RESEARCH ARTICLE OPEN ACCESS

**Zomato Restaurant Analysis Based On Ratings**

Prapti Dubey¹, Femenca Noronha²

¹ H.O.D (Data Science), ² ³ P.G. Student Data Science

¹ ² ³ Thakur College of Science & Commerce, Mumbai, India

¹praptidubey866@gmsil.com, ²femenca1996@gmail.com

----------------------------------------\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*----------------------------------

Abstract:

In this challenge, we are analysing the Zomato Restaurant dataset to find the more insights about the Restaurant business.The restaurant industry is highly competitive, and small factors can greatly influence a restaurant's success. By analyzing the dataset, business owners, restaurant managers, and other stakeholders can identify patterns and trends that might not be obvious at first glance. Understanding customer preferences, optimal pricing strategies, and competitive dynamics can give businesses a competitive edge.The Zomato Restaurant dataset presents a unique opportunity to dwelve into the dynamic and competitive world of the restaurant business. By analyzing the data, we can uncover key insights that could help restaurants make data-driven decisions to improve their operations, customer satisfaction, and profitability. The overarching goal of this challenge is to explore various aspects of the restaurant industry through the lens of this dataset, which includes valuable information such as restaurant ratings, cuisines, locations, pricing, and customer reviews.

The challenge likely involves various analytical techniques, such as:

* **Descriptive Analytics**: To understand the distribution of restaurants, cuisine types, pricing, and ratings across different regions.
* **Predictive Analytics**: To predict what factors most influence customer ratings and determine potential success factors for restaurants.
* **Clustering and Segmentation**: To identify distinct customer or restaurant segments based on shared attributes such as pricing or location.

Keywords **—** **Online food delivery, Marketing mix strategies, Competitive analysis, Pre-processing, Data Cleaning, Data Mining, Exploratory data analysis , Classification , Pandas , MatPlotL.**

----------------------------------------\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*----------------------------------

1. **INTRODUCTION**

The basic idea of analyzing the Zomato dataset is to get a fair idea about the factors affecting the aggregate rating of each restaurant, establishment of different types of restaurant at different places, 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. Inspite of increasing demand it however has become difficult for new restaurants to compete with established restaurants. Most of them serving 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. 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. These kind of analysis can be done using the data, by studying different factors.

Bangalore(officially known as Bengaluru) is the capital and largest city of the Indian state of Karnataka. With a population of over 15 million, Bangalore is the third largest city in India and 27th largest city in the world.Bangalore is one of the most ethnically diverse cities in the country, with over 51% of the city’s population being migrants from other parts of India. Bangalore is sometimes referred to as the “Silicon Valley of India” because of its role as the nation’s leading information technology (IT) exporter.Bangalore has a unique food culture. Restaurants from all over the world can be found here in Bengaluru, with various kinds of cuisines. Some might even say that Bangalore is the best place for foodies. I will be analyzing the Zomato restaurant data for the city, Bangalore.

The Internet is becoming an pervasively common platform to facilitate searching, choosing, and buying products. Online food ordering companies offer a range of options and conveniences that enable consumers to have their favourite food on their fingertips. Recently, Online Food Delivery become a new trend comforting the foodies and Zomato is the biggest name that come to mind when talking with reference to India. Zomato helps various restaurants to increase their customer base and even the concept of cloud kitchen also finds its way in India having only delivery but not dine in facilities. Many people across the country want to get into this profitable business of food delivery and wants to open restaurants and cloud kitchen in different parts of India.

1. **Literature Review**

In this paper [1], the authors surveyed a bunch of existing research papers which explore generic rating-based recommender models as well as research papers on how the user generated reviews can be utilized as an alternative and valuable source in the recommendation process through merging them with the Multi-Criteria Recommender System (MCRS) to enhance the accuracy of the Recommender System (RS)’s performance. User-generated reviews are used to improve the accuracy of the RSs performance by using text analysis and sentiment analysis to transform the unstructured user reviews into a structured form that can be merged with RSs. Hidden elements can be extracted from the user’s reviews and delivering them to the RSS, tries to solve the problem of inaccurate recommendations caused by relying only on the overall ratings in the recommendation process. The major drawback of this approach is that this approach heavily depends on the the text analysis and sentiment analysis done on users’ reviews. These analytical models cannot fully comprehend or understand the user’s review, therefore utilizing only part of the review.

In this research paper [2], the authors have designed a restaurant recommender system based on a novel model that captures correlations between hidden aspects in reviews and numeric ratings. The authors were motivated by the observation that a user’s preference against an item is affected by different aspects discussed in reviews. Their method first explores topic modelling to discover hidden aspects from review text. The authors use Latent Dirichlet Allocation (LDA) to create profiles for every user who has written a review. For predicting ratings for restaurants any user has not reviewed, Linear/Logistic Regression is employed. The authors use a browsing tool to select representative reviews for a particular restaurant. The model was tested on 1,168,420 reviews written by 316,702 users for 42,274 restaurants in Shanghai. Since the data available is sparse, the majority of the ratings and user profiles are estimates. Hence the accuracy achieved is a mere 50 percent.

In [3], multiple independent input attributes from the Zomato Restaurant data set are used to predict the rating of a restaurant. In this paper, the authors also do a comparative study between the performance of various regression models to arrive at a conclusion as to which one is best for this data set. The metrics used to measure each of the regression models are regression score, absolute error, mean squared error and root mean squared error.

In [4], the authors have presented an approach to combine the advantages of both, collaborative filtering and knowledge based filtering. Collaborative Filtering method of mining data involves filtering of information or patterns done using techniques involving collaboration among users’ viewpoints, data sources, user ratings, etc. A hybrid model combining the advantages of both these methods is implemented. The model will choose either of the techniques based on the situation. The collaborative approach is chosen if the system knows the user well and the knowledge based approach is chosen if the user is new to the system. Although reported to be highly accurate, the drawback with this model is the computation cost and the cost to search in the user database. Methods such as data partitioning and multi threading are discussed to improve on the above-mentioned drawbacks.

In [5], the authors have proposed three different approaches which include TF-IDF based approach, Association Rule Mining Approach, and the Deep-Learning-based approach. In the TF-IDF-based approach, which is a content-based collaborative filtering approach, they feed the cuisine type and cost as inputs to the model, and ’Cosine Similarity’ is used as a metric to determine and return recommendations based on the input restaurant by the user. The second approach is based on the ’Apriori algorithm’ where they recommend food items based on the frequency of the occurrences of the dish in the data set, they calculate a support vector, and association rules are formed. The third approach is the deep neural network approach where they use several of the Natural Language Processing.

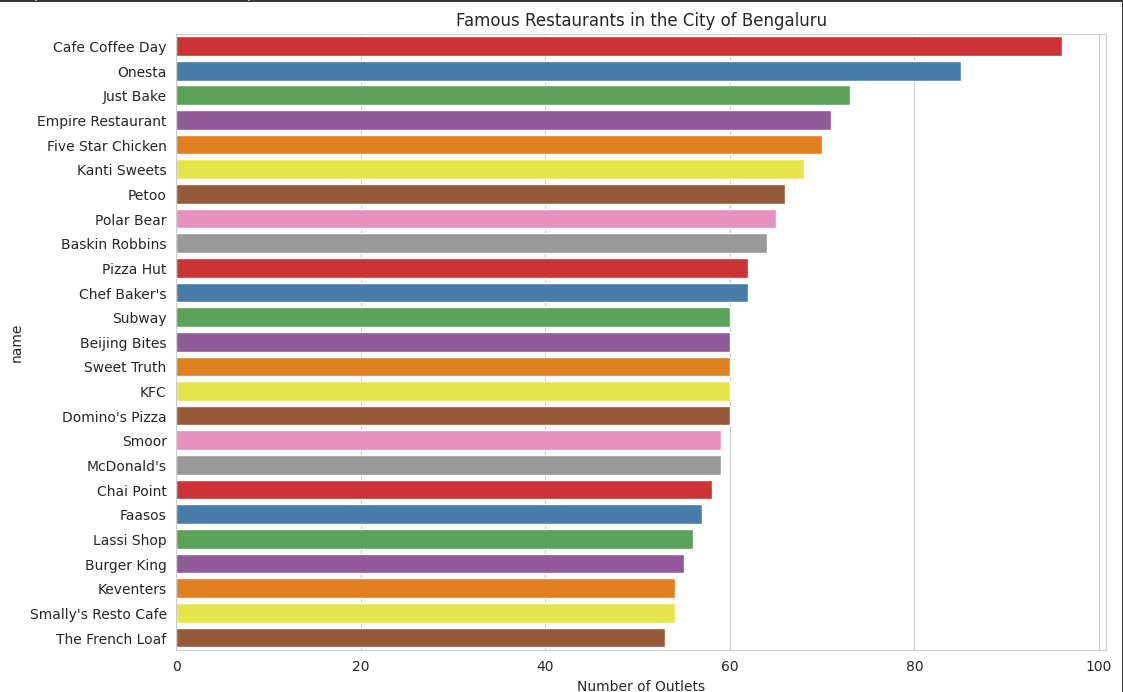
**Methodology**

This research will follow a structured methodology that involves several stages, ranging from data collection and preprocessing to model selection, training, and evaluation. The aim is to explore how machine learning (ML) techniques can be applied to do analysis using relevant demographic and behavioural data, with a focus on improving accuracy, reliability, and scalability of diagnosis tools.

1. ***Loading the dataset*:** Load the data and import the libraries.
2. ***Data Cleaning*:**

* Deleting redundant columns.
* Renaming the columns.
* Dropping duplicates.
* Cleaning individual columns.

1. ***Data Visualization*:** Using plots to find relations between the features.



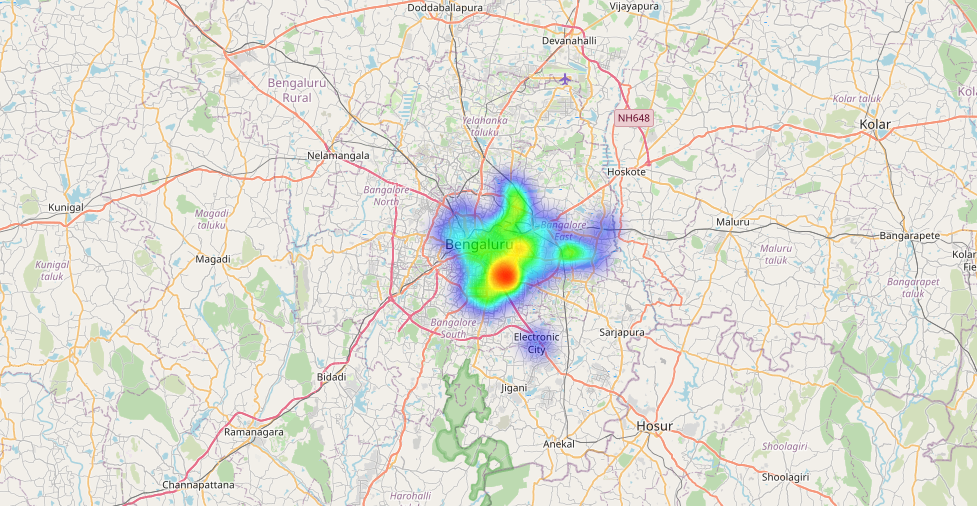
*Fig 1.1: Famous Restaurants*

1. ***Finding the best cheap restaurants*:**

* Cheapest, Highest rated and largely voted.
* Is there a relation between cuisine,location and the cost?

1. ***Exploring the best expensive restaurants*:**

* Restaurants that are expensive, Highest rated and largely voted.
* Is there a relation between restaurant type,location and the cost?



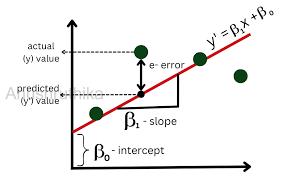
*Fig 1.2:**heatmap of north Indian restaurant in south Bangalore*

1. ***Model Selection*:** Four different machine learning algorithms are explored and compared.

**ALGORITHMS**

1. **Linear Regression (LR):**

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model. Before attempting to fit a linear model to observed data, a modeler should first determine whether or not there is a relationship between the variables of interest. This does not necessarily imply that one variable causes the other (for example, higher SAT scores do not cause higher college grades), but that there is some significant association strength of the relationship between two variables. If there appears to be no association between the proposed explanatory and dependent variables (i.e., the scatter plot does not indicate any increasing or decreasing trends), then fitting a linear regression model to the data probably will not provide a useful model. A valuable numerical measure of association between two variables is the correlation coefficient, which is a value between -1 and 1 indicating the strength of the association of the observed data for the two variables. A linear regression line has an equation of the form Y = a + bX, where X is the explanatory variable and Y is the dependent variable. The slope of the line is b, and a is the intercept (the value of y when x = 0)



*Fig 1.3: Linear Regression*

* **Pros**: Easy to interpret, works well when the relationship between variables is linear, and computationally efficient.
* **Cons**: Limited in capturing complex, non-linear relationships, which could lead to underfitting for certain data patterns.

1. **Decision Tree Regression (DTR):**

Decision Tree is a supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.

The decisions or the test are performed on the basis of features of the given dataset. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions*.* It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further

branches and constructs a tree-like structure.

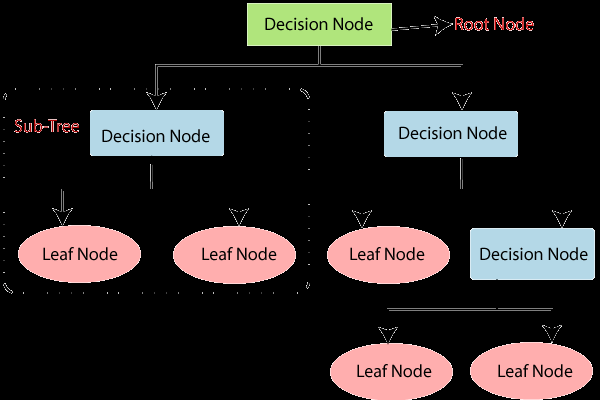
**Types of Decision Trees:**

Types of decision trees are based on the type of target variable we have. It can be of two types:

1. **Categorical Variable Decision Tree:** Decision Tree which has a categorical target variable then it called a Categorical variable decision tree.

2. **Continuous Variable Decision Tree:** Decision Tree has a continuous target variable then it is

called Continuous Variable Decision Tree.

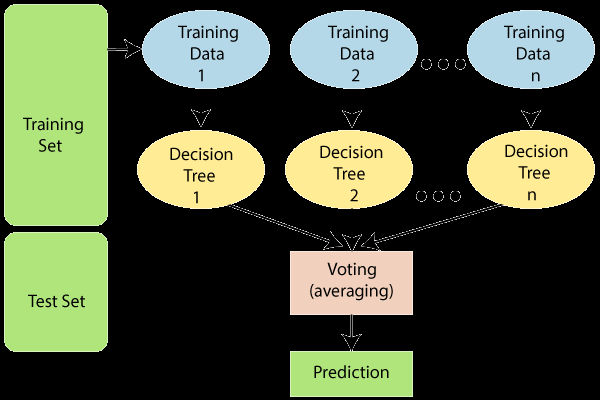


*Fig 1.4: Decision Tree*

* **Pros**: Easy to visualize, can handle both numerical and categorical data, and captures non-linear relationships.
* **Cons**: Prone to overfitting, especially with deeper trees. It may not generalize well without proper pruning or regularization.

1. **Random Forest Regression:**

Random forests or random decision forests are an ensemble learning technique for classification. Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification andRegression problems in ML. It is based on the concept of ensemble learning**,** which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, “Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.” Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of over fitting. The below diagram explains the working of the Random Forest algorithm.



*Fig 1.5: Random Forest*

* **Pros**: Reduces overfitting compared to individual decision trees, works well on large datasets with many features, and is robust to noisy data.
* **Cons**: Requires more computational resources and can be less interpretable due to the ensemble nature of the model.

1. **CatBoost Regressor:**

CatBoost (Categorical Boosting) is a powerful gradient boosting algorithm that excels at handling categorical data, which is often prevalent in real-world datasets like Zomato’s. It automatically handles categorical features without needing extensive preprocessing (like one-hot encoding), making it highly suitable for datasets with variables such as restaurant type, cuisine, and location.

* **Pros**: Excellent at handling categorical data, robust to overfitting, and performs well even with default hyperparameters. It also tends to be faster than other boosting algorithms.
* **Cons**: Requires more computational power compared to linear models and can be harder to interpret due to the complexity of the model.

**VALIDATION OF MODELLING TECHNIQUE**

Each of the selected models will be trained using the training dataset, and hyperparameter tuning will be performed to optimize their performance. Techniques such as grid search and cross-validation will be applied to find the best set of parameters for each algorithm. For instance, in Random Forest, the number of trees and maximum depth will be tuned. To avoid overfitting, use K-fold cross-validation (typically K=5 or 10). This involves dividing the training data into K subsets, training the model K times, and using a different subset for validation each time.

**Performance Evaluation**: - Once the models are trained and validated, evaluating their performance on the test set is crucial. Here’s a common set of metrics and methods for performance evaluation across all models:

**Mean Absolute Error (MAE):**MAE measures the average magnitude of errors in predictions, without considering their direction. It's useful for understanding how far predictions are from actual values.

**Mean Squared Error (MSE):**MSE calculates the average of the squared differences between the predicted and actual values. It gives more weight to larger errors, which is useful when large errors are particularly undesirable.

## Root Mean Squared Error (RMSE):RMSE is the square root of MSE and gives an estimate of how large the error is, in the same units as the target variable. It’s commonly used for regression tasks.

## R-squared (R²):R² explains the proportion of variance in the dependent variable that can be predicted from the independent variables. It ranges from 0 to 1, with higher values indicating better model fit.Adjusted R² is also useful for linear regression models, as it adjusts for the number of predictors in the model, helping to avoid overfitting.

## Comparison and Analysis: - Linear regression provides a baseline for modeling relationships in a straightforward, interpretable manner.

## CatBoost is highly efficient for handling the categorical data typical in the restaurant industry, making it a strong candidate for capturing intricate relationships between factors like location and cuisine type.

## Decision trees offer clarity by visualizing the decision-making process, helping identify the key features that influence restaurant success.

## Random forests enhance decision trees by reducing overfitting, improving the model’s ability to generalize from the dataset to unseen data.

Each of these models offers a unique approach to analyzing the Zomato dataset, providing valuable insights into pricing strategies, customer satisfaction, and restaurant success. By comparing their performance and interpreting their results, you can obtain a holistic understanding of what drives restaurant performance and customer preferences.

1. **Results & Discussion**

Based on the R² scores I’ve obtained from regression models, here's a breakdown of their performances:

**1. Linear Regression**

* **R² Score:** 0.2419
* **Interpretation:** This score indicates that only about 24.19% of the variance in the target variable is explained by the model. This is relatively low, suggesting that the linear regression model is not capturing the underlying relationship well. There may be non-linear relationships or other important features not included.

**2. CatBoost Regressor**

* **R² Score:** 0.3489
* **Interpretation:** With 34.89% of the variance explained, the CatBoost model shows some improvement over linear regression but still leaves a substantial amount of variance unexplained.

**3. Random Forest Regressor**

* **R² Score:** 0.8754
* **Interpretation:** This model performs significantly better, explaining 87.54% of the variance in the target variable. The Random Forest algorithm is robust to overfitting (to a degree) and can capture complex interactions in the data, which is likely why it's performing well.

**4. Decision Tree Regressor**

* **R² Score:** 0.8062
* **Interpretation:** The Decision Tree model explains 79.42% of the variance, which is also a good score but lower than that of the Random Forest model. Decision trees can be prone to overfitting, especially with complex datasets.

1. **Conclusions**

This analysis will provide valuable insights that can help restaurants and stakeholders make informed decisions to optimize their operations, improve customer experiences, and stay competitive in a crowded marketplace. The findings could also guide future investments, marketing strategies, and operational improvements for businesses in the restaurant industry.

With the implementation it is possible to deliver useful information for Business areas and also for Zomato customers on choosing the best restaurant for ordering (specially the new ones). This is an example of what predictive models can do in practice. This project helps us to gain more insights about the restaurant business.

1. **Reference**
2. Sumaia Mohammed Al-Ghuribi and Sharul Azman Mohd Noah, ” Multi-Criteria Review-Based Recommender System - The State of the Art ”.
3. Gao, Yifan, et al. ”A restaurant recommendation system by analyzing ratings and aspects in reviews.” International Conference on Database Systems for Advanced Applications. Springer, Cham, 2015.
4. J. Priya, ”Predicting Restaurant Rating using Machine Learning and comparison of Regression Models,” 2020 International Conference on Emerging Trends in Information Technology and Engineering (icETITE), 2020, pp. 1-5, doi: 10.1109/ic-ETITE47903.2020.238.
5. Dwivedi, Prerna, and Nikita Chheda. ”A hybrid restaurant recommender.” International Journal of Computer Applications 55.16 (2012).
6. A. Sarkar, A. Baksy and V. Kirpalani, ”Analysis of Zomato Services using Recommender System Models,” 2021 International Conference on Intelligent Technologies (CONIT), 2021, pp. 1-5, doi: 10.1109/CONIT51480.2021.9498534.