Architecture Design

Restaurant Ratings Prediction

Written By	Shivansh Jayara, Priyanka Garach, Abhinai Alavelli, Anubhav Srivastav
Document Version	1.0
Last Revised Date	06-Nov-2021

Document Control

- Change Record:

Version	Date	Author	Comments
1.0	29-Oct-21	Abhinai Alavelli,	Initial Draft
2.0	02-Nov-21	Priyanka Garach	Updated Architecture workflow chart.
3.0	06-Nov-21	Shivansh Jayara, Anubhav Srivastav	Added Model Training / Validation workflow chart. Restructure and Re-format entire document.

Approval Status:

Version	Review date	Reviewed By	Approved By	Comments

Index

Abstract

Machine Learning is a category of algorithms that allows software applications to become more accurate in predicting outcomes without being explicitly programmed. The basic premise of machine learning is to build models and employ algorithms that can receive input data and use statistical analysis to predict an output while updating outputs as new data becomes available. These models can be applied in different areas and trained to match the expectations of management so that accurate steps can be taken to achieve the organization's target. In this paper, the case of Restaurant ratings Prediction model, a one place to get the ratings of your restaurant based on certain criteria. Also, this will help the Restaurant owner to know what changes need to be made to get the good ratings and good business. Nowadays, things are moving towards the online business, people are avoiding to dine in and try to get the food delivered at home as per their convenience. so, business becomes more challenging; restaurant owners are required to put more focus towards the client requirements.

As customers are more centric towards online orders, and ordered based on ratings, ratings calculated based on some factors so we try to put these factors together in the form of data and build a use case to help Restaurant's owners to get good ratings and attract business. Taking various aspects of a dataset collected for Restaurant Rating Predication, and the methodology followed for building a predictive model, results with high levels of accuracy are generated, and these observations can be employed to make decisions to improve ratings for business purpose.

1. Introduction

1.1. What is Architecture Design?

The goal of Architecture Design (AD) or a low-level design document is to give the internal design of the actual program code for the `Restaurant Ratings Prediction`. AD describes the class diagrams with the methods and relation between classes and program specification. It describes the modules so that the programmer can directly code the program from the document.

1.2. Scope

Architecture Design (AD) is a component-level design process that follows a step-by-step refinement process. This process can be used for designing data structures, required software, architecture, source code, and ultimately, performance algorithms. Overall, the data organization may be defined during requirement analysis and then refined during data design work. And the complete workflow.

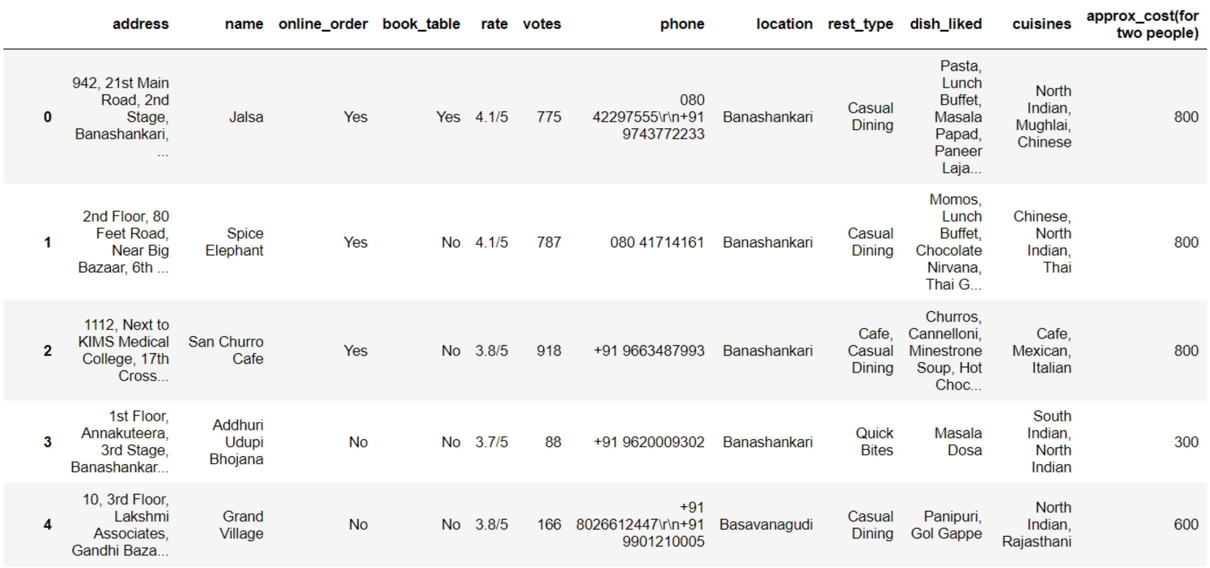
1.3. Constraints

We only predict the expected casual and registered customers based on the weather condition and date information.

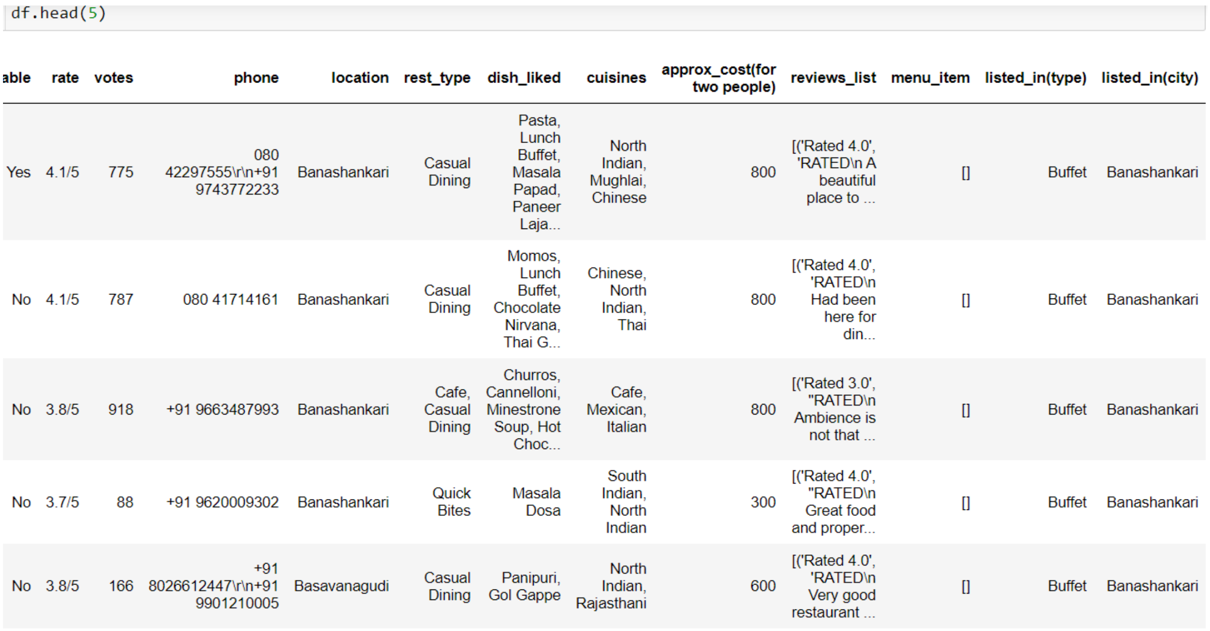
2. Technical Specification

2.1. Dataset

For Restaurant Rating Prediction, we have consolidated the data of different 56521 Restaurant, available at different locations along with the ratings, number of votes as per the Zomato website. Each having different number cuisines, cost, votes, rate and location.as per data collection. Using all the observations it is inferred what role certain properties of an item play and how they affect their Ratings. The dataset looks like as follow:



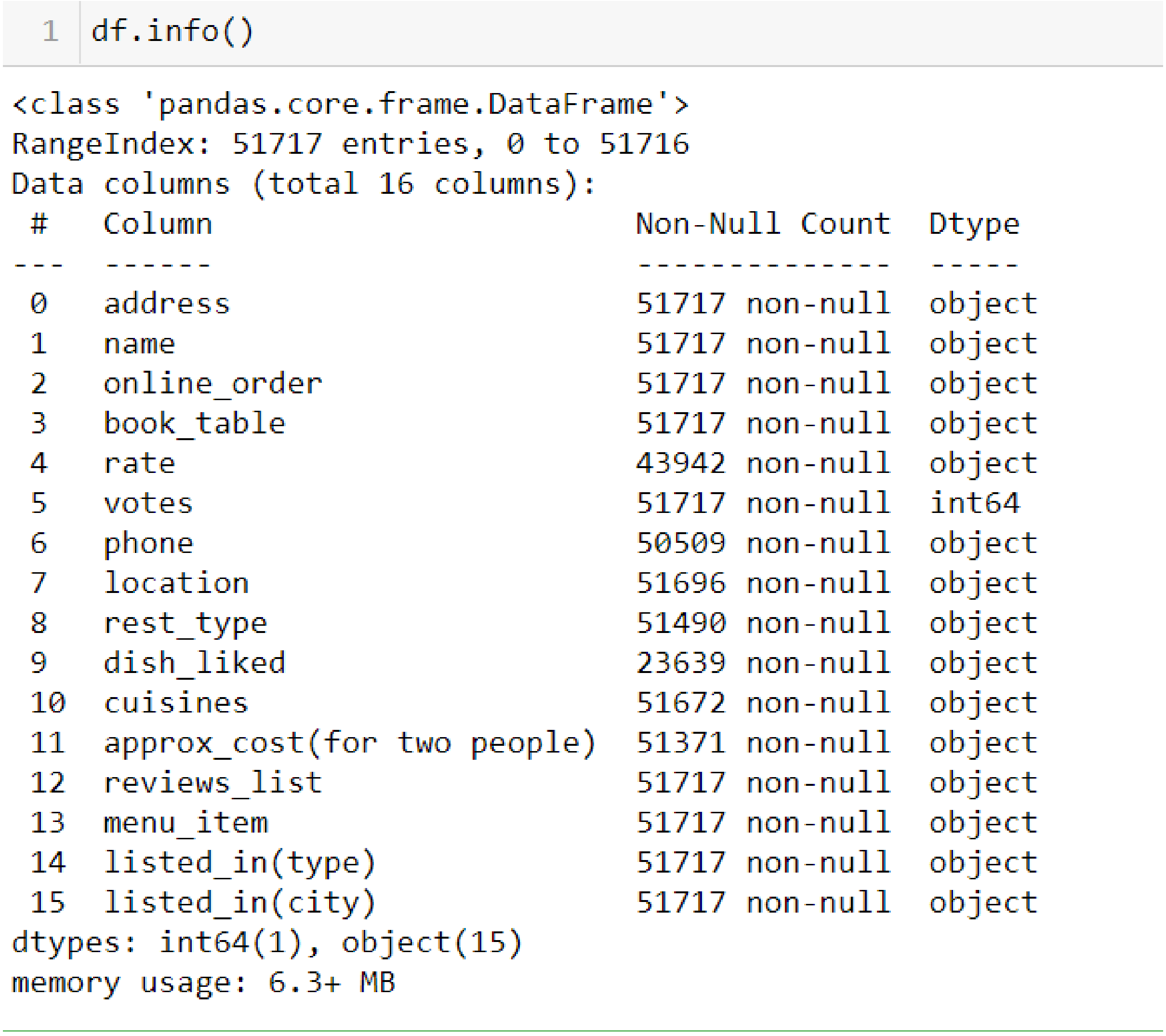
	address	name	online_order	book_table	rate	votes	phone	location	rest_type	dish_liked	cuisines	approx_cost(for two people)
0	942, 21st Main Road, 2nd Stage, Banashankari, ...	Jalsa	Yes	Yes	4.1/5	775	080 42297555\r\n+91 9743772233	Banashankari	Casual Dining	Pasta, Lunch Buffet, Masala Papad, Paneer Laja...	North Indian, Mughlai, Chinese	800
1	2nd Floor, 80 Feet Road, Near Big Bazaar, 6th ...	Spice Elephant	Yes	No	4.1/5	787	080 41714161	Banashankari	Casual Dining	Momos, Lunch Buffet, Chocolate Nirvana, Thai G...	Chinese, North Indian, Thai	800
2	1112, Next to KIMS Medical College, 17th Cross...	San Churro Cafe	Yes	No	3.8/5	918	+91 9663487993	Banashankari	Cafe, Casual Dining	Churros, Cannelloni, Minestrone Soup, Hot Choc...	Cafe, Mexican, Italian	800
3	1st Floor, Annakuteera, 3rd Stage, Banashankar...	Addhuri Udupi Bhojana	No	No	3.7/5	88	+91 9620009302	Banashankari	Quick Bites	Masala Dosa	South Indian, North Indian	300
4	10, 3rd Floor, Lakshmi Associates, Gandhi Baza...	Grand Village	No	No	3.8/5	166	+91 8026612447\r\n+91 9901210005	Basavanagudi	Casual Dining	Panipuri, Gol Gappe	North Indian, Rajasthani	600



```
df.head(5)
```

able	rate	votes	phone	location	rest_type	dish_liked	cuisines	approx_cost(for two people)	reviews_list	menu_item	listed_in(type)	listed_in(city)
Yes	4.1/5	775	080 42297555\r\n+91 9743772233	Banashankari	Casual Dining	Pasta, Lunch Buffet, Masala Papad, Paneer Laja...	North Indian, Mughlai, Chinese	800	[('Rated 4.0', 'RATED\n A beautiful place to ...	[]	Buffet	Banashankari
No	4.1/5	787	080 41714161	Banashankari	Casual Dining	Momos, Lunch Buffet, Chocolate Nirvana, Thai G...	Chinese, North Indian, Thai	800	[('Rated 4.0', 'RATED\n Had been here for din...	[]	Buffet	Banashankari
No	3.8/5	918	+91 9663487993	Banashankari	Cafe, Casual Dining	Churros, Cannelloni, Minestrone Soup, Hot Choc...	Cafe, Mexican, Italian	800	[('Rated 3.0', "RATED\n Ambience is not that ...	[]	Buffet	Banashankari
No	3.7/5	88	+91 9620009302	Banashankari	Quick Bites	Masala Dosa	South Indian, North Indian	300	[('Rated 4.0', "RATED\n Great food and proper...	[]	Buffet	Banashankari
No	3.8/5	166	+91 8026612447\r\n+91 9901210005	Basavanagudi	Casual Dining	Panipuri, Gol Gappe	North Indian, Rajasthani	600	[('Rated 4.0', 'RATED\n Very good restaurant ...	[]	Buffet	Banashankari

The data set consists of various data types from integer to floating to object as shown in Fig.

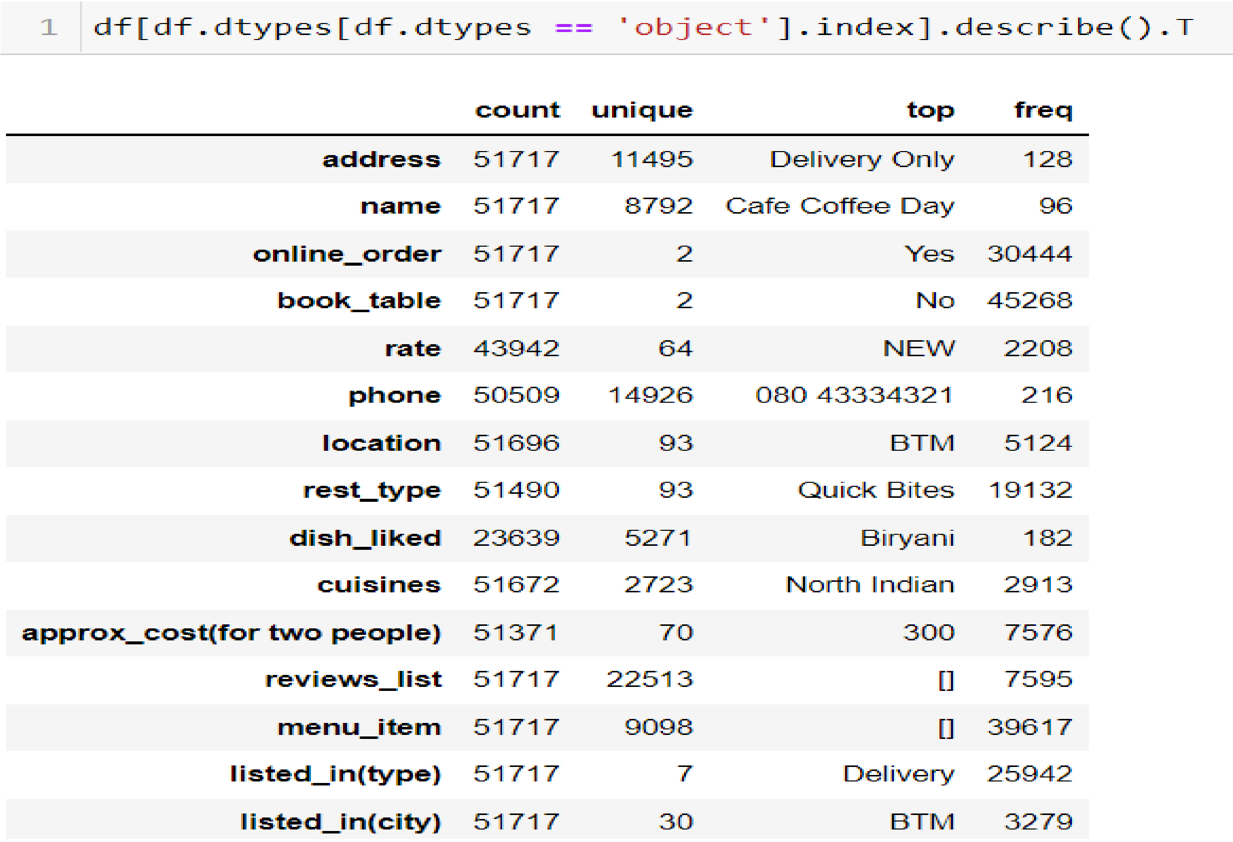


```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 51717 entries, 0 to 51716
Data columns (total 16 columns):
 #   Column                       Non-Null Count  Dtype 
---  ------                       --------------  ----- 
 0   address                      51717 non-null  object
 1   name                         51717 non-null  object
 2   online_order                 51717 non-null  object
 3   book_table                   51717 non-null  object
 4   rate                         43942 non-null  object
 5   votes                        51717 non-null  int64 
 6   phone                        50509 non-null  object
 7   location                     51696 non-null  object
 8   rest_type                    51490 non-null  object
 9   dish_liked                   23639 non-null  object
 10  cuisines                     51672 non-null  object
 11  approx_cost(for two people)  51371 non-null  object
 12  reviews_list                 51717 non-null  object
 13  menu_item                    51717 non-null  object
 14  listed_in(type)              51717 non-null  object
 15  listed_in(city)              51717 non-null  object
dtypes: int64(1), object(15)
memory usage: 6.3+ MB
```

In the raw data, there can be various types of underlying patterns which also gives an in-depth knowledge about the subject of interest and provides insights into the problem. But caution should be observed with respect to data as it may contain null values, or redundant values, or various types of ambiguity, which also demands pre-processing of data. The dataset should therefore be explored as much as possible.

Various factors important by statistical means like mean, standard deviation, median, count of values and maximum value, etc. are shown below for numerical attributes.



```python
df[df.dtypes[df.dtypes == 'object'].index].describe().T
```

	count	unique	top	freq
address	51717	11495	Delivery Only	128
name	51717	8792	Cafe Coffee Day	96
online_order	51717	2	Yes	30444
book_table	51717	2	No	45268
rate	43942	64	NEW	2208
phone	50509	14926	080 43334321	216
location	51696	93	BTM	5124
rest_type	51490	93	Quick Bites	19132
dish_liked	23639	5271	Biryani	182
cuisines	51672	2723	North Indian	2913
approx_cost(for two people)	51371	70	300	7576
reviews_list	51717	22513	[]	7595
menu_item	51717	9098	[]	39617
listed_in(type)	51717	7	Delivery	25942
listed_in(city)	51717	30	BTM	3279

Preprocessing of this dataset includes doing analysis on the independent variables like checking for null values in each column and then replacing or filling them with supported appropriate data types so that analysis and model fitting is not hindered from their way to accuracy. Shown above are some of the representations obtained by using Pandas tools which tell about variable count for numerical columns and model values for categorical columns. As most of the data types are Object, we need to look at it and try to fix them with the help of encoding and other technique. So data plays an important factor in deciding which value to be chosen at priority for further exploration tasks and analysis. Data types of different columns are used further in label processing and a one-hot encoding scheme during the model building.

2.2. Logging

We should be able to log every activity done by the user

- The system identifies at which step logging require.
- The system should be able to log each system flow.
- Developers can choose logging methods. Also, can choose database logging.
- The system should not be hung even after using so much logging. Logging just because we can easily debug issuing so logging is mandatory to do.

2.3. Deployment

For the hosting of the project, we will use Heroku

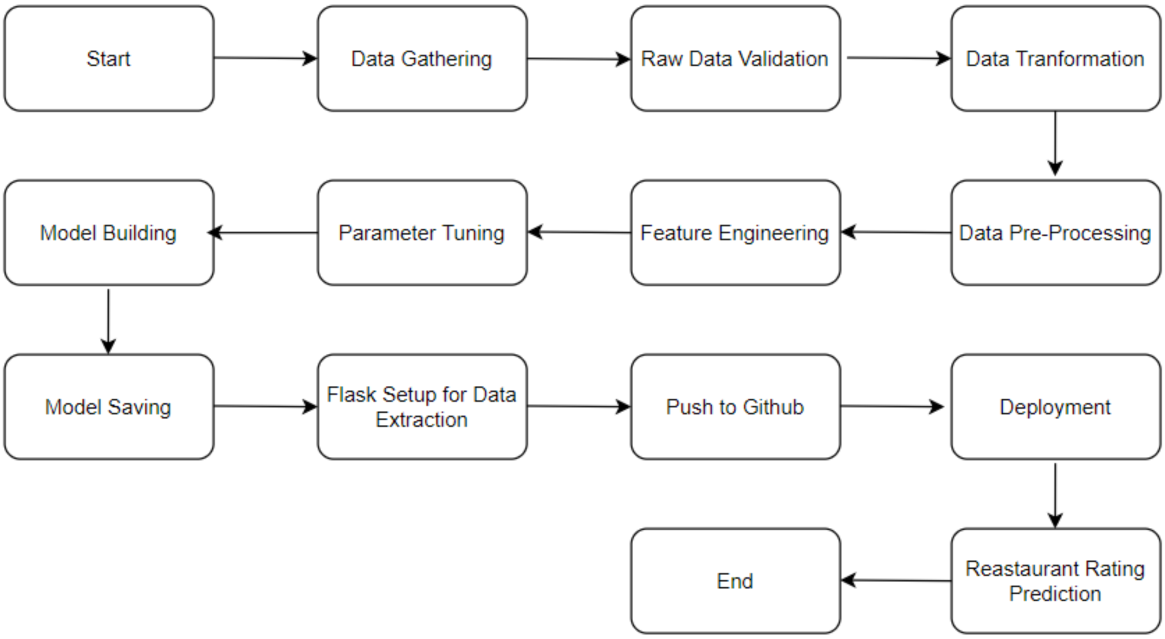
3. Technology Stack

Front End	HTML/CSS
Backend	Python Flask
Deployment	Heroku

4. Proposed Solution

We will use performed EDA to find the important relation between different attributes and will use a machine-learning algorithm to predict the ratings for the restaurant. The Owner of the restaurant will be filled the required feature as input and will get results through the web application. The system will get features and it will be passed into the backend where the features will be validated and preprocessed and then it will be passed to a hyperparameter tuned machine learning model to predict the final outcome.

5. Architecture



5.1. Data Gathering

Data source: https://www.kaggle.com/himanshupoddar/zomato-bangalore-restaurants

Train and Test data are stored in .csv format.

5.2. Raw Data Validation

After data is loaded, various types of validation are required before we proceed further with any operation. Checking for complete missing values in any columns, etc. These are required because The attributes which contain these are of no use. Removing those are not contributing and won't play role in contributing to the predication of the ratings of the restaurant.

If any attribute is having 50% and above missing values, then there is no use in taking that attribute into an account for operation. It's unnecessary increasing the chances of dimensionality curse.

5.3. Data Transformation

Before sending the data into preprocessing, data transformation is required so that data are converted into such form with which it can be more insightful, and machine can easily understand the pattern. Here, the 'Dish Liked attributes contain the missing values. Also, there are features those are not participating like 'Phone Number' and 'City' So, they are filled in both the train set as well as the test set with supported appropriate data types.

5.4. Data Preprocessing

In data preprocessing all the processes required before sending the data for model building are performed. Like,

1) Renaming the columns names as the names available in dataset are not easily understandable. E.g. ("approx. Cost (for two people)" to "Cost, "listed_in(type) to type, "listed_in(city) to "city").
2) Dropping the unwanted columns. E.g. ('Dish_liked', 'Cuisines', 'City').
3) Dropping duplicate values as they are not required in model building suppose. When you duplicate the same data, the training data will have same instances again and again. i.e., there will be multiple same values of dependent and independent variable combinations. Hence, when your model learns from this data, you will get very high accuracy on in-sample testing but out of sample testing will be much lesser. i.e., you will have over fitted model.
4) Handling Null values by removing them as they are not participating.

5.5. Feature Engineering

After preprocessing it was found that some of the attributes are not important, so we required Feature engineering, Data transformation and label encoding is required, as the dataset which we are using have lots of object data types and without encoding them we won't achieve the accuracy on important features, So those attributes are removed. Even one hot encoding is also performed to convert the categorical features into numerical features.

5.6. Parameter Tuning

Parameters are tuned using RandomizedSearchCV. Two algorithms are used in this problem, ExtraTree Regressor and RandomForest Regressor. The parameters of these 2 algorithms are tuned and passed into the model.

5.7. Model Building

After doing all kinds of preprocessing operations mention above and performing scaling and hyperparameter tuning, the data set is passed into all two models, ExtraTree Regressor and Random Forest Regressor. It was found that Extra Tree regressor performs best with train accuracy: 90% and Test accuracy: 77% along with the smallest RMSE value i.e. 21.7 'Extra Tree Regressor' performed well in this problem.

5.8. Model Saving

Model is saved using pickle library in `.pkl` format.

5.9. Flask Setup for Data Extraction

After saving the model, the API building process started using Flask. Web application creation was created here. Whatever the data user will enter and then that data will be extracted by the model to predict the rating of restaurant, this is performed in this stage.

5.10. GitHub

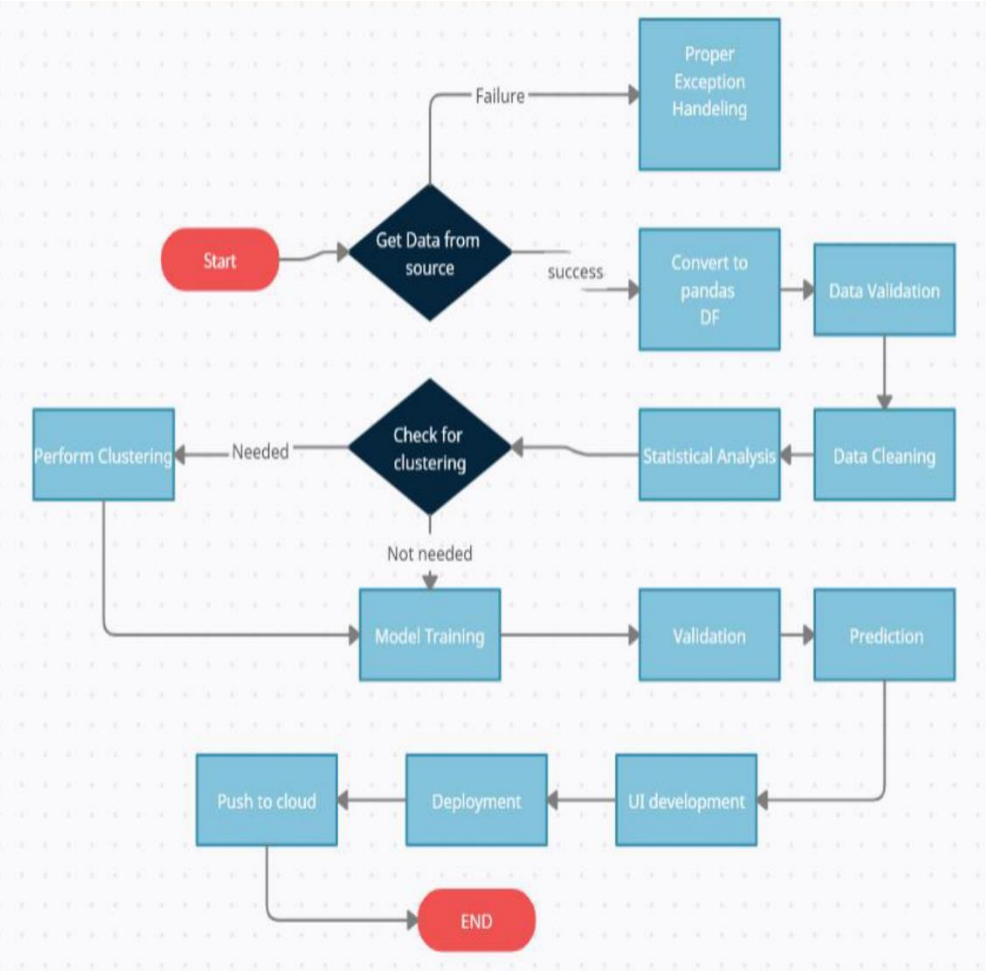
The whole project directory will be pushed into the GitHub repository.

5.11. Deployment

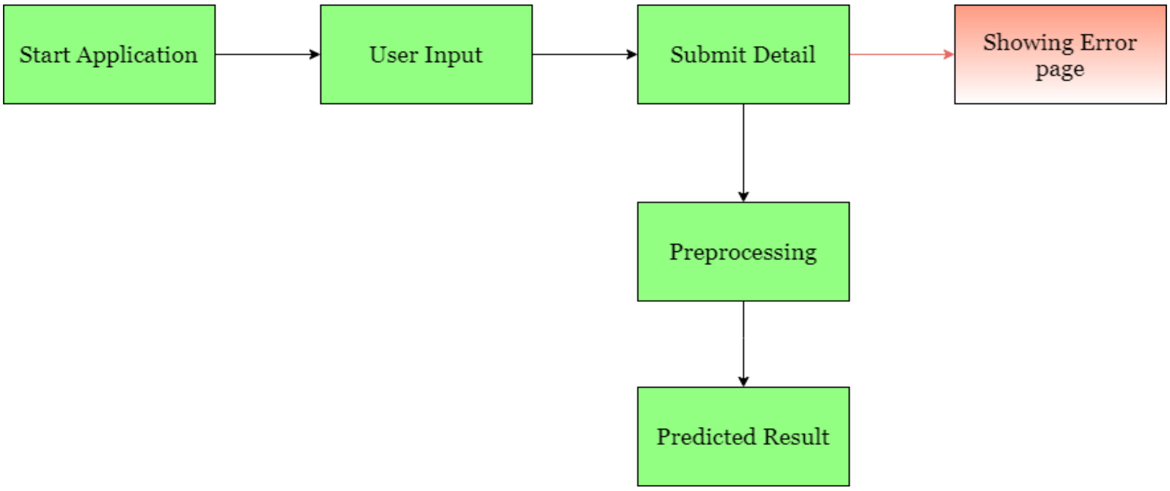
The cloud environment was set up and the project was deployed from GitHub into the Heroku cloud platform.

App link- **https://zomato-rrp.herokuapp.com/**

6. Model Training / Validation workflow



7. User Input / Output Workflow.


Start Application
User Input
Submit Detail
Showing Error page
Preprocessing
Predicted Result