MA678 project Report

Jinzhe Zhang

Abstract

For this project, I converted the Yelp data Jpson data set by using Panda package in Python, cleaning and organizing the data in R. My data set was loaded from the Yelp academic data set , the subset of data contain 5775 restaurants with 13 business attributes, which is used to figure out what factors influence the rating and tried to predict the rating by review text. After that I made some EDA and several models.

Introduction

The word_frequency.csv contain the words that are frequently used when reviewing a Japanese Restaurant. The rating.csv contain the information of 5775 restaurants with some attributes of each of them. The meaning of business attributes can be interpreted very intuitively.

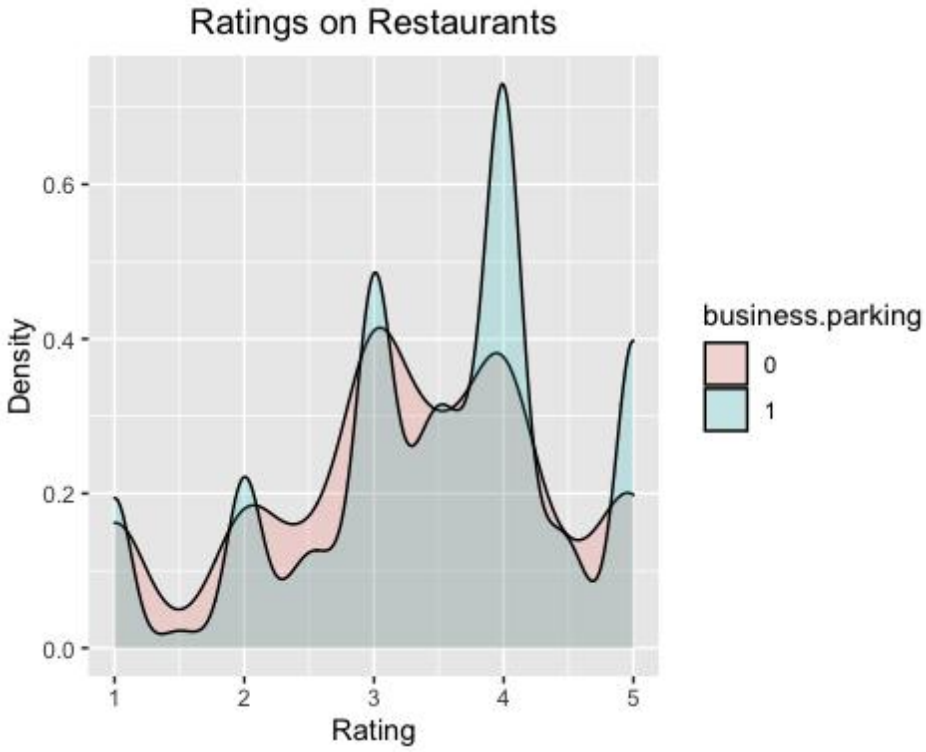
1.Due to the original Yelp data Jpson files are too large to upload, I can only upload the link of downloading path of Yelp data files and the .csv files which are used to EDA and Model. I randomly took the 5775 restaurants from the Yelp files and and extracted the attributes of each of them. 2.I tried to do the EDA part in many different angels such as price range and Noise level. 3.I made a text analysis to show the hot words of reviewing the Japanese restaurants. 4.I fit a linear regression and logistic regression model to find which factors are more crucial to get the higher rating. 5.The validation of the model

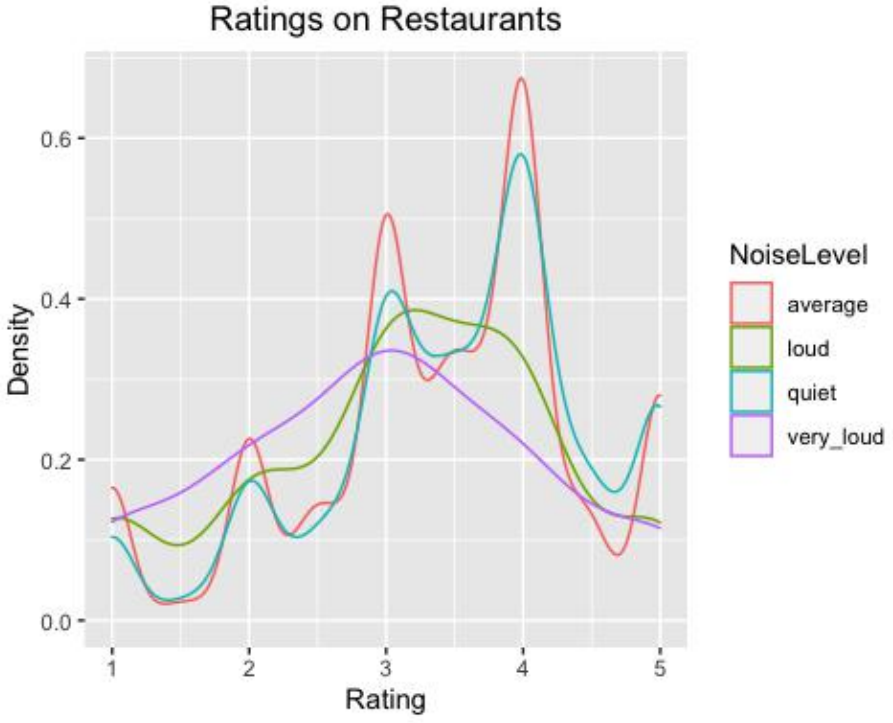
Restaurants EDA

First of all, we are interested in the ratings for these restaurants. I used the ggplot to should the distribution of those ratings for the restaurants.

Next, the distribution of price ranges for these restaurants are shown below,

Then, what happens if we divided these restaurants into different groups? We plan to divide them into disparate groups based on free parking and the noise level


Ratings on Restaurants
Density
0.0
0.2
0.4
0.6
1
2
3
4
5
Rating
business.parking
0
1


Ratings on Restaurants
Density
0.0
0.2
0.4
0.6
1
2
3
4
5
Rating
NoiseLevel
average
loud
quiet
very_loud

Text analysis

We also investigate what words are frequently used when reviewing a Japanese Restaurant?

Of course, we expected SUSHI and SASHIMI. However, no RAMEN! It surprises me.

Model analysis & Method

I want to investigate the factors affecting the rating of restaurants. We have totally 5775 different restaurants, but we have to delete 2016 observations due to the lack of essential variables.When I tried to use the logistic regression, I re-scaled the rating to the (0,1) by dividing rating by 10. We select our model using AIC and Adjusted R-squared as criteria.

Result

Based on our regression result, we can see the model did a bad job on fitting and predicting by summarizing the fitting information. The Adjusted R-squared is much lower than the expectation. However, some sign of coefficients follow our intuition, I find that the parking and noise level may be a more critical factor.

The interpretation of linear regression： The intercept shows that the average of the rating is about the 3.3 which aligns my EDA of rating distribution. On average, the restaurant with business parking will have a 0.137 higher rating. The restaurants with higher noise levels may get a rating penalty, while the quiet restaurants will win a rating bonus. However, the restaurant's rating is not determined by its price range because its coefficient is not significant. In summary, we can conclude that if a restaurant wants to get a higher rating, it should provide a quiet place and free business parking. Besides, it seems that the prices will not influence the rating of a restaurant too much.

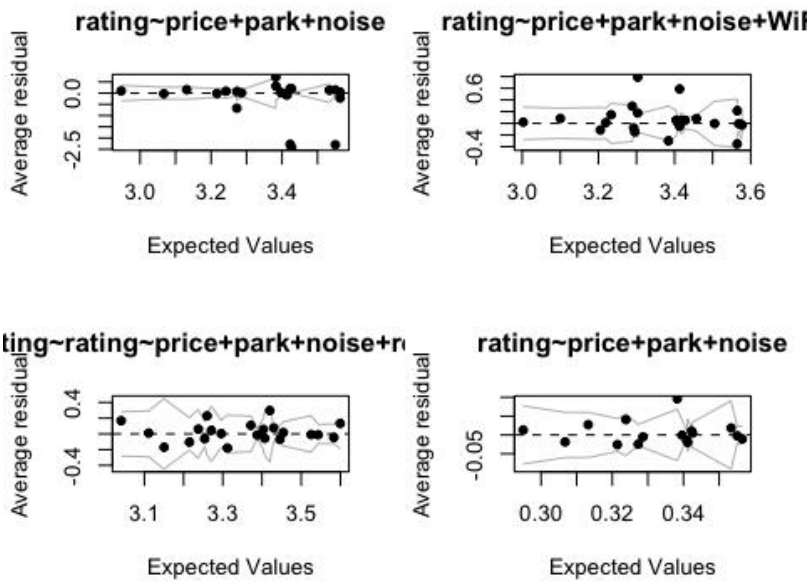
The interpretation of logistic regression: To use the logistic regression, I convert the data of rating into (0.1-0.5) while it is original on 1-5 scale. By looking at the information of the fitted model the result is similar with the linear regression.The coefficients of parking and NoiseLevelloud quiet show the positive response to the rating while the restaurantsPriceRange2 does not has significant influence on the rating.

Discussion:

Although the model do not perform good, the sign of the coefficients align our EDA. To improve the model, I may need to include more categories of the business in the data set.

Appendix

The residual plot



The summary of fitted model

```
##
## Call:
## lm(formula = rating ~ RestaurantsPriceRange2 + business.parking +
##     NoiseLevel, data = rating)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## −2.5638 −0.4549  0.1041  0.5906  2.0752
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               3.29942    0.05969  55.277  < 2e−16 ***
## RestaurantsPriceRange2 −0.01347    0.02827  −0.477 0.633730
## business.parking          0.13692    0.04046   3.384 0.000721 ***
## NoiseLevelloud           −0.19382    0.06606  −2.934 0.003368 **
## NoiseLevelquiet           0.14093    0.03969   3.551 0.000389 ***
## NoiseLevelvery_loud      −0.36115    0.12588  −2.869 0.004142 **
## −−−
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.014 on 3753 degrees of freedom
##   (2016 observations deleted due to missingness)
## Multiple R−squared:  0.01215,    Adjusted R−squared:  0.01084
## F−statistic: 9.234 on 5 and 3753 DF,  p−value: 9.497e−09

##
## Call:
## lm(formula = rating ~ RestaurantsPriceRange2 + business.parking +
##     NoiseLevel + WiFi, data = rating)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## −2.5734 −0.4585  0.1171  0.5909  2.0501
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               3.308220   0.070878  46.675  < 2e−16 ***
## RestaurantsPriceRange2 −0.009094   0.029846  −0.305 0.760620
## business.parking          0.119037   0.044352   2.684 0.007311 **
```

```
## NoiseLevelloud          −0.193852   0.070756  −2.740 0.006181 **
## NoiseLevelquiet          0.155202   0.041878   3.706 0.000214 ***
## NoiseLevelvery_loud     −0.353333   0.141554  −2.496 0.012603 *
## WiFino                   0.004141   0.037095   0.112 0.911120
## WiFipaid                −0.366778   0.294285  −1.246 0.212725
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.013 on 3451 degrees of freedom
##   (2316 observations deleted due to missingness)
## Multiple R-squared:  0.01176,    Adjusted R-squared:  0.009757
## F-statistic: 5.867 on 7 and 3451 DF,  p-value: 8.62e-07

##
## Call:
## glm(formula = rating ~ RestaurantsPriceRange2 + business.parking +
##     NoiseLevel + RestaurantsReservations, data = rating)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## −2.5992  −0.4654   0.1125   0.6125   1.9993
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   3.27997    0.06130  53.505  < 2e-16 ***
## RestaurantsPriceRange2        0.01596    0.03293   0.485 0.627993
## business.parking              0.13384    0.04153   3.223 0.001280 **
## NoiseLevelloud               −0.17529    0.06702  −2.616 0.008940 **
## NoiseLevelquiet               0.15350    0.04033   3.806 0.000144 ***
## NoiseLevelvery_loud          −0.29520    0.13396  −2.204 0.027608 *
## RestaurantsReservationsTRUE −0.05818    0.03894  −1.494 0.135191
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 1.026865)
##
##     Null deviance: 3803.7  on 3668  degrees of freedom
## Residual deviance: 3760.4  on 3662  degrees of freedom
##   (2106 observations deleted due to missingness)
## AIC: 10518
##
## Number of Fisher Scoring iterations: 2

##
## Call:
## glm(formula = rating ~ RestaurantsPriceRange2 + business.parking +
##     NoiseLevel, family = "binomial", data = rating)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.59125  -0.09763   0.02194   0.12308   0.43380
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -0.708491   0.124675  -5.683 1.33e-08 ***
## RestaurantsPriceRange2 -0.006006   0.058916  -0.102    0.919
## business.parking         0.061366   0.084692   0.725    0.469
## NoiseLevelloud          -0.088093   0.139738  -0.630    0.528
## NoiseLevelquiet          0.062452   0.082229   0.759    0.448
## NoiseLevelvery_loud     -0.166766   0.270732  -0.616    0.538
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 187.07  on 3758  degrees of freedom
## Residual deviance: 184.94  on 3753  degrees of freedom
##   (2016 observations deleted due to missingness)
## AIC: 3132.5
##
## Number of Fisher Scoring iterations: 3
```