**Architecture Design**

**Restaurant Ratings Prediction**

|  |  |
| --- | --- |
| Written By | Manu Raj |
| Document Version | 1.0 |
| Last Revised Date | 20-Aug-2021 |

# Document Control

1

Change Record:

|  |  |  |  |
| --- | --- | --- | --- |
| **Version** | **Date** | **Author** | **Comments** |
| 1.0 | 22-Aug-21 | Manu Raj | Created Architecture workflow chart |
|  |  |  |  |
|  |  |

**Approval Status:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Version** | **Review date** | **Reviewed By** | **Approved By** | **Comments** |
|  |  |  |  |  |

**Index**

|  |  |
| --- | --- |
| **Content** | **Page No** |
| Abstract | 4 |
| 1. Introduction | 4 |
| 1.1 What is Architecture Design? | 4 |
| 1.2 Scope | 4 |
| 1.3 Constraints | 4 |
| 2. Technical Specification | 5 |
| 2.1 Dataset | 5 |
| 2.2 Logging | 7 |
| 2.3 Deployment | 7 |
| 3. Technology Stack | 8 |
| 4. Proposed Solution | 8 |
| 5. Architecture | 8 |
| 5.1 Data gathering | 9 |
| 5.2 Raw data validation | 9 |
| 5.3 Data Transformation | 9 |
| 5.4 Data Preprocessing | 9 |
| 5.5 Feature Engineering | 10 |
| 5.6 Parameter Tuning | 10 |
| 5.7 Model Building | 10 |
| 5.8 Model Saving | 10 |
| 5.9 Flask Setup for Data Extraction | 10 |
| 5.10 Github | 10 |
| 5.11 Deployment | 10 |
| 6. Model Training / Validation workflow | 11 |
| 7. User Input/Output Workflow | 11 |

3

# 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.

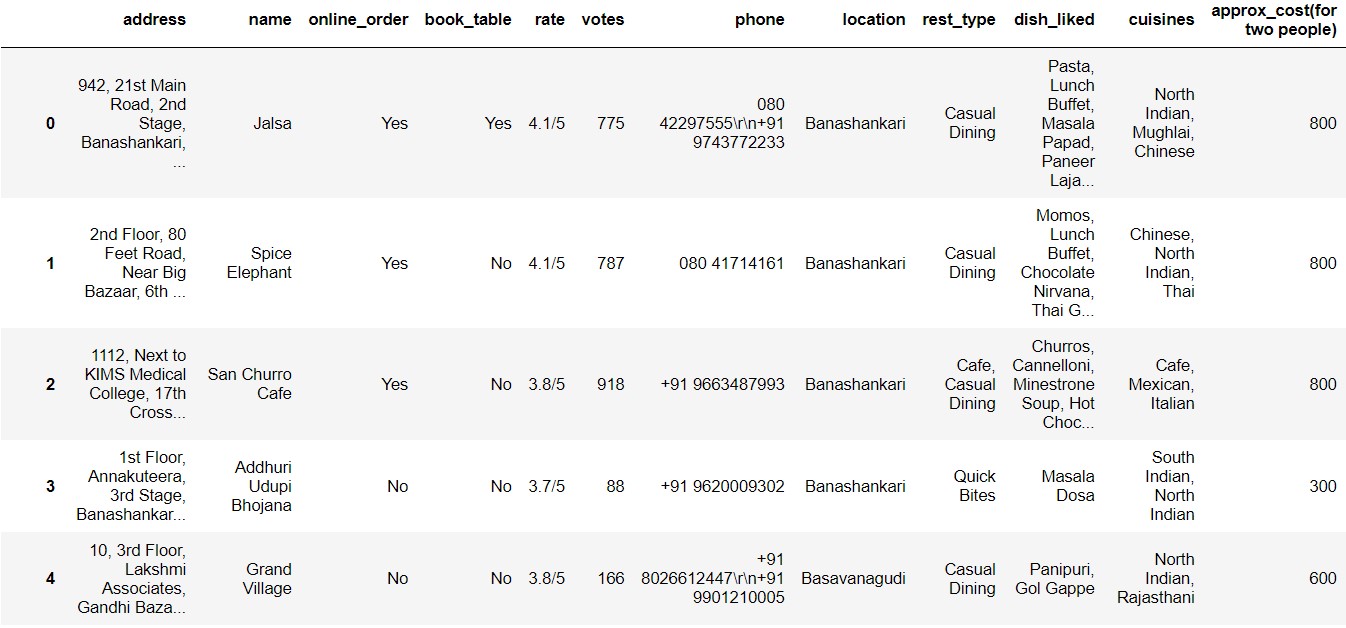
## 1.3. Constraints

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

# 2. Technical Specification

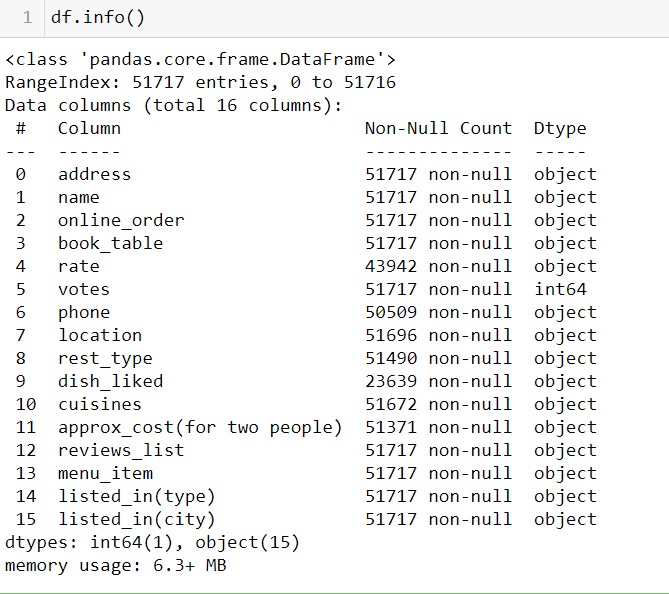
## 2.1. Dataset

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:



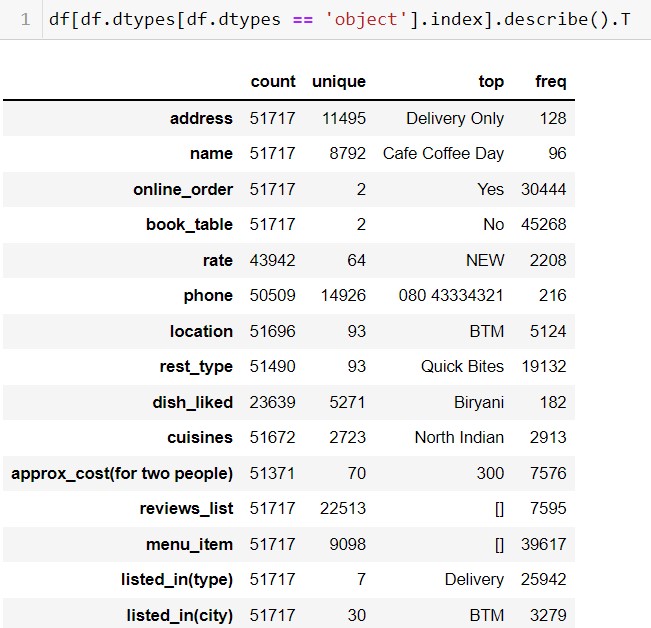


The data set consists of various data types from integer to floating to object as shown in Fig.



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.



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/AWS |

## 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

Diagram

Description automatically generated

### 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-

# 6. Model Training / Validation workflow

Diagram

Description automatically generated

**7. User Input / Output Workflow.**

