**Zomato Restaurant Clustering & Sentiment Analysis**

**Technical documentation**

**Team Member**

**Srijan Rai**

**Abstract:**

Zomato is an Indian restaurant aggregator and food delivery start-up founded by Deepinder Goyal and Pankaj Chaddah in 2008. Zomato provides information, menus and user-reviews of restaurants, and also has food delivery options from partner restaurants in select cities.

India is quite famous for its diverse multi cuisine available in a large number of restaurants and hotel resorts, which is reminiscent of unity in diversity. Restaurant business in India is always evolving. More Indians are warming up to the idea of eating restaurant food whether by dining outside or getting food delivered. The growing number of restaurants in every state of India has been a motivation to inspect the data to get some insights, interesting facts and figures about the Indian food industry in each city. So, this project focuses on analysing the Zomato restaurant data for each city in India.

The Project focuses on Customers and Company, we have to analyse the sentiments of the reviews given by the customer in the data and make some useful conclusions. Also, cluster the Zomato restaurants into different segments. The data is visualized as it becomes easy to analyse data at instant. The analysis also solves some of the business cases that can directly help the customers finding the Best restaurant in their locality and for the company to grow up and work on the fields they are currently lagging in.

**Keywords:**

Restaurant, Sentiment, Machine Learning (ML), Clustering, Data Frames, Data Analysis.

**Problem Statement:**

Two datasets (Zomato Restaurant names and Metadata and Zomato Restaurant reviews), are provided in this project were various details about the restaurants and the reviews provided by the costumers are available

The goal of this project is to build an unsupervised clustering model and a supervised machine learning model which will help us to cluster these restaurants based on different parameters and help us to predict the sentiment of the review provided by the customers.

After performing the above processes, necessary conclusion must be obtained which can be visually presented, that help us to solve some of the important business needs of the company.

**2. Introduction:**

Founded in 2008 Zomato is a major food delivery aggregator with a markdown cap of 1 Trillion INR. It started as Foodiebay, a restaurant recommendation product, at its peak, it has 35000 menus and Rs 60 Lakh monthly revenue. Foodiebay.com reroutes to zomato.com now.

Zomato is an Indian restaurant aggregator and food delivery start-up founded by Deepinder Goyal and Pankaj Chaddah in 2008. Zomato provides information, menus, and user reviews of restaurants, and also has food delivery options from partner restaurants in select cities.

India is quite famous for its diverse multi-cuisine available in a large number of restaurants and hotel resorts, which is reminiscent of unity in diversity. The restaurant business in India is always evolving. More Indians are warming up to the idea of eating restaurant food whether by dining outside or getting food delivered. The growing number of restaurants in every state of India has been a motivation to inspect the data to get some insights, interesting facts, and figures about the Indian food industry in each city. So, this project focuses on analysing the Zomato restaurant data for each city in India.

The Project focuses on Customers and Company, we have to analyse the sentiments of the reviews given by the customer in the data and made some useful conclusions in the form of Visualizations. Also, cluster the Zomato restaurants into different segments. The data is visualized as it becomes easy to analyse data in an instant. The Analysis also solve some of the business cases that can directly help the customers find the Best restaurant in their locality and for the company to grow up and work in the fields they are currently lagging in.

**Python**

Most data scientists use python due to the good built-in library functions and therefore the decent community Specifically, for data scientists, the foremost popular data-inbuilt open-source library is named pandas. We can plot scatterplots, heat maps, graphs, and 3-dimensional data, using the python built-in library. Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modelling including classification, regression, clustering, and dimensionality reduction via a consistency interface in Python.

Unsupervised learning could also be performed using many clustering algorithms provided in python and Sklearn library. We also have Natural language Toolkit (NLTK) library which provides us various algorithms which helps in various natural language processing (NLP) problems.

**Zomato Restaurant Clustering & Sentiment Analysis datasets**

Two datasets are provided in this project to achieve our objective of the project. The first dataset i.e. Zomato Restaurant names and Metadata is a dataset provides us various details of different restaurants. The second dataset i.e. Zomato Restaurant reviews dataset provides us various reviews and other customer details for the restaurants.

**Zomato Restaurant names and Metadata:**

Name : Name of Restaurants

Links : URL Links of Restaurants

Cost : Per person estimated Cost of dining

Collection : Tagging of Restaurants w.r.t. Zomato categories

Cuisines : Cuisines served by Restaurants

Timings : Restaurant Timings

**Zomato Restaurant reviews:**

Restaurant : Name of the Restaurant

Reviewer : Name of the Reviewer

Review : Review Text

Rating : Rating Provided by Reviewer

MetaData : Reviewer Metadata - No. of Reviews and followers

Time: Date and Time of Review

Pictures : No. of pictures posted with review

**3. Steps Involved:**

* **Reading and Understanding data frame:**

The first step in our project is to create the data frames. This can be achieved by loading the given CSV files and converting them to Data frames using Pandas Library provided by Python. These data frames are now the source of raw data for our project. From this data frame, we get to know various statistical information about the features in the data frame. This initial analysis will help us to perform the necessary data preprocessing required to build a good regression model.

From the initial understanding of the data, we got to know that the metadata dataframe contains105 entries and 6 features. And the reviews dataframe contains 10000 entries and 7 features. Some of the features in both the dataframes contains null values and we need to also check the duplicates in the dataframes.

* **Data Pre-Processing:**

Handling Duplicates:

There were no duplicates present in the metadata dataframe. And for the reviews dataframe there were some duplicates, but all the values were null in these duplicates, hence we would be dropping these duplicate values

Handling Null Values:

There were a lot of null values under Collections feature in the metadata dataframe. Since the null values are large in number, dropping these data would result in loss of data, hence we would be replacing the null values under collections as “No Collection”.

There was a very small amount of null values in reviews dataframe, and these entries were dropped from the dataframe.

Handling Outliers:

From the boxplots and distribution plots we got to know that Metadata dataframe do not have any outliers, but the reviews dataframe had some outliers which were dropped from the dataframe

* **Feature Engineering:**

Metadata dataframe: Converted the datatype of Cost feature to float, Created a new feature called the number of cuisines, to obtain the number of cuisines served in a restaurant, dropped the Links and Timings features which are not necessary for our further analysis.

Reviews dataframe: Converted the datatype of Rating Feature to float, Converted the datatype of "Time" to datetime and created new features to get the day and hour from it, Obtain the number of reviews and followers of the restaurant reviews.

* **Data Visualization:**

The data in the dataframes were visualized using matplotlib and seaborn libraries in python. From the data visualization, we got to know:

* Bar plots were obtained to visualize various parameters for all the restaurants and top 5 restaurants in different parameters were obtained.
* Majority of the restaurants serve North Indian cuisines.
* Most of the people order food on weekends.
* Most of the people order food for their lunch and dinner.

**4. Restaurant Clustering:**

**Initial Hypothesis:**

* Restaurants with similar kinds of ratings can be clustered together. Ratings are done by people on the basis of food quality, service, packaging among other things.
* Restaurants with high ratings would also probably be expensive and would be having a similar pricing strategy as well. They can be clustered according to the costs. Hence we will create clusters using the price and ratings
* Restaurants having similar cusines can be clustered together.

**Data Pre-processing:**

One-Hot encoding was performed on the metadata dataframe to create the new cusines features which will help us to cluster the restaurants based on the cusines they serve.

Since many of our clustering algorithms depends on calculating the distance between each cluster and among the data points within the clusters. It is better to obtain a uniform scaling across the dataframe to achieve a better model performance. We would be using the MinMax scaler offered by scikit-learn to perform scaling on our dataframe.

**Kmeans Clustering on Cost and Ratings:**

5 Clusters were obtained when we consider the Cost and Ratings of the restaurants to perform the clustering, the clusters are shown in the below scatter plot.

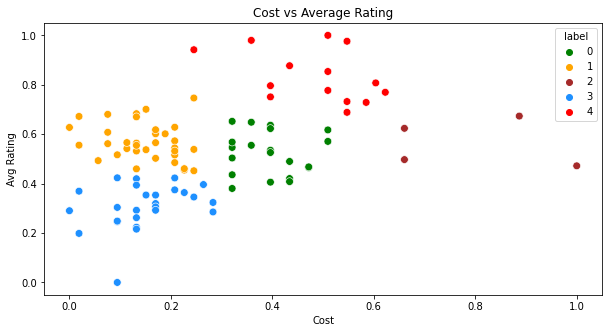


Fig1: Ratings vs Cost Clusters

* Label 0 are the restaurants with good rating and average pricing.
* Label 1 are the restaurants with low pricing and good rating.
* Label 2 are some expensive restaurants with good rating.
* Label 3 are the restaurants with low rating and low pricing.
* Label 4 are the restaurants whith high rating and average pricing.

**Kmeans Clustering on clustering dataframe:**

3- Clusters were obtained when we performed the clustering on the dataframe by considering the cuisines served in the restaurants.

* Cluster 0 are the type of restaurants that serve healthy food and hence these restaurants can be targeted to highly health-conscious people.
* Cluster 1 consists of restaurants that serve moderately healthy food, these restaurants would be recommended for casual dining.
* Cluster 2 consists of restaurants that serve unhealthy food such as Junk foods, bakeries, street food, etc. Most of the fast-food
* Cluster 0 restaurants have a high average rating and a high average cost. Also, the number of restaurants is very limited.
* Cluster 1 restaurants have a low rating and average cost and have a considerable number of restaurants.
* Cluster 2 restaurants have a moderate rating and a very low price and have a greater number of restaurants compared to other clusters.

**Hierarchical Clustering on dataframe:**

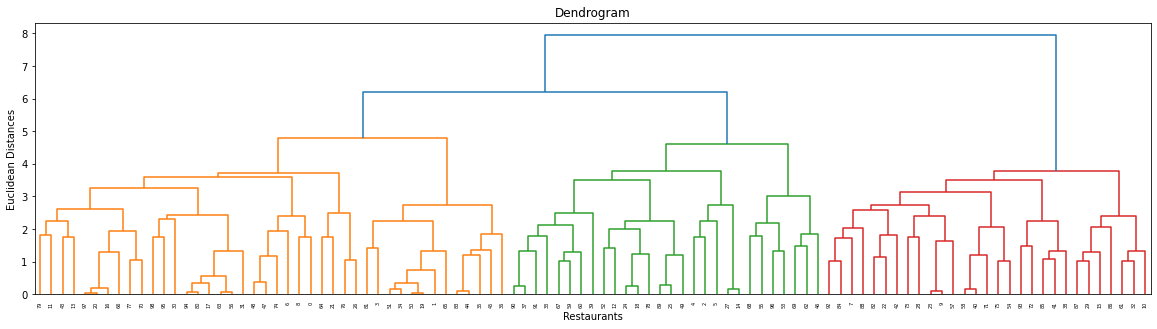
****

Fig2: Hierarchical Clustering

From the above dendrogram, we could see that there are 3 prominent clusters of restaurants in the dataframe.

**5. Sentiment Analysis:**

**Data Pre-Processing:**

Created the new "Sentiment" feature, where it will return the sentiment as positive when the rating is equal to or above 3.5, and negative when it is below 3.5.

We will apply a Supervised Machine learning approach, here the dependent variable are the reviews by the customers, and the dependent variable is the sentiment feature that we created for the analysis.

* **Text Pre-Processing:**

The following Text Preprocessing we considered in this project:

* Remove newline(\n) tags
* Remove html tags
* Transform the emojis to its respective text form
* Remove punctuation
* Remove numbers from the reviews
* Remove brackets from the reviews
* Converting the reviews to lower case
* **Stopwords Removal:**

Stop words are available in abundance in any human language. By removing these words, we remove the low-level information from our text in order to give more focus to the important information.

* **Stemming:**

Stemming is a natural language processing technique that lowers inflection in words to their root forms, hence aiding in the preprocessing of text, words, and documents for text normalization.

We would we using the snowball stemmer offered by nltk.

* **Train Test Split:**

75% of the data is utilized for training and 25% of data is utilized for testing.

* **TFID Vectorization:**

TF-IDF is an abbreviation for Term Frequency Inverse Document Frequency. This is very common algorithm to transform text into a meaningful representation of numbers which is used to fit machine algorithm for prediction

**6. Machine Leaning Models:**

As we mentioned earlier, we would be applying the supervised machine learning approach on the dataframe to obtain a model to predict the sentiment of the review. Since the dependent variable is dichotomous in nature, we would require classification models to achieve our objective.

Before selecting the models, let us determine some useful evaluation metrics that will be used in our project, which will be required for model selection.

**Evaluation Metrics:**

* **Accuracy**: Accuracy simply measures how often the classifier correctly predicts. We can define accuracy as the ratio of the number of correct predictions and the total number of predictions.
* **Precision**: Precision explains how many of the correctly predicted cases actually turned out to be positive. Precision is useful in the cases where a False Positive is a higher concern than a False Negative.
* **Recall**: Recall explains how many of the actual positive cases we were able to predict correctly with our model. It is a useful metric in cases where a False Negative is of higher concern than a False Positive.
* **F1 Score**: The F1 Score is the harmonic mean of precision and recall.
* **AUC-ROC**: The Receiver Operator Characteristic (ROC) is a probability curve that plots the TPR(True Positive Rate) against the FPR(False Positive Rate) at various threshold values and separates the ‘signal’ from the ‘noise'. The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes.

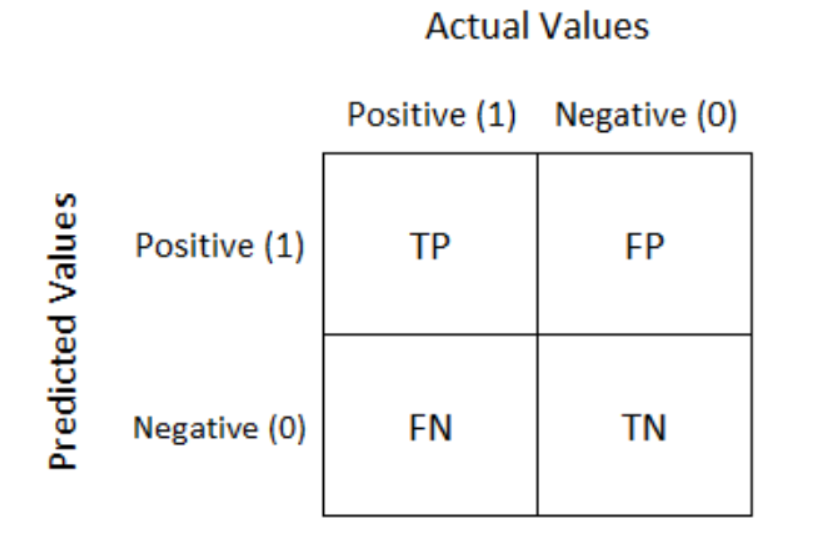


Fig3: Confusion Matrix

**Classification Models:**

* **Logistic regression:** Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary).  Like all regression analyses, logistic regression is a predictive analysis.
* **Decision Tree:** A decision tree is a type of supervised learning algorithm that is mostly used in classification problems. It works for both categorical and continuous input and output variables. Decision trees use multiple algorithms to decide to split a node in two or more sub-nodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes.
* **GradientBoosting:** Gradient Boosting is a popular boosting algorithm. In gradient boosting, each predictor corrects its predecessor’s error.
* **XGBoost:** XgBoost is one of the fastest implementations of gradient boosting. trees. It does this by tackling one of the major inefficiencies of gradient boosted trees: considering the potential loss for all possible splits to create a new branch (especially if you consider the case where there are thousands of features, and therefore thousands of possible splits).
* **RandomForest:** Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as bagging.
* **K-nearest neighbor (KNN):** KNN algorithm stores all the values in the data frame during its training and classifies the new data points based on the similarity among the previous data points.
* **Support Vector Machine (SVM):** Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well it’s best suited for classification. The objective of the SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points.
* **Naïve Bayes’:** Naive Bayes’ classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

**Model Results:**

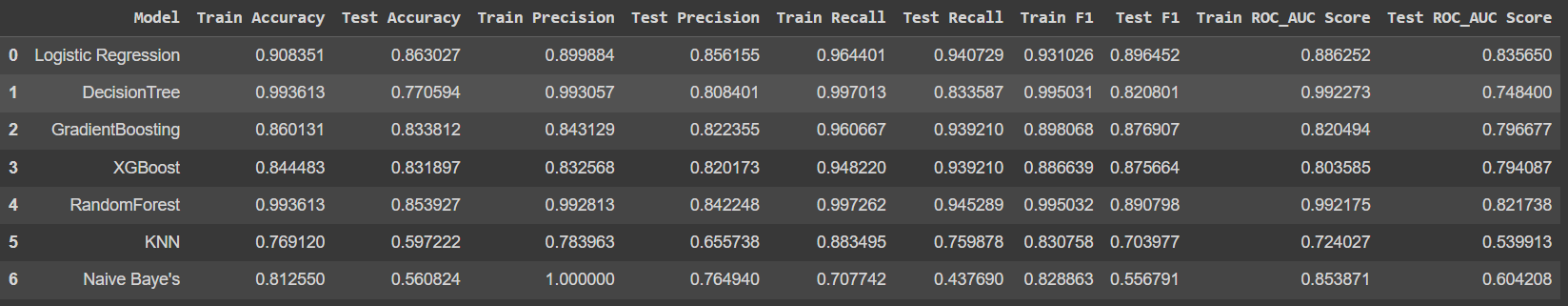
****

Fig4: Evaluation Metrics Results

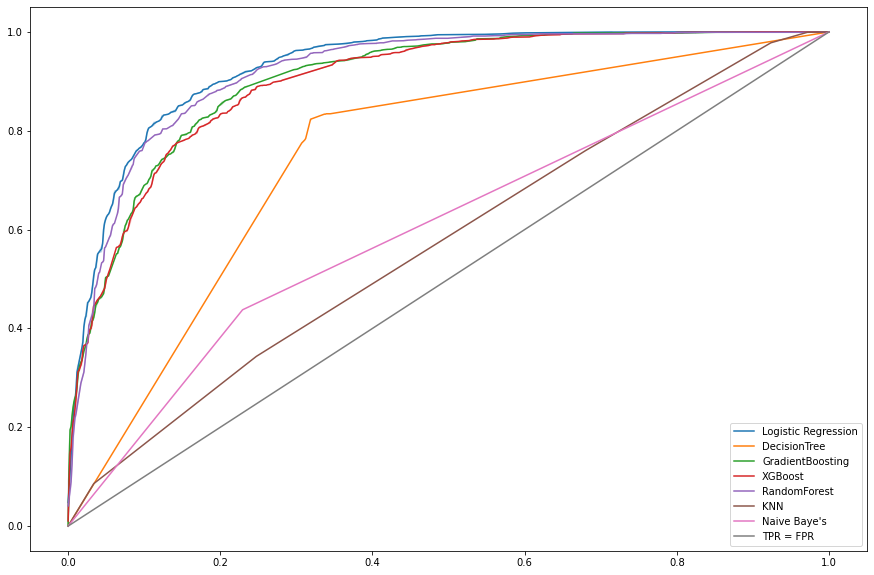


Fig5: AUC-ROC

From the above figures, we could see that Logistic Regression model outperform the rest of the models with good evaluation metrics values and a better area under the roc-curve.

Decision tree and Random Forest have a good training performance but no testing performance, hence indicating a potential overfitting.

**Hyperparameter Tuning:**

From the model results, we have shortlisted the machine learning models to Logistic Regression model. To further improve its performance, we would tune its hyperparameters.

We would be using the GridSearchCV cross-validation method to achieve this objective, there will be 5 cross-validations performed, with the scoring parameter as the ROC-AUC score.

From the hyperparameter tuning the Logistic Regression model’s performance was not raised further, hence we would conclude that the present obtained model is the best machine learning model among all the models selected in this project.

We would carry out some additional testing on this model by passing some reviews which are not present in the dataframe and confirm its performance.

**5. Conclusion:**

From the above data analysis, clustering and sentiment analysis on various restaurants and their reviews, the following observations were acquired:

* Barplots were obtained to visualize various parameters for all the restaurants and top 5 restaurants in different parameters were obtained.
* Majority of the restaurants serve north indian cuisines.
* Most of the people order food on weekends.
* We have obtained 5 different clusters of restaurants when we considered their ratings and cost.
* 3 different clusters of restaurants were obtained that serve healthy, moderately heathy and unhealthy food respectively.
* Sentiment analysis was carried out on the reviews posted by the users.
* Logistic Regression model was the best suited model to carry out the sentiment analysis on the given data.

**6. References-**

<https://www.analyticsvidhya.com>, <https://www.kaggle.com>, <https://towardsdatascience.com>,

<https://www.geeksforgeeks.org>