Exploration Analysis of Yelp Star Ratings

Xiang XU

Content

I. Abstract 2

II. Introduction 2

A. Previous Work 2

B. Data Source & Description 2

C. Exploratory Data Analysis 5

D. Model used 12

E. Result 12

1. Categorical model 12

2. Logistic model 17

III. Discussion 20

A. Implication, Limitation & Future direction 20

IV. Reference 21

I. Abstract

This exploration analysis is based on the restaurants and user information coming from *Yelp Dataset Challenge.* The goal is trying to explore the relationship between different categories restaurants star ratings and restaurants features, further build an appropriate model make predictions. This has been done with multinomial model, logistic model. Based on the prediction result, the logistic model looks better. Also, in the exploration process, I also explore the review text from different star ratings and do some sentiment analysis.

II. Introduction

So much information in Yelp dataset! Here we mainly try to explore the difficulty in predicting restaurants' stars given these different categories restaurants' feature. And at the same time, we know restaurants' average stars come from customers review star! So we also do some analysis on users' review text and stars.

A. Previous Work

After talking to Angela(TA) and scan her Predict Ratings for Chinese Restaurants using Sentiment Analysis [8], I got some inspirations and decide to extend the study subjects to different categories restaurants. Also, I read some of the past winners' paper. The papers has been listed in the Reference.

B. Data Source & Description

All data came from Yelp Dataset Challenge Round 12.

In this paper, we mainly focus on the `review`, `buisness_info` and `user_info` dataset, and filter all the information about restaurants.

Table 1 Business info (restuarant)

	Example
business_id	Apn5Q_b6Nz61Tq4XzPdf9A
name	Minhas Micro Brewery
neighborhood	NA
address	1314 44 Avenue NE
city	Calgary
state	AB
postal_code	T2E 6L6
latitude	51.09181
longitude	-114.0317
stars	4

review_count	24
is_open	TRUE
categories	Tours, Breweries, Pizza, Restaurants, Food, Hotels & Travel
categ.temp	American
BikeParking	False
BusinessAcceptsCreditCards	True
BusinessParking	{‘garage’: False, ‘street’: True, ‘validated’: False, ‘lot’: False, ‘valet’: False}
GoodForKids	True
HasTV	True
NoiseLevel	average
OutdoorSeating	False
RestaurantsAttire	casual
RestaurantsDelivery	False
RestaurantsGoodForGroups	True
RestaurantsPriceRange2	2
RestaurantsReservations	True
RestaurantsTakeOut	True
Alcohol	NA
Caters	NA
DogsAllowed	NA
DriveThru	NA
GoodForMeal	NA
RestaurantsTableService	NA
WheelchairAccessible	NA
WiFi	NA
Ambience	NA
BYOB	NA
BYOBCorkage	NA
BestNights	NA
CoatCheck	NA
Corkage	NA
GoodForDancing	NA
HappyHour	NA
Music	NA
Smoking	NA

ByAppointmentOnly	NA
AcceptsInsurance	NA
BusinessAcceptsBitcoin	NA
HairSpecializesIn	NA
AgesAllowed	NA
RestaurantsCounterService	NA
Open24Hours	NA
DietaryRestrictions	NA

Table 2 Review (restuarant)

	Example
review_id	x7mDIiDB3jEiPGPHOmDzyw
user_id	msQe1u7Z_XuqjGoqhB0J5g
business_id	iCQpiavjjPzJ5_3gPD5Ebg
stars.x	2
date	2011-02-25
text	The pizza was okay. Not the best I've had. I prefer Biaggio's on Flamingo / Fort Apache. The chef there can make a MUCH better NY style pizza. The pizzeria @ Cosmo was over priced for the quality and lack of personality in the food. Biaggio's is a much better pick if youre going for italian - family owned, home made recipes, people that actually CARE if you like their food. You dont get that at a pizzeria in a casino. I dont care what you say…
useful	0
funny	0
cool	0
name	Secret Pizza
neighborhood	The Strip
address	3708 Las Vegas Blvd S
city	Las Vegas
state	NV
postal_code	89109
latitude	36.10984
longitude	-115.1742
stars.y	4
review_count	4078
is_open	TRUE

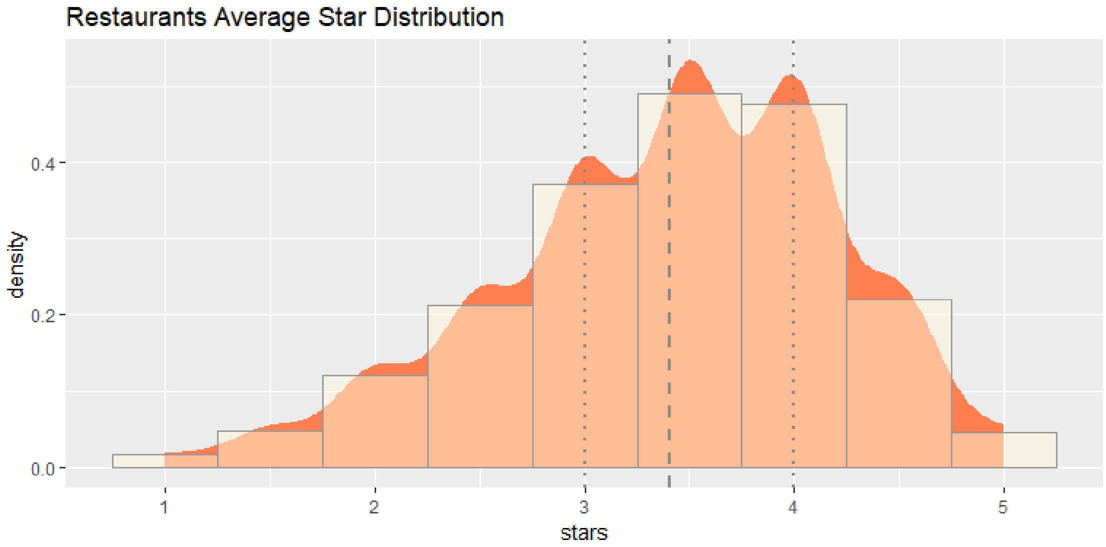
categories	Pizza, Restaurants
categ.temp	American

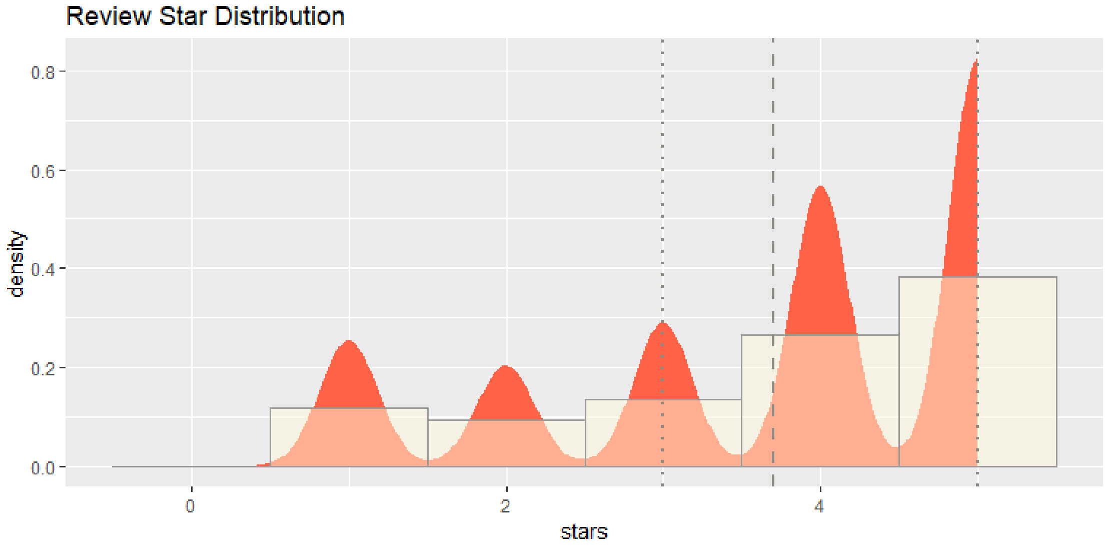
Table 3 User info

	Example
user_id	lzlZwIpuSWXEnNS91wxjHw
name	Susan
review_count	1
yelping_since	2015-09-28
useful	0
funny	0
cool	0
fans	0
average_stars	2
compliment_hot	0
compliment_more	0
compliment_profile	0
compliment_cute	0
compliment_list	0
compliment_note	0
compliment_plain	0
compliment_cool	0
compliment_funny	0
compliment_writer	0
compliment_photos	0
n_friends	1
n_elite	1
yelping_year	3
yelping_year_level	3-5 yrs
yelping_year_level_factor	2

C. Exploratory Data Analysis

Let’s take a glimpse of the distribution of the reviews’ and restuarants’ stars to start our Yelp analysis.





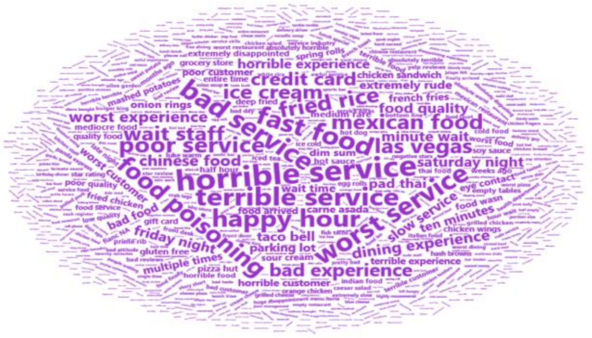
We are initially surprised to see the difference between distributions between reviews' and restaurants' stars category. The review distributions in the restaurants category were skewed to the 4 and 5 star categories heavily, with the average of 3.7 and median 4 stars, while the resturants average star distributions have less skewness, with the average of 3.4 stars and median of 3.5 stars.

The skewness is confirmed by a separate analysis by Max Woolf [7] on 1 and 5 star reviews which showed, excellent visualization aside, that Yelp reviews have started to appear more optimistically biased as time passes.

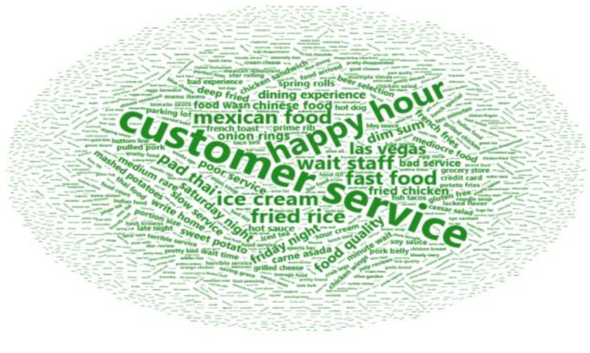
The difference between reviews star and restaurants average star will be taken into consideration in our further analytics task.

Out of curiosity, I take a further look at the review text. what kind of words, mainly bigrams, are characteristic of different star categories so I threw in some quick wordcloud visualizations.

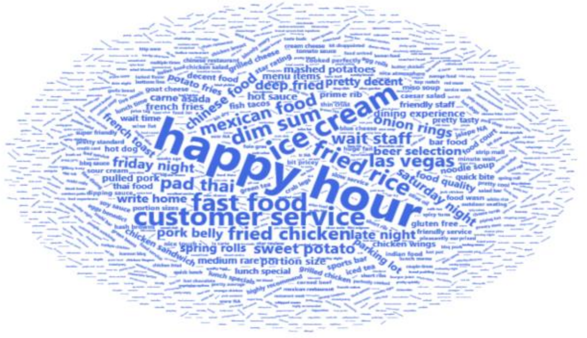
Wordcloud of 1 Star Reviews



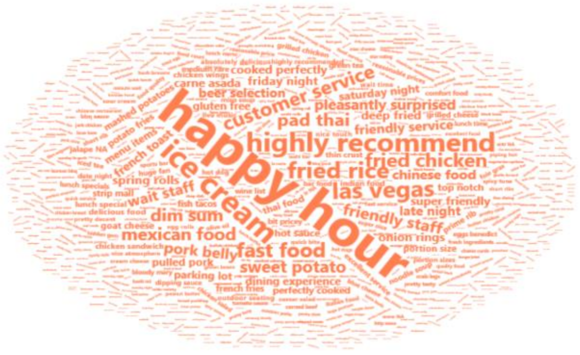
Wordcloud of 2 Star Reviews



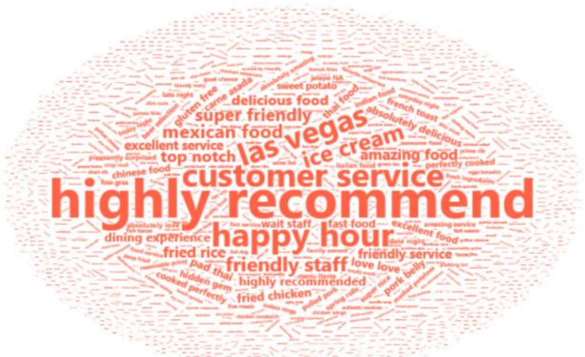
Wordcloud of 3 Star Reviews



Wordcloud of 4 Star Reviews

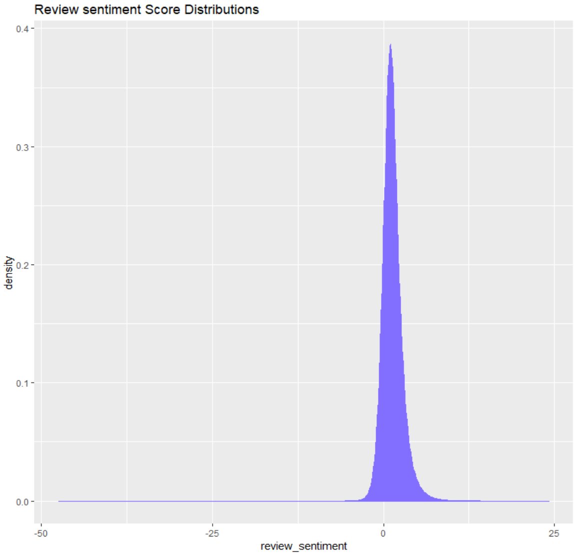


Wordcloud of 5 Star Reviews

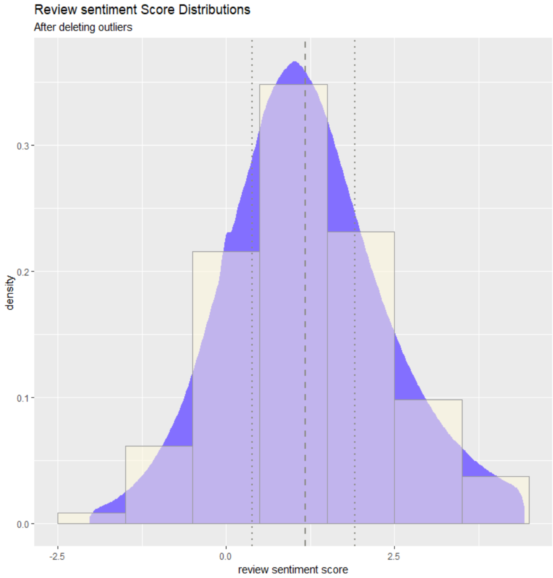


Lesson learned: if one were to start a successful business, then open a Mexican-Chinese-thai buffet in Sin City with free Wifi, convenient parking, icecream on the menu, and make sure to have loving and friendly staff members!

And just out of curiosity, I use `sentiment` function to calculate the sentiment scores of the review text, and look at its distributions.



It seems that there are some extreme values there. Move the outliers and look at it again.



Based on simple analysis if restaurants categories, recategorize them into several main types. Here I have 17 main restuarant categories: “Italian”, “French”,“Vietnamese”, “Chinese”, “Mediterranean”,“Korean”, “Greek”, “Middle Eastern” ,“Canadian”,“German”,“Irish”,“Indian”, “Thai” , “Mexican”, “Portuguese”,“Japanese”,“American”, and their counts.

Table 4 Main Type Restaurants Counts

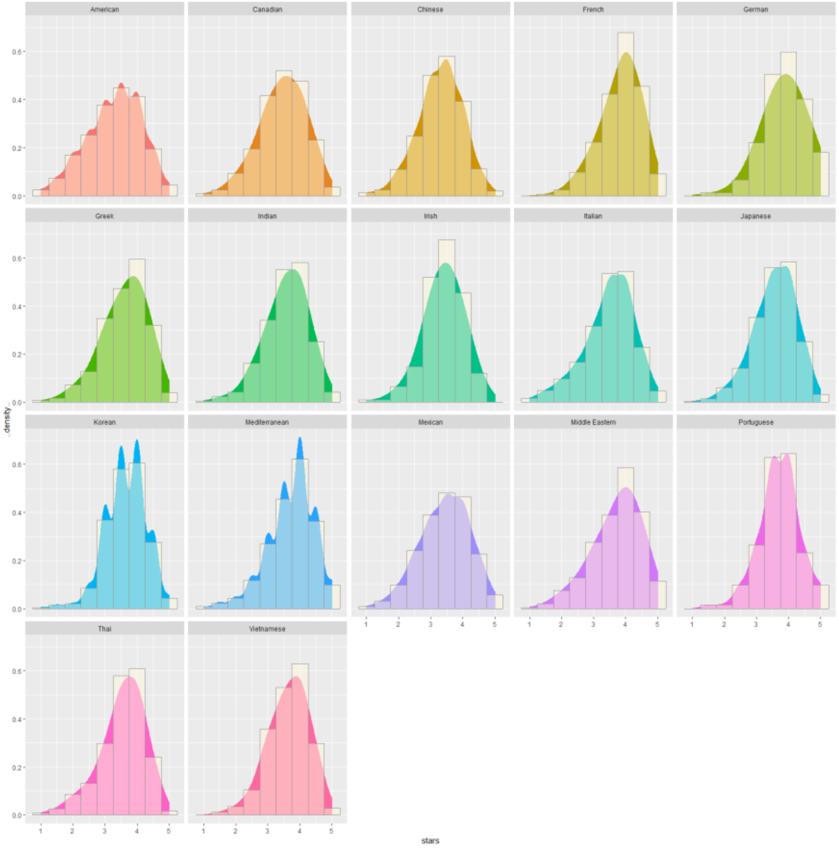
X	categ.temp	n
1	American	22091
2	Mexican	4428
3	Italian	4174
4	Chinese	3833

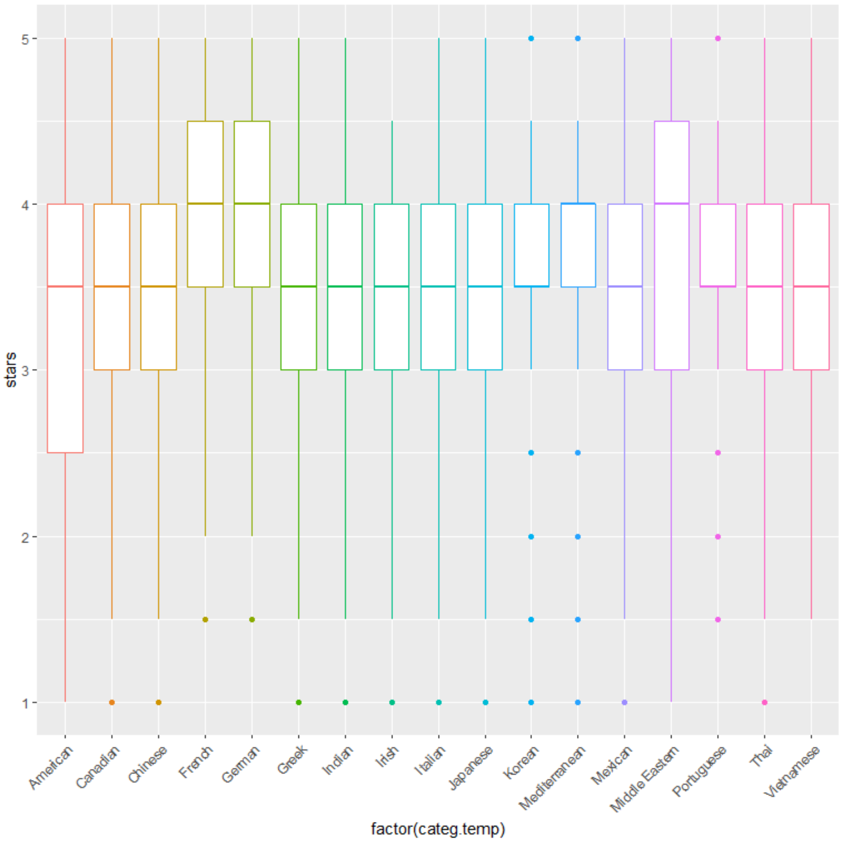
5	Japanese	2559
6	Canadian	1771
7	Indian	1389
8	Thai	1389
9	Middle Eastern	1081
10	Vietnamese	935
11	Greek	900
12	Korean	813
13	French	809
14	Mediterranean	755
15	Portuguese	242
16	Irish	216
17	German	154

How’s star distributions of different category restaurants?

Table 5 stars summary by category

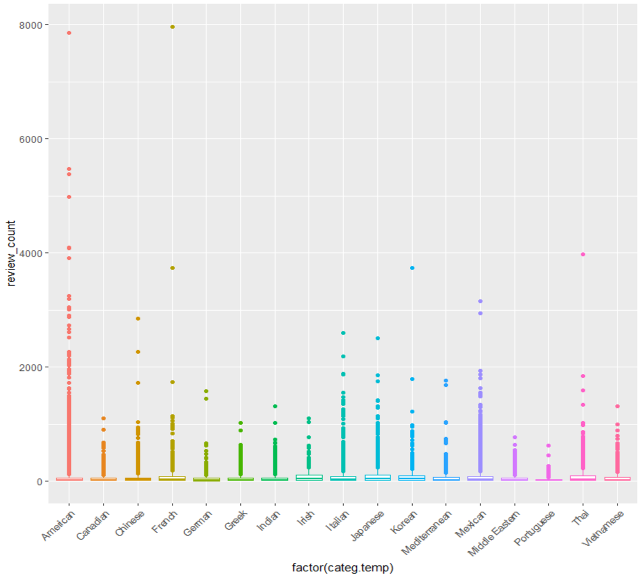
categ.temp	mean	sd	Q1	Q3
German	3.877	0.665	3.5	4.5
French	3.837	0.646	3.5	4.5
Middle Eastern	3.719	0.770	3.0	4.5
Mediterranean	3.719	0.731	3.5	4.0
Portuguese	3.707	0.633	3.5	4.0
Korean	3.651	0.630	3.5	4.0
Vietnamese	3.647	0.634	3.0	4.0
Greek	3.609	0.715	3.0	4.0
Indian	3.577	0.687	3.0	4.0
Japanese	3.575	0.672	3.0	4.0
Thai	3.542	0.705	3.0	4.0
Italian	3.483	0.784	3.0	4.0
Canadian	3.456	0.736	3.0	4.0
Mexican	3.438	0.777	3.0	4.0
Irish	3.400	0.607	3.0	4.0
Chinese	3.295	0.696	3.0	4.0
American	3.282	0.862	2.5	4.0


American
Canadian
Chinese
French
German
Greek
Indian
Irish
Italian
Japanese
Korean
Mediterranean
Mexican
Middle Eastern
Portuguese
Thai
Vietnamese
0.0
0.2
0.4
0.6
density
1
2
3
4
5
stars


5
4
3
2
1
stars
American
Canadian
Chinese
French
German
Greek
Indian
Irish
Italian
Japanese
Korean
Mediterranean
Mexican
Middle Eastern
Portuguese
Thai
Vietnamese
factor(categ.temp)

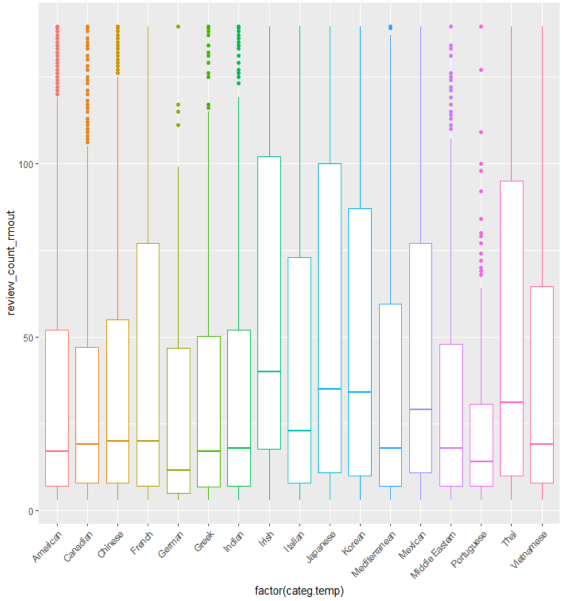
There’are not big distinctions among the stars category distribution of differnet category restaurants. But still we can see German restaurant has higher skewness than others, which means higher star level here.

We also want to get a sense of how the restaurants’ review counts distribute.



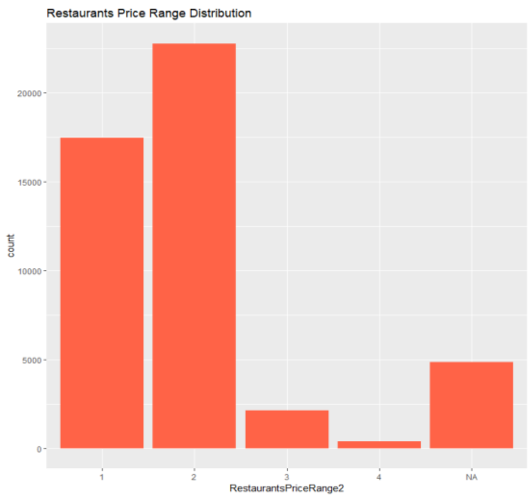
So many outliers!

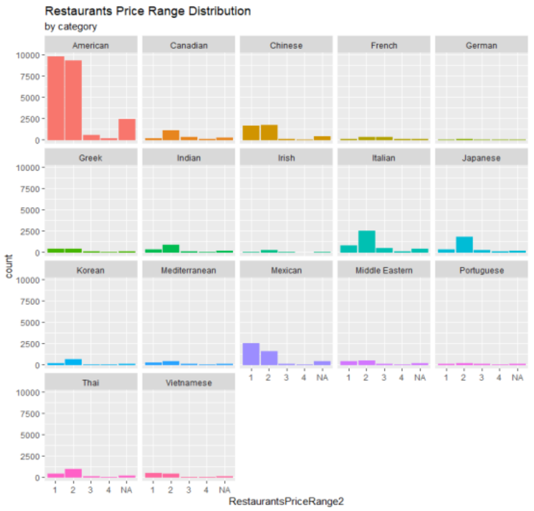
Try to replace outliers by using “outlier rule” q +/- (1.5 * H), so as to view the distribution of review_count more clearly.



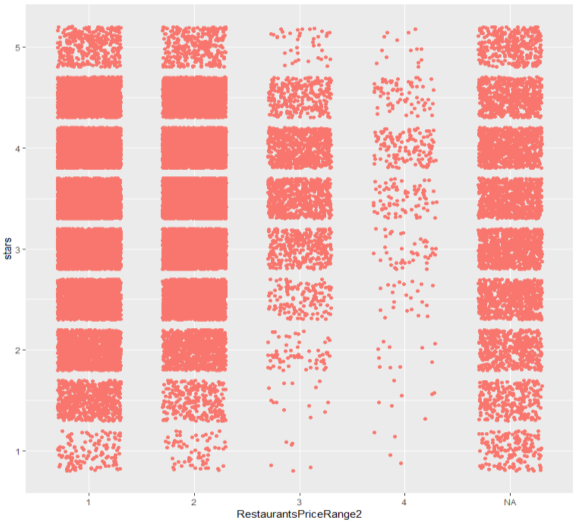
Still some large outliers, but much better now.

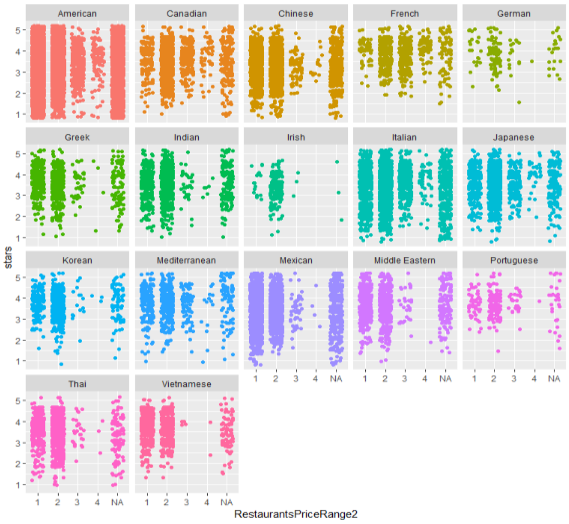
Then let’s look at the restaurant-price-range distributions.





Is there relations between price range and stars?





D. Model used

During the analysis process, first I used a five-level categorical model (I have 5 stars response), later switched to logistic model, recatogorizing reponse(stars) to “above average” and “below average”.

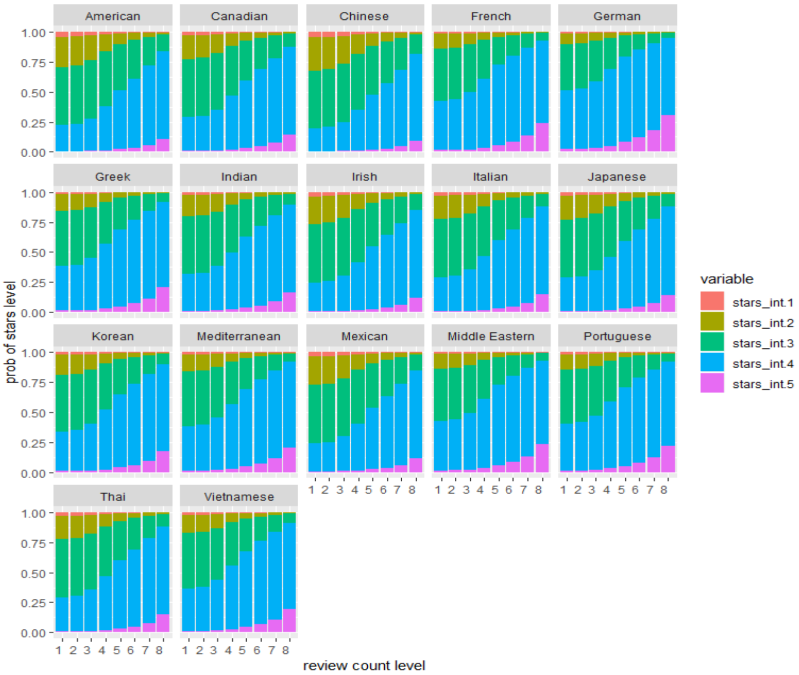
E. Result

1. Multinomial model

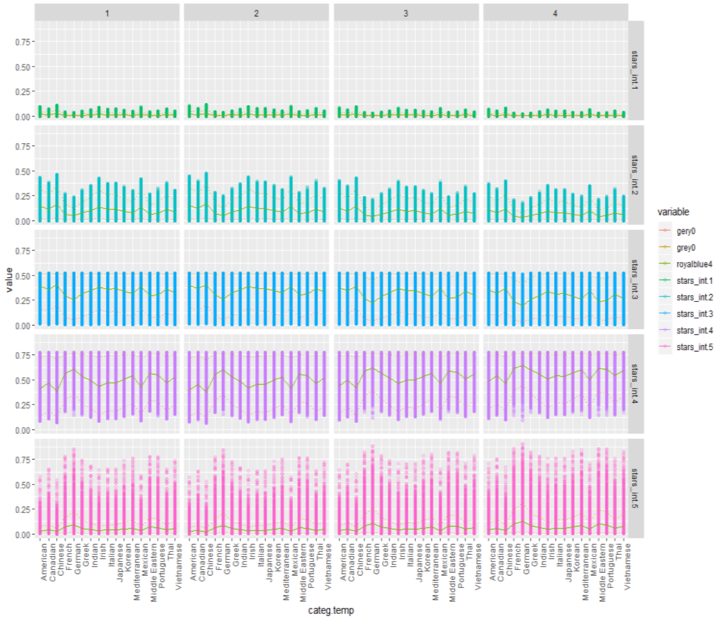
a) Fitting model by polr function

```
## polr(formula = ordered(stars_int) ~ review_cat + categ.temp +
##     RestaurantsPriceRange2 + NoiseLevel, data = restaurant_info)
##                              coef.est coef.se
## review_cat2                   0.06     0.04
## review_cat3                   0.31     0.04
## review_cat4                0.83     0.04
## review_cat5                1.40     0.05
## review_cat6                1.83     0.09
## review_cat7                2.35     0.15
## review_cat8                3.09     1.07
## categ.tempCanadian         0.37     0.05
## categ.tempChinese         -0.16     0.04
## categ.tempFrench           1.01     0.09
## categ.tempGerman           1.40     0.21
## categ.tempGreek            0.82     0.08
## categ.tempIndian           0.50     0.06
## categ.tempIrish            0.14     0.13
## categ.tempItalian          0.37     0.04
## categ.tempJapanese         0.35     0.05
## categ.tempKorean           0.61     0.08
## categ.tempMediterranean    0.82     0.09
## categ.tempMexican          0.10     0.04
## categ.tempMiddle Eastern   1.01     0.07
## categ.tempPortuguese       0.92     0.15
## categ.tempThai             0.39     0.06
## categ.tempVietnamese       0.76     0.07
## RestaurantsPriceRange22   -0.07     0.02
## RestaurantsPriceRange23    0.21     0.05
## RestaurantsPriceRange24    0.50     0.12
## NoiseLevelloud            -0.56     0.04
## NoiseLevelquiet            0.39     0.03
## NoiseLevelvery_loud       -1.06     0.06
## 1|2                       -3.50     0.05
## 2|3                       -1.10     0.04
## 3|4                        1.20     0.04
## 4|5                        5.21     0.06
## ---
## n = 34461, k = 33 (including 4 intercepts)
## residual deviance = 74543.1, null deviance is not computed by polr
```

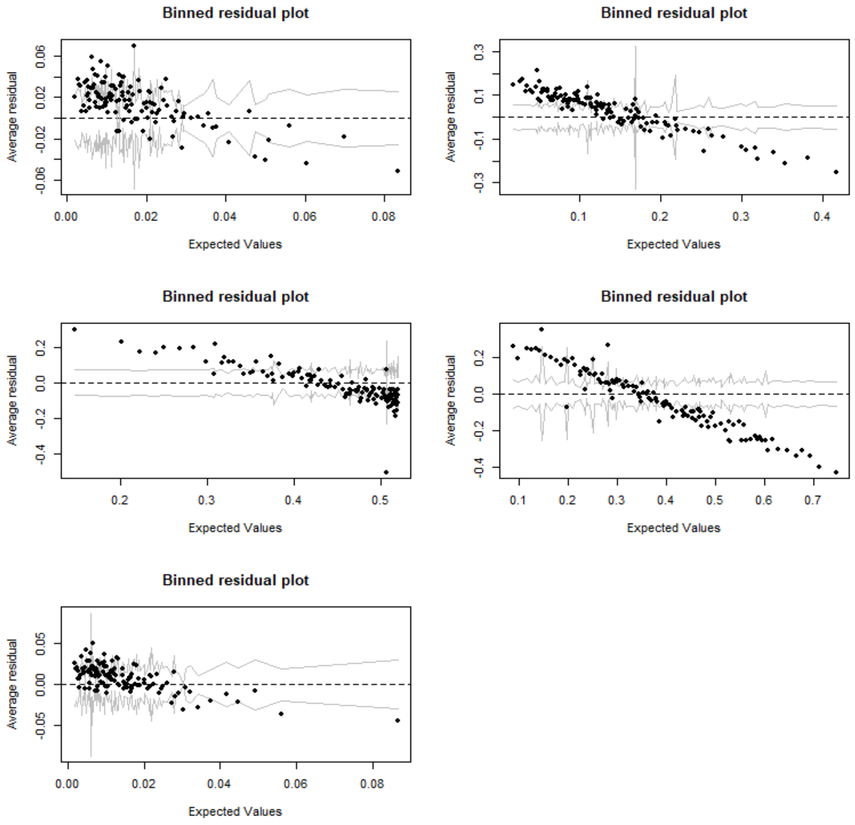
(1) Predictor: “review_cat”,“categ.temp”,“RestaurantsPriceRange2”,“NoiseLevel”



The above result is a point estimate and lacks information regarding the uncertainty of our estimates. We can add the uncertainty in the parameter estimate using the sim function.

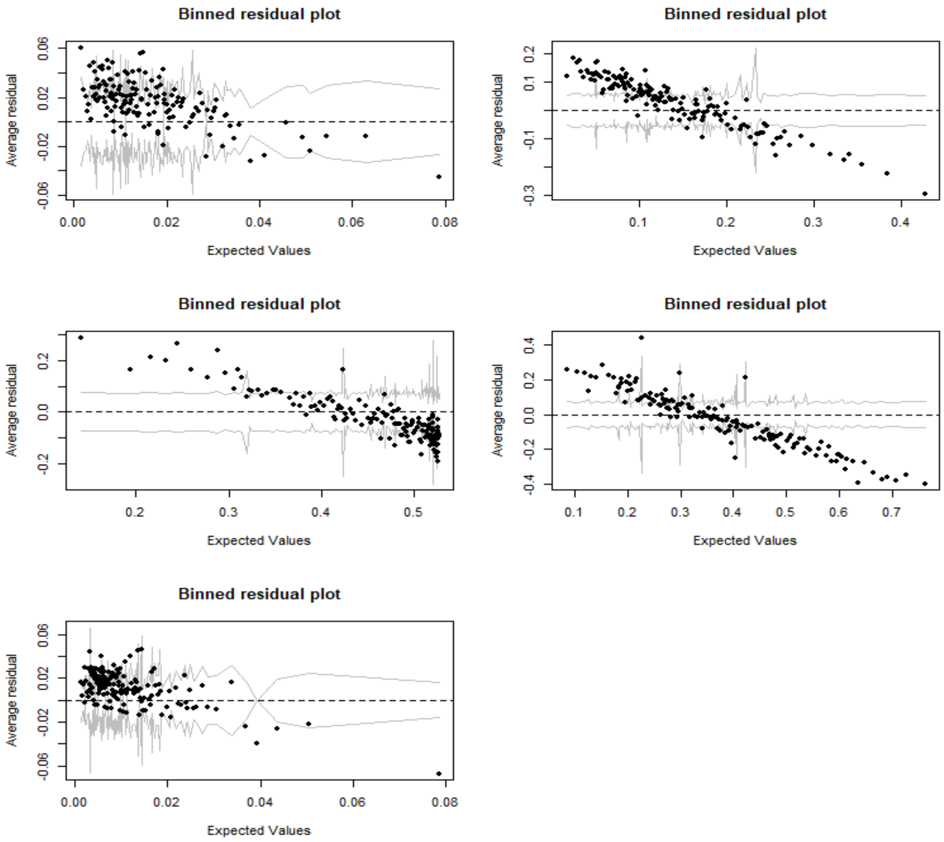


(2) Binned plot to check the model



(3) Consider more predictors: Parking(car), Delvery, Takout etc

```
## polr(formula = ordered(stars_int) ~ review_cat + categ.temp + 
##     RestaurantsPriceRange2 + NoiseLevel + Parking + RestaurantsTakeOut + 
##     RestaurantsDelivery, data = restaurant_info)
##                           coef.est coef.se
## review_cat2                0.00     0.04  
## review_cat3                0.13     0.04  
## review_cat4                0.60     0.05  
## review_cat5                1.17     0.06  
## review_cat6                1.61     0.10  
## review_cat7                2.17     0.16  
## review_cat8                2.86     1.09  
## categ.tempCanadian         0.34     0.06  
## categ.tempChinese         -0.17     0.04  
## categ.tempFrench           0.88     0.09  
## categ.tempGerman           1.26     0.22  
## categ.tempGreek            0.84     0.08  
## categ.tempIndian           0.52     0.07  
## categ.tempIrish            0.08     0.14  
## categ.tempItalian          0.36     0.04  
## categ.tempJapanese         0.37     0.05  
## categ.tempKorean           0.61     0.08  
## categ.tempMediterranean   0.79     0.09
## categ.tempMexican         0.10     0.04
## categ.tempMiddle Eastern  1.05     0.08
## categ.tempPortuguese      0.97     0.16
## categ.tempThai            0.37     0.06
## categ.tempVietnamese      0.77     0.08
## RestaurantsPriceRange22  -0.11     0.02
## RestaurantsPriceRange23   0.12     0.06
## RestaurantsPriceRange24   0.35     0.13
## NoiseLevelloud           -0.55     0.04
## NoiseLevelquiet           0.42     0.03
## NoiseLevelvery_loud      -0.99     0.07
## ParkingTRUE               0.40     0.03
## RestaurantsTakeOutTrue   -0.29     0.06
## RestaurantsDeliveryTrue   0.07     0.03
## 1|2                      -3.78     0.08
## 2|3                      -1.30     0.07
## 3|4                       1.04     0.07
## 4|5                       5.24     0.09
## ---
## n = 31768, k = 36 (including 4 intercepts)
## residual deviance = 67361.0, null deviance is not computed by polr
```

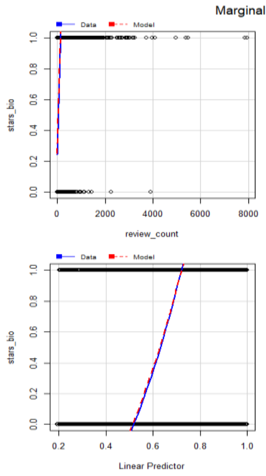


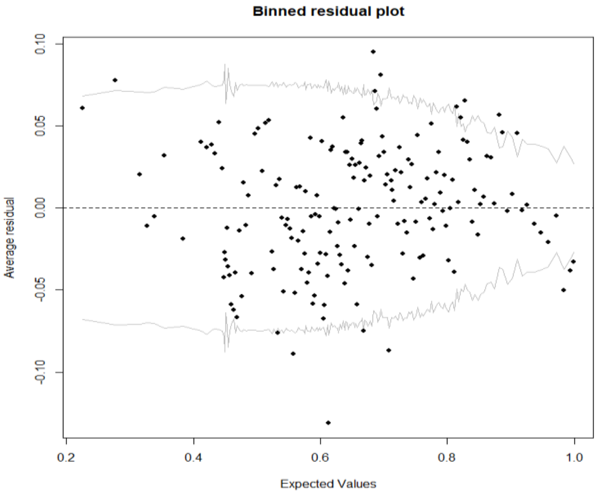
2. Logistic model

For the prediction of categorical model is not that good, switch to the logistic model.

Cut restaurants into two levels, below average(stars equal or less than 3), and above average(stars more than 3)

```
## glm(formula = stars_bio ~ review_count + categ.temp + Parking + 
##     RestaurantsPriceRange2 + NoiseLevel + RestaurantsTakeOut + 
##     RestaurantsDelivery, family = binomial(link = "logit"), data = restaur
ant_info_nona)
##                             coef.est coef.se
## (Intercept)                  0.07     0.07  
## review_count                 0.01     0.00  
## categ.tempCanadian           0.36     0.07  
## categ.tempChinese           -0.07     0.05  
## categ.tempFrench             1.01     0.13  
## categ.tempGerman             1.58     0.34  
## categ.tempGreek              0.78     0.10  
## categ.tempIndian             0.77     0.08  
## categ.tempIrish              0.15     0.17  
## categ.tempItalian            0.48     0.05  
## categ.tempJapanese           0.46     0.06  
## categ.tempKorean             0.72     0.11  
## categ.tempMediterranean      0.87     0.12  
## categ.tempMexican            0.06     0.04  
## categ.tempMiddle Eastern     0.93     0.10  
## categ.tempPortuguese         1.36     0.23  
## categ.tempThai               0.64     0.08  
## categ.tempVietnamese         0.85     0.10  
## ParkingTRUE                  0.47     0.03  
## RestaurantsPriceRange22     -0.12     0.03  
## RestaurantsPriceRange23      0.31     0.08  
## RestaurantsPriceRange24      0.37     0.17  
## NoiseLevelloud              -0.53     0.04  
## NoiseLevelquiet              0.31     0.03  
## NoiseLevelvery_loud         -1.04     0.08  
## RestaurantsTakeOutTrue      -0.30     0.07  
## RestaurantsDeliveryTrue      0.12     0.03  
## ---
##   n = 31768, k = 27
##   residual deviance = 37285.8, null deviance = 40902.6 (difference = 3616.
8)
```



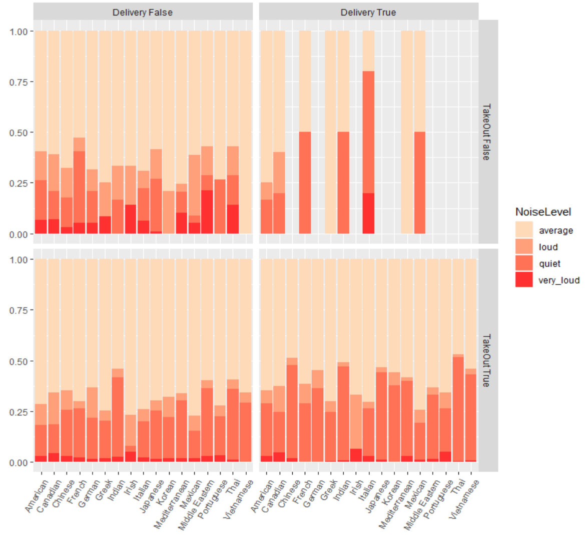


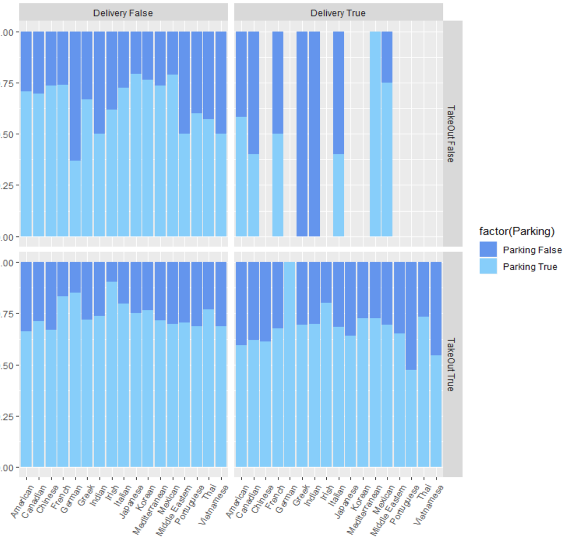
The binned residual looks better than five-stars-level categorical model.

```
## the ommission rate as a proportion of true occurrences misidentified 0.11
.

## The sensitivity is  0.89 .

## The specificity is  0.28 .

## The proportion of the presence and absence records correctly identified is
  0.68 .
```

The results look OK.

Is there Interactions?





As mentioned above, there are both similarity and distinction between restaurants average star distribution and customers review stars distributions. So here let’s look at their correlations.

```
## Correlation between restaurants average stars and customers review stars i
s  0.42 .
```

The correlation seems to be not that high. What if we add customers review stars into predictor?

```
##
## Call:
## glm(formula = stars_bio ~ review_count + categ.temp + RestaurantsPriceRang
e2 +
##     NoiseLevel + Parking + RestaurantsTakeOut + RestaurantsDelivery +
##     stars.x, family = binomial(link = "logit"), data = restaurant_all)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.0726   0.1608   0.4057   0.6092   2.3669
##
## Coefficients:
##                            Estimate Std. Error  z value Pr(>|z|)
## (Intercept)              -1.605e+00  1.045e-02 -153.705   <2e-16 ***
## review_count              1.704e-03  6.994e-06  243.666   <2e-16 ***
## categ.tempCanadian        1.783e-01  1.015e-02   17.566   <2e-16 ***
## categ.tempChinese        -1.845e-01  6.306e-03  -29.255   <2e-16 ***
## categ.tempFrench          8.174e-01  1.689e-02   48.403   <2e-16 ***
## categ.tempGerman          1.917e+00  4.840e-02   39.597   <2e-16 ***
## categ.tempGreek           1.013e+00  1.705e-02   59.376   <2e-16 ***
## categ.tempIndian          9.403e-01  1.381e-02   68.101   <2e-16 ***
## categ.tempIrish           1.948e-01  1.963e-02    9.922   <2e-16 ***
## categ.tempItalian         6.275e-01  6.580e-03   95.363   <2e-16 ***
## categ.tempJapanese        5.652e-01  7.062e-03   80.036   <2e-16 ***
## categ.tempKorean          2.995e-01  1.150e-02   26.042   <2e-16 ***
## categ.tempMediterranean   1.123e+00  1.955e-02   57.406   <2e-16 ***
## categ.tempMexican         1.185e-01  5.358e-03   22.120   <2e-16 ***
## categ.tempMiddle Eastern  1.104e+00  1.688e-02   65.432   <2e-16 ***
## categ.tempPortuguese      1.113e+00  4.331e-02   25.691   <2e-16 ***
## categ.tempThai            9.704e-01  1.132e-02   85.743   <2e-16 ***
## categ.tempVietnamese      6.331e-01  1.375e-02   46.057   <2e-16 ***
## RestaurantsPriceRange22  -1.925e-01  4.064e-03  -47.373   <2e-16 ***
## RestaurantsPriceRange23   7.666e-01  1.077e-02   71.170   <2e-16 ***
## RestaurantsPriceRange24   1.332e+00  2.752e-02   48.406   <2e-16 ***
## NoiseLevelloud           -7.085e-01  5.663e-03 -125.108   <2e-16 ***
## NoiseLevelquiet           1.827e-01  5.915e-03   30.894   <2e-16 ***
## NoiseLevelvery_loud      -1.471e+00  1.230e-02 -119.565   <2e-16 ***
## ParkingTRUE               6.874e-01  4.555e-03  150.908   <2e-16 ***
## RestaurantsTakeOutTrue    1.659e-01  8.344e-03   19.878   <2e-16 ***
## RestaurantsDeliveryTrue   8.858e-02  4.479e-03   19.775   <2e-16 ***
## stars.x                   5.164e-01  1.164e-03  443.675   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2789349  on 2922026  degrees of freedom
## Residual deviance: 2246479  on 2921999  degrees of freedom
## AIC: 2246535
##
## Number of Fisher Scoring iterations: 7

##     obs
## pred      0       1
##    0 140475   80666
##    1 396966 2303920
## attr(,"class")
## [1] "confusion.matrix"

## the ommission rate as a proportion of true occurrences misidentified 0.03
.

## The sensitivity is  0.97 .

## The specificity is  0.26 .

## The proportion of the presence and absence records correctly identified is
  0.84 .
```

It seems the prediction accuracy has been improved by adding the predictor – customers'' review star category.

III. Discussion

A. Implication, Limitation & Future direction

From the exploration process, fitted model and prediction results, it's not accurate to say restaurants average star can be predicted based on the customers' review star, restaurants category, noise level elements etc. But we can see some relations between the average star category and these features. To some extent it can be a reminder for operators how to improve their business and star category.

Analyzing customers' reviews, stars and restuarants' star category, and exploring the relation among them is fun. But due to time and my capability limitation ><, my model and prediction need further improvement.

Another thing I have to mention, there is also some drawback in the Yelp information collection system. It helps a lot if updates were encouraged and even mandatory, and information and records were complete.

IV. Reference

[1] CORALS:Who are My Potential New Customers? Tapping into the Wisdom of Customers' Decisions

[2] Clustered Model Adaption for Personalized Sentiment Analysis

[3] Personalizing Yelp Star Ratings: a Semantic Topic Modeling Approach

[4] Improving Restaurants by Extracting Subtopics from Yelp Reviews

[5] Inferring Future Business Attention

[6] Oversampling with Bigram Multinomial Naive Bayes to Predict Yelp Review Star Classes

[7] The Statistical Difference Between 1-Star and 5-Star Reviews on Yelp

[8] Predict Ratings for Chinese Restaurants using Sentiment Analysis