INSTITUTE FOR ADVANCED COMPUTING AND SOFTWARE DEVELOPMENT AKRUDI, PUNE

DOCUMENTATION ON

"Zomato Restaurant Success Prediction"

e-DBDA May-2021

Submitted by:

Group No.: 04

Katkar Amol J. (1304)

Kharat Ganesh B. (1359)

Mr. Prashant Karhale
Centre Coordinator

Mr.Akshay Tilekar
Project Guide

INDEX

Table of Contents

Chapter 1
Introduction

Zomato is the most reputed company in the field of food reviews. Founded in 2008, this company started in India and now is in 24 different countries. It is so big that the people now use it as a verb. “Did you know about this restaurant? Zomato it”. The rating is the most important feature of any restaurant as it is the first parameter that people look into while searching for a place to eat. It portrays the quality, hygiene and the environment of the place. Higher ratings lead to higher profit margins. Notations of the ratings usually are stars or numbers scaling between 1 and 5. Zomato has changed the way people browse through restaurants. It has helped customers find good places with respect to their dining budget. Different machine learning algorithms like SVM, Linear regression, Decision Tree, Random Forest can be used to predict the ratings of the restaurants.

1.1 Purpose

Zomato is an Online food ordering service, serving worldwide in which users can order food from the website or from mobile based applications. For this business problem, we are restricting only to the Bangalore region and Bangalore based restaurants. Dataset was created by extracting (web scraping) the information such as Approx. Price of food, Theme based restaurant or not, aggregate rating of each restaurant etc. about the existing established restaurants serving through Zomato and made available on Kaggle March 2019. The basic idea of analyzing the Zomato dataset is to get a fair idea about the factors affecting the establishment of different types of the restaurant at different places in Bangalore.

1.2 Related Work

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.

[1] Shina, Sharma S. and Singha A. have used Random forest and decision tree to classifying restaurants into several classes based on their service parameters. Their results say that the Decision

Tree Classifier is more effective with 63.5% of accuracy than Random Forest whose accuracy is merely 56%.

[2] Chirath Kumarasiri's and Cassim Faroo's focuses on a Part-of-Speech (POS) Tagger based NLP technique for aspect identification from reviews. Then a Naïve Bayes (NB) Classifier is used to classify identified aspects into meaningful categories.

[3] I. K. C. U. Perera and H.A. Caldera have used data mining techniques like Opinion mining and Sentiment analysis to automate the analysis and extraction of opinions in restaurant reviews.

1.3 Terms and Definitions

Terms	**Definitions**
Dataset	Data for training and testing for the model
Variance	Difference between the training and testing accuracy
Bias	Both learning and training accuracy is low
Overfitting	Model is very complex
Underfitting	Model is bias
Developer	Who is developing the model
Review	A written recommendation about the appropriateness of an Product for selling and buying may include suggestions for improvement.
Reviewer	A person that examines an Product and has the ability to recommend approval Product for buying or to request that changes be made in the Product.
Software Requirement Specification	A document that completely describes all of the functions of a proposed system and the constraints under which it must operate. For example, this document
User	Reviewer

Table 1.2.2.1: Terms and Definitions.

Chapter 2
Overall Description

2.1 Machine Learning Problem Formulation

The given problem can be solved either by binary classification problem (0 as failure or 1 as success), or Regression problem (for predicting scores) based on the given features.

- Approx. Price of food.
- Theme based restaurant or not.
- Which locality of that city serves that cuisines with maximum number of restaurants.
- The needs of people who are striving to get the best cuisine of the neighbourhood.
- Is a particular neighbourhood famous for its own kind of food etc.

Here, the objective is to predict the success of a restaurant. Based on the prediction from the Model, a new investor can make the decision on whether to establish the restaurant or not.

2.2 Performance Metric

Since we are solving this problem as a Binary classification task, where we need to predict whether a new restaurant will be successful of not, we will take classification metrics such as F1score/AUC into consideration.

2.3 Exploratory Data Analysis

Once understood the business problem and formulated the machine learning problem statement. We should be having a good knowledge of the datasets and most of the features. Now, we will do Exploratory Data Analysis on this dataset and to get more insights into the features.

From all the Data available, we can bring out some neat insights or conclusions such as

- *Which franchise has the highest number of Restaurants?*
- *How many Restaurants are accepting online orders?*
- *How many have a book table facility?*
- *Which location has the highest number of Restaurants?*
- *How many types of Restaurant types are there?*
- *and so on.*

Chapter 3

Requirement Specification

3.1 Detailed Non-Functional Requirements:

3.1.1 Functional Requirement:

First of all, model should be successfully trained by developer. Download and install Anaconda (windows version).

3.1.2 Hardware Requirement:

- **Processor:** Intel Dual Core
- **RAM:** Minimum 1GB
- **OS:** Windows, Linux. MacOS

3.1.3 Software Requirement:

Anaconda Navigator

Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda® distribution that allows you to launch applications and easily manage conda packages, environments and channels without using command-line commands. It is available for Windows, macOS and Linux.

Following libraries of python should be install:

1. **NumPy:** pip install numpy
2. **Pandas:** pip install pandas
3. **scikit-learn:** pip install scikit-learn
4. **WordCloud:** pip install wordcloud
5. **Folium:** pip install folium

Chapter 4

System Design

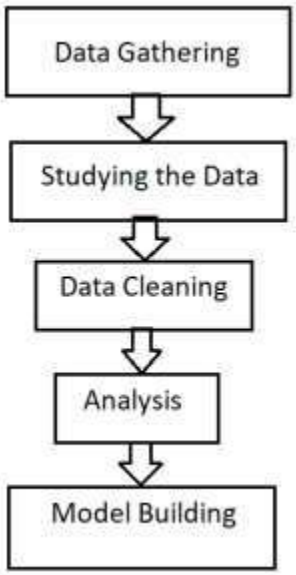


Figure.4.1.1: Flowchart of system

4.1 Data Gathering:

This is a kaggle dataset.

(https://www.kaggle.com/himanshupoddar/zomato-bangalore-restaurants).

It Represents information of Restaurants in the City of Bangalore. It contains 17Columns and 51,000 Rows

4.2 Studying the Data:

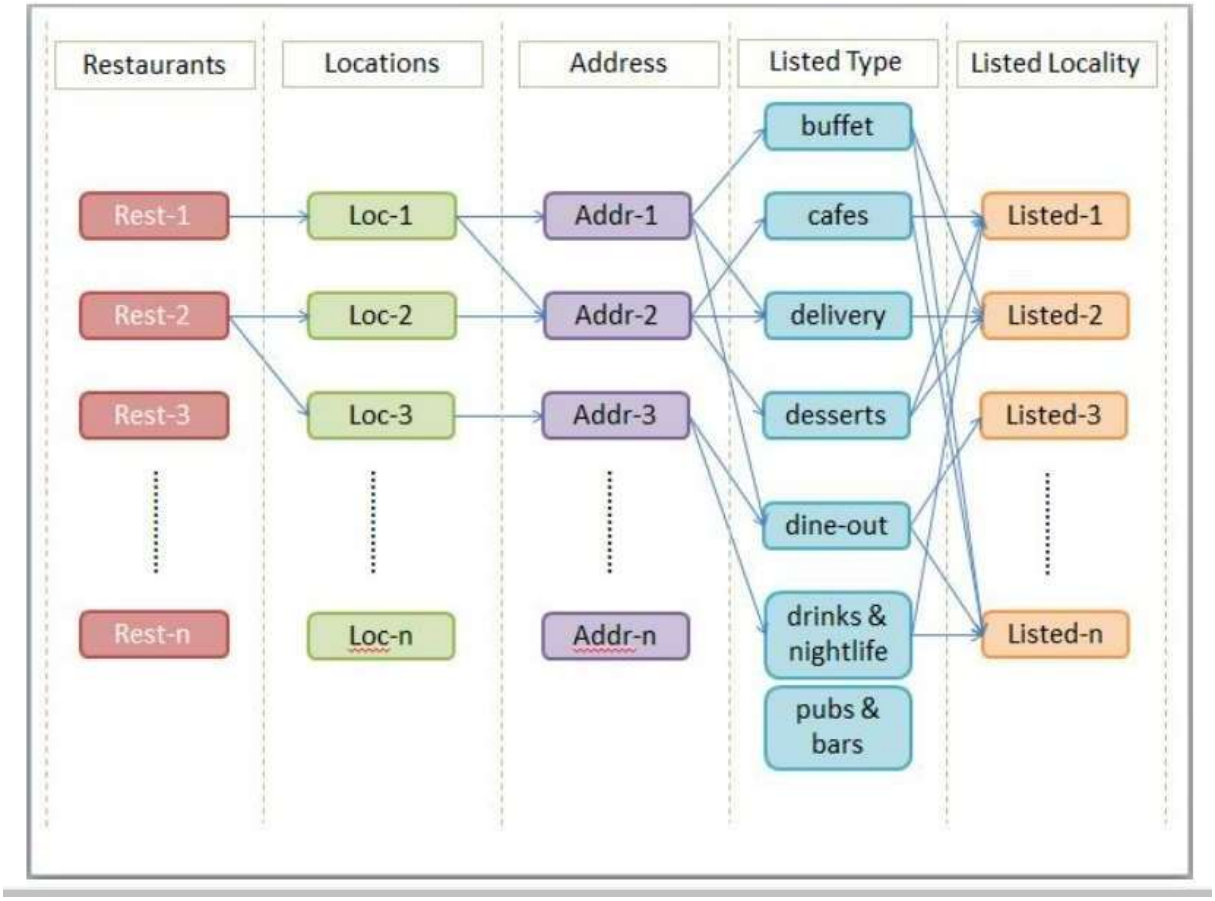
The dataset provided has approx. 50K records. Its features allow analyzing a restaurant from multiple dimensions. Following table provides the snapshot of all the columns.

Column	Description
url	contains the url of the restaurant in the zomato website
address	contains the address of the restaurant in Bengaluru
name	contains the name of the restaurant
online_order	whether online ordering is available in the restaurant or not
book_table	table book option available or not
rate	contains the overall rating of the restaurant out of 5
votes	contains total number of rating for the restaurant as of the above mentioned date
phone	contains the phone number of the restaurant
location	contains the neighborhood in which the restaurant is located
rest_type	restaurant type
dish_liked	dishes people liked in the restaurant
cuisines	food styles, separated by comma
approx_cost	contains the approximate cost for meal for two people
reviews_list	list of tuples containing reviews for the restaurant, each tuple consists of two values, rating and review by the customer
menu_item	contains list of menus available in the restaurant
listed_in	type of meal
listed_in(city)	contains the neighborhood in which the restaurant is listed

4.3 Data Cleaning:

'rating' column contains values such as 'NEW', which is for new restaurants and '-', for those restaurants which are not rated. We will remove these records along with nan. 'approx_cost (for two people)' column contains values with comma; hence this is considered as object, we will convert it back to float. We will also delete Unnecessary Columns 'phone', 'url'. Analysis shows we a have total of 8385 restaurants listed in Zomato Bangalore. Restaurant 'nu.tree' is situated in 3 locations. Every outlet of this restaurant for a specific location is having a separate entry based

on the features restaurant have such as Dine-out, Buffet & Delivery. i.e. We can have multiple records for the same restaurant in the dataset.

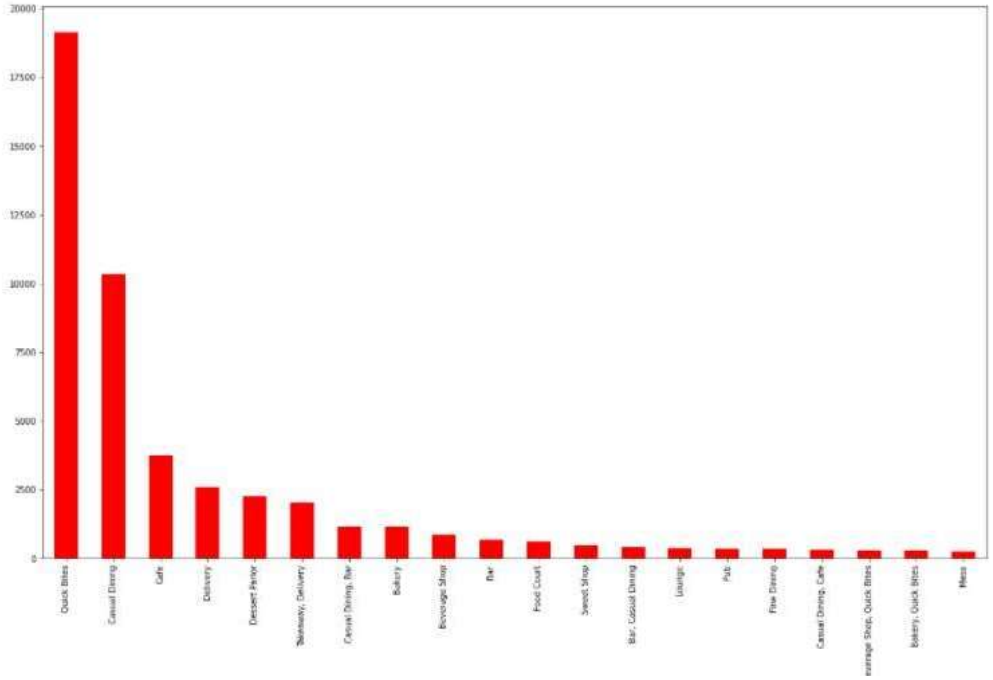


4.4 Analysis:

We started with Univariate analysis on important columns such as name, ‘online_order’, ‘book_table’, ‘rest_type’, ‘cuisines’, ‘approx_cost(for two people)’ etc.

In [33]:
```
#checking how many types of restaurants we have
import matplotlib.pyplot as plt
plt.figure(figsize=(20,12))
df['rest_type'].value_counts().nlargest(20).plot.bar(color='red')
```

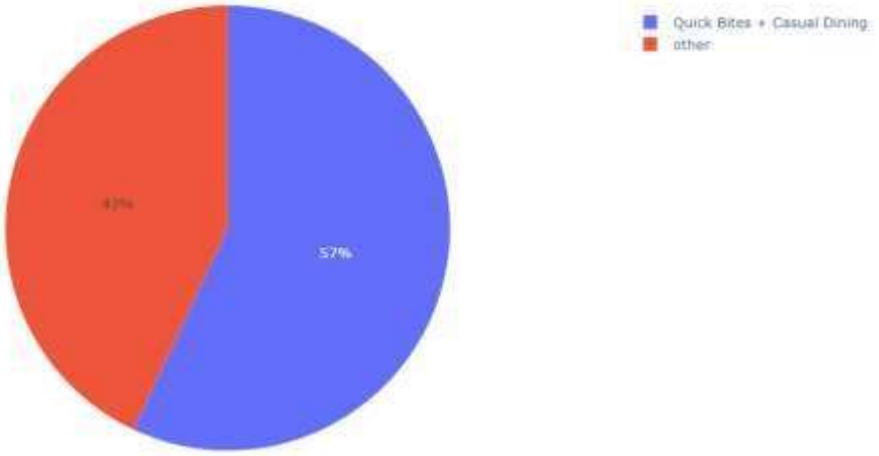
Out[33]: <AxesSubplot:>



In [36]:
```
#finding new feature using previous rest_types
#checking all restaurants who servicing Quick Bites as well as Casual Dining
def mark(x):
    if x in ('Quick Bites','Casual Dining'):
        return 'Quick Bites + Casual Dining'
    else:
        return 'other'
```

In [37]:
```
df['Top_types']=df['rest_type'].apply(mark)
```

In [44]:
```
fig=px.pie(df,names=labels,values=values)
fig.show()
```



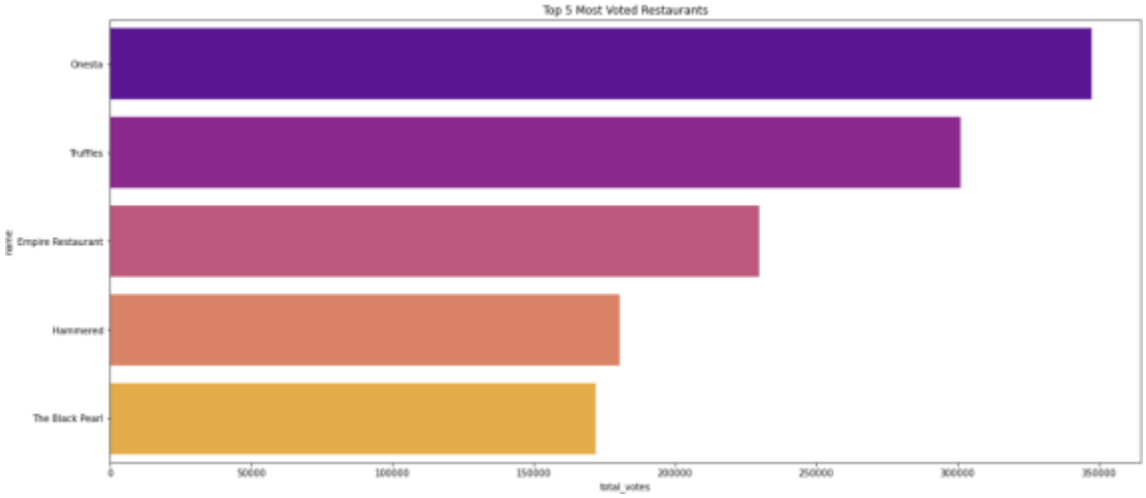
```
In [49]: #creating new datset popular for performing operation
popular=rest.sort_values(by='total_unities', ascending=False)
popular
```

Out[49]:

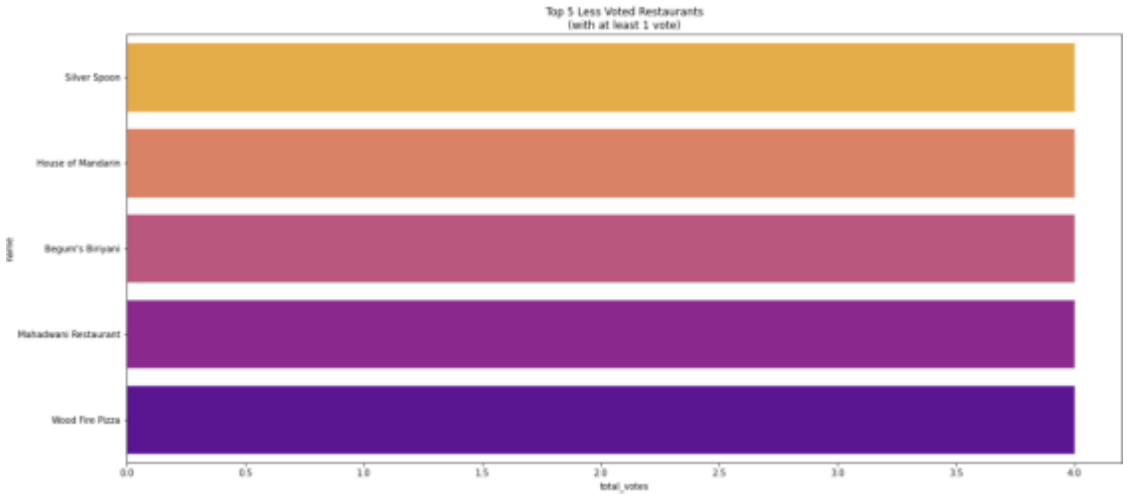
	name	total_votes	total_unities	avg_approx_cost	mean_rating	votes_per_unity
1320	Cafe Coffee Day	3089	96	844.791667	3.147191	32.177083
5549	Onesta	347520	85	600.000000	4.410588	4088.470588
3788	Just Bake	2898	73	400.000000	3.356882	39.698630
2446	Empire Restaurant	229808	71	685.211268	3.916901	3236.732394
2577	Five Star Chicken	3134	70	257.857143	3.425000	44.771429
...	...	...	...	...	...	...
1900	Cool Break	11	1	150.000000	3.400000	11.000000
8076	The Shake Factory Originals	8	1	200.000000	3.300000	8.000000
5216	Nethravathi Military Hotel	0	1	200.000000	NaN	0.000000
7426	Swadisht North Indian Restaurant	23	1	200.000000	3.200000	23.000000
5375	Night Punjabi Folk	0	1	200.000000	NaN	0.000000

8792 rows × 6 columns

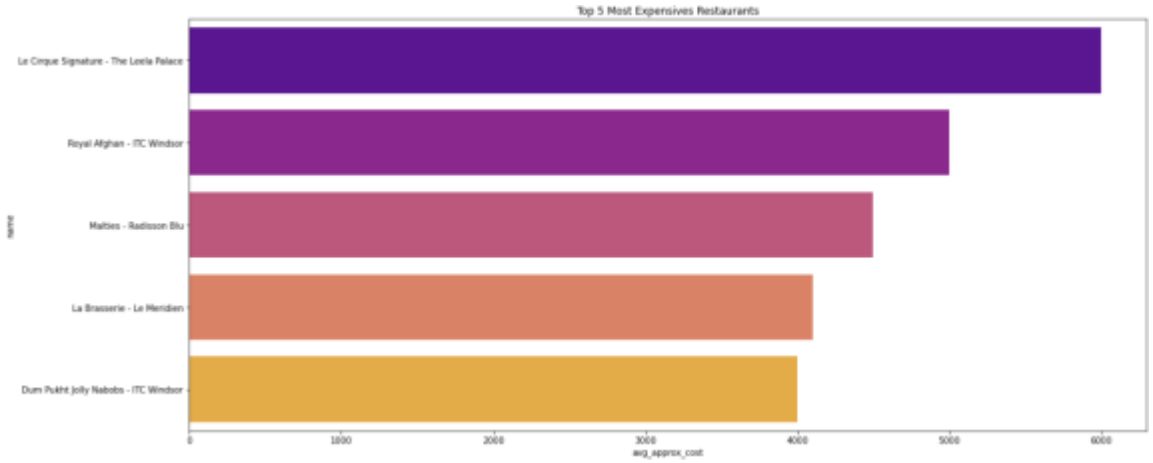
Finding Top 5 Most Voted Restaurants in Bangalore City



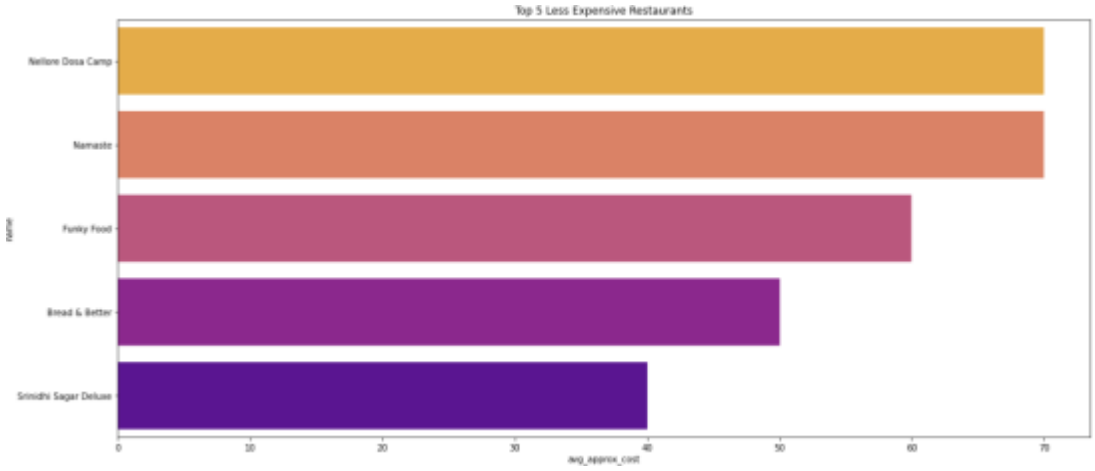
Finding Top 5 Least Voted Restaurants in Bangalore City



Finding Top 5 Most Expensive Restaurants in Bangalore City



Finding Top 5 Least Expensive Restaurants in Bangalore City

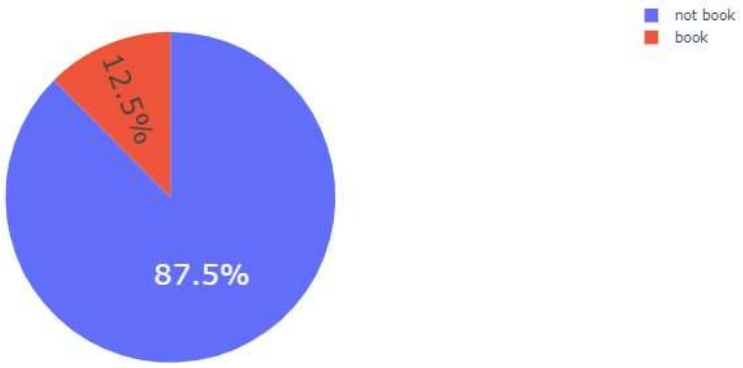


How many restaurants provide online ordering and booking table facility

Lets analyze on columns such as ‘online_order’ and ‘book_table’. From the count plot we can see that majority of the restaurants(15,000+) have online ordering facility. But most of the restaurants don’t have table booking facility.

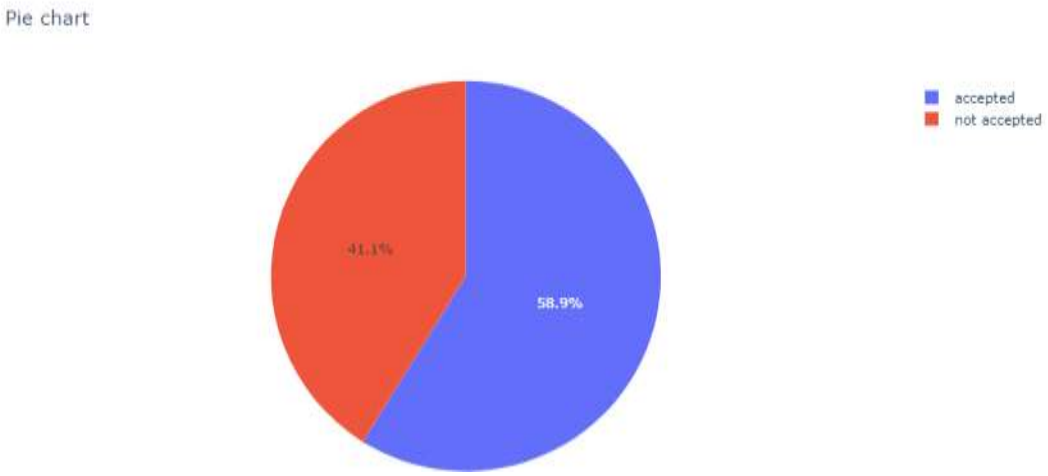
In [58]:
```
import plotly.graph_objs as go
from plotly.offline import iplot
x=df['book_table'].value_counts()
labels=['not book','book']
```

In [59]:
```
trace=go.Pie(labels=labels, values=x,
             hoverinfo='label+percent', textinfo='percent',
             textfont=dict(size=25),
            pull=[0, 0, 0,0.2, 0]
             )
iplot([trace])
```

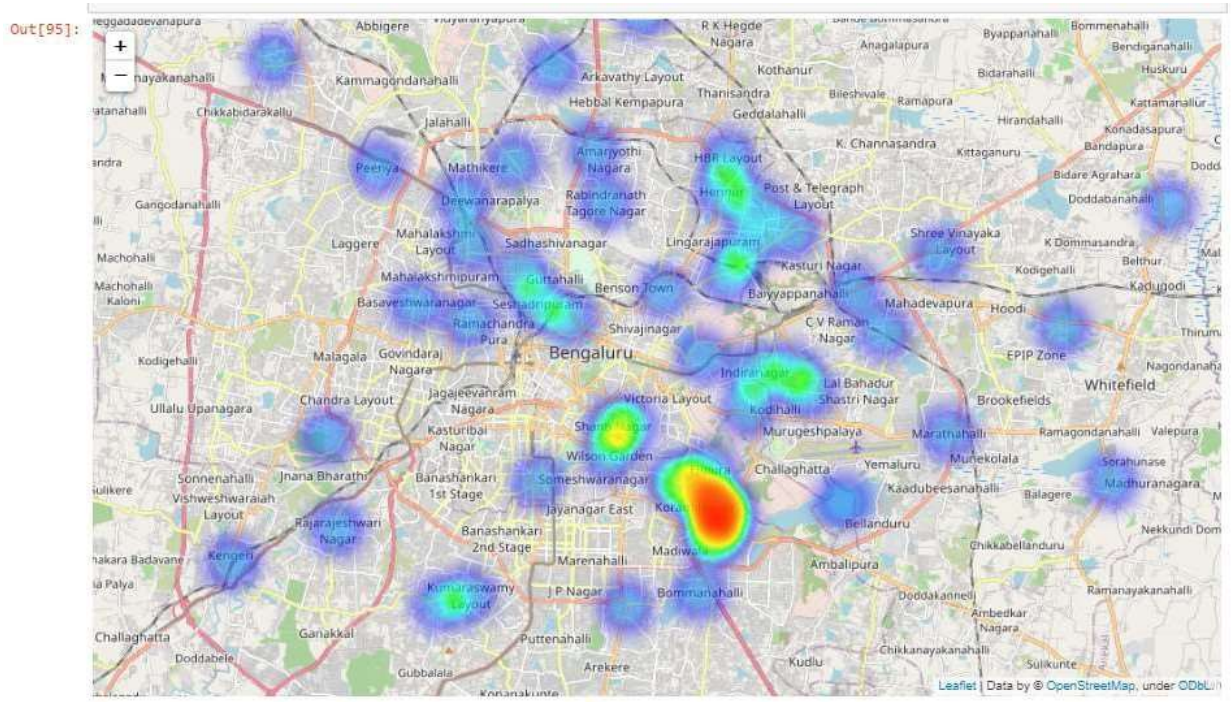


In [60]:
```
import plotly.express as px
x=df['online_order'].value_counts()
labels=['accepted','not accepted']
```

In [61]:
```
fig = px.pie(df, values=x, names=labels,title='Pie chart')
fig.show()
```



Geo Analysis: where are the restaurants located in Bengaluru?



This map vision session have low precision because we get the latitude and longitude information using the cities instead of the address. Again, we did this way because it was difficult to parse all the addresses in this dataset and also because of API's free tier time consuming and total requests limitations.

Even so, it's fair to say that some good information was delivered on the maps above. You can improve it by bringing precision geolocation information to this Zomato dataset.

4.5 Model Building:

Assumptions: We won't use the 'votes', 'reviews', 'dishes_liked' feature as long as this is a information we only know after launching a restaurant. As we want to be predictive, the idea is to return the probability of success of a restaurant before launching it.

Features:

- online_order
- book_table
- location
- rest_type
- cuisines

- city
- approx_cost

Split Data: We should first split data then apply featurization, to avoid data leakage problems. Divide data into Train, Test part.

Data featurization: We will use one-hot encoding technique for featurization to encode Categorical features such as ‘online_order’, ‘book_table’, ‘location’, ‘rest-type’ and ‘listed_in(type)’, listed_in(city).

```
In [164]: for feature in cat_features:
              print('{} has total {} unique features'.format(feature, data[feature].nunique()))

online_order has total 2 unique features
book_table has total 2 unique features
location has total 47 unique features
rest_type has total 11 unique features
listed_in(type) has total 7 unique features
listed_in(city) has total 30 unique features
```

```
In [165]: import pandas as pd
          data_cat = data[cat_features]
          for col in cat_features:
              col_encoded = pd.get_dummies(data_cat[col],prefix=col,drop_first=True)
              data_cat=pd.concat([data_cat,col_encoded],axis=1)
              data_cat.drop(col, axis=1, inplace=True)
```

We will train following models.

Models:

- Logistic Regression
- RandomForestClassifier
- XGBClassifier
- DecisionTreeClassifier
- KNN

```
In [186]: ### classifier models
          models = []
          models.append(('LogisticRegression', LogisticRegression()))
          models.append(('Naive Bayes',GaussianNB()))
          models.append(('RandomForest', RandomForestClassifier()))
          models.append(('Decision Tree', DecisionTreeClassifier()))
          models.append(('KNN', KNeighborsClassifier(n_neighbors = 5)))
```

In [188]:
```
# Make predictions on validation dataset

for name, model in models:
    print(name)
    model.fit(X_train, y_train)

    # Make predictions.
    predictions = model.predict(X_test)

    # Compute the error.
    from sklearn.metrics import confusion_matrix
    print(confusion_matrix(predictions, y_test))

    from sklearn.metrics import accuracy_score
    print(accuracy_score(predictions,y_test))
    print('\n')
```

```
LogisticRegression
[[3475 1526]
 [ 777 2477]]
0.7210175651120533


Naive Bayes
[[3040 1460]
 [1212 2543]]
0.6763173834039976


RandomForest
[[3513  942]
 [ 739 3061]]
0.7963658388855239


Decision Tree
[[3654  812]
 [ 598 3191]]
0.8291944276196245


KNN
[[3623 1006]
 [ 629 2997]]
0.8019382192610539
```

Chapter 5

Results

Model	Accuracy
LogisticRegression	0.7210
Naive Bayes	0.6763
RandomForest	0.7963
Decision Tree	0.8291
KNN	0.8019

From above table, Random forest classifier and Decision Tree having best performance metrics among the list of models we trained. 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.
Different algorithms were used but in the end we found Random forest classifier and Decision Tree having best performance metrics among the list of models we trained.

Chapter 6

Conclusion

In this project we studied 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.

Chapter 7

Future Scope

There is a lot to explore in this existing project. So, our future goal is to apply others algorithms in order to increase the accuracy score of our predictions regarding restaurants' rating obtained from both the features provided by restaurants and the reviews given by restaurants' customers. We will try to make our system more user friendly by adding a user interface which enables the user to provide features that the user is planning to have for their restaurant as input and our system will provide the user with a rating from 1 to 5 that they will get for their restaurant. Future work includes the use of unsupervised learning algorithms in conjunction to the supervised learning algorithms. This is because with the use of algorithms such as k-means clustering, the model is able to more closely to different geographic regions. Although a large k in this case would cause severe overfitting, a reasonable k value could result in a more accurate model due to differences in customers desires depending on the region. We also can try other classification algorithms.

Chapter 8

References

[1] Chirath Kumarasiri, Cassim Faroo,"User Centric Mobile Based Decision-Making System Using Natural Language Processing (NLP) and Aspect Based Opinion Mining (ABOM) Techniques for Restaurant Selection". Springer 2018. DOI: 10.1007/978-3-030-01174-1_4

[2] Shina, Sharma, S. & Singha ,A. (2018). A study of tree based machine learning Machine Learning Techniques for Restaurant review. 2018 4th International Conference on Computing Communication and Automation (ICCCA)
DOI:/10.1109/CCAA.2018.8777649

[3] I. K. C. U. Perera and H. A. Caldera, "Aspect based opinion mining on restaurant reviews," 2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA), Beijing, 2017, pp. 542-546. doi: 10.1109/CIAPP.2017.8167276

[4] Rrubaa Panchendrarajan, Nazick Ahamed, Prakhash Sivakumar, Brunthavan Murugaiah, Surangika Ranathunga and Akila Pemasiri. Eatery – A Multi-Aspect Restaurant Rating System. Conference: the 28th ACM Conference

[5] Neha Joshi. A Study on Customer Preference and Satisfaction towards Restaurant in Dehradun City.
Global Journal of Management and Business Research(2012)
Link: https://pdfs.semanticscholar.org/fef5/88622c39ef76dd773fcad8bb5d233420a270.pdf

[6] Bidisha Das Baksi, Harrsha P, Medha, Mohinishree Asthana, Dr. Anitha C.(2018) Restaurant Market Analysis.
International Research Journal of Engineering and Technology (IRJET)
Link: https://www.irjet.net/archives/V5/i5/IRJET-V5I5489.pdf