

RESTAURANTS RATING PREDICTION



zomato

Better food for more people



MACHINE LEARNING PROJECT



TEAM MEMBERS: NITHISH BABU, AKHIL VYDYULA, SUDIKSHA PATIL,

Business Problem

1.Problem Description

Restaurants from all over the world can be found here in Bengaluru. From United States to Japan, Russia to Antarctica, you get all type of cuisines here. Delivery, Dine-out, Pubs, Bars, Drinks, Buffet, Desserts you name it and Bengaluru has it. Bengaluru is best place for foodies. The number of restaurant is increasing day by day. Currently which stands at approximately 12,000 restaurants. With such a high number of restaurants. And new restaurants are opening every day. However, it has become difficult for them to compete with already established restaurants. This Zomato data aims at analyzing the demography of the location.

Most importantly it will help new restaurants in deciding their theme, menus, cuisine, cost etc for a particular location.

It also aims at finding similarities between neighborhoods of Bengaluru on the basis of food.

- Does demography of area matters?
- Does location of particular type of restaurant depends on people living in that area>
- Are any neighborhood on similar based on the type of food?
- Is particular neighbors is famous for its own kind of food?
- What kind of food is famous in locality?

Problem Statement

The dataset also contains reviews for each of the restaurants which will help in finding an overall rating for the place. So we will try to predict rating for a particular restaurant.

Real-world/Business Objectives

We need to predict rating based on different parameters like Average_cost for two people, Online Order available, foods, menu list, most liked dishes etc... features.

Machine Learning Formulation

Here we suppose to predict rating of a restaurant, so it is basically a **Regression** problem.

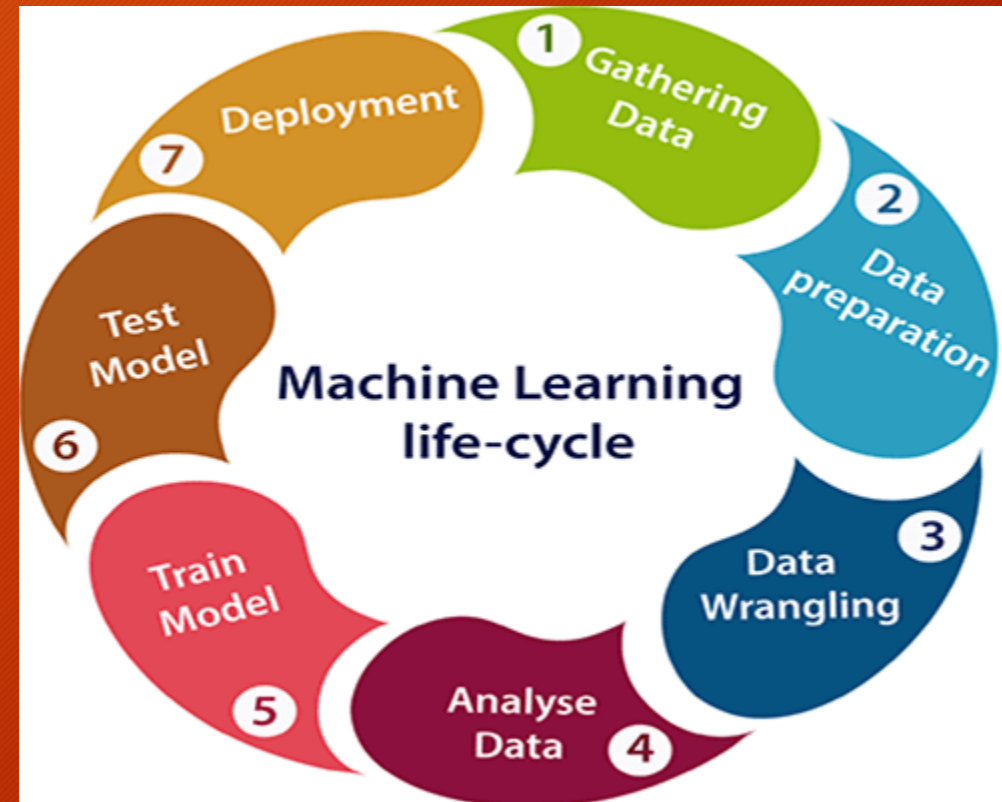
Performance Metric

We will try to reduce Mean Square Error i.e **MSE** as minimum as possible. So it is a **Regression** problem reducing **MSE**.

-Ideal MSE is 0.

STEPS:

1. STRATEGY: Fixing the problem with accurate solution
2. Dataset preparation and preparation and preprocessing
 - Data collection
 - Data visualization
 - Data cleaning
 - Data preprocessing
 - Data transformation
3. Dataset splitting
4. Modelling
5. Model deployment



Data visualization:

- Every machine learning problem solving starts with EDA.
- It is probably one of the most important part of a machine learning project
- In this project also we have clearly visualized all the features.

Data cleaning:

- It is all about cleaning, removing, and removing null values, etc.....
- This dataset contains null values as show below
- And those unwanted columns are removed as shown below.

data.isnull().sum()

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	

dropping the unwanted columns

```
[ ] #data=data.drop(['url','address','phone','dish_liked','menu_item'], axis=1)
    #data.head()
```

DATA PREPROCESSING

```
# Creating a new dataset that has only customer reviews and restaurant ratings
reviews_data = data[["reviews_list", "rate"]]
# Examining the reviews for the first restaurant in the dataset
reviews_data["reviews_list"][0]
# The text needs cleaning up
```

```
'[('\Rated 4.0\',' \RATED\\n A beautiful place to dine in.The interiors take you back to the Mughal era. The lightings are just perfect.We went there on the occ
asion of Christmas and so they had only limited items available. But the taste and service was not compromised at all.The only complaint is that the breads could
have been better.Would surely like to come here again.\'), ('\Rated 4.0\',' \RATED\\n I was here for dinner with my family on a weekday. The restaurant was comp
letely empty. Ambience is good with some good old hindi music. Seating arrangement are good too. We ordered masala papad, panner and baby corn starters, lemon an
d corriander soup, butter roti, olive and chilli paratha. Food was fresh and good, service is good too. Good for family hangout.\\nCheers\'), ('\Rated 2.0\',' \R
ATED\\n Its a restaurant near to Banashankari BDA. Me along with few of my office friends visited to have buffet but unfortunately they only provide veg buffet.
On inquiring they said this place...']'
```

-Remove Duplicate values

-Remove Null Values

We observed that in 'dish_liked' 48.22% data is missing. Similarly in 'Rate' column, 10.22% data is missing. If we directly throw all NULL data out, we have to ignore 48.22% of original data. Can we somehow fill the missing data?

we can have two approaches.

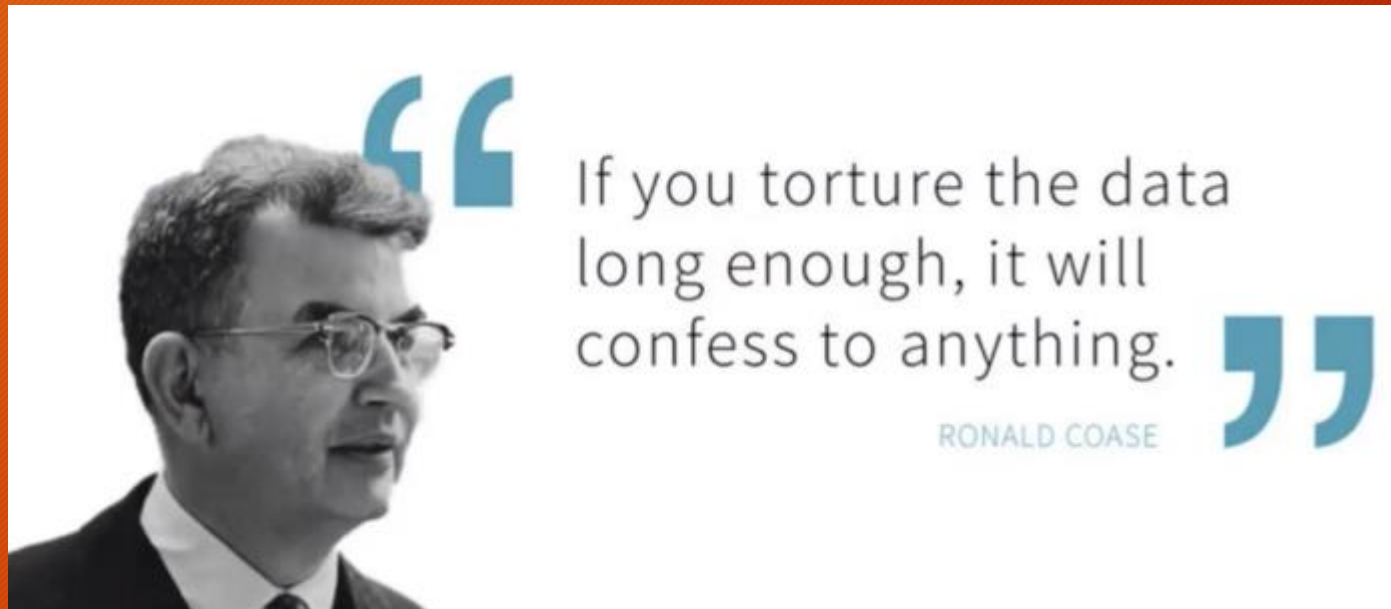
- 1. Filling the missing values with appropriate values then operate.*
- 2. Throw all null values and then operate.*

SPLIT DATA

Always remain we should first split data then apply featurization, to avoid data leakage problems. Divide data into Train, Test part.

DATA FEATURIZATION

We will convert all online_order', 'book_table', 'location', 'rest-type' and 'cuisines' features into Categorical features. Then we will use **one-hot encoding** technique for featurization.



MODEL

we understand the data as well we visualize the data, now the actual machine learning part starts from here.

-After deep-diving into we can clearly say that 'online_order', 'book_table', 'vote', 'location', 'rest-type', 'cuisines' and 'average_cost' are important columns rest, we can drop other columns.

-**'Rate'** is output column

-First featurized review column with BOW, TFIDF, W2V and concatenated these other features and trained all the models with default parameters, the result was not so good.

-Second considered only review column and featurized with BOW,TFIDF,W2V and trained all the models with each of these features.

-Featurized with TFIDF and trained model.

-Trained model with RandomForest Regressor and received MSE values of 0.014 and R2 score of 92.0406 which is good result. Hence finalized this model for deployment.

Q&A

Q1) What is the source of data?

The data for training is taken from Kaggle.

Q2) what was the type of data

The data is combination of numerical and categorical values.

Q3) what is the complete flow you followed in this project?

Refer above slides for better understanding.

Q&A

Q4) What techniques were you used for data pre-processing?

- Removing unwanted attributes
- visualizing relation of independent variables with each other
- removing outliers
- changing the data and imputing the null values
- converting the categorical to numerical values
- scaling the data

Q5) How prediction was done?

It's a regression problem, so we performed some regressor algorithms, based on the customer review it will predict the rating.

Q&A

Q6) What is the final MSE value of model?

MSE=0.5

Q7) What if we didn't clean the data and pre processing?

It won't allow the tokenization, and it doesn't perform model building.

Q8) How does each stats test matter in model building?

We have performed hypothesis testing, chi sq tests. Which tells us how exactly distribution for each feature.

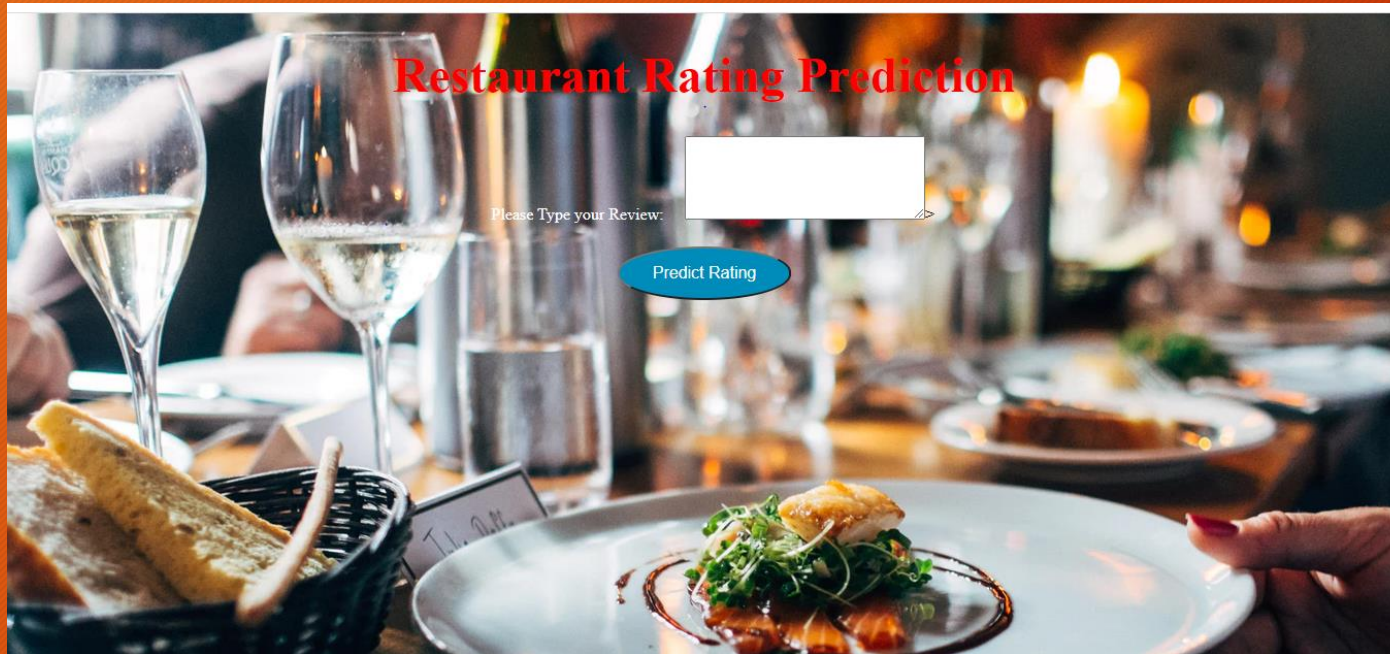
Q9) why do we need flask framework?

For small applications, we mostly use flask which is more reliable and flexible and easy to understand.

Q10) Do we get feature interpretability?

Yes, by observation of feature importance and we can conclude that which feature is dominant for predicting rating.

- Created HTML page which includes a text field to type the review and a predict button to submit the review.
- Once the predict button is clicked the predicted rating will appear in same page.



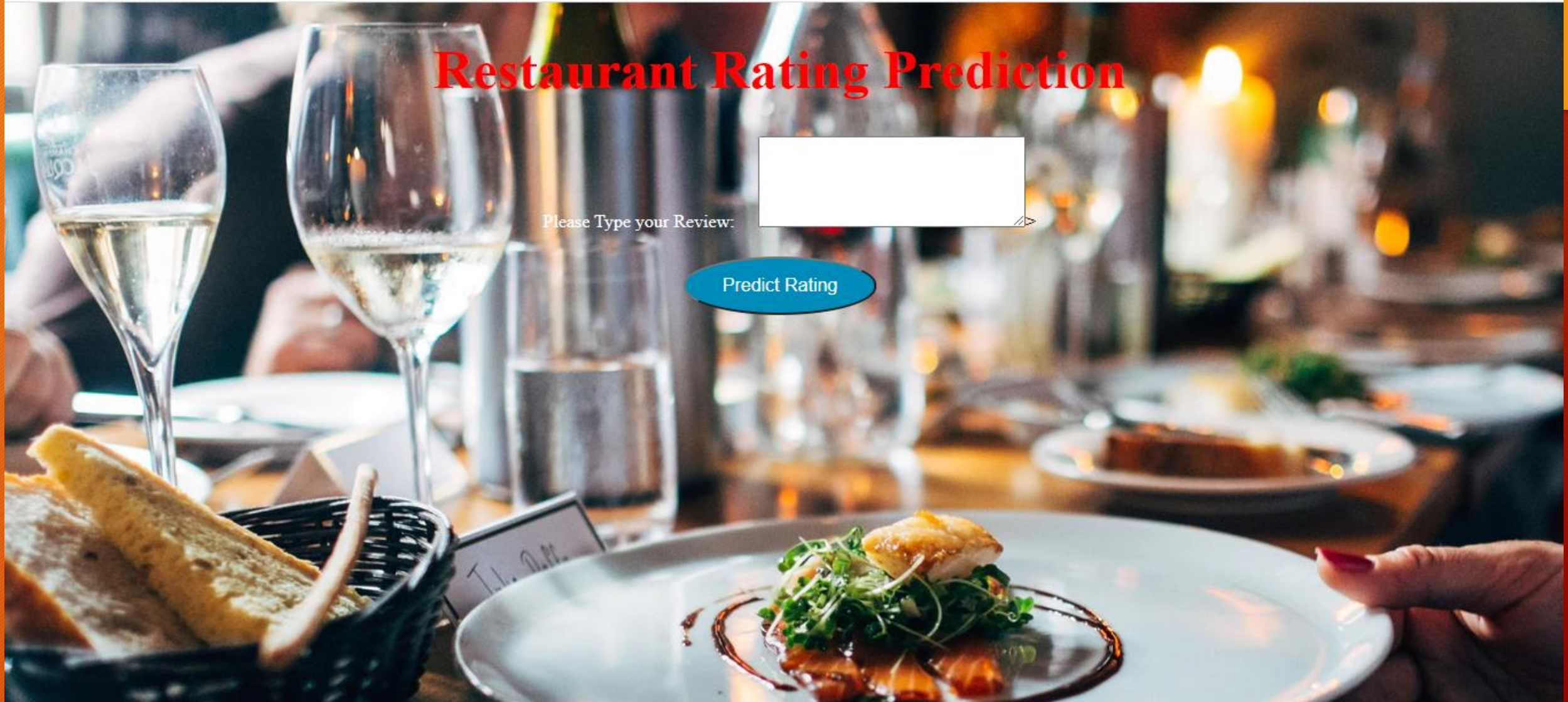
- We have used Flask API and finally it is deployed in Heroku Platform.
- Heroku is a platform as service that enables developers to build, run, and operate applications entirely in the cloud.

WEB PAGE AFTER DEPLOYMENT

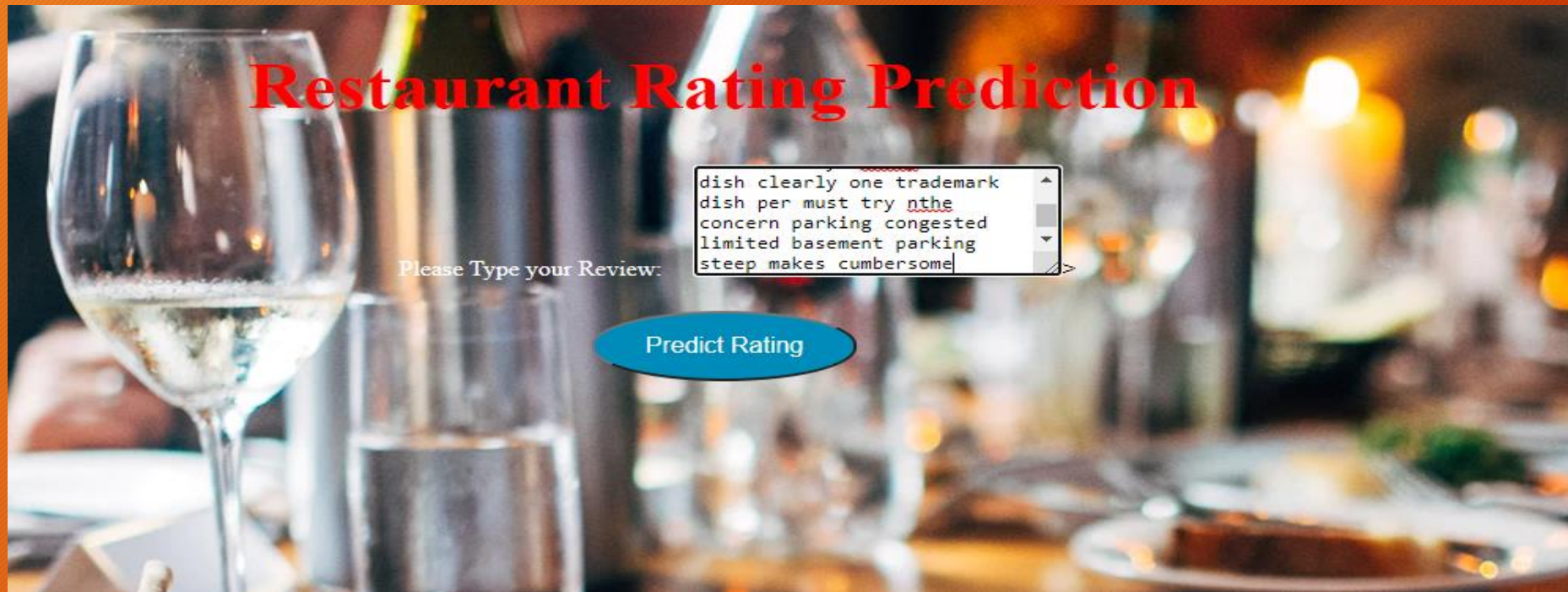
Restaurant Rating Prediction

Please Type your Review:

Predict Rating



Positive Review



Negative Review



CONCLUSION

In this a number of features about existing restaurants of different areas in a city and analyses them to predict rating of the restaurant. This makes it an important aspect to be considered, before making a dining decision. Such analysis is essential part of planning before establishing a venture like that of a restaurant. Lot of researches have been made on factors which affect sales and market in restaurant industry. Various dine-scape factors have been analysed to improve customer satisfaction levels. If the data for other cities is also collected, such predictions could be made for accurate and more helpful.

*** THANK YOU***