Predicting Restaurant Ratings Using Regression Analysis Approach

Sajida Sultana.Sk
Assistant Professor
Department of Computer Science Engineering
Vignan's Foundation for Science, Technology and Research
Guntur, India
sks_cse@vignan.ac.in

G.Joseph Anand Kumar
Department of Computer Science Engineering
Vignan's Foundation for Science, Technology and Research
(Deemed to be University)
Guntur , India
221FA04503@vignan.ac.in

V. Leela Venkata Mani Sai
Department of Computer Science Engineering
Vignan's Foundation for Science, Technology and Research
(Deemed to be University)
Guntur , India
221FA04575@vignan.ac.in

N. Bala Sai
Department of Computer Science Engineering
Vignan's Foundation for Science, Technology and Research
(Deemed to be University)
Guntur , India
221FA04006@vignan.ac.in

E. Sai Naga Lakshmi
Department of Computer Science Engineering
Vignan's Foundation for Science, Technology and Research
(Deemed to be University)
Guntur , India
221FA04591@vignan.ac.in

***Abstract*—The establishment of a restaurant addresses the challenges of establishing a restaurant in a competitive market by introducing a framework for accurately predicting restaurant ratings, a vital tool for attracting customers and assessing venture success. By identifying and analyzing key factors that impact ratings, this research enables potential restaurant owners to make data-driven decisions before launching their business, reducing risks and saving time. The study employs seven regression models to compare performance metrics and determine the most reliable predictive model, ultimately providing a valuable resource that supports informed decision-making and improves the success prospects for new restaurant ventures.**

***Index Terms*—Consumer preferences, Data-driven decision-making, Risk mitigation, Predictive modelling .**

I. Introduction

A restaurant in this competitive hospitality industry will heavily be determined to succeed based on customer ratings and reviews. In a world of such diversity with millions of places to dine at, it leaves the restaurants to be rated out of others and meeting up to the required standards of service. While attracting new customers is inevitable through good reviews, destructive criticism of a business operation will only deter people from coming in. This means that true future rating forecasting is crucial for new companies because the key information it gives them revolves around performance in locational, food-related, and price-sensitive areas. Opening up a new restaurant demands a vast amount of money as well as time. Most entrepreneurs are confused about determining the prospects of the project before the actual day of opening. While more traditional approaches like market research and customer surveys do provide some insight into the situation, they are generally without a data-driven predictive approach. This is where machine learning and data analytics come in.

This paper proposes a system using regression models to predict restaurant rating, making all controllable factors alterable even before its opening. The system analyzes various features of interest such as online ordering, reservations, and average cost for two diners in order to come up with a predicted rating. It allows for comparisons between various restaurant-related aspects and equips business owners with the information needed to make more informed decisions and mitigate risks concerning the opening of a new restaurant.

II. LITERATURE SURVEY

Suhas Somashekar and Suhas Mallesh [1] Restaurants ratings prediction using a number of regression techniques is the scope of the report. The article discusses the processing and analysis of a dataset consisting of more than 43,000 restaurants from Bengaluru, India, while also highlighting the relevance of ratings in the advent of new business ventures. With metrics like R2 score and mean absolute percentage error, seven regression models including XGBoost Regression, decision trees and Linear regression were evaluated Focused more on determination Coefficients of the models with Satisfactory results, the best performing models are Random Forest, ADA Boost and XGBoost regression thus those models remain with the donted capacitation for accurate ratings prediction. The idea of the research is to help new restaurants increase their probability of success by allowing them to make the right decisions. (2021)

The article by J. Priya [2] uses a dataset of Bengaluru restaurants to illustrate the applications of machine learning in predicting restaurant reviews. It includes data preparation, data visualization, as well as building predictive models using 8 different regression algorithms including Bayesian, Random Forest, Ridge and Linear Regression. To evaluate performance of each model, metrics like regression score and error rates are used. In the findings of this analysis, Random Forest Regression performs better than the others, having the lowest error and the maximum accuracy. This analysis is a good way for restaurants to enhance their services and business strategies. (2020)

Ibne Farabi Shihab et al [3] proposed a machine learning-based approach to identify the optimal geographical location for new restaurants by analyzing demand and competition. However, the study concentrated on location suitability and did not integrate sentiment or customer preference data, which could add context to the findings.

Mara-Renata Petrusel and Sergiu-George Limboi [4] developed a restaurant recommendation system that enhanced rating predictions through sentiment analysis, using customer reviews. The study, however, focused solely on text sentiment without considering numeric ratings or other quantitative data that might improve prediction accuracy.

Tugçe Bilen et al.[5] created a smart city application to estimate ideal business locations using machine learning, particularly valuable for urban planning. However, the model was developed for general businesses and not specifically for restaurants, potentially missing industry-specific factors in location suitability.

Ismam Hussain Khan et al. [6] developed a machine learning framework to predict ratings for food recipes based on ingredient sentiment and user preferences, focusing on individual recipe items. The study did not consider external restaurant factors like ambiance or service, which could be important in real-world restaurant settings.

Neha Vaish et al. [7] applied sentiment analysis to hotel reviews to analyze customer feedback, showing that sentiment-driven insights can predict ratings. However, the study centered on hotels, and restaurant-specific considerations such as menu diversity or wait times were not addressed, limiting direct applicability to the restaurant industry.

Sandeep Bhatia et al. [8] used machine learning to predict the success of Zomato restaurants by analyzing attributes like menu, pricing, and location. While effective, the study did not focus on rating prediction itself, which could provide more granular insights for restaurant management.

Yi Luo and Xiaowei Xu [9] explored predicting the helpfulness of restaurant reviews on Yelp using different machine learning algorithms. The focus on helpfulness scores provided insights into review quality rather than directly predicting restaurant ratings, which might add value if combined with other predictive models.

In [10] Joshua [10] emphasized comparing various regression models for restaurant rating prediction, demonstrating that model choice is crucial for accuracy. However, the study did not test ensemble models that could potentially improve predictions by aggregating different regression outcomes.

Xiaochen Wang, Yanyan Shen, Yanmin Zhu [11] proposes a method to predict rating scores for new restaurants by analyzing both restaurant data and urban data, using a model called MR-Net to capture various influential features without relying on customer reviews.

Yifan Chen; Fanzeng Xia [12] predicts future Yelp ratings for restaurants using a combination of non-text and text features, achieving the highest accuracy of 82.5% through Decision Tree and Neural Network models.

Nabiha Asghar [13] addresses Yelp review rating prediction using four feature extraction techniques (unigrams, bigrams, trigrams, Latent Semantic Indexing) combined with four classification algorithms, achieving the best results with Logistic Regression on unigrams and bigrams.

Nanthaphat Koetphrom; Panachai Charusangvittaya; Daricha Sutivong [14] compares filtering techniques—content-based, collaborative, and hybrid filtering—for predicting restaurant satisfaction ratings, concluding that hybrid filtering achieves the highest accuracy.

Sanjukta Saha; A. K. Santra [15] focuses on predicting restaurant ratings in Kolkata by analyzing user feedback on food items using sentiment analysis and collaborative filtering.

Mara-Renata Petrusel; Sergiu-George Limboi [16] presents a restaurant recommendation system that integrates sentiment analysis with collaborative filtering, enhancing rating predictions based on positive and negative reviews.

F. M. Takbir Hossain; Md. Ismail Hossain; Samia Nawshin [17] presents a restaurant recommendation system that integrates sentiment analysis with collaborative filtering, enhancing rating predictions based on positive and negative reviews.

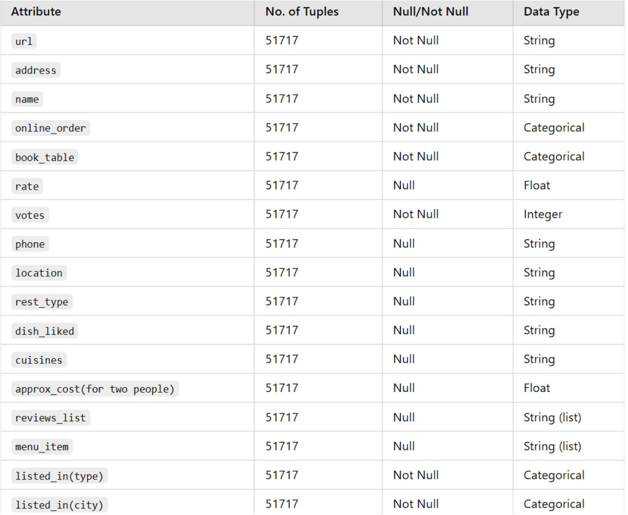
III. Dataset Description

DATASET SOURCE : This dataset is taken from https://www.kaggle.com/datasets/himanshupoddar/zomato-bangalore-restaurants

The dataset, titled Zomato Bangalore Restaurants, contains information about various restaurants in Bangalore, India. This data provides insights into restaurant characteristics, such as location, type, ratings, and more, which are useful for predicting restaurant success and consumer preferences.Below is a breakdown of the dataset's structure :

- General Restaurant Information
- Dining and Service Options
- Ratings and Reviews
- Cuisine and Cost
- Restaurant Type and Listing Category

The dataset consists of multiple attributes related to restaurants, which are listed below along with their descriptions:



Attribute	No. of Tuples	Null/Not Null	Data Type
url	51717	Not Null	String
address	51717	Not Null	String
name	51717	Not Null	String
online_order	51717	Not Null	Categorical
book_table	51717	Not Null	Categorical
rate	51717	Null	Float
votes	51717	Not Null	Integer
phone	51717	Null	String
location	51717	Null	String
rest_type	51717	Null	String
dish_liked	51717	Null	String
cuisines	51717	Null	String
approx_cost(for two people)	51717	Null	Float
reviews_list	51717	Null	String (list)
menu_item	51717	Null	String (list)
listed_in(type)	51717	Not Null	Categorical
listed_in(city)	51717	Not Null	Categorical

- The image shows the attributes, data types, and null/not null values of a dataset with 51717 tuples. The dataset contains various information about restaurants, including their names, addresses, online ordering capabilities, cuisines, average cost, and reviews.

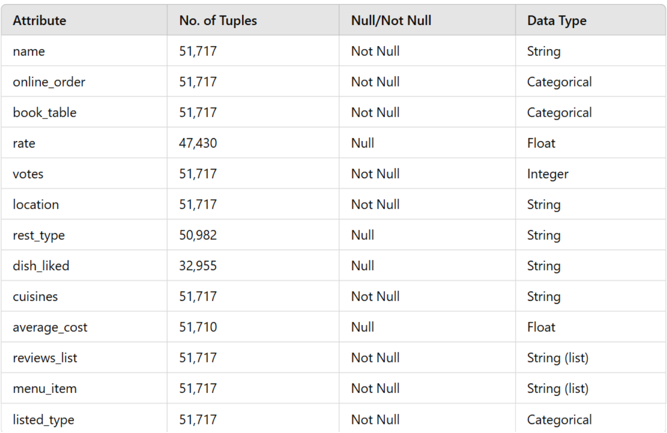
IV. Dataset PreProcessing

Data preprocessing is the essential step of transforming raw, often messy data into a clean and usable format. This process involves handling missing values, converting categorical data into numerical representations, and identifying and rectifying inconsistencies. By removing irrelevant attributes, the focus is sharpened on the most pertinent information for analysis. Through these transformations, the dataset becomes more understandable, reliable, and suitable for extracting meaningful insights.

V. Dataset After PreProcessing

Here, the steps which are involved are :

- Data Cleaning
- Text Processing
- Sentiment Analysis
- Encoding Categorical Variables
- Feature Engineering
- Normalization
- Scaling
- Splitting Data



Attribute	No. of Tuples	Null/Not Null	Data Type
name	51,717	Not Null	String
online_order	51,717	Not Null	Categorical
book_table	51,717	Not Null	Categorical
rate	47,430	Null	Float
votes	51,717	Not Null	Integer
location	51,717	Not Null	String
rest_type	50,982	Null	String
dish_liked	32,955	Null	String
cuisines	51,717	Not Null	String
average_cost	51,710	Null	Float
reviews_list	51,717	Not Null	String (list)
menu_item	51,717	Not Null	String (list)
listed_type	51,717	Not Null	Categorical

Column of Attributes

- The image shows the attributes, data types, and null/not null values of a dataset with 51,717 tuples. The dataset contains various information about restaurants, including their names, online ordering capabilities, cuisines, average cost, and reviews.

VI. Dataset Visualization

Data visualization transforms raw restaurant data into understandable visual representations like charts and graphs. This powerful technique helps uncover hidden patterns and trends within the data, enabling businesses to make informed decisions. By visualizing sales figures, customer preferences, and operational metrics, restaurants can optimize their menus, improve customer experiences, and streamline their operations. This visual approach makes complex data accessible, fostering a deeper understanding of the business landscape and driving strategic growth.

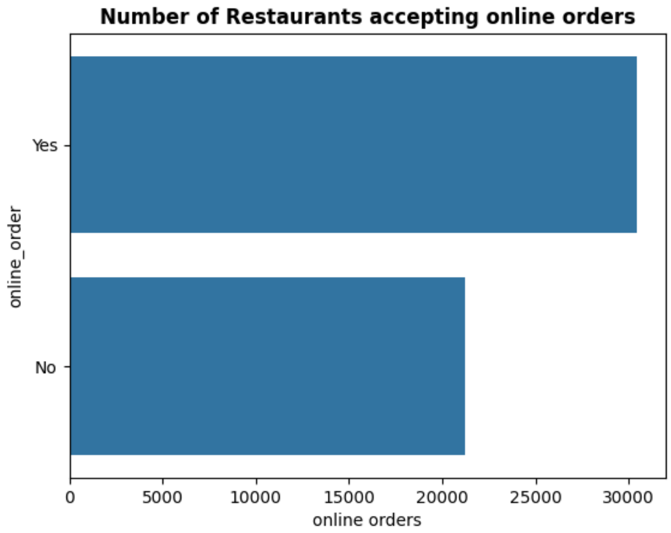


Fig.1:Count of Online Orders

- The bar chart (Fig.1) shows the number of restaurants accepting online orders. A significantly larger number of restaurants accept online orders compared to those that don't.

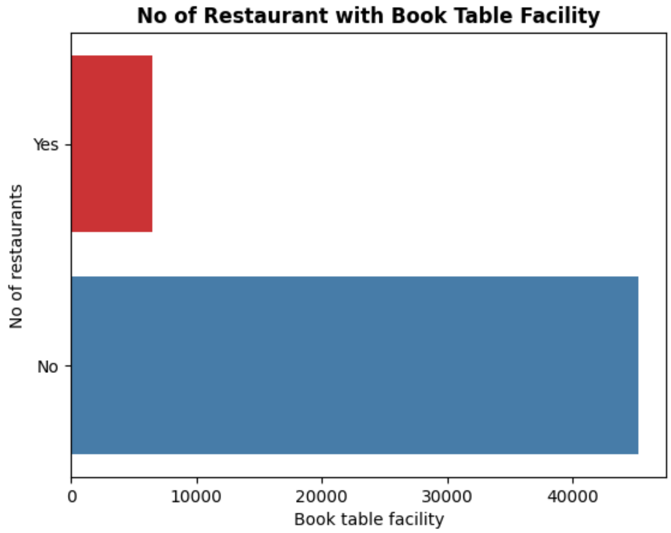


Fig.2:Table Booking Facility

- The bar chart (Fig.2) shows the number of restaurants with and without book table facility. A significantly larger number of restaurants do not have book table facility compared to those that do.

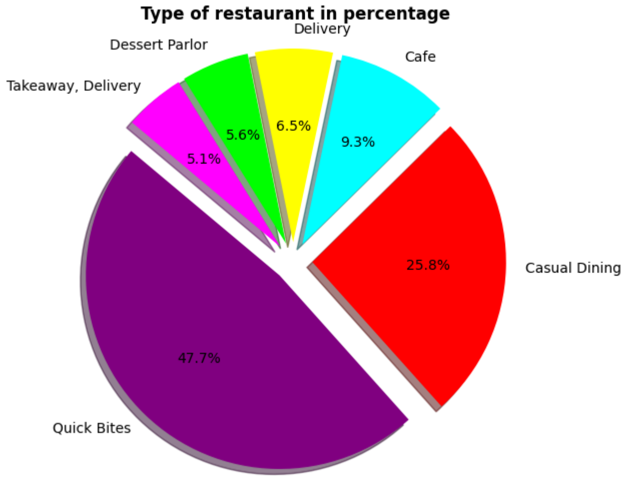


Fig.3:Percentage of Restaurants Types

- The pie chart (Fig.3) shows the percentage distribution of restaurants across different types in Bangalore. Quick Bites has the highest percentage of restaurants (47.7), followed by Casual Dining (25.8) and Cafe (9.3).

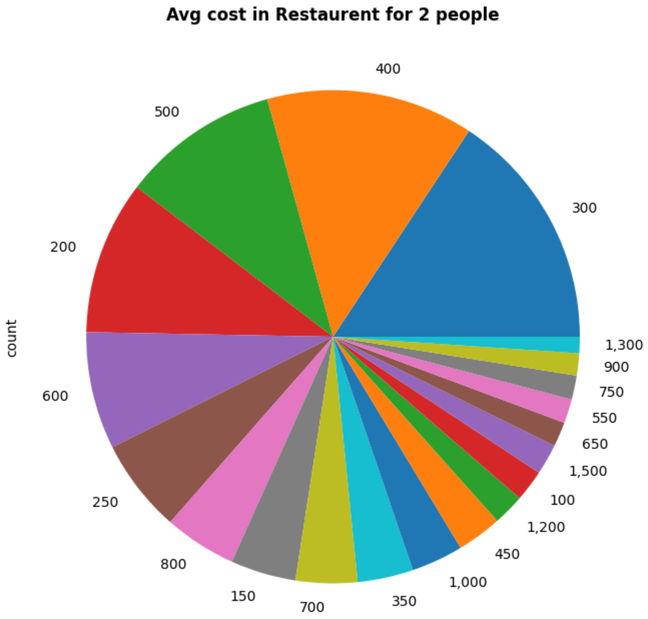


Fig.4:Avg Cost

- The pie chart (Fig.4) shows the distribution of restaurants across different average cost for two people. Most restaurants have an average cost of 300 for two people, followed by 400 and 500.

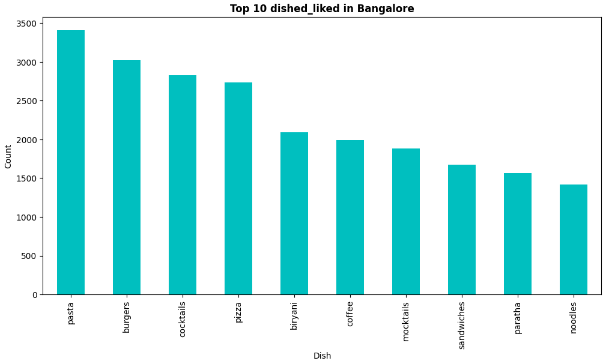


Fig.5: Top 10 Liked dishes in Banglore

- The bar chart (Fig.5) shows the top 10 most liked dishes in Bangalore. Pasta is the most liked dish, followed by burgers and cocktails.

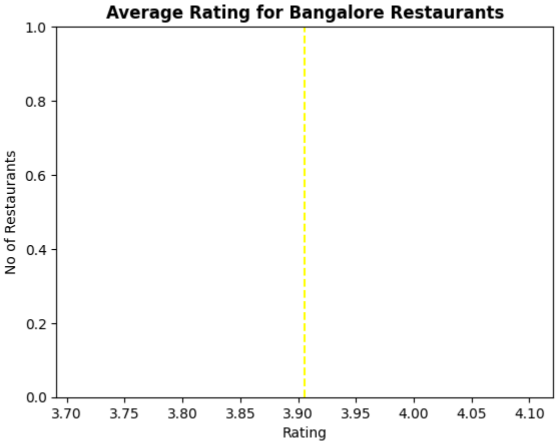


Fig.6:Avg Rating for Banglore Restaurants

- The chart (Fig.6) shows the distribution of restaurants in Bangalore based on their average rating. Most restaurants have an average rating of 3.9, with a few outliers at 3.7 and 4.1.

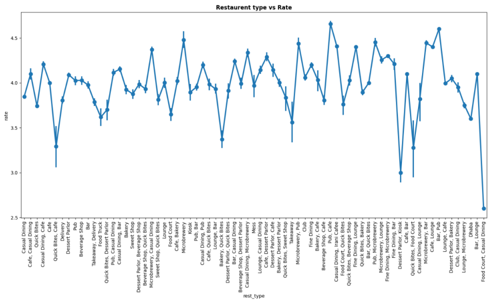


Fig.7:Restaurant Type vs Rate

- The line plot (Fig.7) shows the average rating for different restaurant types in Bangalore. There is a wide range of ratings across different types, with some types having consistently high ratings and others having lower ratings.

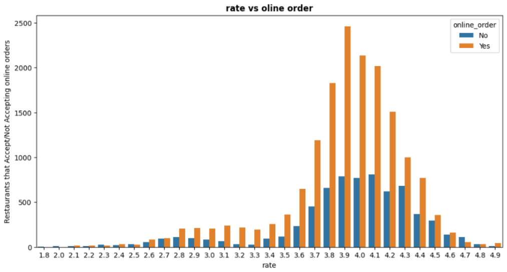


Fig.8:Rate vs Online Order

- The bar chart (Fig.8) shows the number of restaurants that accept or do not accept online orders, grouped by their rating. There is a clear trend that as the rating increases, the number of restaurants accepting online orders also increases.

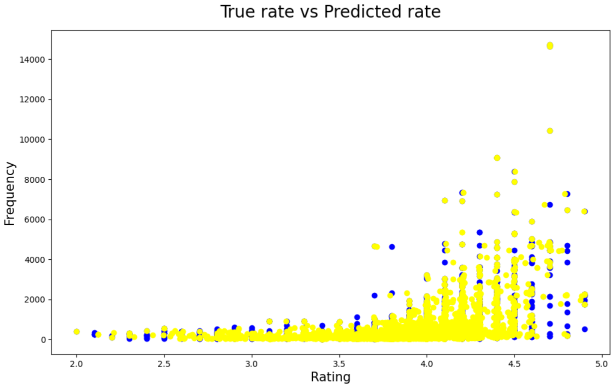


Fig.9 : True vs Predicted Rate

- Plotting the distribution of true and predicted scores in Fig.9 reveals some overlap as well as areas of divergence, which may indicate that the prediction model has flaws.

VII. Model Building

A. Linear Regression

Linear regression shows how a variable relates with other factors that are affecting it by plotting a straight line through the data points. This straight line is chosen in such a way that the least amount of difference (errors) exists between what is being predicted and the actual quantities. It's great at predicting sources of data that are continuous but does not apply if the relationship is not linear.

B. Random Forest

Random Forest is a type of ensemble learning technique that constructs numerous decision trees on different portions of data and combines their outcomes to enhance predictive precision and diminish overfitting. It is resistant to outliers and noise making it appropriate for both classification and regression tasks. Random Forests are useful for managing complicated and high-dimensional data.

C. Ridge Regression

Ridge Regression is a type of linear regression that adds a penalty to the magnitude of coefficients to prevent overfitting, which is helpful when multicollinearity is present. By adding a regularization term, it controls the impact of independent variables and prevents any one variable from overly influencing the model. This technique is useful when predictor variables are correlated.

D. Lasso Regression

Lasso Regression, or Least Absolute Shrinkage and Selection Operator, works by introducing a penalty that is proportional to coefficients' absolute values, shrinking some of these coefficients down to zero. Because of this selection of features property, Lasso is very beneficial when there are many predictors since it minimizes the number of predictors by keeping only the most significant ones.

E. Support Vector Machine (SVM)

SVM is a supervised machine learning algorithm that finds the hyperplane that best separates data points of different classes. In classification, it aims to maximize the margin between classes, while for regression (SVR), it tries to fit data within a certain tolerance margin. SVM is effective for high-dimensional and non-linear data.

F. K-Nearest Neighbors (KNN)

KNN is an uncomplicated non-parametric rule that predicts the class of a data point by k-nearest neighbours or a value by averaging the neighbours. KNN is not difficult to apply in practice, although even in simpler implementations, it may be expensive with respect to time due to the requirement of the computation of distances between all nearby points within the set.

G. Decision Tree

Decision Trees are models that split data into branches based on feature values, ultimately leading to a decision or prediction at each leaf node. They're easy to interpret and can capture non-linear relationships, but they are prone to overfitting. Pruning methods are used to create simpler, more generalizable trees.

H. AdaBoost

AdaBoost (Adaptive Boosting) belongs to the family of ensemble methods that trains weak learners (decision trees most of the times) one after another and modifies the coefficients of the misclassified observations at each stage. It focuses on hard-to-classify points hence leading to an inherently better model over time. AdaBoost can be used successfully for both classification and regression tasks.

I. XGBoost

XGBoost, or Extreme Gradient Boosting, is a more efficient version of the gradient boosting algorithm enabling a combination of several weak learners in a sequential fashion to improve the overall accuracy of the final prediction. It is well regarded due to its speed and the ability to work with queries efficiently for large datasets. For these reasons, it is highly popular in competitive machine learning. XGBoost contains regularization, which is a mechanism that helps to prevent overfitting.

VIII. Approach

The paper improves restaurant rating prediction accuracy through a comparative analysis of various regression models. Each model's performance is assessed to identify the most accurate one based on dataset-specific characteristics like rating distribution and feature relevance. Models are tested and validated using multiple metrics to evaluate accuracy, robustness, and generalizability, helping eliminate those prone to overfitting or underfitting. This rigorous selection process ensures a data-driven approach, focusing on the model with the highest empirical accuracy. Ultimately, the study's model choice balances theoretical soundness with practical effectiveness, optimizing prediction precision for restaurant ratings.

IX. Results

Regression Models Result Before

Model	R2*Score	RMSE
Linear Regression	0.26	0.37
Random Forest	0.04	0.11
Ridge Regression	0.26	0.37
Lasso Regression	0.26	0.37
Svm	0.23	0.35
Knn	0.21	0.30
Desicion Tree	0.88	0.14
ADA Boost	0.01	0.41
XG Boost	0.79	0.18

- The table presents the results of applying various regression models to a dataset. The R2 Score measures the model's ability to explain the variance in the data, while the RMSE (Root Mean Square Error) quantifies the model's prediction accuracy. Decision Tree and XGBoost appear to be the top-performing models based on these metrics.

Regression Models Result After Re-Evaluating Parameters

Model	R2*Score	RMSE	Accuracy
Linear Regression	69.1	26.3	41.6
Random Forest	37.4	10.5	93.3
Ridge Regression	17.5	13.2	89.3
Lasso Regression	25.9	36.5	19.5
Svm	23.4	35.3	27.1
Knn	21.8	30.7	46.5
Desicion Tree	88.3	14.6	42.8
ADA Boost	10.3	41.8	39.1
XG Boost	79.5	18.4	81.2

- The table showcases the performance of various regression models after fine-tuning their parameters. XGBoost and Decision Tree models stand out with the highest R2 scores and accuracy, indicating their effectiveness in predicting the target variable.

X. Conclusion

XGBoost outperforms other models in predicting restaurant ratings due to its higher accuracy and lower error. While Decision Trees show promise, XGBoost's versatility and robustness make it the top choice. To further improve predictions, consider adding features like customer sentiment and seasonal trends. Ensemble techniques can also enhance accuracy. Remember, the best model depends on specific data and business goals, so continuous evaluation is key.

XI. Future Scope

This project includes expanding the model to incorporate additional features, such as customer sentiment from reviews and real-time factors like seasonal trends or promotional events, which could improve prediction accuracy. Integrating other machine learning techniques, such as deep learning or ensemble models, may further enhance performance. Additionally, adapting the model to predict ratings for different regions or cuisines could broaden its applicability, making it useful for various types of restaurants and geographic markets. Finally, creating a user-friendly tool or API based on this model could provide restaurant owners with accessible, data-driven insights to guide business decisions.

References

[1] Somashekar, S., Mallesh, S. (2021, December). Restaurant Rating Prediction Using Regression. In 2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA) (pp. 1139-1144).

[2] Priya, J. (2020, February). Predicting restaurant rating using machine learning and comparison of regression models. In 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE) (pp. 1-5).

[3] Shihab, I. F., Oishi, M. M., Islam, S., Banik, K., Arif, H. (2018, December). A machine learning approach to suggest ideal geographical location for new restaurant establishment. In 2018 IEEE Region 10 Humanitarian Technology Conference (R10-HTC) (pp. 1-5).

[4] Petrusel, M. R., Limboi, S. G. (2019, September). A restaurants recommendation system: Improving rating predictions using sentiment analysis. In 2019 21st International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC) (pp. 190-197).

[5] Bilen, T., Erel-Özçevik, M., Yaslan, Y., Oktug, S. F. (2018, June). A smart city application: Business location estimator using machine learning techniques. In 2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS) (pp. 1314-1321).

[6] Khan, I. H., Khan, M. H. U., Howlader, M. M. (2021, April). An Intelligent Approach for Food Recipe Rating Prediction Using Machine Learning. In 2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA) (pp. 281-283).

[7] Vaish, N., Goel, N., Gupta, G. (2022, January). Machine learning techniques for sentiment analysis of hotel reviews. In 2022 International Conference on Computer Communication and Informatics (ICCCI) (pp. 01-07).

[8] Bhatia, S., Arya, C., Verma, S., Gautam, D., Naib, B. B., Kumar, A. (2023, July). Predict Success of a Zomato Restaurant using Machine Learning. In 2023 World Conference on Communication Computing (WCONF) (pp. 1-7).

[9] Luo, Y., Xu, X. (2019). Predicting the helpfulness of online restaurant reviews using different machine learning algorithms: A case study of yelp. Sustainability, 11(19), 5254.

[10] Wang, X., Shen, Y., Zhu, Y. (2018, November). A Data Driven Approach to Predicting Rating Scores for New Restaurants. In Asian Conference on Machine Learning (pp. 678-693). PMLR.

[11] Chen, Y., Xia, F. (2020, August). Restaurants' rating prediction using Yelp dataset. In 2020 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA) (pp. 113-117). IEEE.

[12] Asghar, N. (2016). Yelp dataset challenge: Review rating prediction. arXiv preprint arXiv:1605.05362.

[13] Koetphrom, N., Charusangvittaya, P., Sutivong, D. (2018, July). Comparing filtering techniques in restaurant recommendation system. In 2018 2nd International Conference on Engineering Innovation (ICEI) (pp. 46-51).

[14] Saha, S., Santra, A. K. (2017, August). Restaurant rating based on textual feedback. In 2017 International conference on Microelectronic Devices, Circuits and Systems (ICMDCS) (pp. 1-5).

[15] Hossain, F. T., Hossain, M. I., Nawshin, S. (2017, December). Machine learning based class level prediction of restaurant reviews. In 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC) (pp. 420-423).