Restaurant Rating Prediction

Megha Sajan

2024.1.7

I. Introduction

In the fast-paced world of online food delivery, user reviews and ratings play a pivotal role in influencing customer decisions.This analysis can be used by online food delivery apps who understands the significance of providing accurate predictions for the quality of restaurants listed on its platform. The ability to predict restaurant ratings can empower the apps to make informed decisions regarding whether to include a restaurant in their app or potentially remove it. This predictive capability can significantly enhance user experience, as customers often rely on ratings to choose the best dining options.

This problem is crucial for food delivery business strategy, as accurate rating predictions contribute to user satisfaction and trust in the platform.Our goal is to leverage machine learning techniques to develop a robust predictive model capable of forecasting restaurant ratings based on various characteristics.A successful machine learning model would potentially save resources by avoiding the inclusion of under-performing restaurants. Overall, this predictive modeling task is relevant and important for for a businesses ongoing commitment to providing users with a seamless and delightful culinary experience through their platform.The research question is: Can we accurately predict restaurant ratings based on features such as cuisine variety, online presence, and location?The task at hand involves leveraging machine learning techniques to develop a model that can forecast the rating of a restaurant based on various characteristics.A Random Forest method is chosen do this regression problem and explained using Local blackbox method shapley.

II. Data

The dataset, sourced from Kaggle, encompasses details from 9551 restaurants worldwide and features 19 variables. Key attributes include 'Votes,' 'Average Cost for Two,' 'Has Table Booking,' 'Has Online Delivery,' and 'Rating.' The objective is to unravel the factors influencing restaurant ratings, aiding food delivery apps in strategic decisions regarding restaurant inclusion or removal. As Rating is continuous, and data has labels, a regression with supervised ML algorithm used.

Duplicate records and missing values were scrutinized, and the dataset was found to be free of such issues, ensuring data integrity. The target variable, 'Rating,' exhibits a non-normal distribution, characterized by negative skewness and a bi-modal pattern. To accommodate this, a tree-based model was chosen, given its resilience to the effects of non-normality.(as shown in *figure 1*)

A new feature, 'Cuisine Count,' was created by tallying the number of cuisines offered by each restaurant.Several categorical variables, including 'Country Code,' 'Currency,' 'Has Table Booking,' 'Has Online Delivery,' 'Is Delivering Now,' and 'Switch to Order Menu,' were identified and converted into factors.Acknowledging the inequity in 'Average Price for Two' due to varying purchasing powers, a PPP adjustment was proposed. This adjustment standardizes the values, allowing for fair cross-country comparisons.Since this variable had negligible correlation with target variable, it was finally left out.

Certain variables ('Country Code,' 'Is Delivering Now,' 'Switch to Order Menu') exhibited significant skewness, with one dominant category. To enhance model efficiency, these columns were deemed less informative and were removed.The remaining categorical variables are shown in *figure 2*

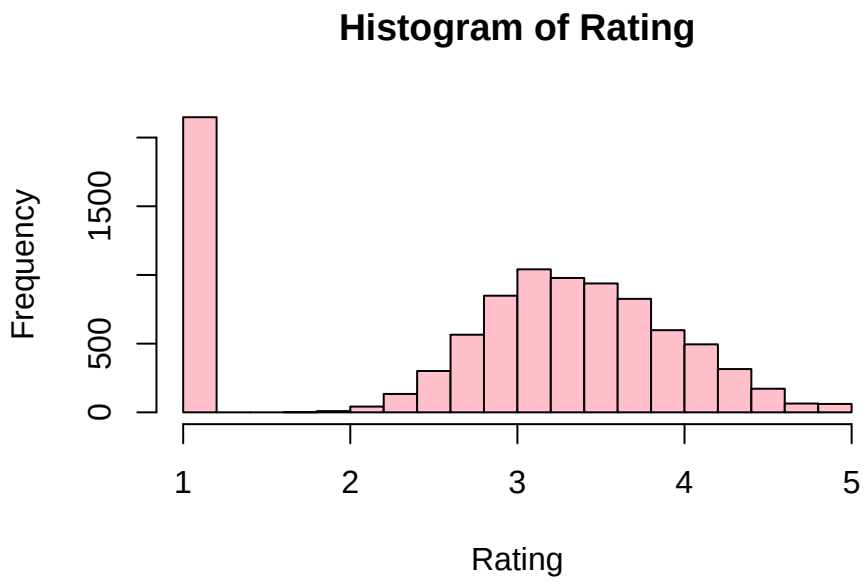

Figure 1: Histogram of Rating

Figure 2: Categorical Variables

Some qualitative variables and some outliers were also removed. Finally,'Votes', 'Has.Table.booking', 'Has.Online.delivery', 'Price.range', 'CuisineCount' , 'Rating' were considered fit for further analysis.

The dataset was randomly split into 80% training and 20% test sets, a standard practice in model development. Random forests, the chosen algorithm, inherently utilizes out-of-bag samples for internal valida-

tion.Numerical variables were not scaled, considering the inherent insensitivity of random forests to differences in scale. A substantial majority (90%) of the dataset comprises observations from Country Code 1 (USA).

II. Method

In this study, three distinct methods were employed to model and interpret the relationship between restaurant characteristics and their ratings: Random Forest with deafult parameters and with tuned parameters,Linear Regression, and Shapley values.

Random Forests, a powerful ensemble learning technique, were employed for regression in this analysis. The method constructs multiple decision trees independently, each based on a random subset of features and data points. This randomness helps reduce bias and variance, making Random Forests resilient and robust. A random forest is a classification ensemble composed of a set of tree-structured classifiers $h(x, k), k = 1, ...$ where the k represent independent and identically distributed random vectors. Each tree in the ensemble independently provides a unit vote, and collectively, they contribute to determining the most popular class for a given input x. The strength of a random forest lies in its ability to aggregate diverse decision trees, promoting robustness and enhancing overall predictive performance.It constructs multiple decision trees and combines their predictions. Each tree is trained on a subset of the data, and the final prediction is determined by aggregating the individual tree outputs. Random Forest was chosen for its ability to handle slightly skewed data,fast for this data, and provide robust predictions. Its inherent capacity for handling complex interactions among features makes it suitable for capturing the intricate patterns within the restaurant data. Out-of-bag data points, those not used in the construction of each tree, are utilized for immediate validation. The final prediction is often determined through majority voting in classification tasks and average for regression problems.

Linear Regression is a traditional statistical method that models the relationship between a dependent variable (rating) and one or more independent variables (restaurant characteristics) by fitting a linear equation.Linear Regression was employed to provide a baseline model for comparison. Despite its simplicity, linear regression can capture linear relationships between variables and offers interpretability, allowing us to understand the direct impact of each feature on the predicted rating.

Shapley values, rooted in cooperative game theory, were used to interpret the black-box nature of the Random Forest model. Shapley values assign a value to each feature, indicating its contribution to a specific prediction by considering all possible feature interactions.Shapley values were chosen to enhance interpretability by providing insights into how each feature influences individual predictions. By understanding the impact of each feature on the model's decision-making process, we gain valuable insights into the factors contributing to restaurant ratings.SHAP values offer a systematic approach to interpreting the output of complex black-box models like Random Forests. The Shapley value for a specific feature is calculated by considering all possible subsets and permutations of features. This accounts for the average contribution of the feature, considering its interactions with other features. SHAP provides both local and global interpretability, allowing for insights into how each feature influences individual predictions.Shapley used from iml package.Here, One observation is considered for shapley evaluation with Votes=314,Has.Table.booking=Yes, Has.Online.delivery=No, Price.range=3 and CuisineCount=3.

For evaluating the model on the test set ,R^2 and Mean square error were employed. A higher R^2 suggests that a larger percentage of the variability in restaurant ratings is explained by the model. A lower MSE implies that, on average, the model's predictions are closer to the actual ratings. In the context of restaurant ratings, a lower MSE signifies that the model provides more accurate predictions, and the deviations between predicted and actual ratings are minimized.

This combination of methods aimed to leverage the strengths of each approach: the predictive power of Random Forest, the simplicity and interpretability of Linear Regression, and the interpretative capabilities of Shapley values.

IV. Result

To optimize the Random Forest model, the number of trees was tuned by assessing the error across different tree counts. From the observed trend (see *Figure 3*), it became evident that the OOB error plateaus and shows no significant improvement beyond approximately 100 trees.

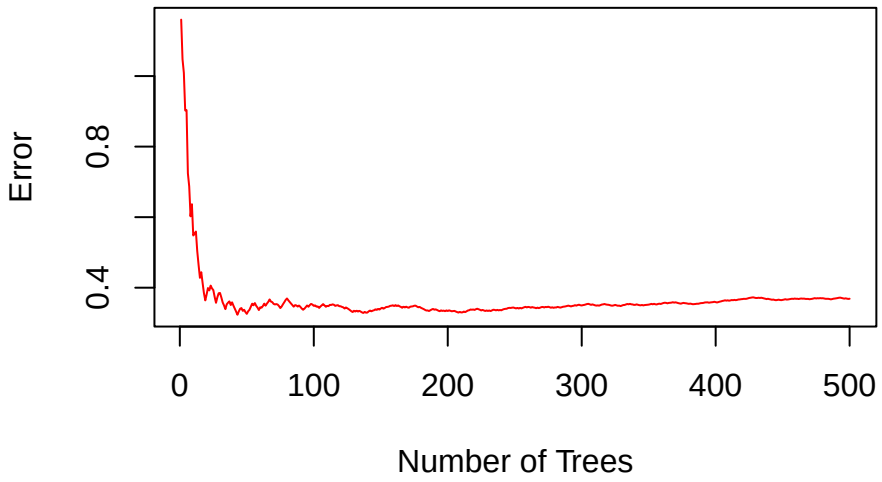

Figure 3: Number of Trees

Next, the tuning process involved determining the optimal number of variables or columns to be considered at each step, known as `mtry`. This was achieved through a systematic evaluation, and the results are visualized in *Figure 4*, where the error is minimized.From the figure an mtry of 3 minimizes the OOB error.

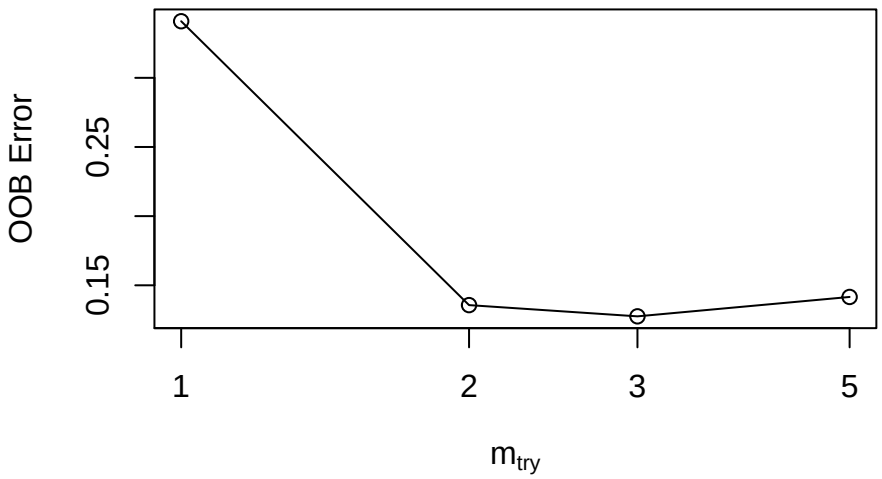

Figure 4: Optimal mtry

Only 3 hyperparameters were tuned due to computational complexity using makeParamset in mlr package.`mtry=3`, `nodesize=5`, and `ntree=129`, was identified as the most effective reducing OOB mean error. A final Random forest model was created with these parameters which yielded MSE 12.8% and R^2 90.2%.

To compare as a base model, linear regression is performed on the dataset and result metrics are drawn.

Table 1 shows a Metrics comparison of Random forest with default parameters and after tuning the parameters along with Linear Regression results :

Metric	Default Result RF	Tuned Result RF	Linear Regression
MSE	36.8%	12.8%	86.7%
R^2	71.0%	90.2%	30.1%

Table 1: Results Random Forest and Linear Regression

Linear Regression shows a poor performance with high error and low R^2 whereas, Random forest with tuned parameters gives the best result with low error.

Using Shapley values to interpret the model predictions, we gain insights into the factors influencing the predicted ratings. The positive Shapley value for 'Votes' (0.78497347) suggests that a higher number of votes contributes significantly to a more favorable predicted rating, indicating popularity and positive feedback. On the contrary, the negative Shapley value for 'Has Table Booking' (-0.19272261) implies that having a table booking has a moderate negative impact on the predicted rating, possibly due to limited seating or exclusive services.

For 'Has Online Delivery,' the positive Shapley value (0.01528998) indicates a small positive impact on the predicted rating, suggesting that offering online delivery contributes slightly to a higher rating. Similarly, the positive Shapley value for 'Price Range' (0.15870087) implies a moderate positive impact, indicating that a higher price range is associated with a more favorable predicted rating, possibly due to perceived better quality in upscale dining experiences. On the other hand, the negative Shapley value for 'Cuisine Count'

(-0.01612846) suggests that a higher count of cuisines has a slight negative impact on the predicted rating. This may indicate that, on average, a moderate number of cuisines is more positively associated with ratings, while an excessively high count might reflect a lack of specialization.

These interpretations provide a glimpse into how changes in each feature influence the model's prediction for the given observation. Positive values contribute positively to the prediction, while negative values have a negative impact. The magnitude of the Shapley value reflects the strength of the impact.

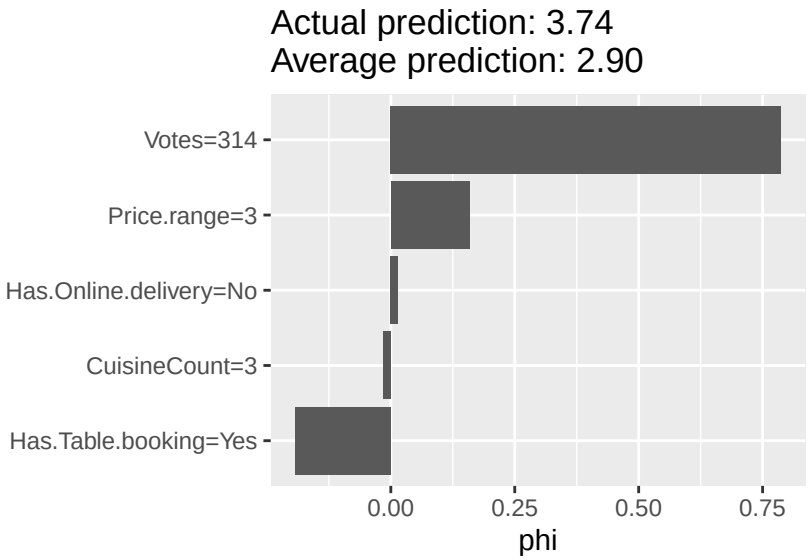

Figure 5: shapley Interpretation

V. Limitations and Conclusion

The research questions aimed to understand the determinants of restaurant ratings and explore the interpretability of a complex model. The Random Forests model, chosen for its ability to handle non-linear relationships and interactions, revealed that features such as Votes, Table Booking, Online Delivery, Price Range, and Cuisine Count play significant roles in predicting restaurant ratings. The application of Shapley values further showed that a company must focus on increasing its votes online for better review.Linear Regression, employed for comparison, offered insights into the linear relationships between the selected features and ratings and proved not to be working for this dataset and a need to an advanced algorithm.Overall, the findings contribute to the understanding of restaurant ratings, aiding decision-making for food delivery apps in selecting or removing restaurants from their platforms. The combined use of machine learning models and interpretability techniques enhances the transparency of the decision-making process, offering valuable insights for stakeholders in the food industry. As a limitation- after data cleaning, from 19 left with only 5 features. More features would have been more insightful and rather than just taking the count of cuisines, each cuisine can be a dummy variable so that cuisine popularity is also known.

VI. References

Bruce, Peter, and Andrew Bruce. 2017. *Practical Statistics for Data Scientists*. O'Reilly Media.

Burzykowski, P. B. a. T. (2020, December 11). 8 *Shapley Additive Explanations (SHAP) for average attributions | Explanatory Model analysis*.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2014. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated.

Shina, Sharma, S. & Singha ,A. (2018). *A study of tree basednmachine learning Machine Learning Techniques for Restaurant review* . 2018 4th International Conference on Computing Communication and Automation (ICCCA)