Consumer Preference Analysis and Rating Prediction Model in the Restaurant Industry Based on Restaurant Information and Consumer Reviews

Li Qing(✉)

Surrey International Institute, Dongbei University of Finance and Economics, Dalian, China
2657919440@qq.com

Abstract. The restaurant industry has increasingly relied on the development of the Internet and mobile apps in recent years. Diner tends to make decisions based on restaurant information and customer reviews on apps, while merchant also focuses on customer reviews to improve the quality of service and attract more customers. It is noteworthy these customer reviews and ratings show that consumers pay more attention to restaurant attributes such as environment, food variety, parking condition and the need for reservation. These obvious tendencies can either serve as positive factors for restaurants, or directly result in consumer dissatisfaction and negative reviews. However, many review apps limit consumer to rating on a scale of 1–5 with a difference of 0.5, which not only restricts customers' ratings, but also affects the rating's authenticity. Therefore, this paper will firstly explore the influence of different factors on consumer ratings by using regression analysis to summarize consumer's preference and selection tendency. Secondly, it will compare the prediction results of consumer ratings by using regression model, decision tree model and random forest model. The result shows that random forest model can effectively predict consumer ratings, reduce rating errors while keeping MSE and R2 within a reasonable range, and reflect consumer's real attitude towards restaurants with greater accuracy.

Keywords: Restaurant · Consumer rating · Machine learning · Random Forest

1 Introduction

Machine learning is a computational process that enables durable alterations in behavior through training [1]. Machine improves its learning performance by organizing existing knowledge [2]. Machine learning and integration algorithms have long been used in the restaurant industry. Algorithms will recommend restaurants that are similar to consumer's preference based on their ratings and browsing history of restaurants [3]. This relies on consumer preferences and tendencies for restaurant attributes. The restaurant industry tends to use regression analysis to explore how different restaurant attributes

S. Tehseen et al. (Eds.): ICEDBC 2023, AEBMR 258, pp. 464–471, 2024.
https://doi.org/10.2991/978-94-6463-246-0_56

affect the average restaurant rating [4]. Under the condition that consumers are increasingly relying on ratings and reviews on review-based apps to make decisions, it is important to capture consumer preferences for merchants in the highly competitive business environment. However, in many review apps, consumers can only rate on a scale of 1–5 with a difference of 0.5 due to different rating requirements and limitations in the accuracy, which reduces the accuracy of consumer ratings. For example, some consumers may want to rate a 4.3 but have to give a 4.5 or a 4 because there is no option for a 4.3, which may not fully reflect the reviewer's true intentions and cause misunderstandings among other consumers and businesses. This article employs real restaurant ratings and reviews datasets to extract the key features of the restaurants and attempt to analyze how these key features affect the average restaurant rating. To achieve this, regression model, decision tree model, and random forest model are utilized to predict consumer ratings. Comparative analysis of the prediction results is conducted to identify a more accurate and suitable model for predicting consumer ratings in the restaurant industry. Therefore, this will not only make a more valuable reference for more consumers to choose restaurants, but also provide a reference basis for entrepreneurs to choose their investment direction. It will provide effective guidance for merchants to improve restaurant management as well, which in turn will enhance service quality and improve the dining environment [5].

2 Empirical Analysis

2.1 Dataset Used in the Study

Two datasets are used in this study. The first dataset is the basic information (including business ID, name, category, attributes, city, state, opening status, opening hours) and average rating of 500 restaurants. The other dataset is the reviews of these 500 restaurants, with a total of 26,013 reviews, including consumer rating, useful rating, funny rating, cool rating, text, and review time.

2.2 Investigating the Impact of Different Factors on the Average Restaurant Rating

First, the raw data situation is cleaned to remove outliers, and the basic situation of the five hundred restaurants is analyzed to explore the factors that affect the average restaurant rating. In terms of number of reviews, the maximum, minimum and average number of restaurant reviews are 875, 5 and 50.17 respectively, and the average restaurant rating is 3.756 stars. After calculating the number of restaurants with different star ratings and the number of reviews, it shows there are 114 four-star restaurants, which are the most. Only 8 Restaurants attain 1.5 stars, which are the least. In addition, the dataset of 500 restaurants does not include one-star restaurants. 3.5-star restaurants have the highest number of reviews, with an average of 72.97 reviews, followed by four-star restaurants, with 67.8 reviews on average. Five-star restaurants have the lowest number of reviews, with 20.92 reviews per restaurant. Continuing to examine the relationship between average ratings and average number of reviews, the line graph in Fig. 1 reflects a positive relationship between these two variables.

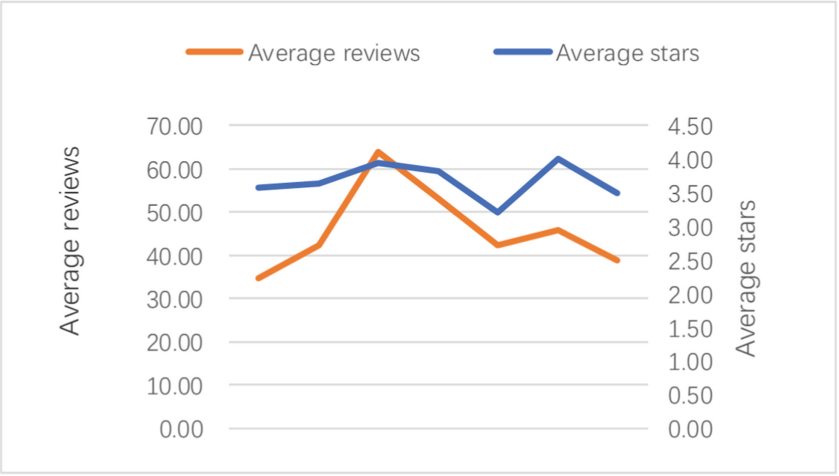

Fig. 1. Relationship between Average Stars and Average Reviews

2.3 Key Factors and Average Restaurant Ratings

Continuing to analyze the extent to how these key factors influence restaurant ratings. The dataset is filtered by using the variables of the access to WiFi, availability of outdoor seating, need for reservation, children friendly and pet allowed to calculate the number and average rating of the corresponding restaurants. According to Table 1, the results show consumer's preference for certain restaurant attributes. People prefer restaurants with outdoor seating as they can enjoy the scenery of street outside. To avoid crowds and lack of seats, consumer tend to reserve seats in advance, which will improve their dining experience. At the same time, restaurants that allow children and pets tend to be noisy and may disturb other diners, which explains their lower average score.

In addition, the study compares the average ratings of different categories of restaurants. The comparison reveals that there are three main types of restaurants. Bar usually

Table 1. Average Star Rating of Restaurants with Different Attributes

	Total number	**Average stars**
No need to reserve	172	3.73
Need to reserve	102	3.93
Indoor seating only	74	3.52
Outdoor seating	84	3.74
No pets allowed	38	3.87
Pets allowed	20	3.78
Kids unfriendly	39	3.88
Good for kids	134	3.65

Table 2. The Proportion of Different Keywords in Restaurants with Different Star Ratings

	5 stars	4 stars	3 stars	2 stars	1 star
WiFi	0.005	0.006	0.006	0.006	0.004
Outside	0.031	0.055	0.042	0.038	0.036
Parking	0.040	0.066	0.056	0.031	0.040
Kid	0.037	0.048	0.041	0.045	0.065
Appointment	0.039	0.032	0.038	0.067	0.075

has the highest rating, followed by the healthcare restaurant and finally the teahouse. Next it looks at the main types of cuisine, which are divided into the following four categories, with local cuisine being the highest rated, followed by Korean and Mexican cuisines, and finally Chinese cuisines. The average ratings are 4.06, 3.65, 3.38 and 3.21. Customers prefer bar restaurant with a good combination of drinks and snacks, which is a good place to relax after work. Local food restaurants are more attractive where people are more accustomed to the taste of local food and service.

2.4 Proportion of Keywords in Restaurant Reviews with Different Average Ratings

Based on the content of restaurant reviews, several keywords are artificially identified such as WiFi, dog, pet, appointment, reservation, child, kid, outside, parking and garage. Then combining similar keywords (e.g., pet and dog, kid and child) to calculate the proportion of these keywords in the reviews of restaurants with different average ratings. Overall, as Table 2 shows, five-star restaurants have the least mention of these keywords in their reviews, while four-star restaurants have the most mention of these keywords. In terms of different average star ratings, parking condition is more important to consumers in restaurants over 3 stars, while appointment is more important to consumers in one-star and two-star restaurants. Five-star restaurants have various kinds of food, better service, and more comfortable dining environment. Most five-star restaurants are similar on the aspects of dining environment and service facilities, so people tend to pay more attention to the food itself and the type of dishes than to the environment. In contrast, there are fewer one-star and two-star restaurants with poorer dining environment and service, which tend to be more crowded, so people will focus on appointment and seat.

2.5 Regression Analysis of Different Variables

The purpose of regression is to predict the target value by creating a set of regression equations between the dependent and independent variables [6]. Regression analysis constructs a function to fit a certain dataset and minimizes the error between them as much as possible [7]. The 500 pieces of restaurant data are matched to 26013 reviews according to restaurant business ID. Defining Y as the ratings of 26013 reviews, regression analysis is performed on different variables such as location, open status, useful

Table 3. Regression Analysis (y is for review rating)

	coef	P > |t|
std	−2.01e-05	0.000
Useful rating	−0.0260	0.000
Average rating	0.7420	0.000
Review count	0.0001	0.002
Dogs allowed: True	−0.0206	0.553
kid	−0.1711	0.000
appointment	−0.4534	0.000
outside	−0.0008	0.985
dog	−0.0161	0.612

ratings, average ratings, number of reviews, restaurant attributes (e.g., dogs allowed) and different keywords in the reviews (e.g., garage, dog). Table 3 shows that the R-squared value and Adj. R-squared value are 0.225 and 0.224 respectively, which are close. Variables like open status, standard deviation of review ratings std., useful ratings, pets allowed, and child friendly have negative correlation coefficients. Variables like standard deviation of review ratings std, useful ratings, average restaurant ratings, number of reviews, and reservations required have significant effects on review ratings. In addition, according to different keywords in the reviews, all of them have a significant effect on the review ratings except for the variables OUTSIDE and DOG. This is in line with the consumer tendency mentioned before. Restaurants that allow children and pets tend to be noisier, which may interfere with other diners. Therefore, variables like kid and pet have a negative correlation with restaurant ratings.

Similar influencing factors are combined (e.g., dog and pet, etc.) according to business ID and keywords. Defining Y as average ratings of 500 restaurants, regression analysis is performed on different variables such as the percentage of keywords in reviews (defined as dfsdog), open status, location, and restaurant attributes. As Table 4 shows, the R-squared value and the Adj. R-squared value for this regression analysis are 0.962 and 0.960 respectively. It can be concluded that variables such as percentage of keywords in reviews, standard deviation of review ratings std, and restaurant location have a significant effect on the average restaurant ratings.

2.6 Continuous Prediction Model

There is a difference of 0.5 points in each level of rating in the dataset. This rating criterion will lead to broad and general results, which will cause a large error in ratings. In order to know the real ratings of diners, the study uses linear regression model, decision tree model and random forest model respectively to make simulations and predictions on the diner rating result by adjusting their model parameters. To improve the accuracy of the prediction models, this study selects the factors that are most relevant to the review ratings [8]. Extracting x as variables like open status, average ratings, useful rating,

Table 4. Regression Analysis (y is for Average rating)

	coef	P> |t|
std	−0.0008	0.000
dfsgarage	−0.0008	0.000
dfswifi	−0.0008	0.000
dfskid	−0.0008	0.000
dfsappointment	−0.0008	0.000
dfsoutside	−0.0008	0.000
dfsdog	−0.0008	0.000

funny rating, cool rating, and the proportion of keywords in reviews, defining y as the review rating. The testing dataset is adjusted to account for 0.3 to obtain MSE and R2 values for the training and testing sets. Firstly, it uses linear regression model. Table 5 shows MSE values are over 1 while R2 values are close to 0, which means there is a large error and this model is not accurate.

Next, this study uses decision tree model. It adjusts the max_depth of the model to 15, min_samples_split to 5, and min_samples_leaf to 100. Table 6 shows its grid search time is 2.409 s. Compared with previous regression model, this model and its corresponding result exhibit a marked improvement, with MSE reduced to less than 1 and R2 gradually approaching 1. However, there is a considerable disparity remaining between the values of training and test datasets, which means the model still needs further improvement.

Finally, it uses a random forest model to make prediction. It is a typical example of ensemble algorithm [9]. A random forest is an oft-used ensemble technique that employs a forest of decision-tree classifiers on various sub-samples of the dataset, with random subsets of the features for node split [10]. It shows the grid search time is 203.6 s by adjusting the max_depth of the model to 15, min_samples_split to 5, and

Table 5. Linear Regression Model

	MSE	**R2**
Training dataset	1.5872	0.2520
Testing dataset	1.6849	0.2188

Table 6. Decision Tree Model

	MSE	**R2**
Training dataset	0.4522	0.7869
Testing dataset	0.8170	0.6212

Table 7. Random Forest Model

	MSE	**R2**
Training dataset	0.4592	0.7836
Testing dataset	0.6764	0.6864

min_samples_leaf to 175. Comparing the following data, according to Table 7, MSE values for these two sets are relatively low and close as well. It is obvious the MSE of the training set is closer to 0. Despite it has a longer grid search time, R2 values of these two sets are closer to 1, which indicates that the model is more suitable and can make more accurate predictions.

3 Conclusion

This paper firstly calculates the average ratings of restaurants based on the review dataset and the restaurant information dataset. Then it summarizes the common attributes of restaurants with different ratings and analyses the influence of different factors on restaurant ratings. Besides, it identifies some preferences of diners in choosing restaurants by using regression model and analyses potential reasons behind the trends. To reduce the variance between ratings and to develop a prediction model with a wide range of applicability, this paper uses different prediction models to compare their parameter values and prediction accuracy. It finally uses random forest model for continuous rating prediction, which improves the accuracy of consumer ratings. Compared with separate decision tree model, random forest model will have more accurate prediction results and is less prone to over-fitting phenomenon. It has better generalization ability due to the integration of multiple decision trees [11]. The random forest model reduces the MSE value of testing set to 0.6764 and improves the R2 value of testing set to 0.6864, making it the best performing forecasting model of the three models. Despite the improved prediction accuracy, the model reveals some disadvantages such as a longer runtime of 203.6 s to traverse all grids. In addition, the MSE and R2 values of the prediction model are still not optimal. Overall, the article focuses on the consumer concerns about the restaurants and improves the accuracy of the regression model in predicting consumer ratings of restaurants. Future research may be carried out in the following 1) Extracting and studying the impact of other factors on consumer ratings such as restaurant environment, types of cuisine, and quality of service, etc., 2) Focusing on the attitudes, emotions, praise, and advice that appear in consumer reviews, 3) Improving the accuracy of the prediction models, trying to reducing MSE and increasing R2 values within a reasonable range, and also reducing the time it takes to traverse the grid.

References

1. Shi, Z. (2019) Cognitive Machine Learning. International Journal of Intelligence Science, 9, 4, pp.111-121.

2. Jordan, M.I., Mitchell T.M. (2015) Machine learning: Trends, perspectives, and prospects. Science, 349, pp.255-260.
3. Mahajan, K., Joshi, V., Khedkar, M., Galani, J., Kulkarni, M. (2021) Restaurant Recommendation System using Machine Learning. International Journal of Advanced Trends in Computer Science and Engineering, 10, 3, pp.1671-1675.
4. Priya, J. (2020) Predicting Restaurant Rating using Machine Learning and Comparison of Regression Models. 2020 International Conference on Emerging Trends in information Technology and Engineering, pp.1–5.
5. Xia, Y., Ha, H. (2023) The role of online reviews in restaurant selections intentions: A latent growth modeling approach. International Journal of Hospitality Management.
6. Maulud, D.H., Abdulazeez, A.M. (2020) A Review on Linear Regression Comprehensive in Machine Learning. Journal of Applied Science and Technology Trends, 1, 4, pp.140-147.
7. Breiman, L. (1984) Classification and Regression Trees. Belmont, CA: Wadswprth International Group.
8. Blum, A.L., Langley, P. (1997) Selection of relevant features and examples in machine learning. Artificial Intelligence, 97, pp.245-271.
9. Dietterich, T.G. (2002) Ensemble Learning. The handbook of brain story and neural networks, 2, pp.110-125.
10. Sipper, M., Moore, J.H. (2021) Conservation machine learning: a case study of random forests. Scientific Reports, 11:3629.
11. Breiman, L. (2001) Random Forests. Machine Learning, 45, pp.5-32.