

RESTAURANT RATING PREDICTION

A PROJECT REPORT

for

BIG DATA ANALYTICS(ITE2013)

in

B.Tech – Information Technology and Engineering

by

Bhaumik Tandan (19BIT0292)

Yash Renwa (19BIT0125)

Under the Guidance of

Dr. D. LOPEZ

SITE



VIT[®]

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Information Technology and Engineering

Nov, 2021

DECLARATION BY THE CANDIDATE

We hereby declare that the project report entitled “**Restaurant Rating Prediction**” submitted by us to Vellore Institute of Technology University, Vellore in partial fulfilment of the requirement for the award of the course **Big Data Analytics (ITE2013)** is a record of bonafide project work carried out by us under the guidance of **Dr. Daphne Lopez**. We further declare that the work reported in this project has not been submitted and will not be submitted, either in part or in full, for the award of any other course.

Place : Vellore

Signature: **Bhaumik Tandan, Yash Renwa**

Date : 10/12/2021



VIT[®]

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Information Technology & Engineering [SITE]

CERTIFICATE

This is to certify that the project report entitled “**Restaurant Rating Prediction**” submitted by **Bhaumik Tandan (19BIT0292)**, **Yash Renwa (19BIT0125)** to Vellore Institute of Technology University, Vellore in partial fulfilment of the requirement for the award of the course **Big Data Analytics (ITE2013)** is a record of bonafide work carried out by them under my guidance.

Dr. Daphne Lopez

GUIDE

SITE

Restaurant Rating Prediction

Abstract

The project aim is to predict the restaurant rating based on the various factors. Online customer feedbacks are taken as an integral part for making a decision before purchasing any product. Present and future generation is the journey of e-commerce platform is booming in world.

Sharing on the internet is something we usually do. Giving a review is a useful activity through which other people on the internet can find out something else and see opinions about things. The usual things reviewed by someone in the form of experiences, places, objects, and others. Give a review we usually use text to explain something that we experience with an item, place, or event that we normally experience. Customer satisfaction is an opinion between expectation and reality obtained by consumers. Giving a review is also a useful activity so that other customer on the internet can find out something else and see opinions about things and its satisfaction.

Commonly, most people express their opinion through social media on the review platform like Zomato, Yelp, etc. Customer reviews on online media like Zomato become important as it might increase the popularity of something. Zomato is a site where someone can give a review of a restaurant, how the restaurant is and someone's opinion about the restaurant. Restaurant customer satisfaction can be analysed by their review on Zomato. Sometimes, restaurants see the reviews in Zomato, but they didn't get if the reviews are positive or negative to their restaurants. Review on Zomato is still in the form of text and can be classified with positive, negative, or neutral with their ratings.

Keywords –Reviews, Classification, Prediction, Random Forest

I. INTRODUCTION

India is popular for its assorted multi-cuisine prepared in a huge number of restaurants and hotel resorts, which is implicative of unity in diversity. The food chain industry and the restaurant business in India is a very competitive one and lack of research and knowledge about the competition usually leads to the failure of many such enterprises. The principal issues that continue to produce difficulties to them include high real estate expenses, escalating food costs, fragmented supply chain, over-licensing, and even after that restaurateur does not know whether the business will develop or not. This project aims to solve this problem by analysing ratings, reviews, cuisines, restaurant type, demand, online ordering service, table booking, availability of the restaurant and make the machine learning model learn these and predict ratings of new restaurant and how positive and negative reviews should be expected. This project considers the data of the restaurant from all over India from Zomato as an example for showing how our model works and can help a restaurateur choose the location and cuisine which will give it better ratings, reviews, and make the business more profitable.

II. BACKGROUND

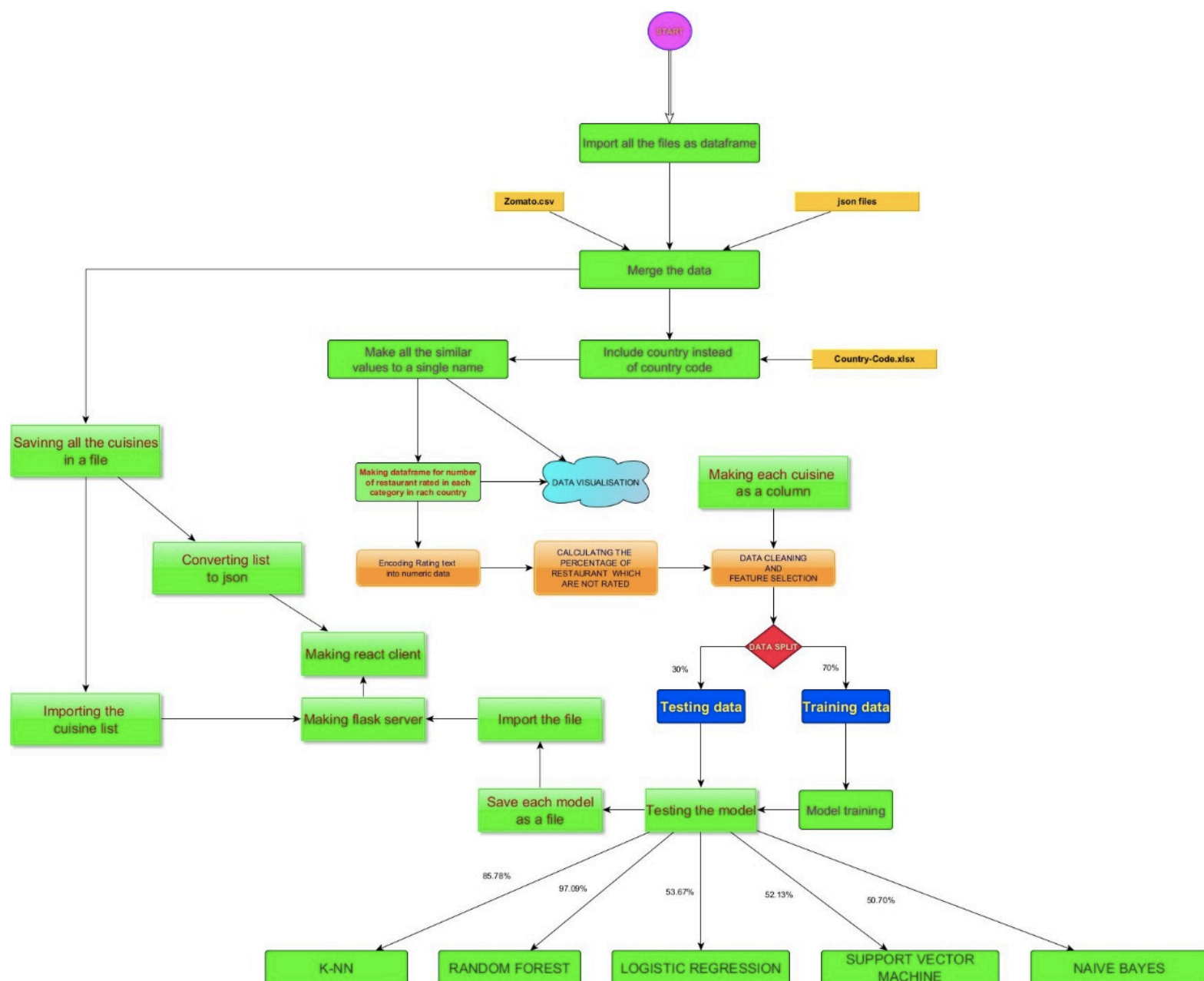
Zomato has changed the way people browse through restaurants. It has helped customers find good places with respect to their dining budget. Various researches and students have published related work in national and international research papers, thesis to understand the objective, types of algorithm they have used and various techniques for pre-processing and feature selection.

For the pre-processing stage we need to apply basic statistical functions as they might be needed for data cleaning purposes. also, we need to know about Exploratory data analysis, Feature engineering through introduction of various functions using python modules, outlier analysis and removal, text cleaning and encoding.

User reviews are an integral part of web services like Tripadvisor, amazon, Zomato and Yelp, where users can post their opinions about restaurants and services through reviews consisting

of free-form text and a numeric star rating, usually out of 5. These online reviews function as the ‘online word-of-mouth’ and a criterion for consumers to choose between similar restaurants. Studies show that they have a significant impact. The process of predicting a relationship for a generic user (but for a specific product/business) is called Review Rating Prediction. We aim to create a model to accurately predict the ratings of the restaurants based off this dataset.

Architecture Diagram



III. LITERATURE SURVEY

Research Paper Title: Predict User Ratings based on Review Texts

Date of Publishing: December, 2016

Survey: a new trend has emerged with emergence of e-commerce website: feedback and reviews, providing platform to market analysis. This project makes use of such review texts in order to perform an inference-based analysis and thereby, train the model to be able to predict the ratings based on these texts. The objective of the project is to implement a model that predicts user ratings based on the review texts. The obtained data-set is in the JSON format containing feature vectors and, KNN algorithms are used. The model has predicted in some cases the exact star rating associated with the review rating online. as the goal of the project is to predict user ratings based on review texts, the prediction includes the review class identification to fall into three broad categories as mentioned earlier. They are, bad (1 star), average (3 stars) and good (5 stars).

Research Paper Title: Restaurant Revenue Prediction using Machine Learning Date of Publishing-June,2016

Review: This paper explores how opening a new restaurant outlets is subjective in nature based on personal judgement and how to predict across geographies and cultures. The model employs a supervised learning algorithm will construct complex features using simple features such as opening date for a restaurant, city that the restaurant is in, type of the restaurant (Food Court, Inline, Drive Thru, Mobile), Demographic data (population in any given area, age and gender distribution, development scales), Real estate data (front facade of the location, car park availability), and points of interest including schools, banks. The paper applies concepts of machine learning such as support vector machines and random forest on these parameters, it predicts the annual revenue of a new restaurant which would help food chains to determine the feasibility of a new outlet. The dataset used in this paper is obtained from Kaggle and consists of a training and test set with 137 and 100,000 samples respectively. The paper experimented with SVM and Random forest and compared their accuracy in terms of estimation.

Research Paper Title: Service Rating Prediction and Recommender System-a Survey

Date of Publishing: July, 2017

Review: Due to increasing access to mobile phones in India, amounting to 300 million(approx.) users has resulted into increase online opinions, reviews and suggestions. We need this huge data to be recommended to users with the help of geographical location and social network. We use location-based services (LBS), recommended system (RS) and finally, location-based service recommended system (LBSRS). additionally, the personalisation and personal interest factors are identified with the user latent features. This combines the individual preference and social interaction to rank the services. as a result, this project provides a review on problems like data- overloading, appropriate service recommendation with help of LBS.

Research Paper Title: Data Mining for Predicting Customer Satisfaction in Fast- Food Restaurants**Date of Publishing: May, 2017**

Review: a fast-food restaurant in Indonesia has been chosen as case study. Multiple data mining techniques are used. Some of the algorithms used are: regression analysis, co-joint analysis, descriptive statistic, correlation analysis, factor analysis and structural equation modelling. This study provides some findings that are useful for the marketers, policy makers who have interest in customer satisfaction research. For practical implication, an understanding of underlying determinant factors of customer satisfaction could strengthen the competitive advantage of the fast- food restaurant.

Research Paper Title: Predicting Yelp Food Establishment Ratings Based on Business attributes**Date of Publishing-2018**

Review: The authors aimed to utilise script analysis to predict helpfulness rating of online reviews. Human annotators were asked to highlight important phrases of reviews along with that phrases were added to a script lexicon. The paper used a Yelp dataset to identify a correlation between rating and the social factors of a user's physical location. They worked to identify a relationship between rating and the distance a user is from their home; they also investigate the connection between the ratings of similar items by users that are friends. They initially inferred that users 5 tend to rate items higher the further away they are from their home, or "activity centre". Text regression was performed to predict review helpfulness based on words in the

lexicon. Compared against a Baseline model and a bag-of-words model, the scripts enriched model was found to predict helpful reviews quicker and more accurately.

Research Paper Title: Predicting the Growth of Restaurants Using Web Data

Date of Publishing: May, 2018

Review: With the rise of data mining techniques in risk and growth field, researchers have also turned their focus towards identifying the reasons of failure of restaurants. Many factors such as competition, business network, characteristics of entrepreneurs etc... are considered in this research. Use of random forest classifier (RFC), multi-layer perception (MLP), neural networks, and also logistic regression (LR) is stated. Therefore, RFCs, MLPs and LRs are tested and challenged with the goal to estimate a binary outcome, i.e. non growing v/s growing restaurants. The results suggest, that RFCs with an average accuracy of 68% outperform MLPs and LRs.

Research Paper Title: Review Rating Prediction Using Yelp

Dataset Date of Publishing: February, 2019

Review: Yelp is a review website that allows users to rate various business categories. Since there are hundreds and millions of reviews, we need to prevent customers from reading a random review and creating bias decision by using ML algorithms. To train our prediction model, linear classification algorithms such as logistic regression, support vector classification and naïve Bayes multi-class classification algorithm are used. The result show that Naïve Bayes' and SVMs provide output with better accuracy. The output are received as form of star rating with 94% when Naïve Bayes' is used.

Literature Summary

Name of Techniques	Advantages	Disadvantages
LDA	LDA is a probabilistic model with interpretable topics.	It is hard to know when LDA is working -topics are soft-clusters so there is no objective metric to say "this is the best choice" of hyper-parameters.
KNN	KNN is very easy to implement. There are only two parameters required to implement KNN i.e. the value of K and the distance function (e.g. Euclidean or Manhattan etc.)	KNN is sensitive to noise in the dataset. We need to manually impute missing values and remove outliers.
Random Forest	Random Forest can be used to solve both classification as well as regression problems.	Random Forest require much more time to train as compared to decision trees as it generates a lot of trees (instead of one tree in case of decision tree) and makes decision on the majority of votes.
SVM	SVM works relatively well when there is a clear margin of separation between classes.	as the support vector classifier works by putting data points, above and below the classifying hyperplane there is no probabilistic explanation for the classification.
Multilayer Perception	Can be applied to complex non-linear problems	It is not known to what extent each independent variable is affected by the dependent variable. Computations are difficult and time consuming.

Name of Techniques	Advantages	Disadvantages
Naive Baiyes	When assumption of independent predictors holds true, a Naive Bayes classifier performs better as compared to other models.	If categorical variable has a category in test data set, which was not observed in training data set, then model will assign a 0 (zero) probability and will be unable to make a prediction. This is often known as Zero Frequency. To solve this, we can use the smoothing technique. One of the simplest smoothing techniques is called Laplace estimation.
Logistic Regression	Logistic regression is less inclined to over-fitting but it can overfit in high dimensional datasets. One may consider Regularisation (L1 and L2) techniques to avoid over-fitting-in these scenarios.	In Linear Regression independent and dependent variables are related linearly. But Logistic Regression needs that independent variables are linearly related to the log odds ($\log(p/(1-p))$).
AdaBoost	Adaboost is less prone to overfitting as the input parameters are not jointly optimised. The accuracy of weak classifiers can be improved by using adaboost. Nowadays, adaboost is being used to classify text and images rather than binary classification problems.	The main disadvantage of adaboost is that it needs a quality dataset. Noisy data and outliers have to be avoided before adopting an adaboost algorithm.

IV. PROPOSED ALGORITHM

Zomato has changed the way people browse through restaurants. It has helped customers find good places with respect to their dining budget. Various researches and students have published related work in national and international research papers, thesis to understand the objective, types of algorithm they have used and various techniques for pre-processing and feature selection.

For the pre-processing stage we need to apply basic statistical functions as they might be needed for data cleaning purposes. also, we need to know about Exploratory data analysis, Feature engineering through introduction of various functions using python modules, outlier analysis and removal, text cleaning and encoding.

User reviews are an integral part of web services like Tripadvisor, amazon, Zomato and Yelp, where users can post their opinions about restaurants and services through reviews consisting of free-form text and a numeric star rating, usually out of 5. These online reviews function as the ‘online word-of-mouth’ and a criterion for consumers to choose between similar restaurants. Studies show that they have a significant impact. The process of predicting a relationship for a generic user (but for a specific product/business) is called Review Rating Prediction. We aim to create a model to accurately predict the ratings of the restaurants based off this dataset.

The proposed system for predictive analysis consists of language features and embedded classification models. The dataset (Zomato Dataset) is split into test and training data. The first step is to generate the ground truth, and is performed based on the training reviews and ratings. We perform Exploratory data analysis and gain an insight into the contents of dataset and their relation to each other through various graphs and charts. The next step in the model was to perform Data Cleaning, Feature Selection and Feature Engineering. In this step we also deal with missing data. In the process of building models, we extract:

(a) Statistical features of users and reviews

- (b) Language features extracted from the review corpus. The next step was to test and evaluate the various classifiers which were used. We also decided to test and check for overfitting and find out the trend of the training vs true error to pick the best model.

The dataset used in the project gives a fair idea about the factors affecting the establishment of different types of restaurant at different places in country aggregating the rating of each restaurant. The Dataset contains all details of the restraint listed on the Zomato website as of March 2019. The dataset contains 9554 rows. In order to build the final model, four phases were followed (1) exploratory data analysis, (2) data cleaning, (3) feature engineering and (4) model creation.

- (1) The data analysis and exploration of the dataset consisted of Displaying the ratings of restaurants by country code and sorting them according to that criteria. We explored the data further by calculating the percentage of restaurants that were not rated and compared these restaurants in terms of percentage to whole of the dataset. From this analysis we explored and analysed that lower priced food was often not rated. Therefore, by similar case imputation, we replaced the 'not rated' with the average rating of lower priced food. We also noted that lower priced food is considered average for the most part, while more expensive food is rated more highly. Expensive food seems rarer than low cost food. We further explored the data with the help of various charts and compared the relationships of the various attributes of the dataset.
- (2) Data Cleaning step of the model consisted of dealing with the attributes not important for the prediction of the final output and hence these attributes were dropped. We analysed the dataset for aSCII variable and URL's and other noise but found out that the dataset was clean enough to be used for prediction without much data cleaning.
- (3) Feature Engineering and Feature selection: In this step we analysed all the restaurants that were not rated and were given a rating based on their price range. as an aggregate rating of 2.5-3.5 translates to a rating of "average", the "Not Rated", the text can be replaced with "average" in the dependent variable. The next step in feature engineering was dealing with the 'Rating Text' as it is given as a string. This needed to be encoded into numeric data so

that the algorithm can perform calculations. This was done using the LabelEncoder tool from sklearn. We classified the rating text into the form where excellent = 1, very good = 4, good = 2, average = 0, and poor = 3.

(4) Having pre-processed the data, We now provided the data to different classifiers and see which one performs better in creating a model of classification for this data. We compared performance using the cross-validation. Python has the cross_val_score class from the sklearn.model_selection library to perform cross validation. We utilised supervised learning techniques for classifying high-dimensional data in order to predict.

- (a) Logistic regression is a classification algorithm, used when the value of the target variable is categorical in nature.
- (b) K nearest neighbours is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). KNN has been used in statistical estimation and pattern recognition as a non-parametric technique.
- (c) Support vector machines (SVM) represents examples as points in space for non- linear classification. Different kernels lead to differences in the performance based on the dataset.
- (d) Naïve Bayes (NB) is a probabilistic model with maximum likelihood to classify categories. We use NB to compare the results of different distributions.
- (e) Random forest (RF) is an ensemble technique that fits a number of decision tree classifiers on various sub-samples of the dataset. We have experimented with different number of decision trees as estimators.

V. MOTIVATION

The motivation for this project came from my own bakery shop, my father wants to open another restaurant so for this we need to know about all the factors like location, price range, food, online booking availability, etc which will affect the rating of the restaurant and hence our profit. Most of the people when start a business of this kind, they face many losses initially because they are not able to satisfy their customers. They don't know about those factors which will affect their rating directly. So to solve this problem of my father and all other businessmen who are in this field, we predicted the rating of restaurants (3star, 5 star, etc) by considering all the important factors which will affect the rating.

Our model will predict the restaurant rating by considering factors like types of cuisines available, online booking availability, price range, average cost for two person, etc.

VI. EXPERIMENTAL RESULTS

To have an overview of how different methods perform, we have summarised the accuracy for all the previous methods in Table 1. For the rating prediction, we can see that linear regression perform comparatively worse than the Native Bayes and Random Forest. Clearly from the bar plot we can see how the accuracy for Random Forest algorithm is the highest achieved and is significantly higher than Naïve Bayes as well. One possible explanation is that the Random Forest method apply to a wider range of problems, and are more robust to problematic model specifications. For example, in the Naive Bayes model, we need to assume conditional independence among the independent variables, which clearly cannot be the case. another possible explanation is that we don't have enough input features to make a decent prediction. a more complicated method might involve more noise and thus yield low accuracy when the information is limited. With more features that can capture the factor affecting ratings, and also more training samples, the last two methods should improve their performance significantly.

Screenshots of Codes

Data Warehousing :

```
In [1]: from pandas import *
from numpy.random import *
from numpy import *
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')

data_frame= read_csv('archive/zomato.csv',engine='python',index_col=0,encoding = 'unicode_escape')
country=read_excel('archive/Country-Code.xlsx',index_col=0)
```

In [2]:

country

Out [2]:

Country	
Country Code	
1	India
14	Australia
30	Brazil
37	Canada
94	Indonesia
148	New Zealand
162	Phillipines
166	Qatar
184	Singapore
189	South Africa
191	Sri Lanka
208	Turkey
214	UAE
215	United Kingdom
216	United States

In [3]:

data_frame.head(5)

Out [3]:

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
6317637	Le Petit Souffle	162	Makati City	Third Floor, Century City Mall, Kalayaan Avenu...	Century City Mall, Poblacion, Makati City	Century City Mall, Poblacion, Makati City, Mak...	121.027535	14.565443	French, Japanese, Desserts	1100	Botswana Pula(P)	Yes	No	No
6304287	Izakaya Kikufuji	162	Makati City	Little Tokyo, 2277 Chino Roces Avenue, Legaspi...	Little Tokyo, Legaspi Village, Makati City	Little Tokyo, Legaspi Village, Makati City, Ma...	121.014101	14.553708	Japanese	1200	Botswana Pula(P)	Yes	No	No
6300002	Heat - Edsa Shangri-La	162	Mandaluyong City	Edsa Shangri-La, 1 Garden Way, Ortigas, Mandal...	Edsa Shangri-La, Ortigas, Mandaluyong City	Edsa Shangri-La, Ortigas, Mandaluyong City, Ma...	121.056831	14.581404	Seafood, Asian, Filipino, Indian	4000	Botswana Pula(P)	Yes	No	No
6318506	Ooma	162	Mandaluyong City	Third Floor, Mega Fashion Hall, SM Megamall, O...	SM Megamall, Ortigas, Mandaluyong City	SM Megamall, Ortigas, Mandaluyong City, Mandal...	121.056475	14.585318	Japanese, Sushi	1500	Botswana Pula(P)	No	No	No
6314302	Sambo Kojin	162	Mandaluyong City	Third Floor, Mega Atrium, SM Megamall, Ortigas...	SM Megamall, Ortigas, Mandaluyong City	SM Megamall, Ortigas, Mandaluyong City, Mandal...	121.057508	14.584450	Japanese, Korean	1500	Botswana Pula(P)	Yes	No	No


```
In [19]: error_load
```

[illegible]

Data Analysing:

```
In [1]: #importing all the libraries and csv file
from pandas import *
from numpy.random import *
from numpy import *
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')
df= read_csv('DataWarehouse.csv', engine='python')
```

```
In [2]: #Sorting the ratings of restaurants by country code
#We are rating the restrants of a certain country code
e = df[df['Rating text']=='Excellent']['Country'].value_counts()
v = df[df['Rating text']=='Very Good']['Country'].value_counts()
g = df[df['Rating text']=='Good']['Country'].value_counts()
a = df[df['Rating text']=='Average']['Country'].value_counts()
p = df[df['Rating text']=='Poor']['Country'].value_counts()
nr = df[df['Rating text']=='Not rated']['Country'].value_counts()

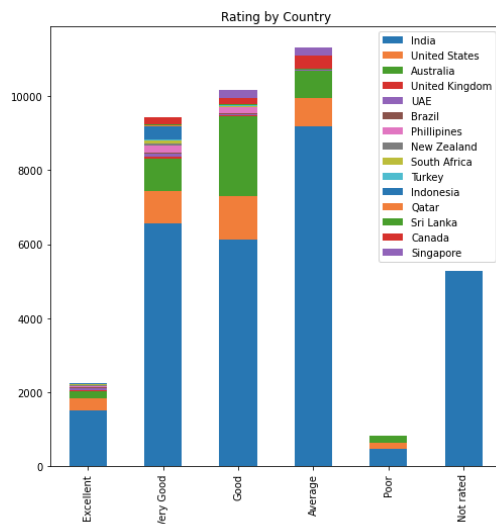
#display distinct country codes
df["Country"].unique()
```

```
Out[2]: array(['Philippines', 'Brazil', 'United States', 'Australia', 'Canada',
        'Singapore', 'UAE', 'India', 'Indonesia', 'New Zealand',
        'United Kingdom', 'Qatar', 'South Africa', 'Sri Lanka', 'Turkey',
        nan], dtype=object)
```

```
In [3]: dfr = DataFrame([e,v,g,r,a,r, pr,nr])
dfr.index = ['Excellent','Very Good','Good','Average','Poor','Not rated']
dfr.fillna(0) #while displaying replaces all null values with 0
```

[illegible]

```
In [5]: #DATA EXPLORATION AND VISUALISATION
dfr.plot(kind='bar',stacked=True, figsize=(8,8), title="Rating by Country")
plt.show()
```

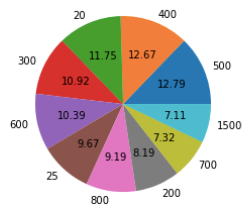


```
In [6]: #1.How many Restaurant accepting online orders
ax = df['Has Online delivery'].value_counts().plot(kind='bar')
plt.title('Number of Restaurants accepting online orders', weight='bold')
plt.xlabel('online orders')
plt.ylabel('counts')
plt.show()
```

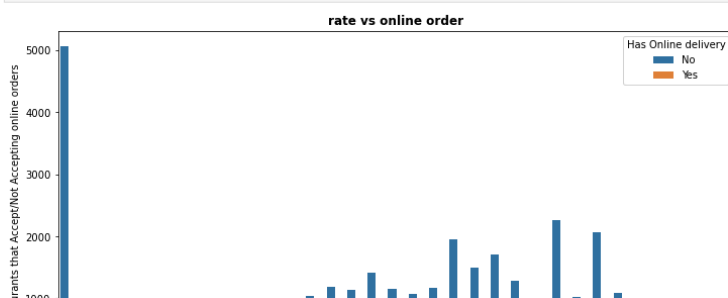


```
In [9]: #Average cost for 2 people to dine
values = df['Average Cost for two'].value_counts()[0:10]
labels = df['Average Cost for two'].value_counts()[0:10].index
plt.pie(values, labels=labels, autopct='%1.2f')
plt.title('Average cost for two person(in %)', weight='bold')
plt.show()
```

Average cost for two person(in %)



```
In [10]: #Rating vs Online orders
plt.figure(figsize = (12,6))
sns.countplot(x=df['Aggregate rating'], hue = df['Has Online delivery'])
plt.ylabel("Restaurants that Accept/Not Accepting online orders")
plt.title("rate vs online order",weight = 'bold')
plt.show()
```



Data Cleaning :

```
In [1]: #importing all the libraries and csv file
from pandas import *
from numpy.random import *
from numpy import *
import seaborn as sns
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings('ignore')
df= read_csv('DataWarehouse.csv',engine='python')
```

```
In [2]: # the lowest priced restaurants that are not rated
mean_lowest_price = df[df['Price range']== 1]['Aggregate rating'].replace(0.0, nan).mean()
print(f"Mean of aggregate rating with price range 1: {mean_lowest_price}")
```

Mean of aggregate rating with price range 1: 3.334984844952209

```
In [3]: # filling with their average
df[df['Price range']== 1]['Aggregate rating'].replace(0.0,nan, inplace=True)
df['Aggregate rating'].replace(nan, mean_lowest_price, inplace=True)
```

```
In [4]: #Next the lower priced restaurants that are not rated
mean_lower_price = df[df['Price range']== 2]['Aggregate rating'].replace(0, nan).mean()
print(f"Mean of aggregate rating with price range 2: {mean_lowest_price}")
```

Mean of aggregate rating with price range 2: 3.334984844952209

```
In [5]: # filling with their average
df['Aggregate rating'].replace(0, mean_lower_price, inplace=True)
```

```
In [6]: #Checking
print(df['Aggregate rating'].mean())

#Checking the new scatterplot, no restaurant has a rating of 0 now.
sns.scatterplot(x='Aggregate rating',y='Votes',data=df)
y=df["Rating text"]

#As an aggregate rating of 2.5-3.5 translates to a rating of "Average", the "Not Rated"
#text can be replaced with "Average" in the dependent variable.

y.replace('Not rated','Average', inplace=True)

3.6490085496296096
```

```
In [10]: df['Latitude'].isnull().unique()
```

```
Out[10]: array([False])
```

```
In [11]: df['Longitude'].isnull().unique()
```

```
Out[11]: array([False])
```

```
In [12]: df["Has Online delivery"].replace(["Yes","No"],[1,0],inplace=True)
df["Is delivering now"].replace(["Yes","No"],[1,0],inplace=True)
df["Has Table booking"].replace(["Yes","No"],[1,0],inplace=True)
df['Switch to order menu'].replace(["Yes","No"],[1,0],inplace=True)
```

```
In [13]: df = df.drop(['Restaurant Name','Locality','Country','City','Address','Currency','Rating color','Aggregate rating'], axis=1)
```

```
In [14]: df.head()
```

```
Out[14]:
```

	Longitude	Latitude	Average Cost for two	Has Table booking	Has Online delivery	Is delivering now	Switch to order menu	Price range	Rating text	Votes	...	Kebab	Turkish Pizza	Izgara	World Cuisine	Dinner	Restaurant Cafe	Beer	Döner
0	121.027535	14.565443	1100	1	0	0	0	3	1	314	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	121.014101	14.553708	1200	1	0	0	0	3	1	591	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	121.056831	14.581404	4000	1	0	0	0	4	4	270	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	121.056475	14.585318	1500	0	0	0	0	4	1	365	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	121.057508	14.584450	1500	1	0	0	0	4	1	229	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

5 rows x 158 columns

```
In [15]: df.to_csv('CleanedData.csv',index=False)
```

Model Training and Testing :

```
In [6]: #PART THREE: TESTING DIFFERENT CLASSIFIERS

#-----K-NN -----

# Fitting K-NN to the Training set
from sklearn.neighbors import KNeighborsClassifier
classifier = KNeighborsClassifier(n_neighbors = 9, metric = 'minkowski', p = 2)

classifier.fit(x_train,y_train)
knn_score=classifier.score(x_test,y_test)
print(f"K-NN:\n Accuracy: {knn_score*100}%")
```

K-NN:
Accuracy: 85.78697421981005%

```
In [7]: #-----Logistic Regression-----

# Fitting Logistic Regression to the Training set
from sklearn.linear_model import LogisticRegression
classifier = LogisticRegression(penalty='l2',random_state = 0)

classifier.fit(x_train,y_train)
lr_score=classifier.score(x_test,y_test)
print(f"Logistic Regression:\n Accuracy: {lr_score*100}%")
```

Logistic Regression:
Accuracy: 53.67198100407056%

```
In [8]: # Fitting SVM to the Training set
from sklearn.svm import SVC
classifier = SVC(kernel = 'rbf', random_state = 0)

classifier.fit(x_train,y_train)
svm_score=classifier.score(x_test,y_test)
print(f"SVM:\n Accuracy: {svm_score*100}%")
```

SVM:
Accuracy: 52.1370420624152%

```
In [9]: #-----Naive Bayes-----

# Fitting Naive Bayes to the Training set
from sklearn.naive_bayes import GaussianNB
classifier = GaussianNB()

classifier.fit(x_train,y_train)
nb_score=classifier.score(x_test,y_test)
print(f"Naive Bayes:\n Accuracy: {nb_score*100}%")
```

```
In [9]: #-----Naive Bayes-----

# Fitting Naive Bayes to the Training set
from sklearn.naive_bayes import GaussianNB
classifier = GaussianNB()

classifier.fit(x_train,y_train)
nb_score=classifier.score(x_test,y_test)
print(f"Naive Bayes:\n Accuracy: {nb_score*100}%")
```

Naive Bayes:
Accuracy: 50.703867028493896%

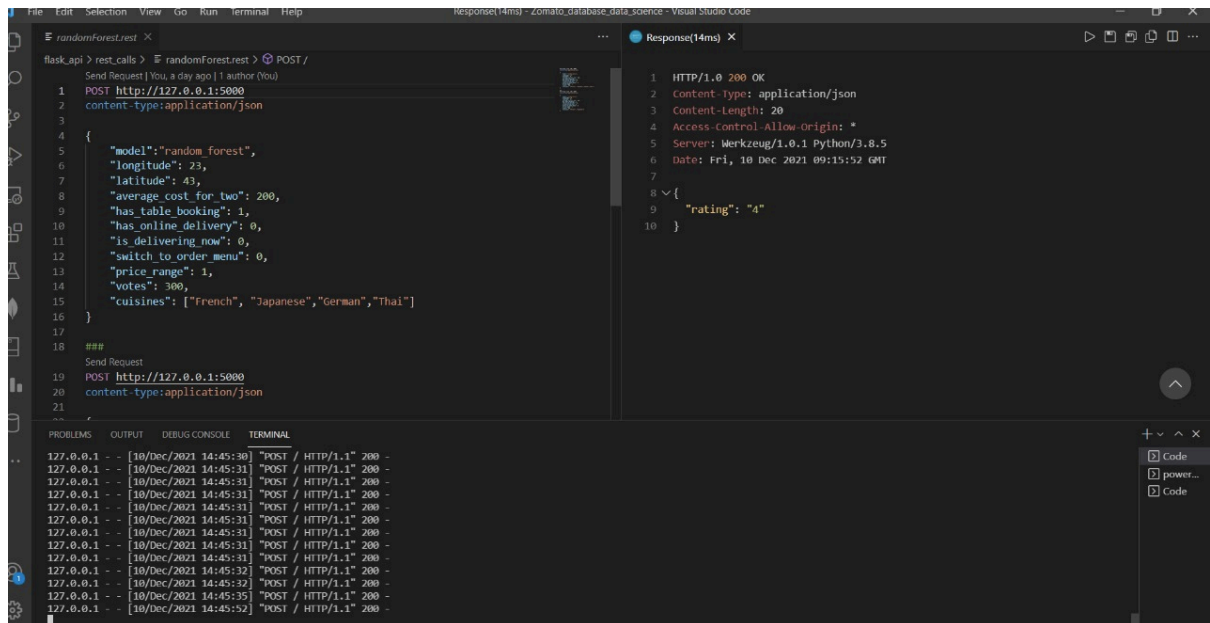
```
In [10]: #-----Random Forest-----

# Fitting Random Forest Classification to the Training set
from sklearn.ensemble import RandomForestClassifier
classifier = RandomForestClassifier(n_estimators = 100, criterion = 'entropy', random_state = 0)

classifier.fit(x_train,y_train)
rf_score=classifier.score(x_test,y_test)
print(f"Random Forest:\n Accuracy: {rf_score*100}%")
```

Random Forest:
Accuracy: 97.09972862957937%

Flask API



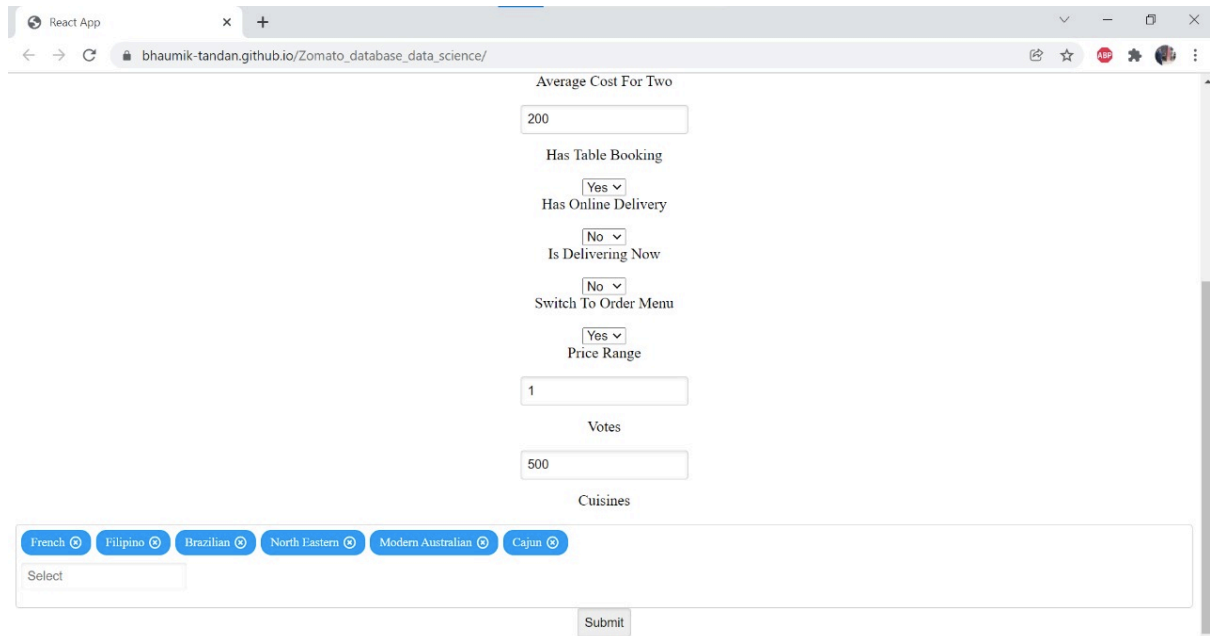
The screenshot shows a REST client interface in Visual Studio Code. The request is a POST to `http://127.0.0.1:5000` with a JSON body. The response is a 200 OK status with a JSON body containing a rating of 4.

```
1 POST http://127.0.0.1:5000
2 content-type:application/json
3
4 {
5   "model": "random forest",
6   "longitude": 23,
7   "latitude": 43,
8   "average cost for two": 200,
9   "has_table_booking": 1,
10  "has_online_delivery": 0,
11  "is_delivering_now": 0,
12  "switch_to_order_menu": 0,
13  "price_range": 1,
14  "votes": 300,
15  "cuisines": ["French", "Japanese", "German", "Thai"]
16 }
17
18 ###
19 Send Request
20 POST http://127.0.0.1:5000
21 content-type:application/json
```

```
1 HTTP/1.0 200 OK
2 Content-Type: application/json
3 Content-Length: 20
4 Access-Control-Allow-Origin: *
5 Server: Werkzeug/1.0.1 Python/3.8.5
6 Date: Fri, 10 Dec 2021 09:15:52 GMT
7
8 {
9   "rating": "4"
10 }
```

Frontend :

Inserting Values -



The screenshot shows a web application interface with a form for inserting values. The form includes input fields for 'Average Cost For Two', 'Has Table Booking', 'Has Online Delivery', 'Is Delivering Now', 'Switch To Order Menu', 'Price Range', 'Votes', and 'Cuisines'. The 'Cuisines' field is a multi-select dropdown with options: French, Filipino, Brazilian, North Eastern, Modern Australian, and Cajun. A 'Submit' button is at the bottom.

Average Cost For Two
200

Has Table Booking
Yes

Has Online Delivery
No

Is Delivering Now
No

Switch To Order Menu
Yes

Price Range
1

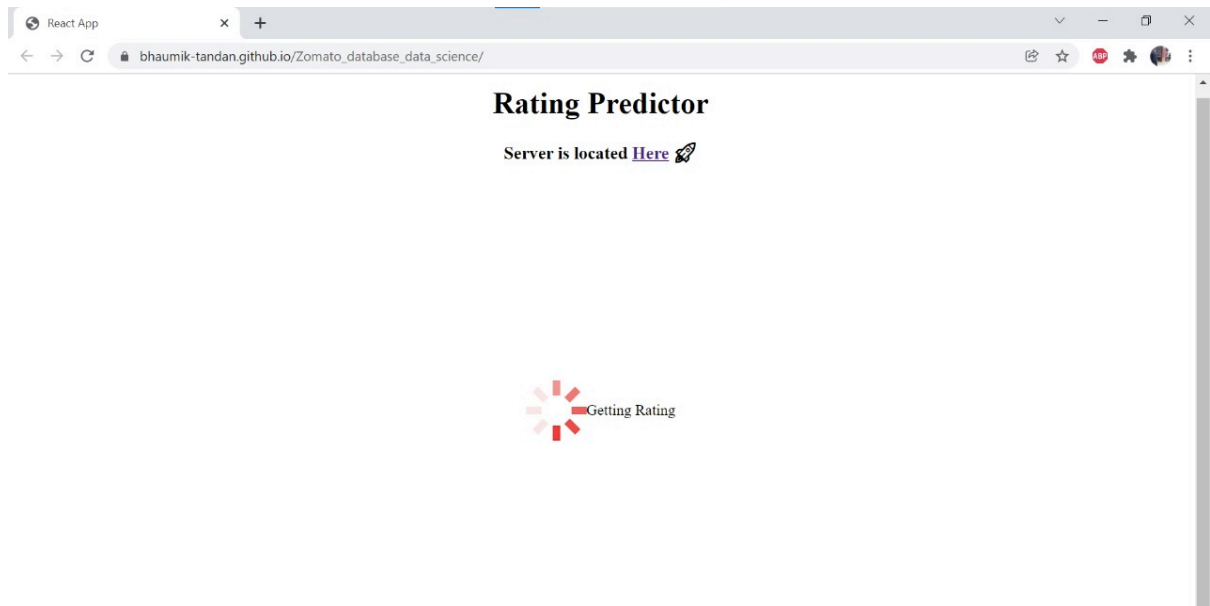
Votes
500

Cuisines
French Filipino Brazilian North Eastern Modern Australian Cajun

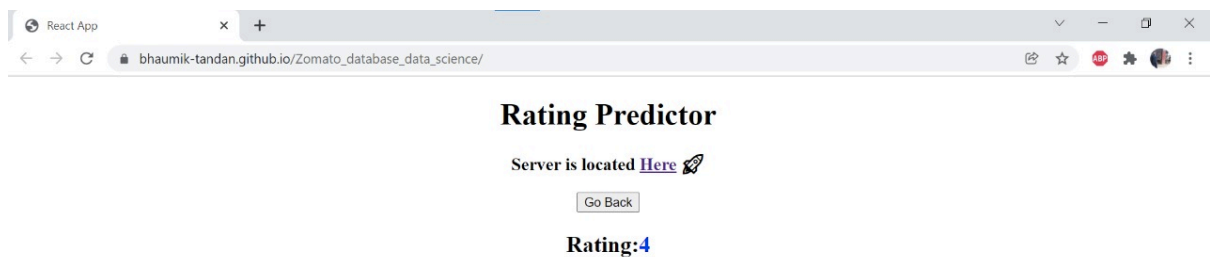
Select

Submit

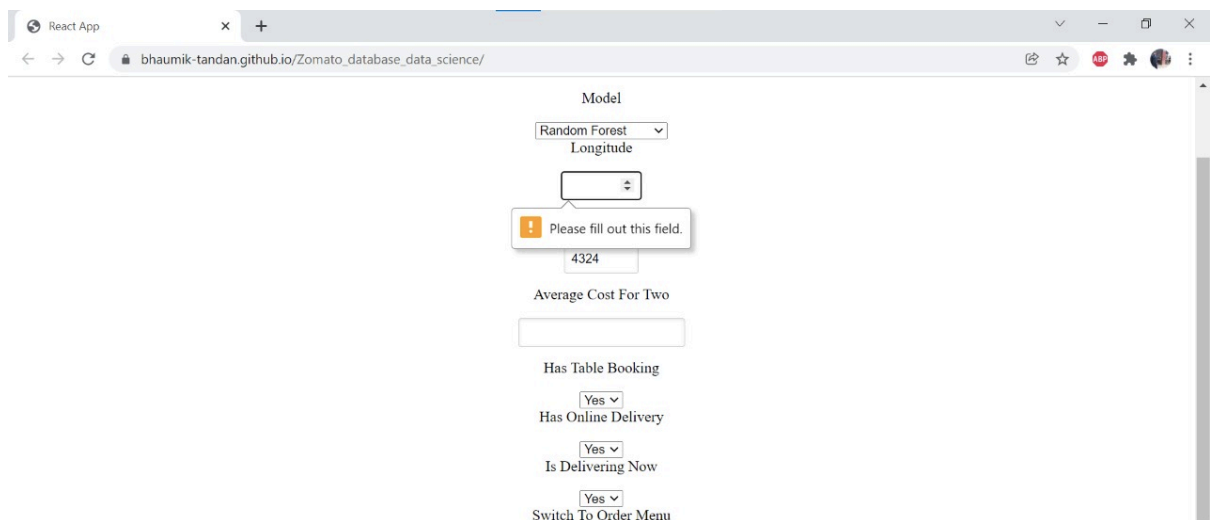
Loading Page -



After getting response from the API -



Form Validation -



React App

bhaumik-tandan.github.io/Zomato_database_data_science/

Rating Predictor

Server is located [Here](#)

Model
Random Forest

Longitude
32

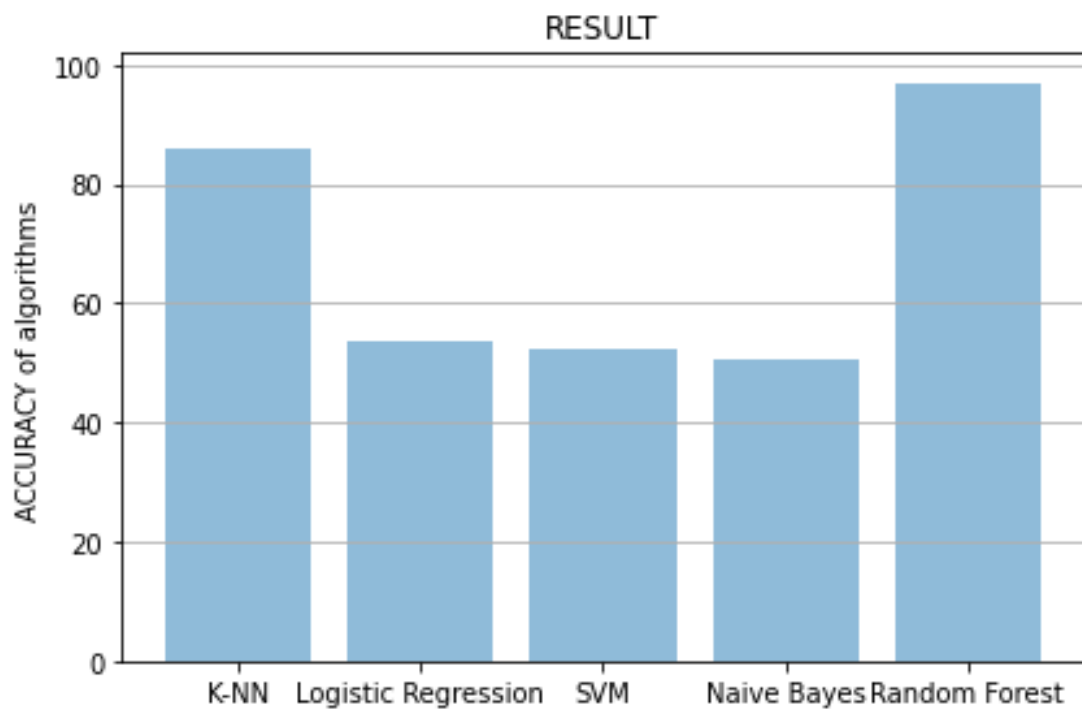
Latitude
4324

Value must be less than or equal to 90.

Has Table Booking
Yes

Has Online Delivery
Yes

RESULT



VII. COMPARATIVE STUDY

Due to the rich information contained in the Yelp dataset, many past research and projects tried to use it to predict ratings of restaurants and to evaluate the future development. For example, Kong, Nguyen and Xu in the paper Predicting International Restaurant Success with Yelp, Stanford University classified restaurants based on cultural categories and analysed international restaurants success mostly with Gaussian Discriminant analysis (GDa). Several other previous papers focused on the sentiment analysis with text content from Yelp. a few papers combined the customer reviews and ratings together to conduct sentiment analysis, while others mainly used matrix factorisation to analyse text information and predict Yelp ratings. Other than Yelp reviews, Tang, Qin, Liu and Yang in their paper User Modelling with Neural Network for Review Rating Prediction. IJCal. 2015 introduced neural network to predict restaurant reviews. The paper claimed that matrix-vector multiplication would be more effective than vector concatenation when considering text analysis. So far, most research works on text analysis of customer reviews, but leaves out other features in Yelp Dataset Challenge.

VIII. CONCLUSION AND FUTURE WORK

In this model, we have considered various restaurants records with features like the name, average cost, locality, whether it accepts online order, can we book a table, type of restaurant. This model will help business owners predict their rating on the parameters considered in our model and improve the customer experience. Future studies involve neural networks to undergo pruning to remove unwanted excess nodes. We can create more columns for cuisine types to separate the comma-separated cuisine. XGBoost and other algorithms can be used to check if they perform better than neural networks. Creating off unit tests for testing stages at the time of training the model, processing and prediction is also to be considered for future works. This model and the flask API provided by us can be used by many developers to integrate in their websites, which in turn will help the businessmen to know about their restaurants and help them gain more profit.

IX. REFERENCES

1. Uncovering Business Opportunities from Yelp, Gianni Passerini
2. Predicting Restaurants' Rating and Popularity Based On Yelp Dataset, Yiwen Guo, ICME, anran Lu, ICME, and Zeyu Wang
3. Reviews, Reputation, and Revenue: The Case of Yelp.com, Michael Luca
4. Restaurant Revenue Prediction using Machine Learning, Prof. Nataasha Raul, Yash Shah, Mehul Davganiya
5. YELP RATING PREDICTION WITH SENTIMENT AND TOPIC MODELS, Ying Liang
6. Service Rating Prediction and Recommender System-a Survey, T. Sivakumar , Sneha Prakash
7. Data mining for predicting customer satisfaction in fast-food restaurant, Bayu adhi Tama
8. Predicting Yelp Food Establishment Ratings Based on Business attributes,Peter Mark Shellenberger Jr
9. PREDICTING THE GROWTH OF RESTAURANTS USING WEB DATA, Yiea-Funk Te, Daniel Müller, Sebastian Wyder, Dwian Pramono
10. Predictive analytics Using Text Classification for Restaurant Inspections, Zhu Wang, Booma Sowkarthiga Balasubramani, Isabel F. Cruz
11. Restaurants Rating Prediction using Machine Learning algorithm, atharva Kulkarni, Divya Bhandari, Sachin Bhoite
12. Prediction of rating based on review text of Yelp reviews, Sasank Channapragada, Ruchika Shivaswamy
13. Data mining of restaurant review using WEKA, Gayathri.T
14. Yelp Dataset: Review Rating Prediction, Nabiha asghar
15. Machine learning based class level prediction of restaurant reviews, F. M. Takbir Hossain; Md. Ismail Hossain; Samia Nawshi