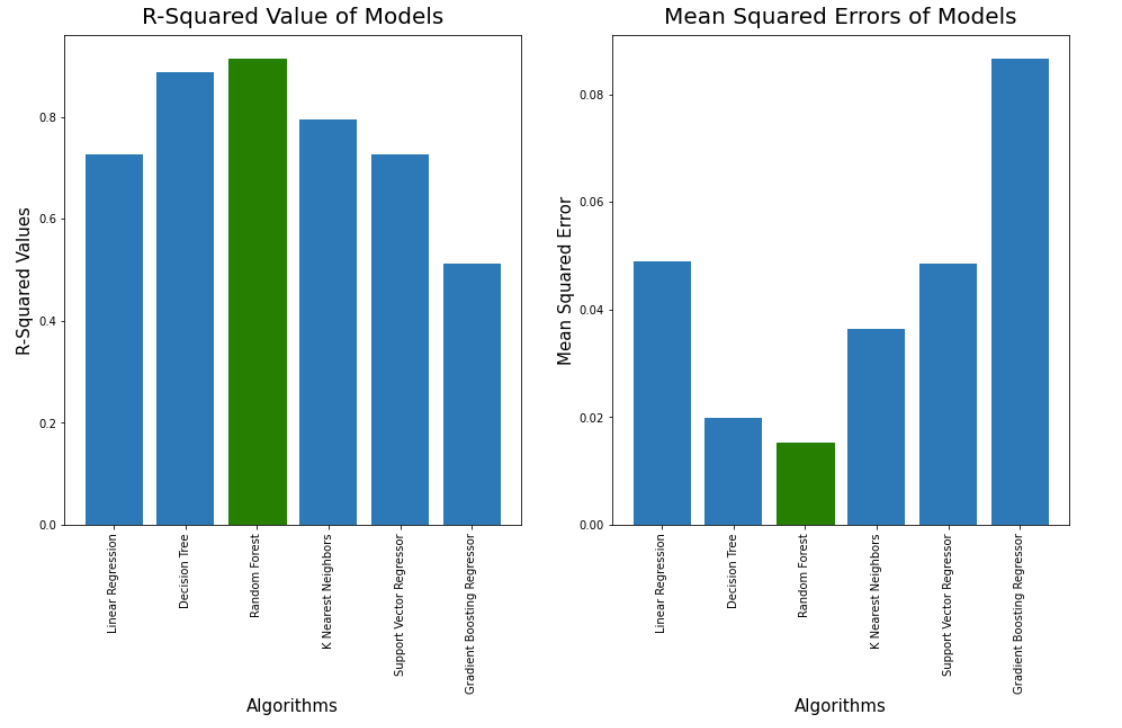
**Actual Rating prediction after applying Random forest regression:**

**Chart, bar chart

Description automatically generated**

**Conclusion :**

From the above models applied we conclude that **Random Forest Regression** gave the maximum accuracy of 91% along with KNN regression with hyperparameter tuning.The lowest among them was gradient boosting with 51% which is not conventional method for the chosen dataset.



The better model can be chosen based on how close the mean square error to 0. In the above applied machine learning models its evident that Random Forest has the lowest mean squared error and its close to 0 than any other models.

This concludes that

* The most famous cuisine in banglore is North Indian Cuisine.
* BTM layout has the most number of restaurant.
* We analyzed totally 8792 resturants reviews in the used dataset.
* Based on the votes and reviews provided we can conclude Koramangla 5th block has the best rated restaurants.
* On an average it costs 1500-2000 Indian rupees for two people dining, online order or take away experience through Zomato.
* Almost 59% restaurant undertake order from Online booking and the rest has only walk-in or take away facilities.
* Only 12% of the restaurants allow table booking facility, which is really important moving forward when we have busy schedule or for a popular restaurant.
* And Biriyani is the most favourite dish for people in Banglore. So any new resturants which is planning to open should consider having this dish.

These are the above analysis that one should consider before opening any restaurant in banglore location and the kind of cuisine that should they offer to customers.

After going trough all the tasks we can conclude that it was possible to predict the success (in terms of rate) for new restaurants put in Zomato service. The goal of this kernel is to predict the rate of the restaurants in Bangalore. We tried different models and different feature engineering. The best model we obtained is the random forest regressor which outperfomed boosting models (big surprise!).

This is very interesting for business areas to evaluate the main concepts of restaurants and customers preferences in Bengaluru region. With this information, it's clearly possible to take a look at restaurant's features before launching it! The business men and the customers would appreciate it!

**References**

* https://www.kaggle.com/himanshupoddar/zomato-bangalore-restaurants
* https://towardsdatascience.com/exploratory-data-analysis-with-pandas-profiling-de3aae2ddff3 <https://scikit-> learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html
* <https://stackoverflow.com/>
* International Journal of Computer Applications Technology and Research Volume 8–Issue 09, 375-378, 2019, ISSN:-2319–8656
* [Decision Tree Regression With Hyper Parameter Tuning In Python (nbshare.io)](https://www.nbshare.io/notebook/312837011/Decision-Tree-Regression-With-Hyper-Parameter-Tuning-In-Python/)
* <https://scikit-learn.org/stable/>
* <https://matplotlib.org/>