MA 678

Name: He Guo

Date: 12/05/2019

Yelp Project

**Introduction:**

Yelp is a famous business review website in America. The business information they offered has consisted of restaurants, shopping malls and hotels, etc. Users can comment on some businesses they have visited before. They also can give stars to those businesses and communicate their experience with other users. There is a data challenge page on Yelp, and it offer real data from Yelp.

This study is interested in business files and Review file. The business file contains information about businesses. It includes the unique business id for each business, and their GPS information, the stars information, etc. The photo file contains the photo id and their corresponding business id and label. Business id indicates which businesses are contained in this photo file. The label explains the photo condition. The review file has the review information of businesses and restaurants from users.

In the business file, the current study finds out that there is one data file store open and close time for each business. In this study, we fit a model to predict the probability of businesses having a high star. Next, this study focuses on information from American restaurants. By building a model to predicted American restaurants having a high star rate given information in American restaurants data file.

Users love to get some useful reviews on Yelp. Since they can use those useful reviews to help them make some choices. Yelp also wants to have more useful reviews. Since they want to attract more users to make more profit. The study wants to build a model, which allows yelp to guide the user to write a useful review. The second file has been used in this study which is a Review file. In the Review file, this study is designed to detect the odds of getting useful information.

**Methods and Material:**

To analysis business files. This study read the JSON file into R studio. There are 17 variables in business and two data frames, which are open hours for business and attributes for each business. Data frame is a table in which each column contains values of one variable and each row contains one set of values from each variable. This study uses a data frame called hours to get the open hours information. There are seven variables in hours, which are the business open time and close time for each day in a week. The data type for each variable is character. This study changes the data type into numeric. Using close time subtract open time to gets the business hours. There are a lot of NA in this data frame. Since this study is more interesting in using the business hour to predict the stars for business. The study deletes NA in hours data frame. According Figure 1, there is multicollinearity between business hours for each day in a week. To avoid multicollinearity, this study computes the mean of business hour. The mean of the business hour cannot be zero, then the study rescales the mean of the business hour. Instead of using the mean of business hour, this study uses z-score of mean of business hour.

Each business has a unique business id. If some business ids have the same name, this study creates a new column that stores how many branches for corresponding business. For example, Starbucks has 1066 branches in this business data file. This study includes an individual-level predictor to predict the star for each store. The business\_id variable represents each unique business, and state variable indicated which state they are located in. We need to have a binary variable to be the respond variable in the model. The study creates an indicator variable called is.High. When a business has 4 or more than 4 stars, the indicator for this business equals to 1. If a business has 3 or less than 3 stars, the indicator for this business equals to 0.

To build this model, this study creates a data frame called Data.star that are consisted of business id variable, state information, means of business hour, numbers of branches, z-score for means of business hour, and the indicator variable. This study builds a multi-level random slope logistic model to predict the probability of getting 4 or more than 4 stars. Is.High is predicted variable, z-score for means of business hour and numbers of branches are predictors. State is an individual-level predictor. In this study, we called this model as Stars.mod. This study splits Data.star into two parts. The first part is training data called star.training, the second part is testing data called star.testing. This study uses star.training to fit the multi-level random intercept logistic model and uses star.testing to test the multi-level random intercept logistic mode. There are some states only has less than 4 business in this data file. It causes that businesses from state A are contained in the training file and does include in the testing file. For example, one business in NY from testing data does not exist in training data. Then, the study deletes those rows which contain those state. After fitting the model, the study uses the confusion matrix to do model checking. Model-checking is a method for verifying the model.

The second model is designed to predict the probability of American restaurants having a high star. The study uses left join business data frame and photo data frame by business id. We called the new data frame as business. If a business id does not contain in the photo data frame, the value of the photo id variable for this business id will be NA. This study creates a has.photo variable. When photo\_id equals to NA, the corresponding columns in has photo will be 0. Otherwise, it will be 1. In business data frame, the study creates an indicator variable called is.High. When a business has 4 or more than 4 stars, the indicator for this business equals to 1. If a business has 3 or less than 3 stars, the indicator for this business equals to 0. There is a variable in business file called categories, it contains the categories of each business. This study made a subset of business information, which has American restaurants contained in its categories. This study fits the multi-level random intercept logistic model, is.High is the predicted variable and review\_count (how many reviews for this business), has.photo (1 denotes the business has a photo on Yelp, 0 denotes the business have a photo on Yelp) and Branch (number of branches for each business) are predictors. The individual-level predictor is State. Then the study builds a multi-level random intercept logistic model to predict the probability of American restaurants getting 4 or more than 4 stars. For model checking, the study did the same step which is described in the previous paragraph.

The third model is using data from the review file. This study using text mining to analysis reviews from users in Yelp. The study wants to know which phrase people are more like to post on their review. People post different comments for business with different stars. The current study divides the review data file into two data frames. The first data frame is called top.samp. It contains business with more than 3 stars. The second part is called low.samp, it contains business that did not have more than 3 stars. This study uses text variables in a review file and low.samp. Then, examining pairs of two consecutive words, called bigrams. We pick the most frequent phase in top.samp and create a variable that contains the number of matches in the text variable. The new variable called include.freq. For example, if the phase occurs in text review of business id A for once, the include.freq equals to 1. The current study builds a model to predict the number of useful. Useful in the review data frame means how many people think this review is useful. We use stars (how many stars of this review), cool (how many people think this review is cool), funny (how many people think this model is funny) and has.feq (how many high-frequency phase this review has) as predictors. The individual-level predictor is state.

**Result:**

According to Figure 2, Star.mod is multilevel logistic regression and this model is the varied intercept. The multilevel logistic model is a generalization of logistic regression where intercepts allow vary by group. The individual-level predictor in this model is state. The reason for choosing multilevel logistic regression is that stars of each business is varying by state. Then, the study uses multilevel logistic regression to predict the stars of business is high or not. We can look at the estimated model averaging over states, which is called fixed effect. In figure 1, it shows that the fix effect logistic regression is

Pr ( is.High = 1) = logit^-1(-0.07526 -0.168\*Mean.Hours.z -0.0140\*Num\_branch) .

The interpretation for the intercept is that when a business has no brunch and the mean business hour at the average level. The probability of getting a high star is logit^-1(-0.07526). The coefficient for Mean.Hours.z is -0.168. Dividing by 4 gets 1% difference in z-score of mean business hour can have a predicted 4% difference in probability having high stars. The coefficient for Mean.Hours.z is -0.0140. Dividing by 4 gets 1% difference in the number of branches can have a predicted 0.35% difference in probability having high stars.

In order to get the regression coefficients within each state. The Figure 3 shows that the regression for state AZ is  Pr ( is.High = 1) = logit^-1(0.15722-0.168\*Mean.Hours.z -0.0140\*Num\_branch), and Pr ( is.High = 1) = logit^-1(-0.0804-0.168\*Mean.Hours.z -0.0140\*Num\_branch)

in IL, etc…. The study uses the interpretation of the logistic regression in AZ as an example. The interpretation for the intercept is that when a business has no brunch and the mean business hour at the average level. The probability of getting a high star is logit^-1(0.15722). The coefficient for Mean.Hours.z is -0.168. Dividing by 4 gets 1% difference in z-score of mean business hour can have a predicted 4% difference in probability having a high star. The coefficient for Mean.Hours.z is -0.0140. Dividing by 4 gets 1% difference in the number of branches can have a predicted 0.35% difference in probability having a high star.

Figure 4 shows that the confusion matrix for this model. This confusion matrix summarizes the result of testing the model. In this confusion matrix, there are 21795 true negative (the model correctly predicts the number of the low star) and 1017 true positive (the model correctly predicts the number of the high star)

According to Figure 5, mod.A is multilevel logistic regression and this model is the varied intercept. The individual-level predictor in this model is state. The reason for choosing multilevel logistic regression is stars of each American restaurant is varying by state. Then, the study uses multilevel logistic regression to predict the stars of business is high or not. We can look at the estimated model averaging over states, which is called fixed effect. In figure 5, it shows that the fix effect logistic regression is Pr (is.High = 1) = logit^-1(-0.8072-0.00306\*Branch+0.00125\*review\_count+0.5535\*has.photo) . The interpretation for the intercept is that when an American restaurant has no brunch on review and did have a photo on Yelp. The probability of getting a high star is logit^-1(-0.8072). The coefficient for Branch is 0.00306. Dividing by 4 get 1% difference in the number of branches can have a roughly predicted 0.07% difference in probability having a high star. The coefficient for review\_count is 0.00125. Dividing by 4 get 1% difference in the number of branches can have a roughly predicted 0.03% difference in probability having a high star. The coefficient for has.photo is 0.5535. Dividing by 4 get a rough estimation that restaurants without a photo on Yelp were 13.84% more likely to have a high star.

To get the regression coefficients within each state. The Figure 6 shows that the regression for state AZ is Pr(is.High = 1) = logit^-1(-0.419-0.00306\*Branch+0.00125\*review\_count+0.5535\*has.photo), and Pr ( is.High = 1) = logit^-1(-0.6474-0.00306\*Branch+0.00125\*review\_count+0.5535\*has.photo) in IL, etc. The study uses the interpretation of logistic regression in AZ as an example. The interpretation for the intercept is that when an American restaurant has no brunch on review and did have a photo on Yelp. The probability of getting a high star is logit^-1(-0.419). The coefficient for Branch is 0.00306. Dividing by 4 get 1% difference in the number of branches can have a roughly predicted 0.07% difference in probability having a high star. The coefficient for review\_count is 0.00125. Dividing by 4 get 1% difference in the number of branches can have a roughly predicted 0.03% difference in probability having a high star. The coefficient for has.photo is 0.5535. Dividing by 4 get a rough estimation that restaurants in AZ without a photo on Yelp were 13.84% more likely to have a high star.

In Figure 7 shows that the confusion matrix for this model. This confusion matrix summarizes the result of testing the model. In this confusion matrix, there are 13016 true negative (the model correctly predicts the number of low star) and 14476 true positives (the model correctly predicts the number of high star).

According to Figure 8, useful is multilevel linear regression and this model is the varied intercept. The individual-level predictor in this model is state. The reason for choosing multilevel linear regression is stars of each business, which has been reviewed, is varying by state. Then, the study uses multilevel linear regression to predict the use of reviews. We can look at the estimated model averaging over states, which is called fixed effect. Useful = 0.813865 + 0.4518 \* funny + 0.7970 \*cool + 0.2915 \* num\_phrase. The interpretation for the intercept is the useful number of the review is 0.813865 if the review did not include any high-frequency phrase, other users didn’t think this review is funny and cool. If one more user thinks this review is funny, the useful number will increase 0.4518 units, and holding other variables as constant. If one more user thinks this review is cool, the useful number will increase 0.7970 units, and holding other variables as constant. If the review includes one more high-frequency phrase, a useful number will increase by 0.2915 units, and holding other variables as constant.

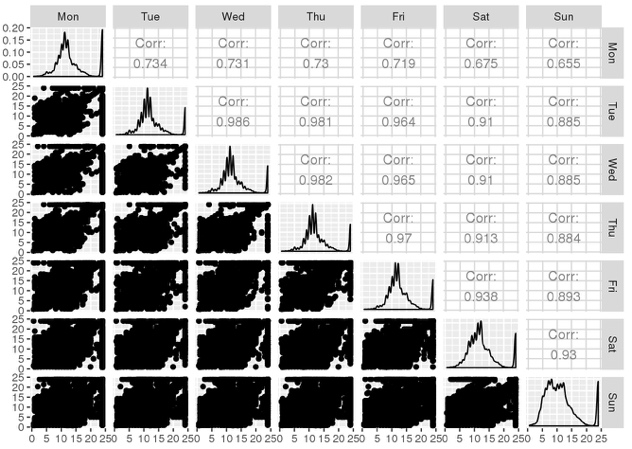
To get the regression coefficients within each state. The Figure 9 shows that the regression for state AZ is Useful = 0.61647 + 0.4518 \* funny + 0.7970 \*cool + 0.2915 \* num\_phrase and Useful = 0.6398 + 0.4518 \* funny + 0.7970 \*cool + 0.2915 \* num\_phrase in IL, etc. The study uses the interpretation of logistic regression in AZ as an example. The interpretation for the intercept is the useful number of the review is 0.61647 if the review did not include any high-frequency phrase, other users didn’t think this review is funny and cool. If one more user thinks this review is funny, the useful number will increase 0.4518 units, and holding other variables as constant. If one more user think this review is cool, the useful number will increase 0.7970 units, and holding other variables as constant. If the review includes one more high-frequency phrase, the useful number will increase by 0.2915 units, and holding other variables as constant.

According to the figure 10. The root of mean square error is 2.1438. The mean absolute error is 1.047.

**Discussion:**

According to the model Stars.mod, if a business wants to have a higher star, they need to reduce their branches and their average business hours. The binned residual plot of this model has some outlier, so the model doesn't fit very well. For further research, we need to find more valuable information as predictors. Based on the summary of Mod.A (the model of predicting the probability of American restaurants having a high star). American restaurants need to reduce their branches and attract users to post more reviews. At the same time, they also need to post some pictures of restaurants on Yelp. By analyzing how to write a useful review, Yelp needs to encourage users to write more funny and cool reviews to have more useful reviews. Including some high-frequency phrases also is an important method to let reviews become more useful.

Appendix

(Figure 1)

A screenshot of a cell phone

Description automatically generated(Figure 2)

A close up of text on a white background

Description automatically generated(Figure 3)

A picture containing object

Description automatically generated(Figure 4)

A close up of a newspaper

Description automatically generated(Figure 5)

A close up of text on a white background

Description automatically generated(Figure 6)

A close up of a clock

Description automatically generated(Figure 7)

A close up of text on a black background

Description automatically generated(Figure 8)A close up of text on a white background

Description automatically generated(Figure 9)

A screenshot of a cell phone

Description automatically generated(Figure 10)

Reference

Gelman, A., & Hill, J. (2018). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.