**Zomato Restaurants Bangalore**

**Context**

I was always fascinated by the food culture of Bengaluru. Restaurants from all over the world can be found here in Bengaluru. From the United States to Japan, Russia to Antarctica, you get all types of cuisines here. Delivery, Dine-out, Pubs, Bars, Drinks,Buffet, Desserts you name it and Bengaluru has it. Bengaluru is the best place for foodies. The number of restaurants is increasing day by day.

Currently which stands at approximately 12,000 restaurants. With such a high number of

restaurants. This industry hasn't been saturated yet. And new restaurants are opening every day. However it has become difficult for them to compete with already established restaurants. The key issues that continue to pose a challenge to them include high real estate costs, rising food costs, shortage of quality manpower, fragmented supply chain and over-licensing. This Zomato data aims at analysing 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 similarity between neighborhoods of Bengaluru on the basis of food. The dataset also contains reviews for each of the restaurants which will help in finding overall ratings for the place.

# Content

The basic idea of analyzing the Zomato dataset is to get a fair idea about the factors affecting the establishment of different types of restaurant at different places in Bengaluru, aggregate rating of each restaurant, Bengaluru being one such city has more than 12,000 restaurants with restaurants serving dishes from all over the world.

With each day new restaurants opening the industry hasn't been saturated yet and the demand is increasing day by day. In Spite of increasing demand, it has become difficult for new restaurants to compete with established restaurants. Most of them serve the same food. Bengaluru being an IT capital of India. Most of the people here are dependent mainly on the restaurant food as they don’t have time to cook for themselves.

With such an overwhelming demand of restaurants it has therefore become important to study the demography of a location. What kind of a food is more popular in a locality? The entire locality loves vegetarian food.

If yes then is that locality populated by a particular sect of people for eg. Jain, Marwaris, Gujaratis who are mostly vegetarian. This kind of analysis can be done using the data, by studying the factors such as

* Location of the restaurant
* 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 neighborhood
* Is a particular neighborhood famous for its own kind of food? “Just so that you have a good meal the next time you step out”

The data is accurate to that available on the Zomato website until 15 March 2019.

The data was scraped from Zomato in two phases. After going through the structure of the website I found that for each neighborhood there are 6-7 categories of restaurants viz. Buffet, Cafes, Delivery, Desserts, Dine-out, Drinks & nightlife, Pubs and bars.

There are two separate files, while the columns are self-explanatory. Below is a brief description:

1. Restaurant names and Metadata - This could help in clustering the restaurants into segments. Also the data has valuable information around cuisine and costing which can be used in cost vs. benefit analysis
2. Restaurant reviews - Data could be used for sentiment analysis. Also the metadata of reviewers can be used for identifying the critics in the industry.

Phase I,

In Phase I of extraction only the URL, name and address of the restaurant were extracted which

were visible on the front page. The URl's for each of the restaurants on the zomato were recorded in the csv file so that later the data can be extracted individually for each restaurant. This made the extraction process easier and reduced the extra load on my machine

Phase II,

In Phase II the recorded data for each restaurant and each category was read and data for each

restaurant was scraped individually. For each of the neighborhood and for each category their online order*, book*table, rate, votes, phone, location, rest type*, disliked*, cuisines, approx cost*(for twopeople), reviews*list, menu\_item was extracted.

# Acknowledgements

The data scraped was entirely for educational purposes only. Note that I don’t claim any copyright for the data. All copyrights for the data are owned by Zomato Media Pvt. Ltd..

# Inspiration

I was always astonished by how each of the restaurants are able to keep up the pace in spite of that cutting-edge competition. And what factors should be kept in mind if someone wants to open new

restaurant. Does the demography of an area matters? Does location of a particular type of

restaurant also depends on the people living in that area? Does the theme and rating of the restaurant matter? Is a food chain category restaurant likely to have more customers than its

counterpart? Are any neighborhoods similar? If two neighborhood are similar does that mean these are related or particular group of people live in the neighborhood or these are the places to it? What kind of a food is more popular in a locality. Do the entire locality loves vegetarian food. If yes then is that locality populated by a particular sect of people for eg. Jain, Marwaris, Gujaratis who are mostly vegetarian.

**About the file:**

It contains the names, links and other metadata of each restaurant, which could help in clustering

the restaurants into segments. Also the data has valuable information around cuisine and costing

which can be used in cost vs. benefit analysis.

**Data Pre-Processing**

As we see that the dataset is not clean and contains redundant data we start by deleting the unnecessary or redundant features. For data analysis, we delete "name", "url", "phone", "listed\_in(city)", "listed\_in(type)\_x", "address", "dish\_liked", "listed\_in(type)\_y", "menu\_item", "cuisines", "reviews\_list" as we are more interested to infer the characteristics of rating with respect to the dataset. We consider the first 3 characters present in the rating column as its string datatype and get an overall distribution of the rating for the whole dataset. We get fair insights of different parameters like restaurant location, highest rated restaurants based on locations and similar insights that motivated to take up the dataset. Further the ratings are segmented and sorted into bins to make the ratings categorical variables.

**Model Building**

The dataset is split in 70:30 division, i.e. 70% comprises of the training dataset and 30% comprisies of the test dataset. The missing values are filled based on their median values for numerical variables like votes and cost, whereas for the categorical variables the most frequently value along the column is used for variables namely location, rest\_type, online\_order and book\_table.

We have made use of Decision Tree Classifier, Random Forest Classifier and XG Boost Classifer in our analysis.

**Model Evaluation**

1) **Decision Tree Classifier:**

Training accuracy : 0.993

Test accuracy: 0.566

2) **Random Forest Classifier:**

Training accuracy : 0.693

Test accuracy: 0.680

3) **XG Boost Classifier:**

Training accuracy : 0.716

Test accuracy: 0.666

**Results summary**

A basic decision tree model is overfitting on the train data with a higher accuracy rate. The Random Forest Classifier is hyper parameter tuned manually. The RF classifier has not predicted any samples for the minority class (0) in the test data, which means it has not learnt that specific class well enough.

We split the restaurant ratings into four ranges (classes) and built tree-based classifiers to predict them. The best performer was a manually tuned XGBoost classifier with 72% accuracy on train and 67% accuracy on test. But since the rating classes are somewhat imbalanced, we can further evaluate our predictions with Cohen Kappa and F1 scores.

Random Forest also gave similar scores but the class counts in RF predictions showed that minority classes were being misclassified, i.e. almost all restaurants with "rare" ratings were being misclassified as having a more "common" rating. This made the classifier overcompensate for the minority classes which incorrectly classified too many restaurants and brought down the scores.

XGBoost with manually tuned depth, learning rate and gamma (regularization) hyperparameters predicted more minority samples correctly than RF. The most important feature for prediction was Votes, followed by rest\_type\_Dessert\_Parlor.