Yelp Data Analysis

Tianzhixi Yin

November 21, 2015

Title

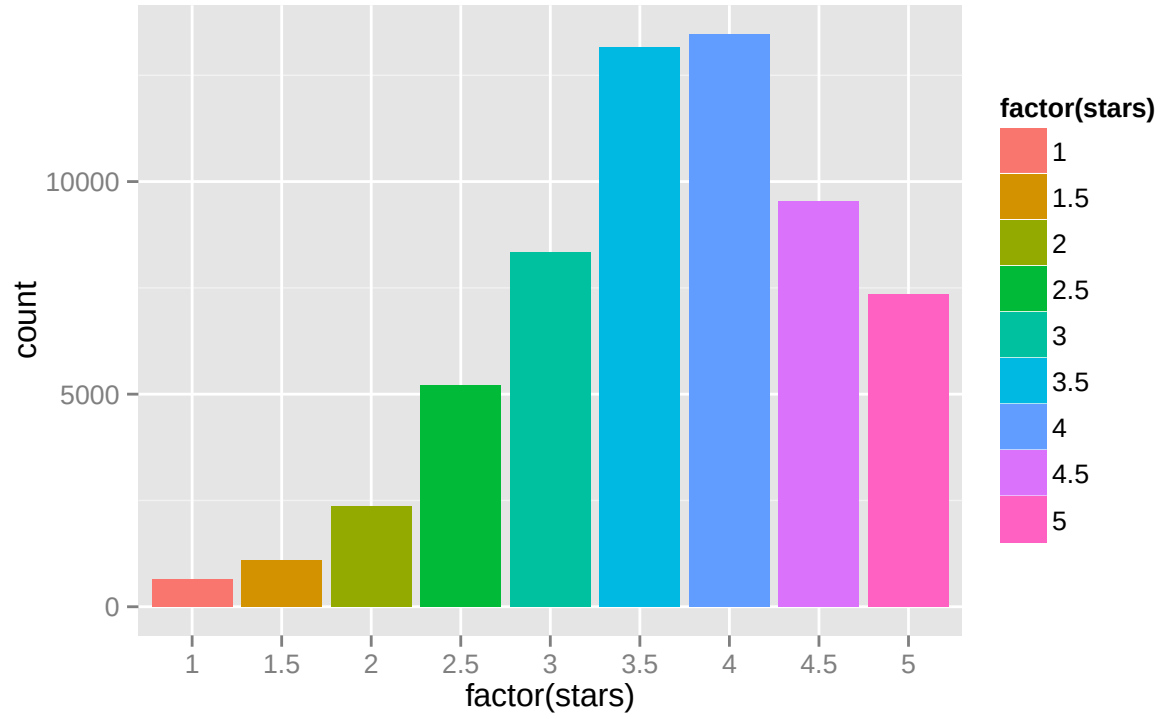
I investigated three of the five datasets (**business**, **review**, and **tip**) to answer my primary question. I did some exploratory data analysis. I built a predictive model for *stars*. I did some text mining to find out what customers are talking about frequently in their comments.

Introduction

My primary question is: what factors influence the rating of a business? In the **business** data, there are several predictor variables that might affect the rating of a business, therefore I built a predictive model using these variables and evaluated the predicting power of this model. For **review** and **tip** datasets, I focused on the *text* data and utilized some text mining techniques to plot word clouds for discovering what customers care about most.

Methods and Data

First I check the distribution of the stars of businesses.



The stars are normally distributed with a negative skew. The columns in **business** are:

```
##  [1] "business_id"   "full_address"  "hours"         "open"
```

```
##  [5] "categories"    "city"          "review_count"  "name"
##  [9] "neighborhoods" "longitude"     "state"         "stars"
## [13] "latitude"      "attributes"    "type"          "newcat"
## [17] "newnei"
```

I do not think the id, address, hours, name, latitude and longitude of a business will have great impact on the ratings. For whether a business is still running or not, I do not want to include it in my predictive model because it will not help a business to improve.

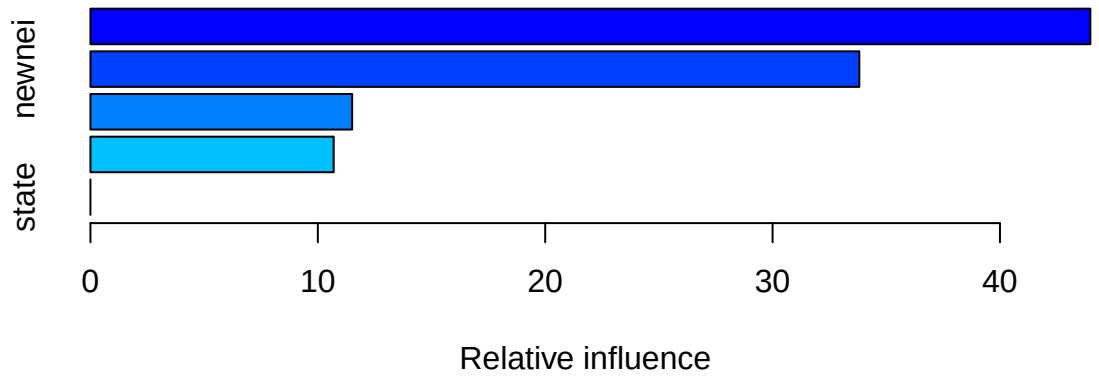
I think that *categories*, *state*, *city*, *neighborhoods*, and *review_count* could be contributing to the rating of a business. I decide to use boosting to build a predictive model for *stars*. Since the original **business** dataset is huge, I randomly select a smaller subset of this dataset (400 observations). I need to transform *neighborhoods* into a character variable. Also, I need to create a new variable for *categories*, to let an individual business have only one category which belongs to the major 23 categories given by Yelp. I randomly select 250 observation for model building, the other 150 observations are left for validation.

Boosting is a supervised machine learning algorithm for reducing bias and variance. It is developed upon the tree-based methods and is one of the ensemble methods. The trees in boosting are grown sequentially: each tree is grown using information from previously grown trees.

The *attributes* variables contain a lot of information. Most of them are binary variables simply saying whether a certain condition is true or not. I decide to compare the ratings for businesses in the two categories for each binary attribue. (Here I only show a few examples due to limited space.) Since the distribution of *stars* is not strictly normal, I use Mann-Whitney U test, which is a nonparametric test for testing the difference between two groups. If the p-value of the test is less than 0.05, we would say this factor has a significant effect on rating. For the attributes that are not binary, I also investigate some. (I only present one boxplot example to illustrate my approach.)

For **review** and **tip** datasets, I mainly utilize the *text* data for text mining. I decide to plot word clouds for 1 star businesses and 5 star businesses, and see what people mention most in these two different cases. Those words should indicate the aspects that affect the ratings for businesses.

Results




```
##                      var  rel.inf
## city                city 43.97568
```

```
## newnei             newnei 33.81501
## newcat             newcat 11.51073
## review_count review_count 10.69858
## state               state  0.00000
```

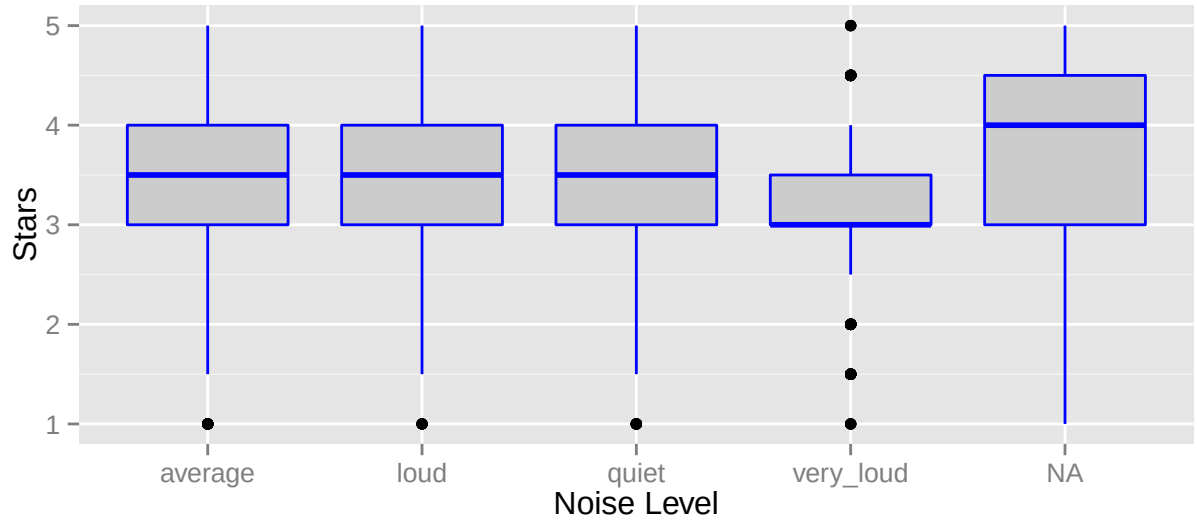
The boosting model reveals the relative influence of the variables. We can see *city* has the greatest influence on rating, then *neiborghhood*. *category* and *review_count* have small influence, while *state* makes absolutely no impact on ratings.

```
## [1] 0.7717972
```

The mean prediction error is 0.77, which I think is acceptable, especially considering the small sample size that I use. For the binary *attributes*, a few examples are shown below from the Mann-Whitney U test. we can see *By Appointment Only*, *Happy Hour*, and *Good For Groups* make a difference in ratings while *Delivery* is not that important.

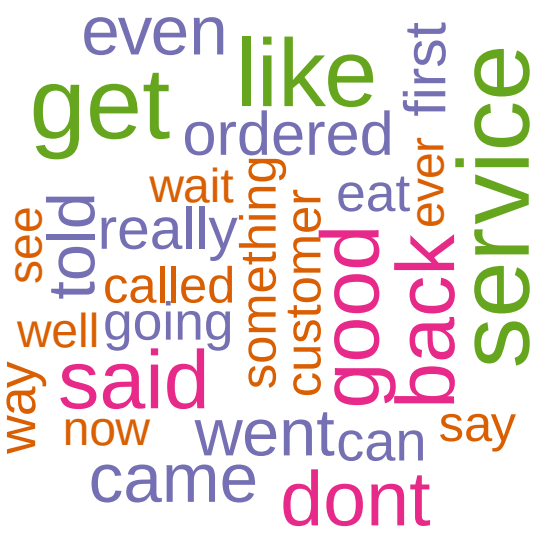
Variables	p-value
By Appointment Only	$<$ 2.2e-16
Happy Hour	9.805e-11
Good For Groups	1.16e-10
Delivery	0.08402

By Appointment Only	Mean Star
Yes	4.097
No	3.823



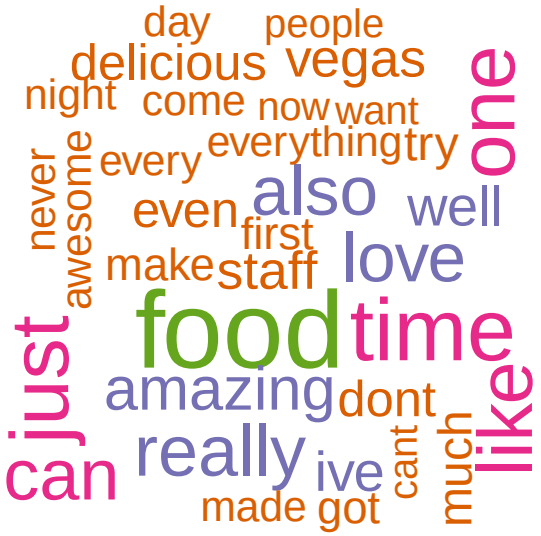
This boxplot shows that very loud noise leads to a lower rating generally than other levels of noise.

1 Star Word Cloud



The word cloud for 1 star ratings shows that poor food quality, long waiting time, and bad service are usually what lead to low ratings.

5 Star Word Cloud



The word cloud for 5 star ratings are quite similar to the 1 star, but with more complimentary words. We can see that people also care about the food, the waiting time and the staff being friendly. Some people mention that they will be back.

Tip Word Cloud

The word cloud for **tip** is pretty much the same as the previous two, with a lot of complimentary comments. I guess people like to give tips when they find out a good place.

Discussion

Since I only use a small subset of the whole data, I would not say that my findings are flawless. But at least they are decent enough for me to have some initial ideas about what influence the ratings of businesses. If I need to write a longer, more detailed report, I could follow my approaches here.

I believe my analysis has answered my primary question of interest adequately. I find out that people in different cities and neighborhoods tend to give different ratings. Moreover, various attributes of a business could be important, for example, places "By Appointment Only" have higher ratings than those don't require appointments and very loud noise will lead to low stars. By looking into the comments, I also realize that food quality, waiting time and service provided are always what customers care most about. Doing bad in these categories will result in bad ratings while performing well earns praises and recommendation.