Advances in Economics, Business and Management Research, volume 53
7th International Conference on Education and Management (ICEM 2017)

How Low-rating Restaurants Crack Business

Bo Gao[1, a], Xinjian Qiang[1, b*] and Shuyu Chen[2, c]

[1]School of Computer Science, Xi'an Shiyou University, Xi'an Shaanxi, China

[2]Financial Shared Department, China General Nuclear Power Company, Shenzhen Guangdong, China

[a]gaobo@xsyu.edu.cn, [b]xasyu@126.com, [c]shuyuchen@163.com

Keywords: Multinomial logistic regression; Random forest; Rating prediction; Restaurant; EDA

Abstract. This work is focusing on "Yelp" businesses. The main idea is to analyze the data from the Yelp web site. These two methods are used for grade prediction and feature selection: multinomial logistic regression and random forest. In conclusion, the results of the two methods are mostly the same with acceptable difference. Although reviews are too sophisticated to generate, some trials are included in the very last section for interest because it is a hot topic these days and it is indeed very effective for rating prediction.

Introduction

As we all know, everyone can register an account on Yelp website so that he or she could search for restaurants corresponding to their preferences, either American or Chinese, close to you or not, pick up or delivery. This is a very convenient application, and the customer can help improve the business rate, they can also publish their reviews, special restaurants. The rate data will be set in the restaurant configuration file, which can be referred directly to the data when he or she chooses. It is clear that a relatively high rating may be more attractive and a greater contribution to the business of the owners.

Objective

This paper aims to build up solutions to help restaurants that with low ratings to improve their ratings on "Yelp" based on reviews and business features like opening hours, noise Level and parking space or Wi-Fi. We choose to implement two different methods, multinomial logistic regression and random forest, to obtain the most important attributions that affect the business' ratings [1, 2]. Therefore, the goal is to find out whether there are some crucial attributions that could make one restaurant obtain a higher rate or better reviews.

Data Source

We collected our data from the Yelp Dataset Challenge. The basic dataset contains 1.6 gigabytes reviews and 500,000 tips by 366,000 users for 6100 businesses, 481,000 business attributes, social network of 366,000 users and aggregated check-ins over time for each of the 61,000 businesses for different cities in four countries: Canada, Germany, the United Kingdom and the United States.

As shown in the Data graph, "Business", "Review", "Tip", "User" and "Check-in" are the five main parts in this "Yelp dataset". For instance, the "Business" sector contains information of each restaurant's full address, city, state, latitude, longitude, stars, categories, etc. The "Check-in" sector memories the place where customers have been to and the corresponding time for this check-in. For "reviews", we have a user id refers to one customer. His or her reviews in text as well as rating in stars, date, votes are also included for different business ID.

In this project, we choose to focus on "Business" and "Reviews" sectors, and then implement analysis on the attributions as well as reviews of each restaurant.

Primary Analysis

EDA. Before we start to construct a model for the dataset, our group analyzes the EDA at the first stage. In order to clean our data, we analyze the average rating between states to see whether there is a difference of average rating between states. Two-sample t-test is involved to test this difference and we choose Illinois and Pennsylvania states as an example pair in this case.

We can do box-plot as well as histograms for the data to see how ratings distribute among these many restaurants.

Methodology. We have a rather large dataset (1625×72) and relatively numerous features (70), a majority of which is categorical variables and the predictors classified. In view of the properties of our dataset and our objective, multinomial logistic regression is an appropriate method to implement because it not only contains a variable selection process that can screen out the most important features for the future use, but it also can determine the coefficients of the selected features, which can show the importance of each variable.

In order to secure our result, we also need to confirm our result by applying another method (sensitivity analysis) [3]. According to the "Law-of-parsimony" which an explanation of a thing or event is made with the fewest possible assumptions. Therefore, nonparametric methodologies are an ideal choice because they are very flexible with fewer assumptions. In this case, random forest would be a powerful tool that can be quick and simple to fit computationally. Even for the large dataset, it can handle category predictor naturally. What is more important, random forest also includes variable selection process, since it could detect important variables and handle the noise very well during the calculation process.

After nail down the methods, we try to divide ratings into nine categories, which are 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5. However, the result of accuracy tends to be not ideal enough due to the big variety of categories for both two methods. Therefore, we consider merging categories into only three categories, which are 'Low-rating' (rating score<=2.5), 'medium-rating' (rating score=3, or 3.5) and 'high-rating' (rating score>=4). By using the predictors selected from the "nine categories", it is not surprising that the accuracy of the merger has indeed increased. In the next section, we will analyze the process details of these two methods.

Process

Multinomial Logistic Regression.

Assumption. There are two main assumptions:

- Each independent variable has a single value for each case.
- If the multinomial logistic regression is used to model choices, the odds of preferring one class before another don not depend on the presence or absence of other "irrelevant" alternatives.

Algorithm. The multinomial logistic regression is just an extension of binary logistic regression [4]. We model the logarithm of the probability of seeing a given output using the linear predictor as well as an additional normalization factor. If we have k classes, then the mode is as follows Eq.1:

$$\ln\Pr(\gamma i=1) = \beta_1 X_i - \ln Z$$
$$\ln\Pr(\gamma i=2) = \beta_2 X_i - \ln Z$$
$$\cdots\cdots \quad (1)$$
$$\ln\Pr(\gamma i=k) = \beta_k X_i - \ln Z$$

Since

$$1=\sum_{k=1}^{k}\Pr(Y_i=k)=\sum_{k=1}^{k}\frac{1}{z_i}e^{\beta_k x_i}=\frac{1}{z}\sum_{k=1}^{k}e^{\beta_k x_i} \quad (2)$$

Therefore

$$\Pr(Y_i = 1) = \frac{e^{\beta_1 x_i}}{\sum_{k=1}^{k} e^{\beta_k x_i}}$$

$$\Pr(Y_i = 2) = \frac{e^{\beta_2 x_i}}{\sum_{k=1}^{k} e^{\beta_k x_i}}$$

……

$$\Pr(Y_i = \mathrm{k}) = \frac{e^{\beta_k x_i}}{\sum_{k=1}^{k} e^{\beta_k x_i}} \quad (3)$$

We add an L1-norm regularized term (LASSO) to the original objective function. There are two main advantages of LASSO. On the one hand, when we use LASSO, we lessen much model variance by just adding a little bias. On the other hand, LASSO give us sparse solution so that we can do feature selection to improve the model accuracy.

Procedure. Firstly, we select 80% of the data as the training set and the rest 20% as the testing set. Then, we use L1-norm regularized (LASSO) multinomial logistic regression to delete redundant variable.

Specifically, the number of non-zero coefficients of features varies with the value change of lambda.

So, we use cross validation to select the best lambda (lambda=0.03167592), whose average training error rate is lowest.

Then, we select nine predictors, whose coefficients are non-zero, out of 79 for our future research.

In the end, we use 9 predictors and our re-grouped 3 classes to train our model and analyze our results.

Random Forest.

Assumption. Random forest is a non-parametric method, so it does not have any specific assumptions about distribution of variables and predictor [5].

Algorithm. Random forests is an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision-trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

Procedure. First, we construct 9 classes of random forest models on the whole data set to calculate the average bit error rate. Then, to improve the model accuracy, we use bootstrap strategy to estimate the random forest error rate. Specifically, we do 100 bootstrapping, every time, 200 items randomly selected from the whole data set as test data sets, and the rest 1425 used as testing dataset to calculate the testing error.

Secondly, In order to increase the accuracy of classification, we divided the data into three categories, which is the same as procedure of part (multinomial logistic regression). Here we do 100 times bootstrap, each time 200 items randomly selected as the testing dataset and the rest as the training dataset to get the mean testing error.

Finally, we select important variables that computed using the mean decrease in Gini index.

Results

EDA. It is useful to test whether there is a difference in the average rating between states.The two-sample t-test shows that the difference between states is truly not equal to 0. For instance, we selected the commentary between the density map state IL and the PA. It is obvious that the densities of two states are significantly different especially at review rate 3.5 and 4. By two-sample t-test, 95 percent confidence interval of mean difference is [-0.0678, -0.0420], which proves that our assumption is right. Therefore, this is why we decide to focus the target on the "Business" and "Reviews" of state Illinois

only. The majority of restaurants have the rate of 3.5 or 4, and very small amount of restaurants have the rate of 1, 1.5 or 5.

According to the test data, about 75% of restaurants are less than 500 for restaurants. Only several restaurants could have more than 20,000 rates.

Multinomial Logistic Regression. First, we select 9 features as a predictor. The characteristics of the selection are as follows in Table 1.

Table 1 9 features of the selection

Alcoholfull_bar(v1)	Ambience.intimate(v2)	Ambience.divey(v3)
Ambience.casual(v4)	Good.for.Kids(v5)	Good.For.lunch(v6)
Good.For.dinner(v7)	Parking.street(v8)	Ambience.intimate(v9)

Then, after re-group our data into 3 classes, we get 3 models for 3 classes, separately.

For high-rating:

-0.023-0.544*v1+0.705*v2+0.065*v4+0.262*v5+0.089*v7+0.414*v8

For medium-rating:

-0.120-0.001*v3-0.111*v4-0.324*v5-0.226*v6-0.507*v7-0.836*v8

For low-rating:

0.143+0.250*v1+0.238*v3+0.272*v6+0.186*v9

Finally, we get the testing accuracy as 51.38%.

Random Forest. Firstly, we divided the rating of restaurants into 9 categories. The mean error rate using random forest is 69.38%, which is relatively high. Then, the mean testing error is 66%. Secondly, we divided ratings into 3 categories: the first category is those whose ratings are less than or equal to 2.5 (there are 211 items), the second category is those whose ratings equal to 3 or 3.5 (there are 751 items), and the third category is those whose ratings are equal to or greater than 4 (there are 663 items). The mean error rate is 49.303% and the mean testing error is 49.175%. There is an obvious decrease of the error rate, which declares the cut-down of category is feasible.

To compare which variable are more important to our model, we can measure variables' importance using mean decrease accuracy. From the plot above, we can visualize the variables' importance. Obviously, the higher a bar, the more importance the according variable has to our model. Further, we can quantify importance index of the first 9 variables as list below in Table 2.

Table 2 9 variables of the selection

Alcoholfull_bar(v1)	Ambience.intimate(v2)	Ambience.divey(v3)
Ambience.casual(v4)	Good.for.Kids(v5)	Good.For.lunch(v6)
Good.For.dinner(v7)	Parking.street(v8)	Ambience.intimate(v9)

Summary

From the high-rating model of multinomial logistic regression, we can see 5 features ('Ambience.intimate', 'Ambience.casual', 'Good.for.Kids', 'Good.For.dinner',and 'Parking.street') positively affect the rating score. According to the value of coefficients of these features, we can order their importance towards the rating: 'Ambience.intimate' > 'Parking.street' > 'Good.for.Kids' > 'Good.For.dinner' > 'Ambience.casual'. From the random forest model, we can select the first 5 variable that are the most important: 'Alcohol' > 'Drive.Thru' > 'Good.for.Kids' > 'Accepts.Credit.Cards' > 'Parking.street'. Although the two groups of important variables are not totally the same, the difference is acceptable. Therefore, we may combine the two groups of important variables and give restaurant guidance, which is to improve the restaurant rating on yelp, it would be better for restaurant to provide alcohol, to create an intimate (casual) ambience, to make it good for kids, to accept payment by credit card, etc..

Further Discussion

Apart from the business attributes, Yelp online reviews are invaluable source of information for users to choose where to visit or what to eat among numerous available options and for business to dig what costumers' focus. Due to overwhelming number of reviews, it is important to organize the data and select out the key words among them. In this part, we analyze the consumer review and create bag of words from all raw text reviews. To further creating a feature vector for prediction model mentioned in previous part, we then apply several ways such as Part-of-Speech analysis and Mutual information to extract the top concerns of consumers from reviews.

According to the Natural Language Processing of reviews, reviews from users may give us more information about other facts that could affect the rating. Based on the selected feature vector, we can conduct further parametric and non-parametric prediction analysis.

In the section above, we discussed performance of multinomial logistic model and random forest model respectively and both of them perform well on business rating prediction problem. However, a more specific problem Yelp would consider of that, how to recommend restaurants to users based on their historical ratings. In such a case, only using business attributes to do recommendation would not be sufficient because different users may have diametrically opposed ratings on one restaurant. We should pay more attention to users' underlining preference and try to build up its relationship with business attributes.

In this section, we discuss another important problem not mentioned in our project-recommendation system, which is one of the most popular problems in machine learning area. The main advantage of this model is, compared to predicting ratings by looking at potential significant business attributes only, NMF recommendation model abstracts topics from both users and restaurants, then uses these topics to weight users' choices when predict their preference on restaurants, resulting in a "user-custom" prediction.

However, one obviously drawback of this model is that, all the 30 feature are "latent"; in another word, although we could perform a good recommendation, it would be very hard to know what these features are. This makes the model harder to interpret than the classification model or logistic model.

References

[1] Yu, Oya, K. Kanamori, and H. Ohwada, Recursive Ensemble Land Cover Classification with Little Training Data and Many Classes, in: Intelligent Information and Database Systems, Springer Berlin Heidelberg, 2016.

[2] Rusu V, Singerman E. : Twenty-Fourth International Florida Artificial Intelligence Research Society Conference (Palm Beach, Florida, USA, May 18-20, 2011).Vol.1999, p.206.

[3] Ledolter, Johannes, Multinomial Logistic Regression, in: Data Mining and Business Analytics with R. John Wiley & Sons, Inc., 2013, p.132.

[4] Almashraee, Mohammed, Feature Extraction Based on Semantic Sentiment Analysis, in: Business Information Systems Workshops, Springer Berlin Heidelberg, 2013, pp. 270-277.

[5] Rodriguez-Galiano V F, Ghimire B, Rogan J, et al, An assessment of the effectiveness of a random forest classifier for land-cover classification, in:Isprs Journal of Photogrammetry & Remote Sensing, Vol. 67 (2012) No.1, p.93.