SCTR's Pune Institute of Computer Technology
Dhankawadi, Pune

DSBDAL MINI PROJECT ON

Zomato Rating Prediction

SUBMITTED BY

Neeraj Bukane 33114
Sanyog Kalantri 33137
Samyak Jain 33169
Gauri Takalikar 33178

Under the guidance of

Dr. Anant Bagade

DEPARTMENT OF INFORMATION TECHNOLOGY

ACADEMIC YEAR 2022-23

Contents

1 Abstract

With the increasing demand for online food delivery services, Zomato has emerged as one of the leading platforms in India, providing users with restaurant ratings and reviews. This project aims to analyze the Zomato dataset to gain insights into the factors influencing the establishment of different types of restaurants in Bengaluru, focusing on the city's diverse culinary landscape.

The dataset contains valuable information, including restaurant ratings, reviews, location, price, and cuisine, which are crucial factors in determining a restaurant's popularity. By exploring these features, we can understand the dynamics of Bengaluru's restaurant industry and identify opportunities for new establishments to thrive.

Bengaluru, as the IT capital of India, witnesses a continuous influx of people who heavily rely on restaurant food due to time constraints. However, the market remains highly competitive, with numerous restaurants serving similar cuisine. Therefore, understanding the local demography and culinary preferences becomes crucial in gaining a competitive edge.

The analysis will delve into various aspects, such as the impact of location on restaurant success, the influence of price range on customer choices, the prevalence of theme-based restaurants, and the distribution of cuisines across different localities. Furthermore, the project aims to identify neighborhoods known for specific types of cuisine and cater to the unique needs of the residents.

By leveraging the real-time Zomato dataset, encompassing data up to March 15, 2019, we will perform comprehensive data mining and analysis techniques. This project will provide valuable insights to restaurant owners, aspiring entrepreneurs, and researchers, helping them make informed decisions regarding the establishment, menu offerings, and target audience of restaurants in Bengaluru.

2 Introduction

The restaurant industry in Bengaluru, often referred to as the Silicon Valley of India, is a vibrant and ever-growing sector. With a population that heavily relies on restaurant food due to time constraints, the demand for dining options continues to rise. In this highly competitive market, it has become increasingly challenging for new restaurants to establish themselves and compete with well-established establishments. The ability to understand and analyze the factors influencing restaurant success has become crucial for entrepreneurs and industry stakeholders.

The Zomato dataset provides a wealth of information that can offer valuable insights into the establishment of different types of restaurants in Bengaluru. By leveraging this dataset, which includes ratings, reviews, location details, pricing information, and cuisine preferences, we can gain a comprehensive understanding of the dynamics at play in the local restaurant landscape. This analysis aims to identify key factors that contribute to a restaurant's popularity and explore the culinary preferences of different neighborhoods within the city.

3 Problem Statement

The restaurant industry in Bengaluru is highly dependent on customer ratings and reviews, which significantly impact a restaurant's reputation and success. Therefore, the focus of this project is to develop a rating prediction model using the Zomato dataset to accurately predict the ratings of restaurants in Bengaluru.

Utilizing the available features such as location, pricing, cuisine, and other relevant factors from the Zomato dataset, we will develop a machine learning model that can accurately predict the ratings of restaurants. The model will consider the relationships between different features and their influence on the overall rating.

Ultimately, by developing an accurate rating prediction model this project aims to assist restaurant owners in Bengaluru in enhancing their offerings, customer satisfaction, and overall success in a highly competitive market.

4 Objectives and Scope

4.1 Objectives

We aim to achieve the following objectives:

1. **Build a reliable and robust rating prediction model:** Utilizing the available features such as location, pricing, cuisine, and other relevant factors from the Zomato dataset, we will develop a machine learning model that can accurately predict the ratings of restaurants. The model will consider the relationships between different features and their influence on the overall rating.

2. **Identify key features that impact restaurant ratings:** Through the rating prediction model, we can analyze the importance and impact of various factors on the overall rating. This analysis will provide valuable insights into the elements that customers consider while rating a restaurant and help identify areas for improvement.

3. **Evaluate the performance of the rating prediction model:** We will assess the accuracy and effectiveness of the developed model by comparing its predicted ratings with the actual ratings from the Zomato dataset. The evaluation will enable us to gauge the model's performance and its ability to generalize to new, unseen data.

4. **Provide actionable recommendations for improving restaurant ratings:** By understanding the factors that significantly influence ratings, we can offer actionable recommendations to restaurant owners and stakeholders on how to improve their ratings. These recommendations may include aspects such as service quality, menu offerings, pricing strategies, customer engagement, and overall dining experience enhancements.

4.2 Scope

The scope of this project encompasses the analysis of the Zomato dataset to understand the factors influencing restaurant establishment and cuisine preferences in Bengaluru, with a specific focus on rating prediction. The project will involve the following activities:

1. **Data Extraction and Preprocessing:** The Zomato dataset will be extracted and preprocessed to ensure data quality and consistency. This includes handling missing values, data cleaning, and feature engineering.
2. **Exploratory Data Analysis:** An exploratory analysis will be conducted to gain insights into the dataset, understand the distribution of variables, and identify patterns and trends related to restaurant establishments, cuisines, and ratings in Bengaluru.
3. **Feature Selection and Engineering:** Relevant features will be selected from the dataset for rating prediction. Additionally, new features may be created or derived to enhance the model's predictive capability.
4. **Rating Prediction Model Development:** Machine learning algorithms will be employed to develop a rating prediction model. The model will be trained using a suitable regression technique and optimized to achieve accurate rating predictions.
5. **Model Evaluation and Validation:** The developed rating prediction model will be evaluated and validated using appropriate metrics and cross-validation techniques. The model's performance will be assessed to ensure its accuracy and reliability.
6. **Interpretation and Insights:** The analysis results will be interpreted to understand the impact of different factors on restaurant ratings in Bengaluru. Insights and recommendations will be provided to restaurant owners based on the findings to improve their ratings and overall customer satisfaction.

The project specifically focuses on the rating prediction aspect, while also considering other relevant factors such as location, pricing,

cuisine offerings, and customer reviews. However, it is important to note that the project's scope does not include real-time data collection or integration with the live Zomato platform.

The project is carried out using the Zomato dataset available until March 15, 2019. The analysis and recommendations provided are based on the data available within this timeframe.

The scope of this project is limited to Bengaluru, and the findings may not be directly applicable to other cities or regions.

5 Feature Description

1. **url** contains the url of the restaurant in the zomato website
2. **address** contains the address of the restaurant in Bengaluru
3. **name** 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 rating for the restaurant as of the above mentioned date
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
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 listed

6 Implementation Details

The steps followed are:

6.1 Importing the libraries:

1. Start by importing the necessary libraries for data analysis and machine learning, such as Pandas, NumPy, Matplotlib, and scikit-learn.
2. Install and import any additional libraries required for specific tasks, such as seaborn for enhanced data visualization.

6.2 Data Cleaning:

1. Load the Zomato dataset into a Pandas DataFrame.
2. Perform data cleaning tasks, such as:
   (a) Handling missing values: Identify columns with missing values and decide how to handle them (e.g., imputation or removal).
   (b) Removing duplicates: Check for and remove any duplicate rows in the dataset.
   (c) Addressing inconsistencies: Look for any inconsistencies or errors in the data and correct them.
3. Convert data types if necessary and ensure data consistency and integrity.

6.3 Data Visualization:

1. Conduct exploratory data analysis (EDA) to gain insights into the dataset.
2. Utilize data visualization techniques to visualize distributions, correlations, and trends in the data.
3. Generate plots, histograms, scatter plots, box plots, or heatmaps to visualize relationships and identify patterns among variables.
4. Analyze the distribution of ratings, explore the impact of different features on ratings, and identify any significant trends or patterns.

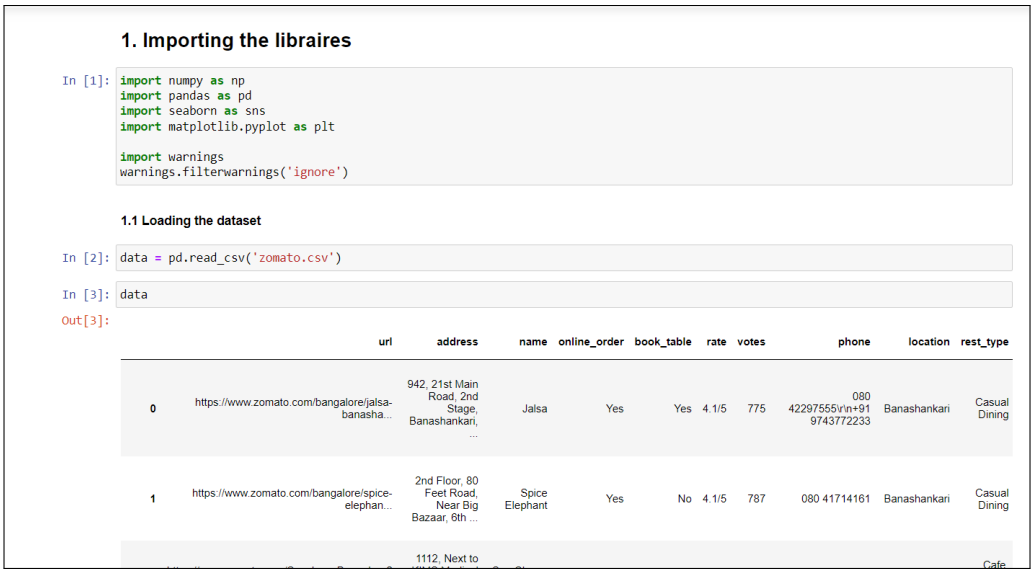
6.4 Data Preparation:

1. Preprocess the data for modeling by transforming categorical variables into numerical representations, such as one-hot encoding or label encoding.
2. Split the dataset into training and testing subsets to evaluate the model's performance.
3. Normalize or scale numerical features if necessary to ensure fair comparisons between different features.

6.5 Modeling:

1. Select an appropriate machine learning algorithm for rating prediction, such as linear regression, decision trees, random forests, or gradient boosting.
2. Train the chosen model on the training dataset and evaluate its performance on the testing dataset using suitable evaluation metrics, such as mean squared error (MSE) or R-squared.
3. Perform hyperparameter tuning to optimize the model's performance, using techniques like grid search or random search.
4. Validate the final model using cross-validation techniques to assess its robustness and generalizability.
5. Interpret the model's results and identify the most important features contributing to rating prediction.

7 Output



1. Importing the libraires

In [1]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

import warnings
warnings.filterwarnings('ignore')
```

1.1 Loading the dataset

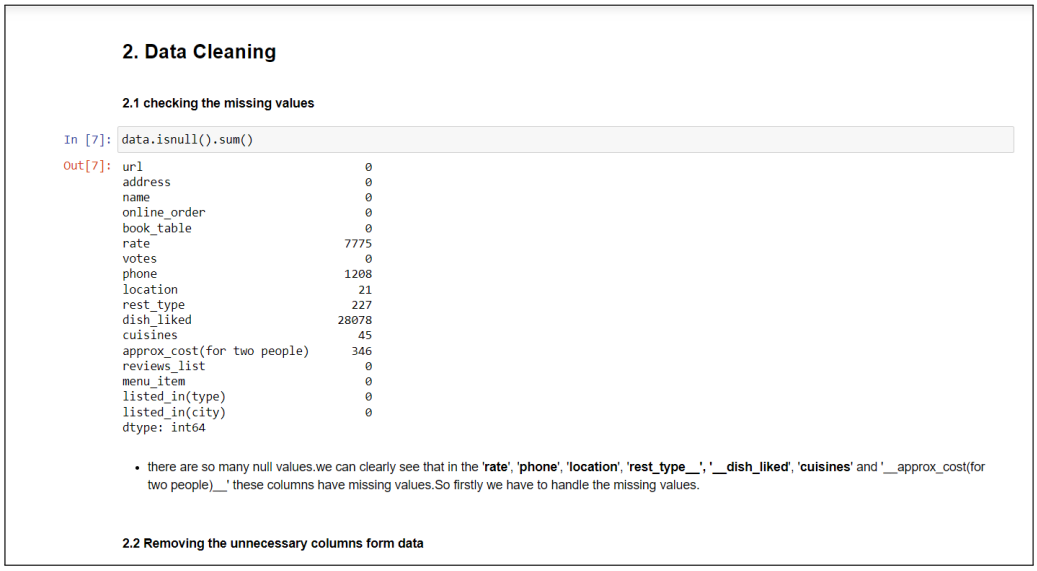
In [2]:
```
data = pd.read_csv('zomato.csv')
```

In [3]:
```
data
```

Out[3]:

	url	address	name	online_order	book_table	rate	votes	phone	location	rest_type
0	https://www.zomato.com/bangalore/jalsa-banasha...	942, 21st Main Road, 2nd Stage, Banashankari, ...	Jalsa	Yes	Yes	4.1/5	775	080 42297555\r\n+91 9743772233	Banashankari	Casual Dining
1	https://www.zomato.com/bangalore/spice-elephan...	2nd Floor, 80 Feet Road, Near Big Bazaar, 6th ...	Spice Elephant	Yes	No	4.1/5	787	080 41714161	Banashankari	Casual Dining
		1112, Next to								Cafe

Figure 1: Importing the libraries



2. Data Cleaning

2.1 checking the missing values

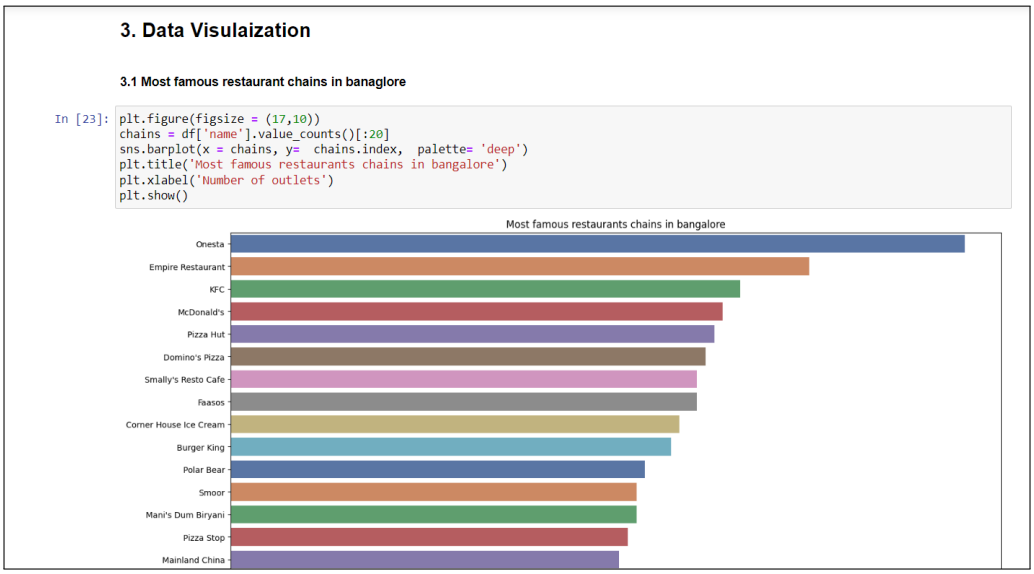
In [7]:
```
data.isnull().sum()
```

Out[7]:
```
url                            0
address                        0
name                           0
online_order                   0
book_table                     0
rate                        7775
votes                          0
phone                       1208
location                      21
rest_type                    227
dish_liked                 28078
cuisines                      45
approx_cost(for two people)  346
reviews_list                   0
menu_item                      0
listed_in(type)                0
listed_in(city)                0
dtype: int64
```

- there are so many null values.we can clearly see that in the '**rate**', '**phone**', '**location**', '**rest_type__**', '**__dish_liked**', '**cuisines**' and '__approx_cost(for two people)__' these columns have missing values.So firstly we have to handle the missing values.

2.2 Removing the unnecessary columns form data

Figure 2: Data Cleaning

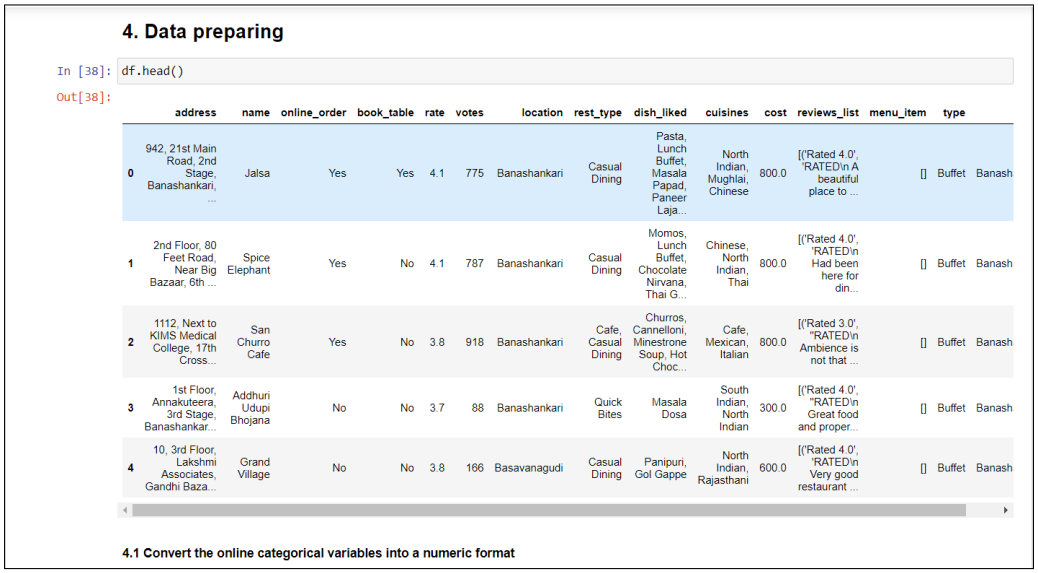


3. Data Visulaization

3.1 Most famous restaurant chains in banaglore

In [23]:
```
plt.figure(figsize = (17,10))
chains = df['name'].value_counts()[:20]
sns.barplot(x = chains, y=  chains.index,  palette= 'deep')
plt.title('Most famous restaurants chains in bangalore')
plt.xlabel('Number of outlets')
plt.show()
```


Figure 3: Data Visualization



4. Data preparing

In [38]:
```
df.head()
```

Out[38]:

	address	name	online_order	book_table	rate	votes	location	rest_type	dish_liked	cuisines	cost	reviews_list	menu_item	type	
0	942, 21st Main Road, 2nd Stage, Banashankari, ...	Jalsa	Yes	Yes	4.1	775	Banashankari	Casual Dining	Pasta, Lunch Buffet, Masala Papad, Paneer Laja...	North Indian, Mughlai, Chinese	800.0	[('Rated 4.0', 'RATED\n A beautiful place to ...	[]	Buffet	Banash
1	2nd Floor, 80 Feet Road, Near Big Bazaar, 6th ...	Spice Elephant	Yes	No	4.1	787	Banashankari	Casual Dining	Momos, Lunch Buffet, Chocolate Nirvana, Thai G...	Chinese, North Indian, Thai	800.0	[('Rated 4.0', 'RATED\n Had been here for din...	[]	Buffet	Banash
2	1112, Next to KIMS Medical College, 17th Cross...	San Churro Cafe	Yes	No	3.8	918	Banashankari	Cafe, Casual Dining	Churros, Cannelloni, Minestrone Soup, Hot Choc...	Cafe, Mexican, Italian	800.0	[('Rated 3.0', 'RATED\n Ambience is not that ...	[]	Buffet	Banash
3	1st Floor, Annakuteera, 3rd Stage, Banashankar...	Addhuri Udupi Bhojana	No	No	3.7	88	Banashankari	Quick Bites	Masala Dosa	South Indian, North Indian	300.0	[('Rated 4.0', 'RATED\n Great food and proper...	[]	Buffet	Banash
4	10, 3rd Floor, Lakshmi Associates, Gandhi Baza...	Grand Village	No	No	3.8	166	Basavanagudi	Casual Dining	Panipuri, Gol Gappe	North Indian, Rajasthani	600.0	[('Rated 4.0', 'RATED\n Very good restaurant ...	[]	Buffet	Banash

4.1 Convert the online categorical variables into a numeric format

Figure 4: Data Preparing

5 Modeling

5.1 Linear Regression

In [52]:
```
from sklearn.linear_model import LinearRegression
lr = LinearRegression()
lr.fit(X_train,y_train)

#predict the test set
y_pred = lr.predict(X_test)

## Evaluate the model
from sklearn.metrics import r2_score
print(r2_score(y_test, y_pred))
```

```
0.2281882852296726
```

In [ ]:

5.2 Decision Tree Regression

In [53]:
```
from sklearn.tree import DecisionTreeRegressor
dtr = DecisionTreeRegressor(min_samples_leaf=0.01)

dtr.fit(X_train,y_train)

# Predict the test ser
y_pred  = dtr.predict(X_test)

# Evaluate the model performance
from sklearn.metrics import r2_score
```

Figure 5: Modeling

8 Conclusion

Based on the analysis of the Zomato dataset, several key findings and insights have been obtained:

1. **Popular Restaurants:** The analysis revealed that 'Onesta', 'Empire Restaurant', and 'KFC' are the most famous restaurants in Bangalore.
2. **Online Ordering and Delivery:** Most restaurants offer options for online ordering and delivery, indicating the increasing trend of food delivery services.
3. **Table Booking:** The majority of restaurants do not offer table booking services, suggesting that walk-in customers are more common.
4. **Rating Distribution:** The analysis showed that most ratings fall within the range of 3.5 to 4.5, indicating that customers generally have positive experiences at Bangalore restaurants.
5. **Restaurant Locations:** The top locations with the highest concentration of restaurants are 'Koramangala 5th Block', 'BTM', and 'Indiranagar', while 'KR Puram', 'Kanakapura', and 'Magadi Road' have the fewest restaurants.
6. **Types of Restaurants:** 'Casual Dining', 'Quick Bites', 'Cafe', and 'Dessert Parlor' are the most common types of restaurants, while 'Food Court', 'Casual Dining', and 'Dhaba' are the least common.
7. **Popular Cuisines:** North Indian cuisines are the most popular among Bangalore restaurants, indicating a strong preference for North Indian flavors.
8. **Service Types:** The two main service types offered by restaurants are delivery and dine-out, catering to customers' preferences for convenience or dine-in experiences.
9. **Highly Voted Restaurants:** 'Onesta', 'Truffles', and 'Empire Restaurant' received the highest number of votes, indicating their popularity and customer satisfaction.

10. **Rating Prediction Model:** For the modeling part, various regression algorithms were tested, including Linear Regression, Decision Tree Regressor, Random Forest Regressor, Support Vector Regressor, and Extra Tree Regressor. Among them, the Extra Tree Regressor performed the best and was selected for the final rating prediction model.

In conclusion, this analysis provides valuable insights into the restaurant landscape in Bangalore, including popular restaurants, cuisines, service types, and customer preferences. The developed rating prediction model can help restaurant owners or potential investors make informed decisions and improve their understanding of factors influencing ratings.