**Article or Blog**

**Select the problem statement: -**

**Zomato Restaurant**

**Project Description**

Zomato Data Analysis is one of the most useful analysis for foodies who want to taste the best cuisines of every part of the world which lies in their budget. This analysis is also for those who want to find the value for money restaurants in various parts of the country for the cuisines.

Additionally, this analysis caters the needs of people who are striving to get the best cuisine of the country and which locality of that country serves that cuisines with maximum number of restaurants.

**Data Storage:**

This problem statement contains two datasets- **Zomato.csv** and **country\_code.csv.**

**Country\_code.csv** contains two variables:

• Country code

• Country name

The collected data has been stored in the Comma Separated Value file **Zomato.csv**. Each

restaurant in the dataset is uniquely identified by its Restaurant Id. Every Restaurant contains the following variables:

• Restaurant Id: Unique id of every restaurant across various cities of the world

• Restaurant Name: Name of the restaurant

• Country Code: Country in which restaurant is located

• City: City in which restaurant is located

• Address: Address of the restaurant

• Locality: Location in the city

• Locality Verbose: Detailed description of the locality

• Longitude: Longitude coordinate of the restaurant&#39;s location

• Latitude: Latitude coordinate of the restaurant&#39;s location

• Cuisines: Cuisines offered by the restaurant

• Average Cost for two: Cost for two people in different currencies

• Currency: Currency of the country

• Has Table booking: yes/no

• Has Online delivery: yes/ no

• Is delivering: yes/ no

• Switch to order menu: yes/no

• Price range: range of price of food

• Aggregate Rating: Average rating out of 5

• Rating color: depending upon the average rating color

• Rating text: text on the basis of rating of rating

• Votes: Number of ratings casted by people

**Problem statement : In this dataset predict 2 things –**

1) Average Cost for two

2) Price range

**Hint :** Use pandas methods to combine all the datasets and then start working on this project.

**Dataset Link-**

* <https://github.com/FlipRoboTechnologies/ML_-Datasets/blob/main/Z_Restaurant/Country-Code.xlsx>
* <https://raw.githubusercontent.com/FlipRoboTechnologies/ML_-Datasets/main/Z_Restaurant/zomato.csv>

1. **Problem Definition: -** As per given problem statement predict the price range.
2. **Data Analysis: -** Import the necessary libraries, extract the data file as per given link and download in the local system. Read the country code download file as a df1. Similarly, Zomato file read using pandas known as df2. After read both the csv files merge in a particular data frame merge called df.

Check the df shape that have 9551 rows and 22 columns. Check the df information df have 3 columns float data type, 5 columns int data type and 14 columns object data type. Cuisine have 9 values null present in the dataset.

To list the all columns are :-

Restaurant ID, Restaurant Name, Country Code, City, Address, Locality,

Locality Verbose, Longitude, Latitude, Cuisines, Average Cost for two,

Currency, Has Table booking, Has Online delivery, Is delivering now,

Switch to order menu, Price range, Aggregate rating, Rating color,

Rating text, Votes and Country.

Impute the null values using dropna / fillna method but I use the dropna method.

After impute the null values recheck the nulls using isnull . sum. Now, no null values present in the dataset.

Check the information of data set there is 9542 rows and 22 clumns.

study the statistical data

* 1. Country code mean = 18.179208 and std= 56.451600
  2. Longitude and Latitude have min. value in negative.
  3. Price for two mean = 1200.441509 and std= 16130.432623
  4. Is delivering now mean = 0.003563 and std = 0.059589
  5. Votes mean = 156.772060 and std= 430.203324
* Above point are clearly shows the mean values of all the columns have less but standard values are more.
* As per my observation above columns have something wrong , I will deal with them later or during the outliers removing / skewness of data.

1. **EDA Concluding Remarks: -**

Check the null values one column cuisines have 9 null values in the df data frame. Deal the null values using dropna and again check the null values exist in the data set or not.

Check the unique values in the data set column wise Restaurant ID = 9542, Restaurant Name = 7437, Country code = 15, city = 140, Address =8910, Locality = 1206, Locality Verbose = 1263, Longitude = 8111, Latitude = 8668, Cuisine = 1825, Average cost for two = 140, Currency = 12, Has table booking = 2, Has online delivery = 2, is delivering now =2, switch to order menu = 1, Price range = 4, Aggregate rating =33, Rating color = 6, Rating text = 6, Votes = 1012, Country = 15.

Rating color and rating text similar to each other.

Separate the numerical columns 8 :-

Restaurant ID, Country Code, Longitude, Latitude, Average Cost for two, Price range, Aggregate rating and Votes.

and categorical columns 14 :-

Restaurant Name, City, Address, Locality, Locality Verbose, Cuisines, Currency, Has Table booking, Has Online delivery, Is delivering now, Switch to order menu, Rating color, Rating text and Country

Check the outliers in numerical columns using boxplot, there are little bit outliers in columns price range and aggregate rating as per my observation.

Check unique vales of price range = 4 i.e. 1,2,3 and 4, let’s visualize the price range column using count plot. 1 = 4438, 2 = 3113, 3 = 1405 and 4 = 586. Price range 1 have highest, second one is 2 , third one is 3 and fourth one is 4.

Similarly, check unique values of currency = 12 , the most one is Indian rupees that have 8652 rows, second one is Dollar = 473 rows, third one is pounds 80 rows, fourth one is Brazilian Real that have 60 rows, fifth one is Emirati Diram that have 60 rows, sixth one is Rand that have 60 rows, seventh one is the New Zealand ($) that have 40 rows, eight one is the Turkish Lira that have 34 rows, nineth one is the Botswana Pula

That have 22 rows, tenth one is the Indonesian Rupiah

That have 21 rows, eleventh one is the Qatari Rial

That have 20 rows and twelfth or last one is Sri Lankan Rupee

That have 20 rows.

Check unique values of has table booking = 2 one is No = 8384 and second one is Yes=1158.

Check unique values of has online delivery = 2 one is No = 7091 and second one is Yes=2451.

Check unique values of Is delivering now = 2 one is No = 9508 and second one is Yes=34.

Check unique values of Rating color = 6, top most orange= 3734, second top most white=2148, third one yellow= 2096, fourth one green = 1078, fifth one dark green = 300 and last one Red = 186.

Check unique values of Rating text = 6, top most average= 3734, second top most not rated=2148, third one good= 2096, fourth one very good = 1078, fifth one excellent = 300 and last one poor = 186.

Check unique values of Country = 15 :-

India 8652

United States 425

United Kingdom 80

Brazil 60

UAE 60

South Africa 60

New Zealand 40

Turkey 34

Australia 24

Phillipines 22

Indonesia 21

Singapore 20

Qatar 20

Sri Lanka 20

Canada 4

Now, get the categorical columns in a particular variable called categorical\_cols.

Encoding all categorical columns using ordinal encoder columns like Restaurant name, city , Address Locality, Locality Verbose, Cuisines, Currency, Has Table booking, Has Online delivery, Is delivering now, Switch to order menu, Rating color, Rating text and Country.

Again, check the data set information all columns have numerical data type, float data type columns=17 and int data type columns = 5.

Remove the column switch to order menu.

Describe the data set using describe method that show us count, mean, std, min, 25% , 50%, 75% and max values of the dataset.

Check the correlation between features & label or features & features using heatmap method.

check the skewness of data. Country code, average cost for two, is delivering now, votes and country having skewness.

Let’s deal with them using np.cbrt method (cube root).

After dealing with skewness, again check the skewness of data. Is delivering still having value is 16.665472. I again try to resolve it using npcbrt method.

Again check the skewness, but value again same 16.665472.

Visualize the correlation between label and features using bar plot.

Positive correlation are Votes is the highest correlation second is has table booking.

Third one is average cost for two, fourth one is aggregate rating , fifth one is country code, sixth one is country and seventh one is rating text, eight one is has online delivery, nineth one is currency, tenth one is restaurant name , eleventh one is delivering now.

The most negative correlation first one is latitude, second one is rating color, third one is restaurant id, fourth one is locality , fifth one is city, sixth one is longitude and seventh one is address.

1. **Pre-processing Pipeline: -**

Separate the features and label x and y.

X variable is holding the whole column except the price range. While y variable holds the only one column that is price range only.

Check the y counts 1 = 4438, 2= 3112, 3=1405, 4=586 as per the y count class imbalance.

Let’s deal with imbalance class using SMOTE, check the y count again 1= 4438, 2=4438, 3= 4438 and 4 = 4438. Now balance the class.

Feature scaling using standard scalarization.

Creating train and test data with test size is 0.2 and random state is 21.

Check the model accuracy / confidence using evaluation metrics , Random forest classifier having highest train result is 100% and Test result is 98.78%

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test Classification Report | | | | |
|  | precision | recall | f1-score | support |
| 1 | 1 | 0.99 | 0.99 | 1358 |
| 2 | 0.98 | 0.99 | 0.98 | 1336 |
| 3 | 0.98 | 0.98 | 0.98 | 1290 |
| 4 | 0.98 | 0.99 | 0.99 | 1342 |
|  |  |  |  |  |
|  |  |  |  |  |
| accuracy |  |  | 0.99 | 5326 |
| macro avg | 0.99 | 0.99 | 0.99 | 5326 |
| weighted avg | 0.99 | 0.99 | 0.99 | 5326 |

I save the final model using .pkl method.

I predict the save model data using .load method.

1. **Building Machine Learning Models:-**

Random forest classifier (), dot fit xtrain , ytrain , predict the xtest print accuracy\_score , confusion matrix and classification report.

accuracy of this model is 98.53%

Logistic Regression (), dot fit xtrain , ytrain , predict the xtest print accuracy\_score , confusion matrix and classification report.

accuracy of this model is 80.65%

support vector machine classifier (), dot fit xtrain , ytrain , predict the xtest print accuracy\_score , confusion matrix and classification report. Accuracy of this model is 91.46%

Gradient Boosting Classifier () , dot fit xtrain , ytrain , predict the xtest print accuracy\_score , confusion matrix and classification report. accuracy of this model is 96.90%

Ada Boost Classifier (), dot fit xtrain , ytrain , predict the xtest print accuracy\_score , confusion matrix and classification report.

accuracy of this model is 72.74%

Bagging Classifier (), dot fit xtrain , ytrain , predict the xtest print accuracy\_score , confusion matrix and classification report.

accuracy of this model is 98.62%

Extra Trees Classifier (), dot fit xtrain , ytrain , predict the xtest print accuracy\_score , confusion matrix and classification report.

accuracy of this model is 97.24%

I select my best model Random Forest classifier, because the train result and test result are highest.

After that find the train and test accuracy we reached the maximum side. So, I am not trying to hyper parameter tunning now.

I again initialize the x train, y train with test size =0.3 and random state = 25.

Train result = 100%

Test result = 98.78%

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Test Classification Report | | | | |
|  | precision | recall | f1-score | support |
| 1 | 1 | 0.99 | 0.99 | 1358 |
| 2 | 0.98 | 0.99 | 0.98 | 1336 |
| 3 | 0.98 | 0.98 | 0.98 | 1290 |
| 4 | 0.98 | 0.99 | 0.99 | 1342 |
|  |  |  |  |  |
|  |  |  |  |  |
| accuracy |  |  | 0.99 | 5326 |
| macro avg | 0.99 | 0.99 | 0.99 | 5326 |
| weighted avg | 0.99 | 0.99 | 0.99 | 5326 |

**6) Concluding Remarks: - As per given problem statement I make a model that is Random Forest classifier which predict the price range have accuracy train result is 100%. Test result is 98.78%**

**Random forest classifier is my best model.**

* + 1. In this data set I find the new location to start new business.
    2. Also set the price range 1 and 2 is the highest sale products.
    3. We also find the cuisine which one is the highest sale (qty and price).
    4. Votes is also a parameter basis of that we decide the number of votes of a particular cuisine to increase the productivity of that product to procure the raw material for that to reduce the time and cost.