**Zomato Data Analysis**



Zomato is a restaurant search and discovery service founded in 2008. It currently operates in 24 countries. It provides information and reviews on restaurants, including images of menus where the restaurant does not have its own website. The primary goal of this project is to perform statistical analysis and twitter sentimental analysis to visualize its presence over the world, types of services it provides, performance in the market as a web application.

The source of this project is Kaggle and the dataset is downloaded from Kaggle (<https://www.kaggle.com/shrutimehta/zomato-restaurants-data/data>).

This data was collected from Zomato API and the source was provided as a CSV file. Each restaurant in the dataset is uniquely identified by its Restaurant Id. Every Restaurant contains the following variables:

* Restaurant Id: Unique id of every restaurant across various cities of the world
* Restaurant Name: Name of the restaurant
* Country Code: Country in which restaurant is located
* City: City in which restaurant is located
* Address: Address of the restaurant
* Locality: Location in the city
* Locality Verbose: Detailed description of the locality
* Longitude: Longitude coordinate of the restaurant's location
* Latitude: Latitude coordinate of the restaurant's location
* Cuisines: Cuisines offered by the restaurant
* Average Cost for two: Cost for two people in different currencies
* Currency: Currency of the country
* Has Table booking: yes/no
* Has Online delivery: yes/ no
* Is delivering: yes/ no
* Switch to order menu: yes/no
* Price range: range of price of food
* Aggregate Rating: Average rating out of 5
* Rating color: depending upon the average rating color
* Rating text: text on the basis of rating of rating
* Votes: Number of ratings casted by people

From this analysis I want to address the relevant business questions using various statistical methods and machine learning methods and packages.

Business questions -

Top 15 cuisines on Zomato

Top 15 restaurant chains with maximum outlets

Restaurants delivering Online or not

Restaurants Has Table booking

What are the top 15 restaurants with Maximum Outlets

Has online delivery vs restaurant rating

Average Price vs restaurant ratings

Interactive maps of Zomato’s presence across the globe

I used linear regression model to derive the relation between our independent and dependent variables like how other factors effect price range. Also to find relation.

I used different classification models to find how factors like avg price, rating etc. effect table booking. Also is there any impact on online delivery parameter based on other features.

These models will let us know how these factors drives ratings and other factors.

Also, I plan to perform sentimental analysis on Twitter data for positive, negative and neutral to derive conclusions. Sentiment analysis or opinion mining is a mechanized method used by businesses to detect positive sentiments or negative sentiments in a text to understand the customers better and keep an eye on its reputation. Sentiment analysis helps us to identify the motive and emotions behind a purchase, gain many more insights and eventually predict the future response.

Key takeaways from this analysis – the implications to the consumer/business owners

* Upon understanding the consumer preferences, the restaurant owners can better target them.
* The customers get better understanding on top cuisines, top restaurant chains amd other factors.
* Predicting what factors effect the prices and other attributes.
* The twitter sentiment analysis for positive, negative and neutral reviews will help the business owners and customers understand where the Zomato business stands in terms of performance and customer reviews.

Libraries used –

Pandas, numpy, matplotlib, seaborn, sklearn, plotly, LinearRegression, DecisionTreeRegressor, RandomForestRegressor, ExtraTreesRegressor, GradientBoostingClassifier, KNeighborsClassifier, LogisticRegression

Data Preparation –

Generally we employ data cleaning steps to derive relevant insights from the data and to get rid of garbage values. Data cleaning was done to check for missing values, dimentionality and feature reduction and creating classifiers etc.

**Visualizations-**

Interactive maps of Zomato’s ratings across the globe

Map

Description automatically generated

Top 15 cuisines on zomato

Chart, bar chart

Description automatically generated

Zomato’s business is mainly concentrated in India. So no wonder that North Indian cuisine takes the prime spot in terms of popular cuisine.

Restaurants delivering online or not -

**Chart, bar chart

Description automatically generated**

5000 restaurants has no online delivery option. Only 2000 restaurants has this option.

Restaurants allowing table booking or not

Chart, bar chart

Description automatically generated

Just like the online delivery option, very few restaurants have table booking option.

Rating vs Online delivery

Chart, box and whisker chart

Description automatically generated

Surprisingly aggregate rating is slightly better for restaurants with no online delivery option. It might be due to the fact that majority restaurants have no online delivery option.

Price vs Rating

Chart, box and whisker chart

Description automatically generated

This box plot shows that high-end restaurants has good ratings compared to low priced ones. Quality of food and ambience might be the reasons behind the high ratings.

Top 15 restaurants with Maximum Outlets

Chart, pie chart

Description automatically generated

Café coffee day is the top beverage/food chain with maximum share of 12.4%. This is the largest coffee chain in India just like Starbucks in the USA. Dominos takes the second place with 11.8% restaurants worldwide.

**Modeling –**

1. Predicting price range based on other features

Regression analysis is a statistical method that is used to analyze the relationship between a dependent variable and one or more independent variables. Regression analysis is most commonly used in forecasting or predicting how a set of conditions will impact an outcome. Here in this case I used Linear regression, Decision Tree regression, Random Forest and Extra Tree regression.

A picture containing text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Random forest regression has the highest r2 score with 0.84. This model is best fit in our analysis. This shows that price range is dependent on other attributes.

2. Predict Table Booking based on other features

A classification model tries to draw some conclusion from the input values given for training. It will predict the class labels/categories for the new data.

Graphical user interface, text, application, email

Description automatically generated

Gradient Boosting Classifier -

Graphical user interface, text, application, email

Description automatically generated

KNeighbors classifier -

Graphical user interface, text

Description automatically generated with medium confidence

Of the 2 Models, Gradient Boosting Classifier has accuracy of around 91.2%. This means that the Gradient Boosting Classifier is predicting nearly 91.2% of the test data accurately. On the other side KNN accuracy is 89.6% which is not bad either.

**Twitter Sentiment analysis -**

I performed sentimental analysis on Twitter data for positive, negative and neutral to derive conclusions. Sentiment analysis or opinion mining is a mechanized method used by businesses to detect positive sentiments or negative sentiments in a text to understand the customers better and keep an eye on its reputation. Sentiment analysis helps us to identify the motive and emotions behind a purchase, gain many more insights and eventually predict the future response.

As part of the sentiment analysis, I amalyzed max tweets, retweeets, favorite tweets etc. wrt Zomato reviews. Also analyzed the percentage of positive, negative and neutral tweets to understand customers views on the same.

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Zomato has approx. 35% positive tweets and 9% negative tweets. Rest are neutral.

Summary

* Major Zomato business is concentrated in India followed by USA, Australia.
* Price range is dependent on other attributes like has table booking, online delivery availability, price etc.
* North Indian cuisine is the most popular cuisine.
* High-end restaurants has good ratings compared to low priced ones. Quality of food and ambience might be the reasons behind the high ratings.
* The twitter sentiment analysis for positive, negative and neutral reviews will help the business owners and customers understand where the Zomato business stands in terms of performance and customer reviews. Zomato has approx. 35% positive tweets and 9% negative tweets. Rest are neutral.