Proceedings of the Fifth International Conference on Electronics, Communication and Aerospace Technology (ICECA 2021)
IEEE Xplore Part Number: CFP21J88-ART; ISBN: 978-1-6654-3524-6

Restaurant Rating Prediction Using Regression

SUHAS SOMASHEKAR
Electronics and Communication Engineer
University Visvesvaraya College of Engineering
Bengaluru, India
ssuhas.28@gmail.com

SUHAS MALLESH
Electronics and Communication Engineer
University Visvesvaraya College of Engineering
Bengaluru, India
sssuhas8@gmail.com

2021 5th International Conference on Electronics, Communication and Aerospace Technology (ICECA) | 978-1-6654-3524-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICECA52323.2021.9675922

***Abstract*—Opening a new Restaurant has been a challenge especially in current times due to tough competition and variety of choices for a consumer. Thus, Rating of a restaurant becomes an important parameter for judging the quality of a Restaurant. A good rating acts as an invitation for new customers. And every new business wishes to know if it will succeed in future or not.**

The purpose of this paper is to provide validation for new Restaurants about their predicted Ratings. By inputting various factors that are characteristics of a planned Restaurant, one is able to predict the average Ratings and if necessary, provide a comparison between various options of a particular characteristic of the Restaurant. It can also help one in making the right choices before the action of opening a new Restaurant is implemented. Thereby securing them against investment losses, saving time and making the whole process a well thought out calculated risk. Seven different types of regression models are used to make the predictions by considering factors that can be easily controlled before setting up a new restaurant and finally model metrics are compared to choose the best regression model for future prediction.

Keywords—Rating Prediction, Regression, Machine Learning, Data Analytics

I. INTRODUCTION

The rating of a Restaurant depends on multiple factors. They include type of restaurant, cuisines offered, average cost of food, location, types of facility available, number of votes etc. These multitude of factors can be quite overwhelming to manage and its usually the combination of these factors that leads to the success of a restaurant. Further analysis into this aspect can not only demystify what customers value the most in a restaurant, but it can also provide suggestions on what combination of features that one should choose before opening a new restaurant and predict how likely this restaurant can succeed.

A statistical method called Regression is used to investigate the relationship between independent variables and the dependent variable. Models which do regression generally finds its utility in forecasting, prediction and establishing causal relationships between independent and dependent variables. And in this specific case, regression can be used to predict ratings and also find which factors influence ratings.

II. LITERATURE REVIEW

In [1] A. Shibata, S. Kamei and K. Nakano used Category-oriented Sentiment Polarity Dictionaries (CSPD) on hotels listed in review database of Rakuten Travel. They mainly focused rating prediction on text review by assigning polarity values for sentiment in their text review. However, they did not consider external factors of a hotel, and this becomes difficult for rating predictions for new hotels with no or less reviews.

In [2] P. Chanwisitkul, A. Shahgholian and N. Mehandjiev used text mining to extract hotel reviews and further analyze features that drive customer satisfaction. The purpose was more driven to guide managers regarding hotel management and marketing of customer experiences. No emphasis was given on rating prediction.

In [3] L. Pradhan, C. Zhang, and P. Chitrakar used collaborative filtering to predict ratings for items that were not rated. They used cluster models to group similar items and similar users and utilized the findings to ratings for unrated items. However, this approach required the availability of ratings data for other items and users which would not exist for a new restaurant.

In [4] S. Saha and A. K. Santra utilized online textual reviews of restaurants to predict ratings. They first calculated user level sentiment and then proceeded to calculate food item level sentiment. They also considered interpersonal sentiment and finally used the findings to denote user preferences and use them in recommendation systems. This paper was only limited to analysis on textual reviews .

In [5] D. Sivabalaselvamani and B. Soorya provided a framework for ratings based on facial expression of customers especially for unmanned and mechanized eateries. They employed convolution neural network to articulate the expressions. The scope was limited to establishment of new rating system based on facial expressions at restaurants.

In [6] I. F. Shihab, M. M. Oishi, S. Islam, K. Banik and H. Arif used multiple machine learning algorithms on Yelp dataset to analyze factors responsible for defining a perfect place to set up a restaurant. They identified 75 features and tried to predict ratings for restaurant based on their location. However, this paper was more inclined towards suggesting the ideal location for setting up a restaurant.

III. DATA PREPARATION

A. Data Source

The dataset is sourced from Kaggle website [7].

B. Data Overview

The data has a list of 43,942 restaurants in Bengaluru city with 14 attributes describing its URL, physical address, online order availability, table booking availability, ratings, number of votes received, contact number, location, restaurant type, number of dishes liked, list of cuisines offered, approximate cost for two people, reviews list, items on the menu.

C. Data Preprocessing

The columns with null values are imputed with appropriate values and rows with more than 50% of the data missing are removed.

The column names are renamed for easier understanding and the data types of corresponding values are converted to their specific format. For example, the column 'Rating' is converted to float and all numerical columns are converted to either integer or float.

Unnecessary punctuations are removed from the data and converted to proper format. For example, one row of 'Ratings' column was represented as '4/5' and it was converted to only '4' by removing the '/5'

D. Data Processing

The categorical columns like online order availability, table booking availability are encoded using label encoder. The column specifying the location of the restaurant is ordinal encoded.

Columns which provided a list of items are replaced with their numerical count. For example, 'list of cuisines offered', 'items on the menu' are replaced with 'count of cuisines offered', 'number of items on menu' respectively.

Unnecessary columns that do not add value to the problem statement like 'URL', 'Address', 'contact number', 'reviews list'.

Finally, the independent variables selected for the model are 'online order availability', 'table booking availability', 'restaurant type', 'location', 'approximate cost for two people', 'count of cuisines offered', 'number of items on menu'.

The dependent variable is the column called 'Ratings'

The whole dataset is divided into training and test dataset in the ratio 80:20. And the training dataset (27,811 rows)and test dataset (6,953 rows)of independent variables is passed through a Minmax Scaler.

IV. DATA VISUALIZATION

Data Visualization is a technique in which the data is represented in the visual format including graphs, charts, or figures. This enables one to handle large volumes of data in a visual format, spot underlying patterns and draw certain conclusions. On the other hand, it acts like a quality check by showing the distribution of data and identifying the outliers present.

Fig. 1. Bar chart showing the frequency of restaurants in top 10 locations

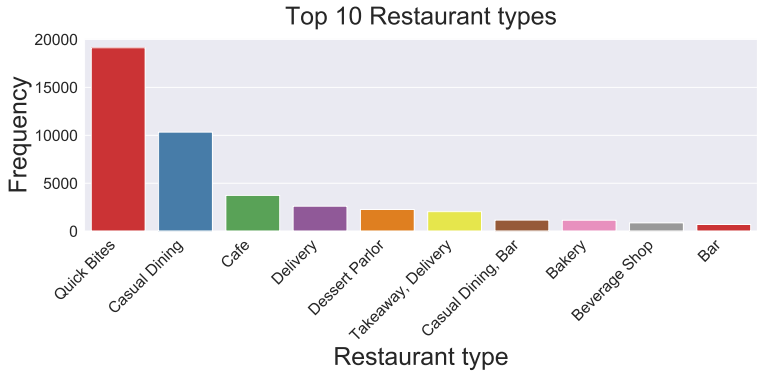

Fig. 2. Bar chart showing the frequency of top 10 restaurant types

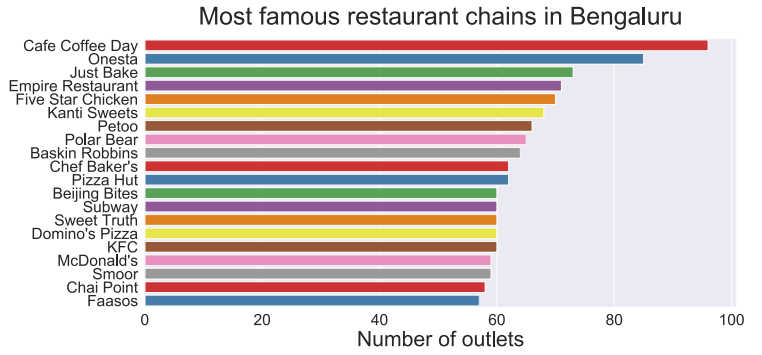

Fig. 3. Bar chart showing the number of outlets of top restaurant chains

Fig. 4. Bar chart showing the count of online order availability in restaurants

Fig. 5. Bar chart showing the count of table booking availability in restaurants

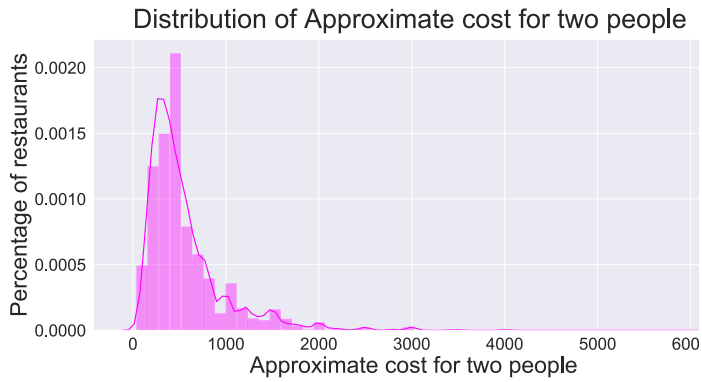

Fig. 6. Histogram showing distribution of Approximate cost for two people in restaurants

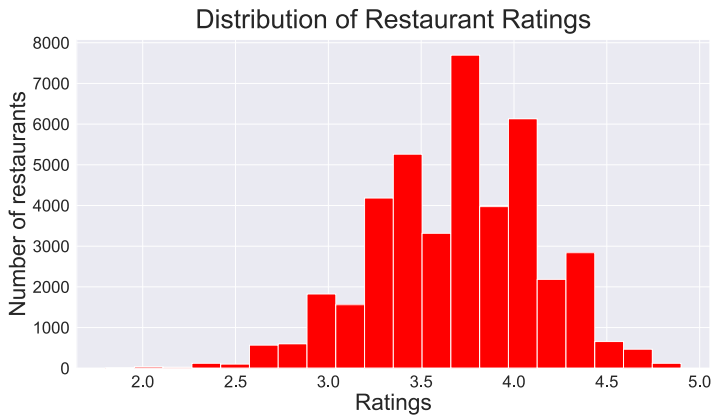

Fig. 7. Histogram showing distribution of Restaurant ratings

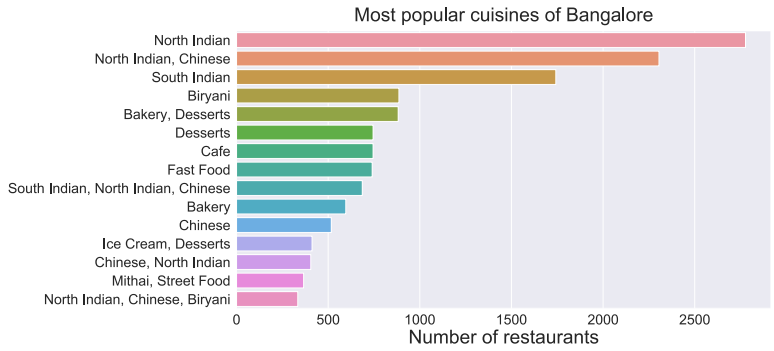

Fig. 8. Horizontal Bar chart showing the most popular cuisines

Above figures can lead to the following inferences. Fig. 1 shows that there are more than 5,000 restaurants in the top location and there are more than 2,000 restaurants each in the top 6 locations. Fig. 2 shows that quick bites and casual dining are marginally the highest types of restaurants present in the dataset. Fig. 3 shows that there are more than 55 number of outlets for each of the top 10 famous restaurant chains. Fig. 4 shows that approximately 30,000 restaurants have the provision on online ordering. Fig. 5 shows that majority restaurants do not have the provision of booking a table in advance. Fig. 6 shows the distribution of the approximate cost for two people. It is evident that the mean of distribution is almost 500 rupees. Similarly, Fig. 7 shows that '3.7' is the most popular ratings for the restaurants in the dataset. Finally, Fig. 8 shows that the popular cuisines are North Indian, Chinese, and South Indian.

V. MODELS USED

A. Linear Regression

This model presumes a linear relationship between the dependent variable (Y) and independent variables (X). Both the input values and output values should be numeric. It assigns one scale factor to every input value or column, called a coefficient. Another bias coefficient is added to give the best fit line an additional degree of freedom. This model estimates the values of coefficients used in the representation with the data. This

model is fitted using least squares approach which tries to minimize the square of errors between the actual value and predicted value.

B. Support Vector Machine

This is a supervised machine learning algorithm that can solve both classification and regression problems. In this model, each point is plotted in a n-dimension space with the value of each attribute being the value of the particular coordinate. In addition to working well on linear equations , it works well on non-linear equations by using the kernel trick. It constructs a set of hyperplanes in a high dimensional space and a good separation is attained by the hyperplane that has highest distance to the nearest training point. It has great regularization capabilities and maintains stability even if there is a small percentage change in the dataset. But this model requires extensive memory, long training time and the results are difficult to interpret.

C. Decision Tree

This model builds a regression or classification model in the form of tree structure. The dataset is broken down into smaller and smaller subsets while an associated decision tree is incrementally developed at the same time. In the result, there is a tree with decision nodes and tree nodes. A decision node , having 2 or more nodes, represents values for the feature tested. Decision on a numerical target is represented by a leaf node. This model can handle both numerical and categorical data. Nonlinear relationship between parameters do not affect the model performance. However, this model is prone to overfitting and can lead unstable results for small variations in dataset. And if some classes dominate others, it can lead to biased trees.

D. Random Forest

This model is a supervised learning algorithm that uses ensemble learning method for regression. It combines predictions from multiple machine learning algorithms to make a prediction that is more accurate than a single model. It actually uses multiple decision trees for making decision and this is what makes this model powerful. It is called 'Random' because uses randomly created decision trees. It can also be utilized to rank the importance of variables in a regression model. Although it can achieve high accuracy, they can lead to overfitting of the data and interpretability of the model is very low. Also, it is unable to discover trends that could enable it to extrapolate values that fall outside the training set.

E. K Nearest Neighbours

This algorithm can be utilized for both classification and regression. It uses feature similarity to predict values for new data points. It means new points are assigned values on how closely they resemble other points in the train dataset. The assigned value is an average of the values of its K nearest neighbors. It is a lazy learner because it follows instance-based learning and does not learn during training period. And new data can be added easily without affecting the accuracy of the model. However, this algorithm does not work well with larger datasets and dataset with high dimensionality. It needs feature scaling and is very sensitive to outliers, noisy data, and missing values.

F. ADA Boost

ADA Boost is a short form for Adaptive Boosting. This can be used in conjunction with other machine learning algorithms to improve performance. It is adaptive because the subsequent weak learners are changed in favor of those instances that were misclassified by previous learners. The individual learners can be weak, but the final model converges as a strong learner. It uses multiple decision stumps which have single node and two leaves, and more weights are assigned in next steps to decision stumps which misclassify. This model needs less tweaking, is theoretically less prone to overfitting, and gives good accuracy. However, it required quality data because it learns progressively and is very sensitive to noisy data. It is also a very slow learner.

G. XG Boost

This is a powerful model for doing supervised regression. It stands for Extreme gradient boosting. Gradient boosting refers to a category of ensemble algorithms and ensemble is constructed from decision tree models. New trees are added one at a time to correct errors of previous models and hence is called boosting algorithm. They use differential loss function and gradient descent optimization for achieving high accuracy. It is computationally efficient and effective algorithm. It works well on small data, big data, data with subgroups and complicated data. However, the disadvantage of this algorithm is its black-box nature. It is more prone to overfitting than other bagging models.

VI. METHODOLOGY

Regression models help in predicting the continuous dependent variable. Thus, the proposed methodology will aid in direct prediction of the numerical values of the restaurant ratings when the input parameters are provided to the model. First, the raw data is collected from its source and then it is put through data cleaning process. Then the cleaned data is visualized to further explore the data for trends, outliers, and its overall composition. In feature selection, the independent variables relevant to the problem are identified. Here, columns like 'online order availability', 'table booking availability', 'restaurant type', 'location', 'approximate cost for two people', 'count of cuisines offered', 'number of items on menu' are chosen because these factors are easily controllable while setting up a new restaurant and predictions using these factors can offer a lot of scope for improving the ratings. All the data attributes are given equal weightage. The selected columns are finally processed by encoding, scaling, and splitting into train and test data. The train dataset is used to train different regression models and the models are evaluated based on certain metrics. Finally, prediction is made on the test dataset.

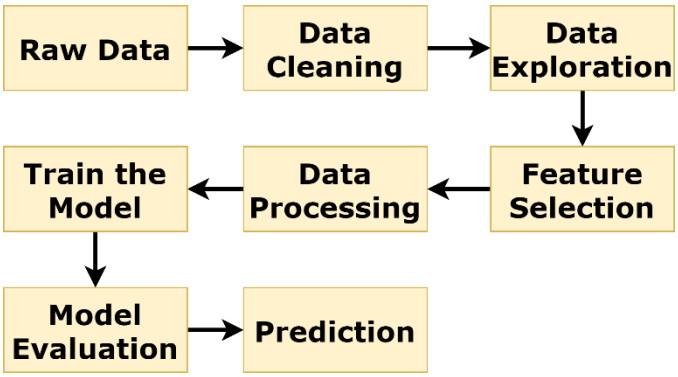

Fig. 9. Block schematic explaining the overall methodology

VII. RESULTS

Predictions are made on the test data and the following metrics are used to measure the performance of the model [8] namely root mean square error, mean absolute percentage error and R^2 score. According to [9],a model is considered good if its R^2 is greater than 0.9. And other metrics like root mean square error and mean absolute percentage error should be as low as possible. Models like Random Forest, XG Boost and ADA Boost perform exceptionally well.

TABLE I. REGRESSION MODELS RESULTS SUMMARY

Model	Root Mean Square Error	Mean Absolute Percentage Error	R^2 Score
Linear Regression	0.34	7.05	0.33
Support Vector Machine	0.23	6.28	0.37
Decision Tree	0.16	1.38	0.84
Random Forest	0.12	1.09	0.91
K Nearest Neighbours	0.15	1.22	0.86
ADA Boost	0.13	1.08	0.90
XG Boost	0.13	1.19	0.91

VIII. CONCLUSION

This paper studied different algorithms and analyzed different features to predict a restaurant's ratings. Predictions were made factors that could be easily controlled. Such an analysis was required to make sound business decisions and aided in planning before setting up a venture like a Restaurant.

Models Random Forest, ADA Boost, XG Boost gave low mean absolute percentage error and high R2 score. These can be used to successfully predict Restaurant's ratings in a reliable manner. However, if more data is made available, the predictions could be more accurate.

IX. REFERENCES

[1] A. Shibata, S. Kamei and K. Nakano, "Category-oriented Sentiment Polarity Dictionary for Rating Prediction of Japanese Hotels," *2020 Eighth International Symposium on Computing and Networking Workshops (CANDARW)*, 2020, pp. 440-444, doi: 10.1109/CANDARW51189.2020.00090.

[2] P. Chanwisitkul, A. Shahgholian and N. Mehandjiev, "The Reason Behind the Rating: Text Mining of Online Hotel Reviews," *2018 IEEE 20th Conference on Business Informatics (CBI)*, 2018, pp. 149-157, doi: 10.1109/CBI.2018.00025.

[3] L. Pradhan, C. Zhang and P. Chitrakar, "Multi-view Clustering in Collaborative Filtering Based Rating Prediction," *2016 IEEE Tenth International Conference on Semantic Computing (ICSC)*, 2016, pp. 250-253, doi: 10.1109/ICSC.2016.40.

[4] S. Saha and A. K. Santra, "Restaurant rating based on textual feedback," *2017 International conference on Microelectronic Devices, Circuits and Systems (ICMDCS)*, 2017, pp. 1-5, doi: 10.1109/ICMDCS.2017.8211542.

[5] D. Sivabalaselvamani and B. Soorya, "Convolution Neural Network based Specialized Restaurant Rating Using Facial Expression Detection," *2020 International Conference on Inventive Computation Technologies (ICICT)*, 2020, pp. 739-744, doi: 10.1109/ICICT48043.2020.9112518.

[6] I. F. Shihab, M. M. Oishi, S. Islam, K. Banik and H. Arif, "A Machine Learning Approach to Suggest Ideal Geographical Location for New Restaurant Establishment," *2018 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, 2018, pp. 1-5, doi: 10.1109/R10-HTC.2018.8629845.

[7] Himanshu Poddar. (2019, April). zomato.csv, [Version 1]. Retrieved June 5, 2021 from https://www.kaggle.com/himanshupoddar/zomato-bangalore-restaurants

[8] A. Botchkarev, "Evaluating performance of regression machine learning models using multiple error metrics in azure machine learning studio," *SSRN Electron. J.*, 2018.

[9] J. Fernando, "R-Squared," *Investopedia.com*, 13-Oct-2021. [Online]. Available: https://www.investopedia.com/terms/r/r-squared.asp. [Accessed: 18-Oct-2021].

[10] Rrubaa Panchendrarajan, Nazick Ahamed, Prakhash Sivakumar, Brunthavan Murugaiah, Surangika Ranathunga, and Akila Pemasiri, 2017, "Eatery: A Multi-Aspect Restaurant Rating System," *In Proceedings of the 28th ACM Conference on Hypertext and Social Media, (HT '17')*, Association for Computing Machinery, New York, NY, USA, 225–234, DOI:https://doi.org/10.1145/3078714.3078737

[11] 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.2017.816727

[12] R. M. Gomathi, P. Ajitha, G. H. S. Krishna and I. H. Pranay, "Restaurant Recommendation System for User Preference and Services Based on Rating and Amenities," *2019 International Conference on Computational Intelligence in Data Science (ICCIDS)*, 2019, pp. 1-6, doi: 10.1109/ICCIDS.2019.8862048.

[13] Q. Gan and Y. Yu, "Restaurant Rating: Industrial Standard and Word-of-Mouth -- A Text Mining and Multi-dimensional Sentiment Analysis," *2015 48th Hawaii International Conference on System Sciences*, 2015, pp. 1332-1340, doi: 10.1109/HICSS.2015.163.

[14] S. Saha and A. K. Santra, "Restaurant rating based on textual feedback," *2017 International conference on Microelectronic Devices, Circuits and Systems (ICMDCS)*, 2017, pp. 1-5, doi: 10.1109/ICMDCS.2017.8211542.

[15] Neha Joshi, " A Study on Customer Preference and Satisfaction towards Restaurant in Dehradun City", *Global Journal of Management and Business Research(2012)* ,

Link: ttps://pdfs.semanticscholar.org/fef5/88622c39ef76dd773fcad8bb5d233420a270.pdf

[16] 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

[17] Haoxiang, Wang, and S. Smys. "Big Data Analysis and Perturbation using Data Mining Algorithm." *Journal of Soft Computing Paradigm (JSCP)* 3, no. 01 (2021): 19-28, doi: 10.36548/jscp.2021.1.003

[18] Chen, Joy Iong-Zong, and Kong-Long Lai. "Deep Convolution Neural Network Model for Credit-Card Fraud Detection and Alert." *Journal of Artificial Intelligence 3*, no. 02 (2021): 101-112, doi: 10.36548/jaicn.2021.2.003

[19] Tripathi, Milan. "Sentiment Analysis of Nepali COVID19 Tweets Using NB, SVM AND LSTM." *Journal of Artificial Intelligence 3*, no. 03 (2021): 151-168, doi: 10.36548/jaicn.2021.3.001