Predicting restaurant ratings using regression analysis approach

*Sajida Sultana.*Sk[1], G. *Joseph Anand Kumar*[2], V. *Leela Venkata Mani Sai*[2], N. *Bala Sai*[2], and E. *Sai Naga Lakshmi*[2]

[1] Assistant Professor in Vadlamudi, Andhra Pradesh, India Department of Computer Science and Engineering

[2] Vignan's Foundation for Science, Technology, and Research, Department of Computer Science Engineering, Vadlamudi, Andhra Pradesh, India

Abstract: Restaurant Builder solves the challenge of building restaurants in a highly competitive market by providing a framework for accurately predicting restaurant prices, a key tool for attracting customers and measuring success. This study identifies and identifies key factors that influence evaluation, allowing restaurant owners to make informed decisions, reduce risk, and save time before starting a business. The study uses seven regression models to compare performance indicators and identify the most reliable predictive models, ultimately providing valuable resources to support informed decision making and increase the likelihood of success for new restaurant ventures.

1 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-

1 Joseph Anand

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.

2.Literature Survey

[1] Suhas Somashekar and Suhas Mallesh 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)

[2] The article by J. Priya uses a dataset of Bengaluru restaurants to illustrate the applications of machine learning in predicting restaurant reviews. Along with creating prediction models using eight distinct regression algorithms—including Bayesian, Random Forest, Ridge, and Linear Regression—it also covers data preparation and presentation. Metrics like regression score and error rates are utilized to evaluate each performance of the model. Based on the findings of the analysis, Random Forest Regression outperforms the others, achieving the highest accuracy and the lowest error. Restaurants can improve their offerings and business plans by using this analysis. (2020)

[3] Ibne Farabi Shihab et al , By examining demand and competition, suggested a technique based on machine learning to select the ideal location for new restaurants. However, mood and consumer preference data, which could provide context for the findings, were not integrated into the study, which focused on geographical suitability.

[4] Sergiu-George Limboi and Mara-Renata Petrusel created a restaurant recommendation system that used customer feedback to improve rate predictions through sentiment analysis. But instead of taking into account numerical ratings or

other quantitative information that could increase prediction accuracy, the study just looked at text sentiment.

[5]Tugc ¸e Bilen et al. developed a smart city software that uses machine learning to predict the best places for businesses 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.

[6]Ismam Hussain Khan et al. created a framework using machine learning to forecast 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.

[7]Neha Vaish et al. 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.

[8]Sandeep Bhatia et al. predicted the success using machine learning 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.

[9] Yi Luo and Xiaowei Xu investigated utilizing several machine learning techniques to forecast how beneficial Yelp restaurant ratings would be. 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.

[10] Yanyan Shen, Yanmin Zhu, and Xiaochen Wang provide a way to forecast new restaurant ratings by analyzing both restaurant data and urban data, using a model called MR-Net to capture various influential features without relying on customer reviews.

[11]Yifan Chen; Fanzeng Xia predicts future Yelp ratings for restaurants Using both text and non-text characteristics, achieving the highest accuracy of 82.5% through Decision Tree and Neural Network models.

[12]Nabiha Asghar addresses Yelp review rating prediction using four classification algorithms with four feature extraction methods (bigrams, trigrams, unigrams, and latent semantic indexing) yielded the best results with Logistic Regression on unigrams and bigrams.

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

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

[15] Samia Nawshin, Md. Ismail Hossain, and F. M. Takbir Hossain offer a restaurant recommendation system that combines collaborative filtering and sentiment analysis. enhancing rating predictions based on positive and negative reviews.

3.Dataset Description

DATASET SOURCE : https://www.kaggle.com/datasets/himanshupoddar/zomato-bangalore-restaurants is where this dataset was gathered.

Information on different restaurants in Bangalore, India, may be found in the dataset called Zomato Bangalore Restaurants. The basis for training and analyzing a model is a prepared dataset. Practitioners can choose, preprocess, and use datasets more effectively if they have a theoretical understanding of their structure, types, and properties. 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

• Scaling

• Ratings and Reviews

• Cuisine and Cost

• Restaurant Type and Listing Category

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

Table 1: Column of Attributes before Pre-Processing

attribute	no.of tuples	null / not Null	data type
link	51717	Not null	String
add	51717	Not null	String

name	51717	Not null	String
onlineOrder	51717	Not null	Categorical
BookTable	51717	Not null	Categorical
cost	51717	null	Float
votes	51717	Not null	Integer
phone	51717	null	String
location	51717	null	String
type_rest	51717	null	String
Dish_liked	51717	null	String
Cuisines	51717	null	String
Approx_cost	51717	null	Float
Reviews	51717	null	String(list)
Menu	51717	null	String(list)
Listin(type)	51717	not null	Categorical
Listin(city)	51717	Not null	Categorical

The table (Table.1) shows the attributes, data types, and null/not null values of a dataset with 51717 tuples. The dataset includes a variety of restaurant-related data, such as names, addresses, online ordering capabilities, cuisines, average prices, and reviews.

4.Dataset Pre-processing

A crucial first step in converting unprocessed, frequently complicated data into a clear, useable format is data preparation. Managing missing numbers, translating category data into numerical representations, and spotting and fixing discrepancies are all part of this process. 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.

5.Dataset after Pre-Processing

Here, the steps which are involved are :

• Data Cleaning

• Text Processing

• Sentiment Analysis

• Encoding Categorical Variables

• Feature Engineering

• Scaling

• Splitting Data

Table.2: Column of Attributes after Pre-Processing

attribute	no.of tuples	null/ not null	data type
name	51717	Not null	String
online_order	51717	Not null	Categorical
book_table	51717	Not null	Categorical
rate	47430	null	Float
votes	51717	Not null	Integer
location	51717	Not null	String
rest_type	50982	Null	String
Dish_liked	32955	Null	String
Cuisines	51717	Not Null	String
Average_cost	51710	Null	Float
Reviews	51717	Not null	String(list)
Menu	51717	Not null	String(list)
Listintype	51717	Not null	Categorical

The Table (Tabel.2) shows the attributes, data types, and null/not null values of a dataset with 51,717 tuples. The dataset contains various information about

restaurants, in cluding their names, online ordering capabilities, cuisines, average cost, and reviews.

6. Dataset Visualization

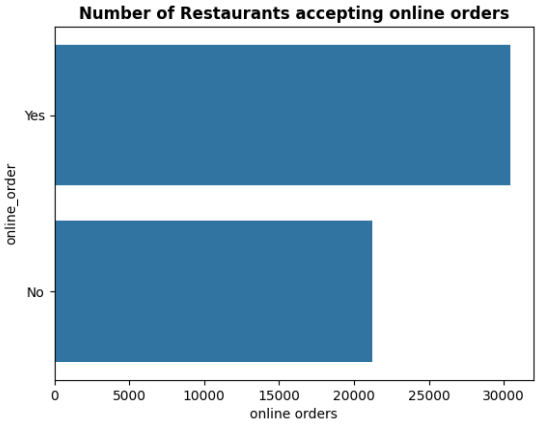

Fig.1: Count of Online Orders

It (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.

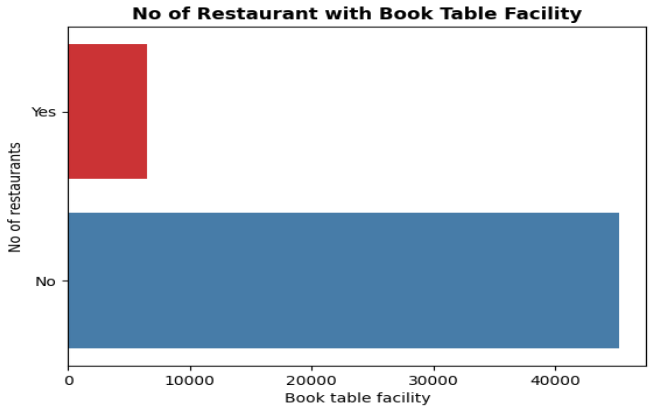

Fig.2: Table Booking Facility

Fig.2 shows the number of restaurants with and without book table facility. A significantly larger lot of eateries lack the ability to reserve tables compared to those that do.

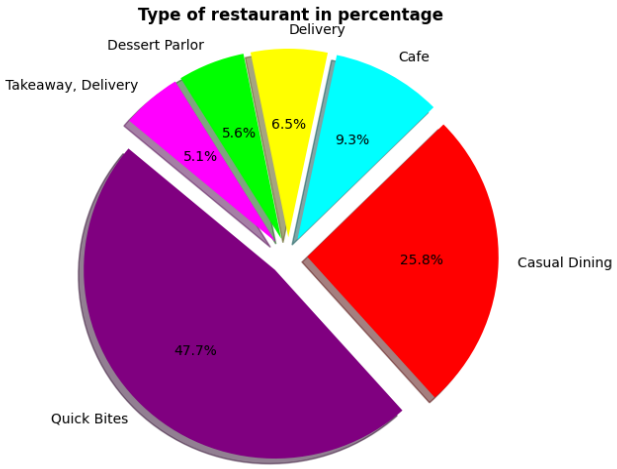

Fig.3: Percentage of Restaurants Types

Fig.3 displays the relative distribution of various restaurant kinds in Bangalore. Quick Bites has the highest percentage of restaurants (47.7), followed by Casual Dining (25.8) and Cafe (9.3).

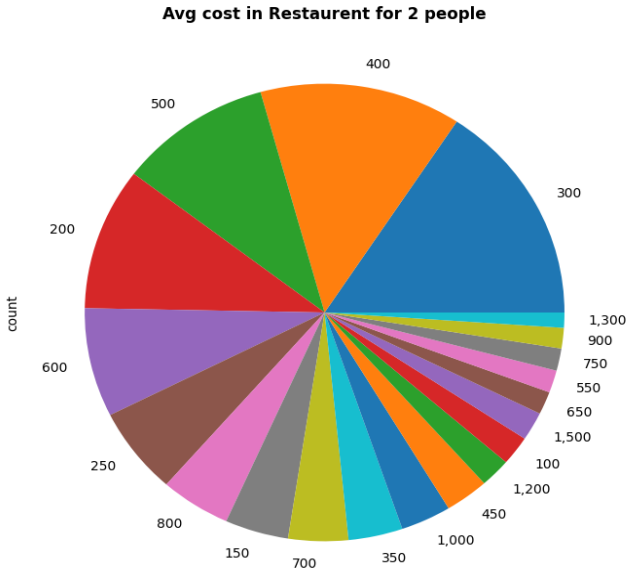

Fig.4: Avg Cost

Fig.4 shows the distribution of restaurants across different average price for a pair. The average cost of dining at most restaurants for two persons is $300 followed by 400 and 500.

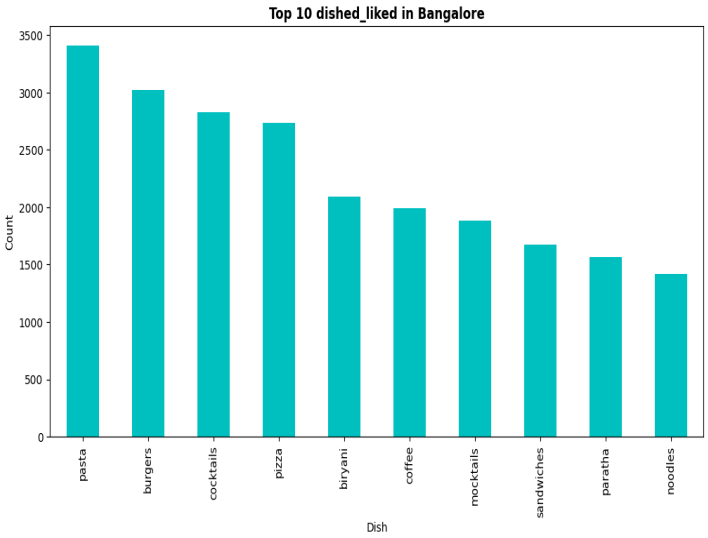

Fig.5: The Top 10 Favorite Recipes in Bangalore

Fig. 5's bar chart displays the ten most popular dishes in Bangalore. Pasta is the most liked dish, followed by burgers and cocktails.

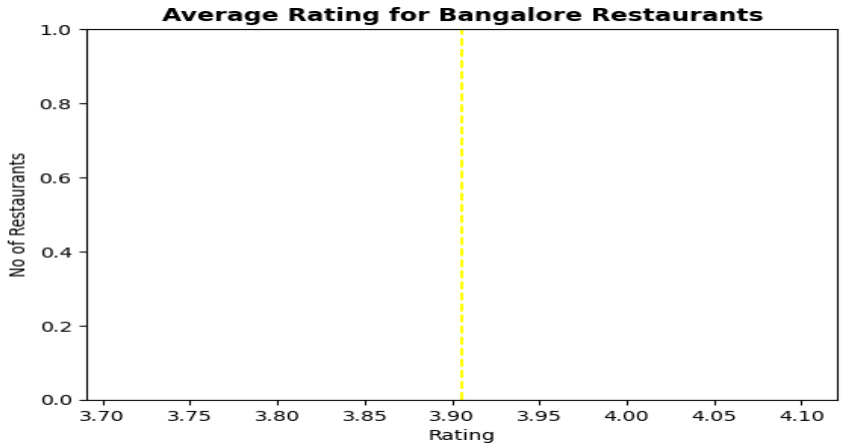

Fig.6: Avg Rating for Banglore Restaurants

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.

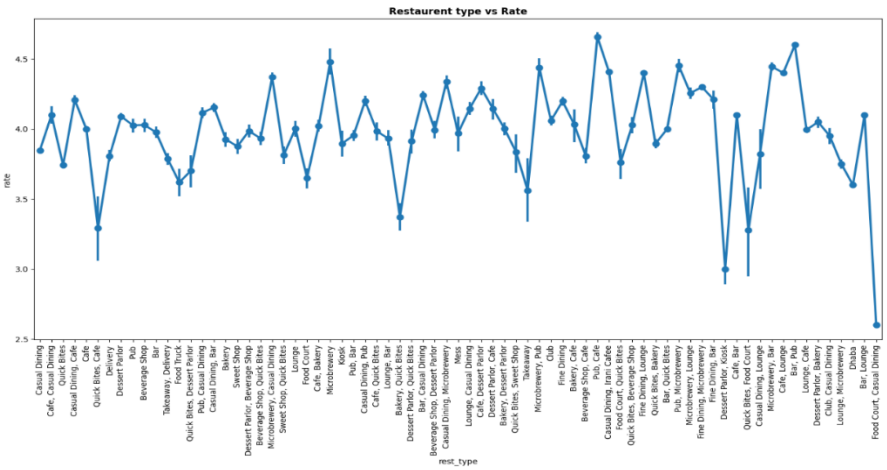

Fig.7: Restaurant Type vs Rate

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

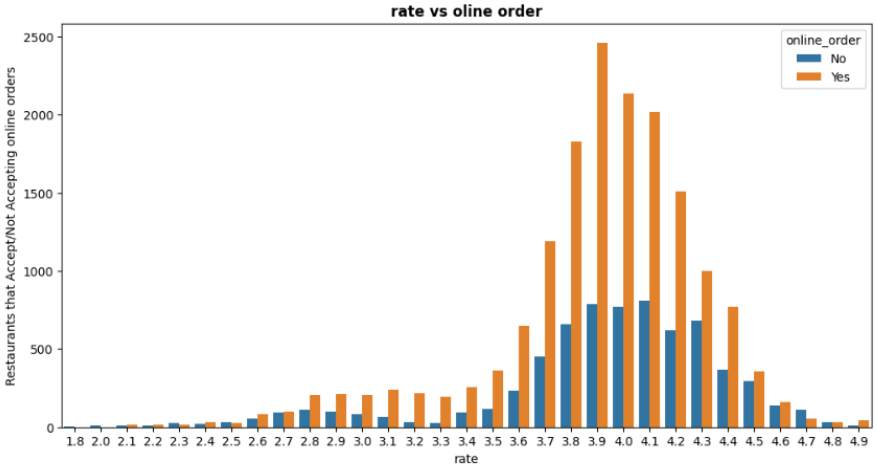

Fig.8: Rate vs Online Order

It (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 the number of restaurants taking online orders rises in tandem with the rating.

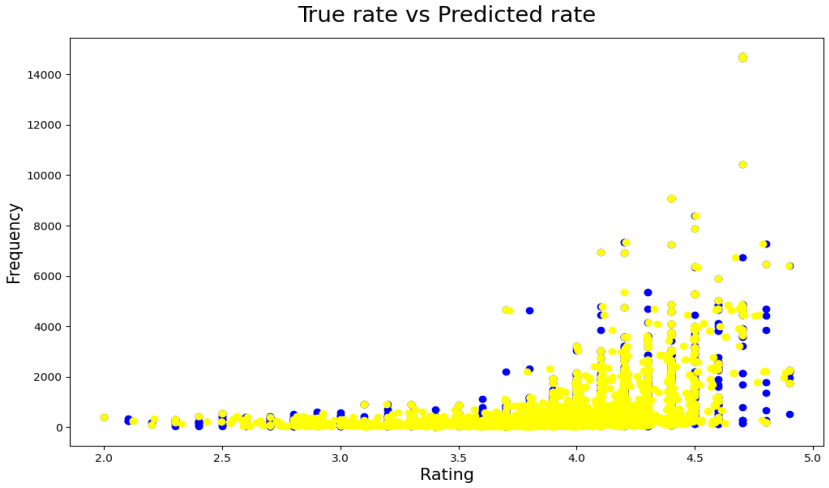

Fig.9 : True vs Predicted Rate

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.

7.Model building

7.1 Linear regression

It 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.

7.2 Random Forest

It is a type of ensemble learning method that builds multiple decision trees on various data sets and aggregates the results to improve prediction accuracy and reduce overfitting. It is suitable for both classification and regression problems due to its resistance to noise and outliers. Random forest are useful for managing complicated and high-dimensional data.

7.3 Ridge Regression

It 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 regu larization 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.

7.4 Lasso Regression

It is also called as 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.

7.5 SVM or support vector machine

The optimal hyperplane to divide data points from several classes is found by this supervised machine learning approach. The goal of categorization is to increase the gap between classes, however for regression (SVR), it tries to fit data within a certain tolerance margin. SVM is effective for high dimensional and non-linear data.

7.6 K-Nearest Neighbors (KNN)

It 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.

7.7 Decision Tree

Based on feature values, these models divide data into branches, which eventually result in a prediction or decision at each leaf node. Although they can capture nonlinear interactions and are simple to read, they are prone to overfitting.. Pruning methods are used to create simpler, more generalizable trees.

7.8 AdaBoost

It (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.

7.9 XGBoost

It is also called 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.

8. Proposed Methodology

The Random Forest Model's Theory

• Random Forest as an Ensemble Model

The Random Forest model was used for this study because of its ensemble structure, which mixes many decision trees to increase forecast accuracy and robustness. The best models for managing complicated, high-dimensional datasets with potentially nonlinear and interdependent feature interactions are ensemble models, such as Random Forest. The model accomplishes this by building a "forest" of decision trees, each of which is trained using a distinct data sample. These trees then aggregate their forecasts to produce a thorough, trustworthy result.

• Ensemble Learning

principle of learning. Several models are individually generated via ensemble approaches, which then combine their predictions. Random Forest is used to separately build each tree using a different set of attributes and data samples. In addition to improving model accuracy, this ensemble technique lowers the possibility of overfitting, which occurs when a model performs remarkably well on training data but is unable to generalize to new data. To arrive at the final regression forecast, the Random Forest model combines the output of every single tree. This average smoothes out extreme forecasts and produces reliable, consistent results because each tree's inaccuracy differs. Because of their capacity to represent intricate feature interactions and handle data noise, Random Forest models are highly regarded.

• Bootstrap Aggregation (Bagging)

Random Forest uses an ensemble approach called bagging (Bootstrap Aggregation) to train its different decision trees on a different bootstrapped subset of the input. Because data samples are chosen at random with replacement, each tree sees a slightly different dataset. By producing trees with varying mistakes, bagging diversifies the model and improves the overall accuracy and generalisability of the outcome. By decreasing the association between individual trees, the Random Forest model's inherent randomness improves its capacity for generalisation. Trees with independent mistakes reduce variance and increase the model's resilience to changes in the data by assisting it in identifying various patterns within the dataset.

• Key Hyper Parametres

•Tree Count (n estimators)

•Depth Maximum (max depth)

•Min samples split (minimum samples per split)

•State of Randomness

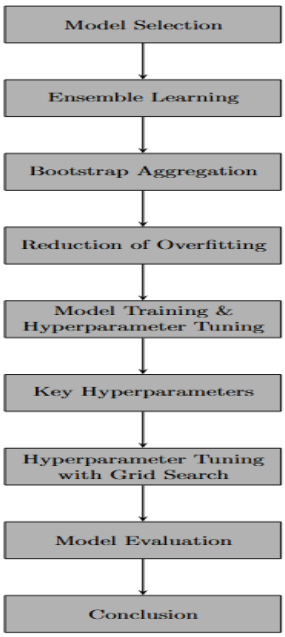

Fig.10. Flow Chart of the Proposed Model(Random Forest)

Fig.10 shows the Model selection is the first step in the sequential flowchart for creating a machine learning model, which then moves on to ensemble learning, hyperparameter tuning, model evaluation, and a summary at the end. It identifies crucial phases to minimise overfitting and maximise model performance.

9.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. By concentrating on the model with the best empirical correctness, this stringent selection procedure guarantees a data-driven strategy. In the end, the study's model selection maximizes prediction accuracy for restaurant evaluations by striking a compromise between theoretical soundness and practical effectiveness.

10.Results

Table.3: Regression models Before parameter tuning

Model	R*2Score	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
Decision Tree	0.88	0.14
AdaBoost	0.01	0.41
XGBoost	0.79	0.18

Table.3 presents the outcomes of applying different regression models to a dataset are shown in a table. While the RMSE (root mean square error) evaluates the model prediction accuracy, the r2 score assesses the model capacity to explain variation in data. Based on these criteria, Decision Tree and XGBoost seem to be the best-performing models.

Table.4: Regression Models Result After Re-Evaluating Parameters

Model	R*2 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
Decision Tree	88.3	14.6	42.8
AdaBoost	10.3	41.8	39.1
XGBoost	79.5	18.4	81.2

table.4 showcases the performance of many regression models after adjusting their parameters is displayed in the table. With the greatest R2 scores and accuracy, the XGBoost and Decision Tree models are particularly noteworthy for their ability to accurately forecast the target variable.

11.Conclusion

Compared to the other models, the Random Forest model performed better in terms of accuracy and error rates when it came to predicting restaurant ratings. The study concludes that Random Forest's robustness and flexibility make it the optimal model for this prediction job. Although Decision Trees also yielded promising results, Random Forest was the best choice in this vestigation based on its performance metrics.

12.Future Scope

In an attempt to increase forecast accuracy, this project entails extending the model to include other information like consumer sentiment from reviews and real-time variables like seasonal trends or promotional events. Performance may be further improved by using other machine learning strategies, such as ensemble models or deep learning. Furthermore, the model's application might be expanded by modifying it to forecast ratings for certain areas or cuisines, which would make it beneficial for a range of restaurant kinds and geographic markets. Last but not least, developing an intuitive tool or API based on this model may give restaurant owners easily accessible, data-driven insights to help them make business decisions.

1 Joseph Anand

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 Engineering and Information Technology (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- ¨ Ozc ¸evik, 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)