Influence of Recent Ratings on New Ratings in Yelp Business Reviews

Coursera Datascience Project

Summary: This study evaluates whether a new star rating is related to the average of older ratings. It also considers whether this relationship depends on the country or the number of reviews at the time of the new rating. Both questions are analysed using an ordered logit model. It's coefficients suggest a significant influence of recent ratings on new ratings, even when controlling for a general business factor as measured by the average of the older (i.e. not 'recent') ratings.

Introduction

Most users of Yelp, when they review a business and judge it with a rating (i.e. a number of "stars" between 1 and 5), have already seen the older reviews of that business, since they are displayed on the same page. Social Psychology suggests that other people's opinions influence ones own judgment (see, among many, for example Asch 1952). So it seems reasonable to assume that older ratings, corresponding to other people's options, could influence a user's new rating.

If that was the case, designers of Yelp or similar services that use any kind of rating might want to look into ways of "blind" voting, so that new raters can see other users' ratings only *after* having stated their own rating.

On obvious problem when comparing previous ratings with new ratings is that we have to assume a "business factor" influencing both previous ratings and older ratings: if the business is, let's say, in general "good", then both previous users and the new user will tend to rate it higher, and vice versa. We therefore need a way to seperate a general "business factor" from a "social factor".

To achieve this, two measures of previous ratings were calculated: the average of the last 20 ratings (named the "recent average rating"), and the average of the rest of the older ratings (named "base average rating"). The base average rating is thought to represent the "general business factor", and the recent average rating is considered the one that has some social influence (assuming that rarely users go further back than one page to look at older ratings, especially on a mobile device).

Those two factors were distinguished with a multivariate model.

Methods and Data

Dataset

I used the data from the Yelp Dataset Challenge: it contains 1.5M reviews of 481 thousand businesses, by 366 thousand users.

To better handle this amount of data, I programmed a small Rails application to seed the data (which is available in JSON format) into a PostgreSQL database.

I then created a table that lists each business in a row, and in separate columns contains:

- the latest rating; the rating of the new user that will be compared to previous ratings
- the recent average rating: the average of the ratings of the most recent 20 reviews (excludes the latest rating)

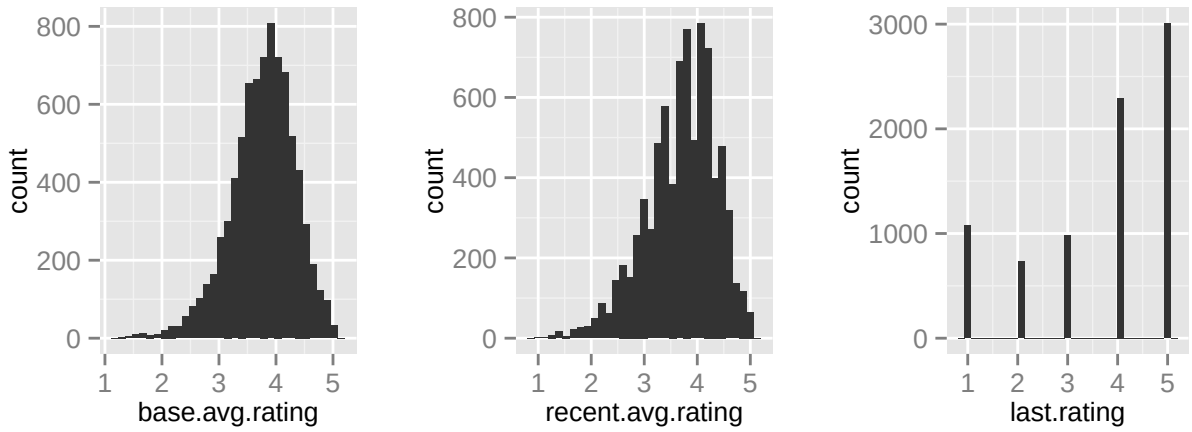
- the base average rating: the average of all older reviews (i.e. all ratings excluding the latest 21 ratings)
- the number of reviews at the time of the latest rating
- the country of the business: either USA, Canada, Great Britain or Germany.

To calculate averages of the latest 20 ratings and meaningful averages of the even older base ratings, we need businesses with at least 40 reviews (2 * 20). Therefore businesses with fewer reviews were filtered out. The top of this table looks like this:

```
    busn.id num.reviews base.avg.rtg recent.avg.rtg latest rtg country
--5jkZ3-n..          52     4.612903           4.25          5      US
--BlvDO_R..          74     3.981132           3.80          5      US
--UE_y6au..          90     3.304348           3.20          4      US
--pOlFxIT..         107     3.976744           4.40          3      US
-0Ackw6MF..         203     3.280220           3.90          1      US
```

Data exploration

The distribution of both `base.avg.rating` and `recent.avg.rating` is close to a normal. Not so the distribution of the latest ratings:

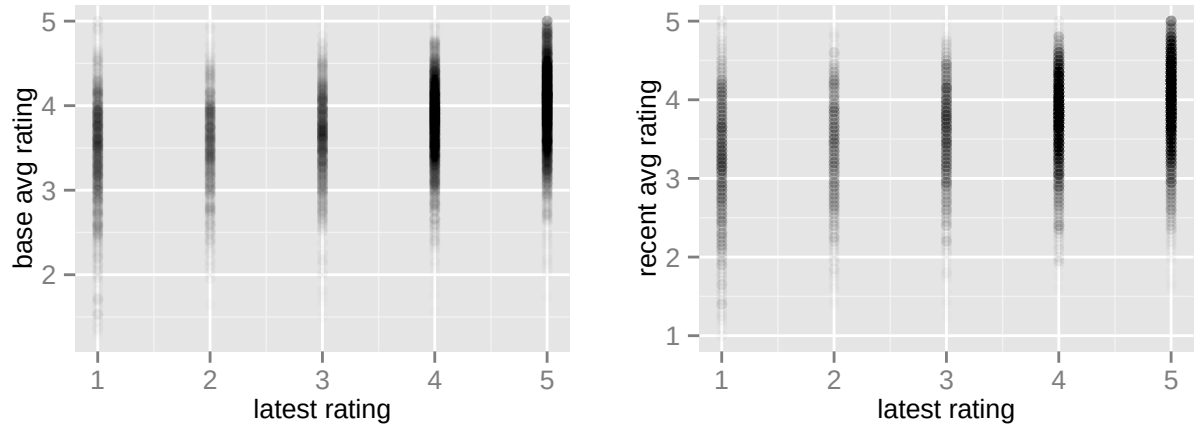


How are the ratings distributed in the three countries? Number of reviews per country and per star:

```
##
##         1    2    3    4    5
##   CA   14   12   27   76   81
##   GB    0    5   11   23   19
##   US 1064  713  942 2194 2905
```

The distribution of the latest rating seems to be roughly similar in the USA, Canada and Great Britain. There are no german businesses any more: all of them had less than 40 reviews and were therefore filtered out.

Plot the `base.avg.rating` and `recent.avg.rating` per stars (latest rating):



There seems to be a small correlation between the latest rating and both the base average rating and the recent average rating.

Let's calculate the Spearman correlation (since the latest rating is not normally distributed):

```
##                   base.avg.rating recent.avg.rating last.rating
## base.avg.rating              1.00              0.75        0.35
## recent.avg.rating            0.75              1.00        0.36
## last.rating                  0.35              0.36        1.00
##
## n= 8086
##
##
## P
##                   base.avg.rating recent.avg.rating last.rating
## base.avg.rating                   0                 0
## recent.avg.rating  0                                 0
## last.rating        0              0
```

To have a preliminary look at the interaction of those average ratings with the latest rating, let's do a multiple regression that predicts the latest rating by recent average rating and controls for base average rating:

```
##
## Call:
## lm(formula = agg.ratings$last.rating ~ agg.ratings$base.avg.rating +
##     agg.ratings$recent.avg.rating)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8578 -0.7812  0.2942  0.9569  3.1951
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     0.21525    0.09724   2.214   0.0269 *
## agg.ratings$base.avg.rating     0.36394    0.04045   8.997   <2e-16 ***
## agg.ratings$recent.avg.rating   0.56458    0.03453  16.349   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1.284 on 8083 degrees of freedom
## Multiple R-squared:  0.1532, Adjusted R-squared:  0.153
## F-statistic: 731.1 on 2 and 8083 DF,  p-value: < 2.2e-16
```

Apparently the variable `recent.avg.rating` does add significantly to the model: it is not redundant even if we include the base average rating.

Methods

Since the dependent variable, the latest rating, is discrete, not normally distributed, and only ranges from 1 to 5, multiple regression is not the ideal method. Therefore an ordered logistic regression model was calculated.

Ordered logistic regression

Ordered logistic regression is a regression model for ordinal dependent variables (as our `last.rating` variable, the "stars" given in a review). It uses maximum likelihood to estimate the coefficients (see Wikipedia; idre).

The function `polr` from the library MASS was used to estimate the coefficients of the ologit model. To compare the influence of the base average rating and the recent average rating with the influence of the number of reviews at the time of the latest rating, `num.reviews` was included in the ordered logit model as well.

```
m <- polr(stars ~ base.avg.rating + recent.avg.rating + num.reviews, data = agg.ratings, Hess=TRUE)
```

The variable `country` was not included in the model since exploratory analysis showed that the distribution of CA and USA looks very similar, and both CA and GB have to small cell sizes for the ologit model to be stable or meaningful.

Results

```
##                         Value   Std. Error   t value       p value
## base.avg.rating   0.5991982608 0.0585859732 10.227674  1.490557e-24
## recent.avg.rating 0.7612636962 0.0501321894 15.185128  4.436986e-52
## num.reviews       0.0001427965 0.0001158496  1.232602  2.177244e-01
## 1|2               2.9911406742 0.1456840851 20.531691  1.122007e-93
## 2|3               3.6899375330 0.1468062387 25.134746 2.075036e-139
## 3|4               4.3764579942 0.1490382098 29.364671 1.552539e-189
## 4|5               5.6987429868 0.1543543624 36.919870 2.218177e-298
```

We see that both the base average rating and the recent average rating exhibit a significant influence on a new rating, whereas the number of reviews seems to be irrelevant.

To easier interpret the coeffient, let's convert them to odds rations and calculate their confidence intervals:

```
##                   OddsRatios     2.5 %   97.5 %
## base.avg.rating     1.820659 1.6233182 2.042434
## recent.avg.rating   2.140980 1.9408026 2.362286
## num.reviews         1.000143 0.9999341 1.000356
```

So we can expect an increase in the base average rating of one star (one 'unit') to increase the *odds* by 82% that the latest rating gets a higher star-rating.

Likewise, an increase in the recent average rating of one star raises the *odds* more than twice that the latest rating gets a (any) higher rating.

Both these odds have a confidence intervall way above one, as opposed to the number of reviews: with 95% probability, one more review does not change the odds of the latest review to have any higher star rating.

Discussion

The main question of this study is whether the recent ratings have an influence on a new rating and whether this influence is not entirely predicted by the 'base rating' for a business. In other words, even if a business is rated in general as, say, "good", do the recent ratings influence a new rating. To control this 'base rating' of a business, the 20 most recent ratings were separated from all the older ratings, i.e. an average was calculated for both the older base rating and the 20 most recent ratings (except the very latest rating). Then a multiple regression model was fitted, using ordered logistic regression, to estimate the influence of the recent rating on the latest rating, controlling for the base rating.

The model coefficients from the fitted model hint at an influence of both the base and the recent average rating on the new (latest) rating. But the recent average rating does have a significant (even bigger) influence on the new rating, even when controlling for the business' base rating. Therefore we can say, yes, recent ratings do influence new ratings, independently of how the business is rated in general.

The number of reviews a business has at the time of the latest rating does not seem to have any relevance on a new rating, and neither does the country.

There are though a couple of drawbacks of this ordered logistic regression:

- This method assumes that the coefficients that describe the relationship between a 1-star rating and any higher rating is the same as between a 1-or-2-star rating and any higher rating (the "proportional odds assumption"). This assumption has not been tested here.
- It is difficult to come up with diagnostics for the model fitted here. I don't know *how* fit it is.

The strong significance of the coefficients related to the main question of this study indicate though, that the hypothesized relationship of recent ratings and new ratings holds, controlled for the business' base rating.

Of course one should keep in mind, that this data is purely 'correlational', not causal.

References

Asch, S. E. 1952. *Effects of Group Pressure on the Modification and Distortion of Judgements. G. E. Swanson, T. M. Newcomb & E. L. Hartley (Eds.), Readings in Social Psychology (2nd Ed., Pp. 2-11).* New York:NY Holt.

idre, UCLA. "R Data Analysis Examples - Ordinal Logistic Regression." http://www.ats.ucla.edu/stat/r/dae/ologit.htm.

Wikipedia. "Ordered Logit." https://en.wikipedia.org/wiki/Ordered_logit.