A Data-Driven Approach to Predicting Restaurant Ratings on Zomato: An Evaluation of Various Machine Learning Techniques

Sudhehan Kaliamurthi[1], Swetha Subramanian[2], Thanseer Hishak[3], Vetrivel M[4], and Vishwajith S.S.[5]

***Abstract*— With the increasing availability of restaurant reviews and ratings data, predicting restaurant ratings has become an interesting task in the restaurant industry. In this paper, we present a comparative study of machine learning models for predicting restaurant ratings based on various user and restaurant features. We evaluate several popular machine learning techniques, including linear regression, decision tree, random forest, and gradient boosting, on a dataset of Yelp reviews and ratings. We also perform feature importance analysis to investigate the impact of different features on rating prediction accuracy.**

I. Introduction

In recent years, consumers' use of restaurant reviews and ratings as a source of information when choosing where to eat has increased significantly. There is a wealth of restaurant review and rating data accessible for research due to the rising popularity of online review platforms like Yelp and Zomato. In this study, we apply machine learning approaches to forecast restaurant evaluations based on numerous user and restaurant characteristics. We leverage the Zomato Restaurants Data, a publicly accessible dataset that includes restaurant details, customer reviews, and ratings for thousands of restaurants across several locations worldwide, to accomplish this purpose.

One of the top websites for finding restaurants and buying meals online, Zomato is present in over 24 countries. Users may find restaurants using Zomato's platform, read reviews and ratings, see menus, and place online meal orders. Additionally, the platform offers information on restaurant attributes including location, pricing range, cuisine type, and user demographics. We use the Zomato Restaurants Dataset, which provides details on hundreds of eateries from various places across the globe.

Machine learning is the most apt tool for our task Machine learning algorithms may discover intricate connections between data and ratings and produce precise predictions based on these connections. In this research, we assess a number of popular machine learning algorithms, including linear regression, decision tree, random forest, and XGboost, on the Zomato Restaurants Data to determine the most crucial elements for rating prediction, we also conduct feature importance analysis.

Our work adds to the body of knowledge on restaurant rating prediction by offering a thorough examination of several machine learning models and feature sets. Our study can help restaurants improve their services and consumers make wise restaurant selections by identifying the most crucial criteria for rating prediction.

II. Related Work

Several studies and students have published similar work in national and international research papers, theses, and books to better understand the purpose, types of algorithms employed, and approaches for pre-processing and feature selection.

[1] I. K. C. U. Perera and H.A. Caldera have used data mining techniques like Opinion mining and Sentiment analysis to automate the analysis and extraction of opinions in restaurant reviews.

[1]205002104 [2]205002111 [3]205002113 [4]205002122 [5]205002127

[2] Chirath Kumarasiri's and Cassim Faroo's focuses on a Part-of-Speech (POS) Tagger based NLP technique for aspect identification from reviews. Then a Naïve Bayes (NB) Classifier is used to classify identified aspects into meaningful categories.

[3]Neha Joshi wrote a paper in 2012 on A Study on Customer Preference and Satisfaction towards Restaurant in Dehradun City which aims to contribute to the limited research in this area and provide insight into the consumer decision making process specifically for the India foodservice industry. She did hypothesis testing using chi-square test.

[4]Shina, Sharma S. and Singha A. have used Random forest and decision tree to classifying restaurants into several classes based on their service parameters. Their results say that the Decision Tree Classifier is more effective with 63.5% of accuracy than Random Forest whose accuracy is merely 56%.

III. Dataset Description

The dataset we have used is collected using the Zomato API. The dataset is a corpus of twitter comments that consists of 9552 records and 25 columns namely Restaurant Id, Restaurant Name, Country Code,City Address, Locality, Locality Verbose: Longitude, Latitude, Cuisines, Average Cost for two, Has Table booking, Has Online delivery, Is delivering, Price range, Aggregate Rating, Rating color, Rating text, Votes.

IV. Preprocessing

The Dataset contained 25 Attributes.

- Records with null values were dropped from ratings columns and were replaced in the other columns with a numerical value.
- Non-Numerical Fields were removed.
- Using LabelEncoding from sklearn library, encoding was done on columns like book_table,online_order,rest_type,listed_in(city).

V. Training the Models

In this paper, 5 models were created and trained. The cleaned and preprocessed dataset was provided as input forthe models. Short notes on the techniques used and the models are given below.

A. Linear Regression

Linear regression is a machine learning technique that uses one or more input variables to predict a continuous output variable. It assumes that the inputs and outputs have a linear connection and finds the best linear function that fits the data by modifying coefficients that weight the input variables. The objective is to minimise the difference between projected and actual output values by minimising a cost function that gauges prediction error.

B. Decision Tree

A decision tree is a popular machine learning technique that creates a tree-like representation of decisions and their potential outcomes. The technique attempts to predict a categorical or continuous output variable based on a set of input variables by recursively splitting the input data into smaller subsets, each time selecting the optimal variable to split on based on a criterion that maximises output variable separation.
The result is a tree-like model where each internal node represents a decision based on an input variable, and each leaf node represents a predicted output value. Decision trees are easy to interpret and can handle both numerical and categorical data, making them a popular choice for classification and regression tasks. However, they are prone to overfitting, which can be addressed using techniques like pruning or ensemble methods.

C. Random Forest

Random Forest is an ensemble learning technique that fuses several decision trees to form a robust and powerful classifier. The algorithm trains each decision tree by randomly selecting subsets of features and data points, which curbs overfitting and enhances generalization. Random Forest can handle both classification and regression problems and estimate feature importance. It finds extensive applications across

different domains, but it can be susceptible to noisy data and needs meticulous parameter tuning to achieve optimal performance.

D. XGBoost

XGBoost is an ensemble learning technique that leverages gradient boosting to merge numerous weak learners. The algorithm incorporates several features, including regularization, early stopping, and feature importance estimation, to enhance its performance. XGBoost employs a specialized loss function that enables it to optimize for specific performance metrics, and it can handle both binary and multi-class classification and regression problems. With its remarkable performance on many benchmark datasets, XGBoost is considered a stateof- the-art technique. However, it can be sensitive to noisy data and necessitates cautious hyperparameter tuning to achieve optimal performance.

E. Support Vector Machine

SVM is a classification and regression analysis machine learning technique. By maximising the margin between the two classes, the SVM classifier attempts to discover the best hyperplane that splits the input data into multiple classes. The margin is defined as the distance between the hyperplane and each class's nearest data points. The ideal hyperplane is determined by locating the hyperplane with the greatest margin while correctly classifying the training data. If the input data cannot be separated linearly, a kernel function can be used to transform the data into a higher dimensional space where it can be separated.

VI. Evaluvation Metrics

A. MSE (Mean Squared Error):

MSE calculates the average squared difference between predicted and actual data. MSE = 1/n(yi - i)2, where n is the number of samples, yi is the actual value of the target variable for the i-th sample, and i is the predicted value of the target variable for the i-th sample.

B. Root Mean Squared Error (RMSE):

The RMSE is the square root of the MSE and represents the residuals' standard deviation. The RMSE formula is: RMSE = (1/n (yi - i)2)^2.

C. R-squared (R^2):

R^2 represents the proportion of the variance in the target variable that is explained by the model.

VII. Results

Algorithms	**Rsquare Value**
Linear Regression	22%
Decision Tree	5.5%
Random Forest	45%
XGBoost	49.62%
SVM	39.61%

Best Performing algorithm is XGBoost and Least performing is Decision Trees algorithm.

VIII. Conclusion

We investigated the use of machine learning approaches for predicting restaurant ratings using the Zomato Restaurants Data in this study. We tested three well-known machine learning algorithms: linear regression, decision trees, and SVM-based classifiers. Our findings indicate that these algorithms can be used to predict restaurant ratings with excellent accuracy. In addition, our research emphasises the significance of data preprocessing, feature selection, and model evaluation in reaching optimal performance. Overall, our findings show that machine learning has the ability to improve restaurant rating forecasts, which might have significant ramifications for the restaurant business and its customers.

References

[1] I. K. C. U. Perera and H. A. Caldera, "Aspect based opinion mining on restaurant reviews," 2017 2nd IEEE International Conference on Computational Intelligence and Applications (ICCIA), Beijing, 2017, pp. 542-546. doi: 10.1109/CIAPP.2017.8167276

[2] Chirath Kumarasiri, Cassim Faroo,"User Centric Mobile Based Decision-Making System Using Natural Language Processing (NLP) and Aspect Based Opinion Mining (ABOM) Techniques for Restaurant Selection". Springer 2018. DOI: 10.1007/978-3-030-01174-1_4

[3] Shina, Sharma, S. Singha ,A. (2018). A study of tree based machine learning Machine Learning Techniques for Restaurant review. 2018 4th International Conference on Computing Communication and Automation (ICCCA) DOI:/10.1109/CCAA.2018.8777649

[4] Shina, Sharma, S. Singha ,A. (2018). A study of tree based machine learning Machine Learning Techniques for Restaurant review. 2018 4th International Conference on Computing Communication and Automation (ICCCA) DOI:/10.1109/CCAA.2018.8777649”