INTRODUCTION TO CONNECTIVE MEDIA - ASSIGNMENT 3

Exploring anchoring theory with the Yelp dataset

by Jing Dai, Alap Parikh

Questions

In this assignment we looked to explore how consistent users' ratings of businesses are based on sentiment analysis of review texts, and consequently attempt to predict ratings solely based on the text content. Furthermore, we looked to see if the error in these predictions arose due to an anchoring bias caused by the existent rating of the business. This business rating is the most prominently seen feature associated with a business profile or business page. In addition to these core concepts, we tried to gain general insight into user behavior based on a couple of simple plots.

Data

We used a Yelp dataset available to us as part of an ongoing Yelp challenge. This dataset, among other things, contained a collection of reviews which contained attributes such as user ID, business ID, rating in stars, text, and date. We performed our analysis using this and the corresponding rating for every business from the collection of businesses. We used a total of 1238 users, and analyzed 60 reviews per user.

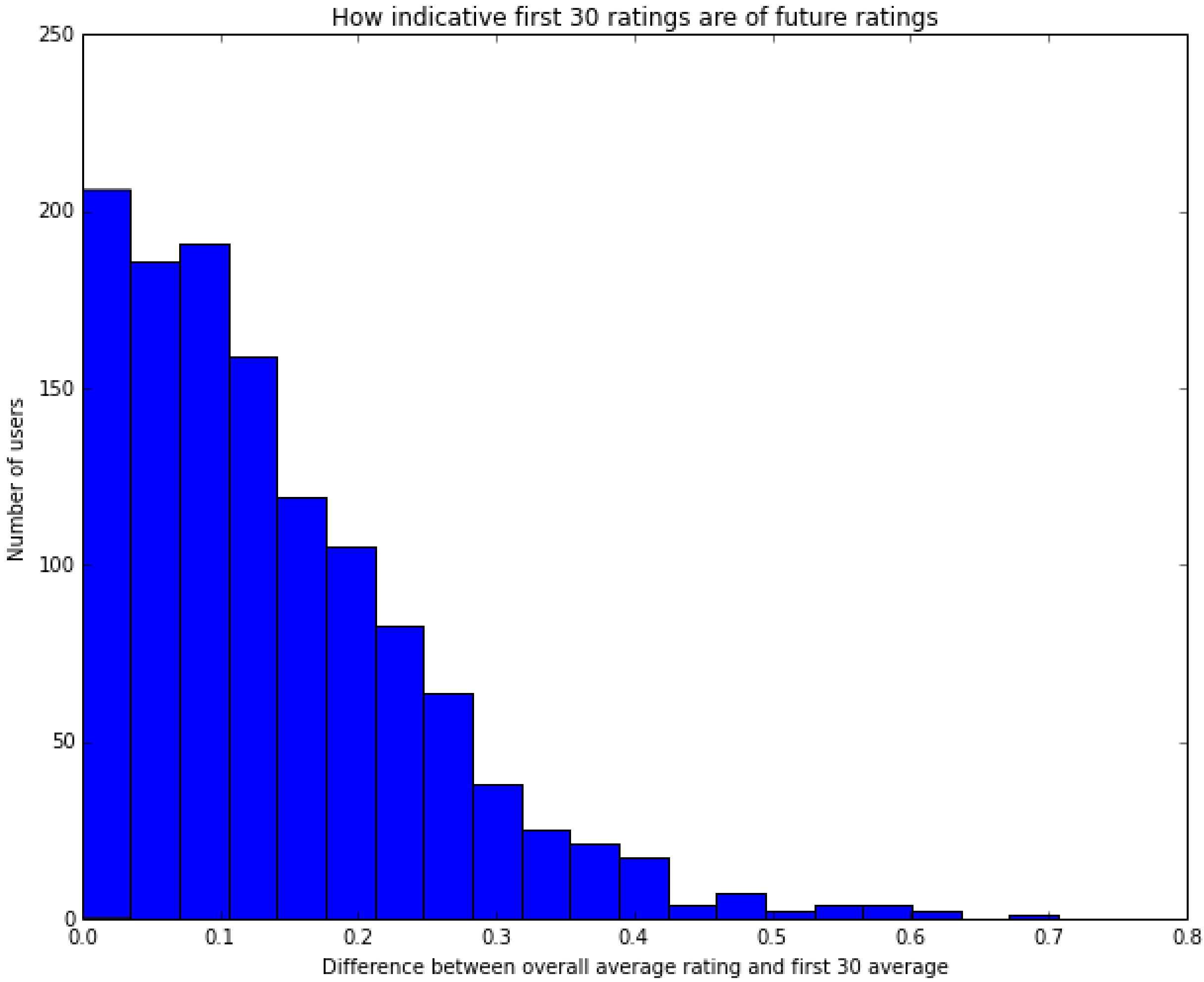
Analysis

Selection of Users

We scanned the reviews to find users with the most number of reviews, and selected the 1238 users that had more than 100 reviews. The assumption here was that users with a higher number of reviews would be more thorough and loyal, and therefore definitely tend to have long reviews.

Indicativeness of initial ratings

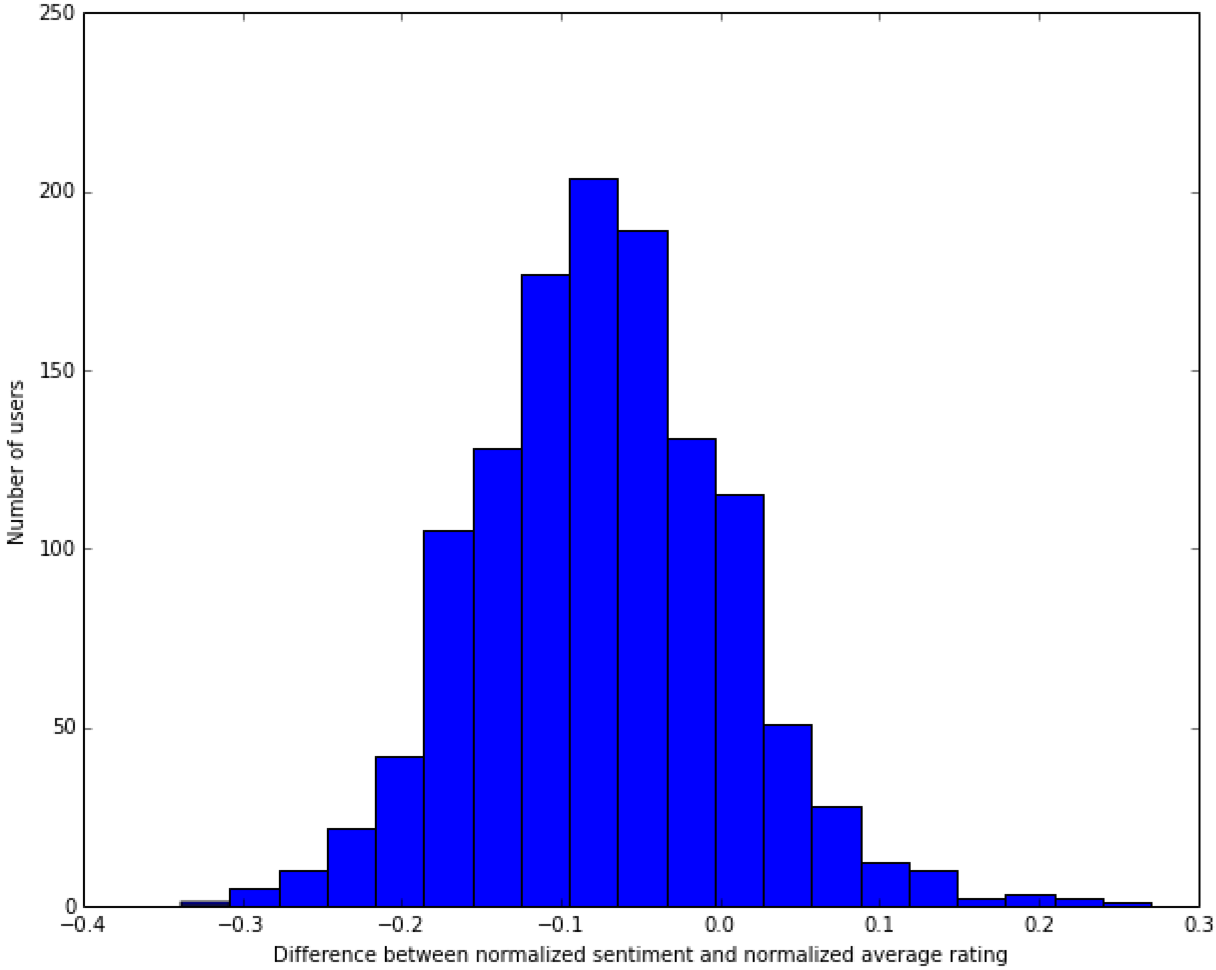
We then used the first 30 ratings (sorted by date of posting) for each of these users as the base for predicting all future ratings, and tested how indicative these ratings are of all future ratings.



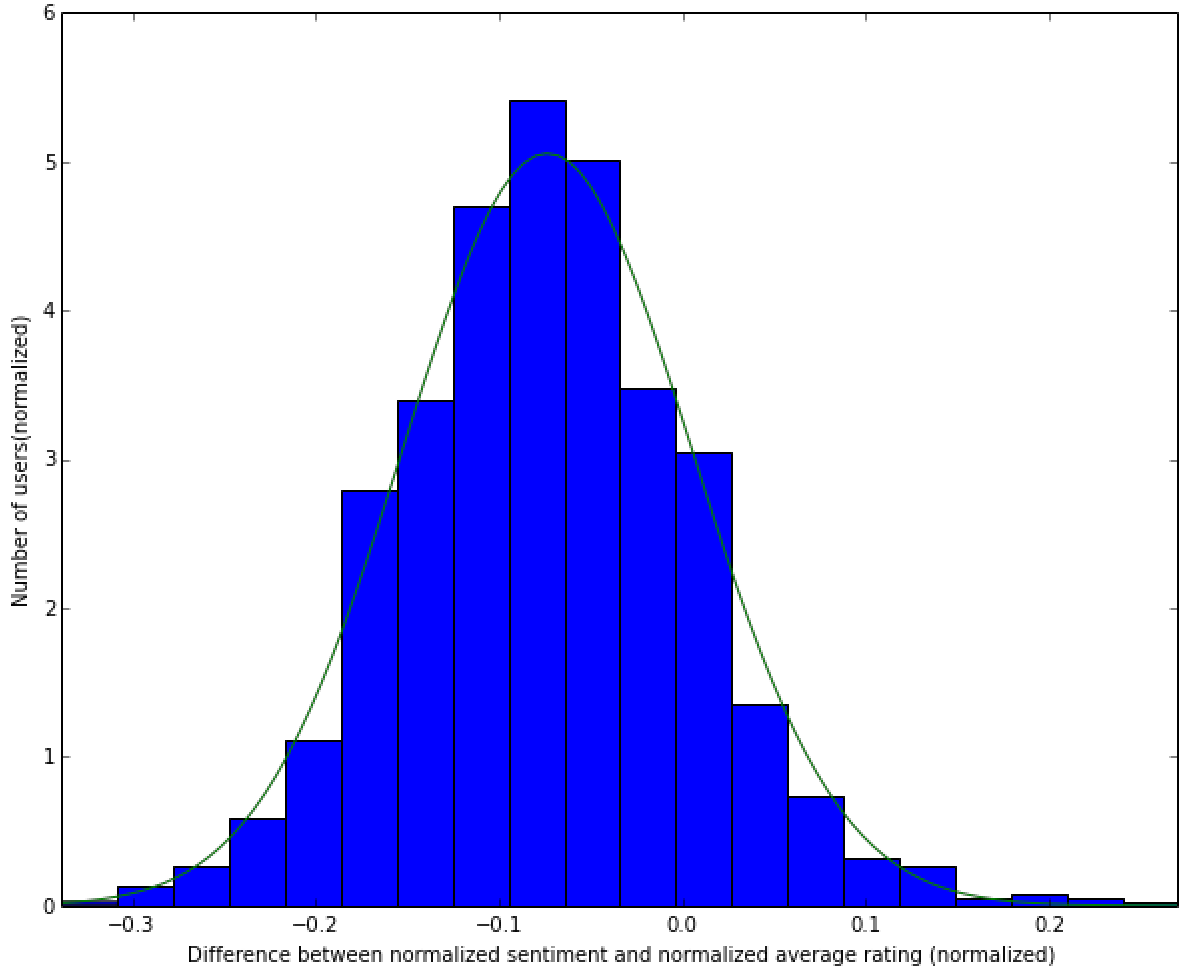
In this figure, the x-axis represents bins divided by the difference between overall average rating and average rating of first 30 reviews, and the y-axis is a count of the number of users in that bin. We can see that for the most part, the first 30 reviews are indicative of future reviews. The difference between overall average rating and average rating for first 30 reviews is less than 0.1 for 47% of users, and less than 0.14 for 60% of users.

Test of Sentiment Analysis

We tested how closely an average sentiment polarity corresponds to an average rating for each user. For sentiment analysis we used a Python library called TextBlob.



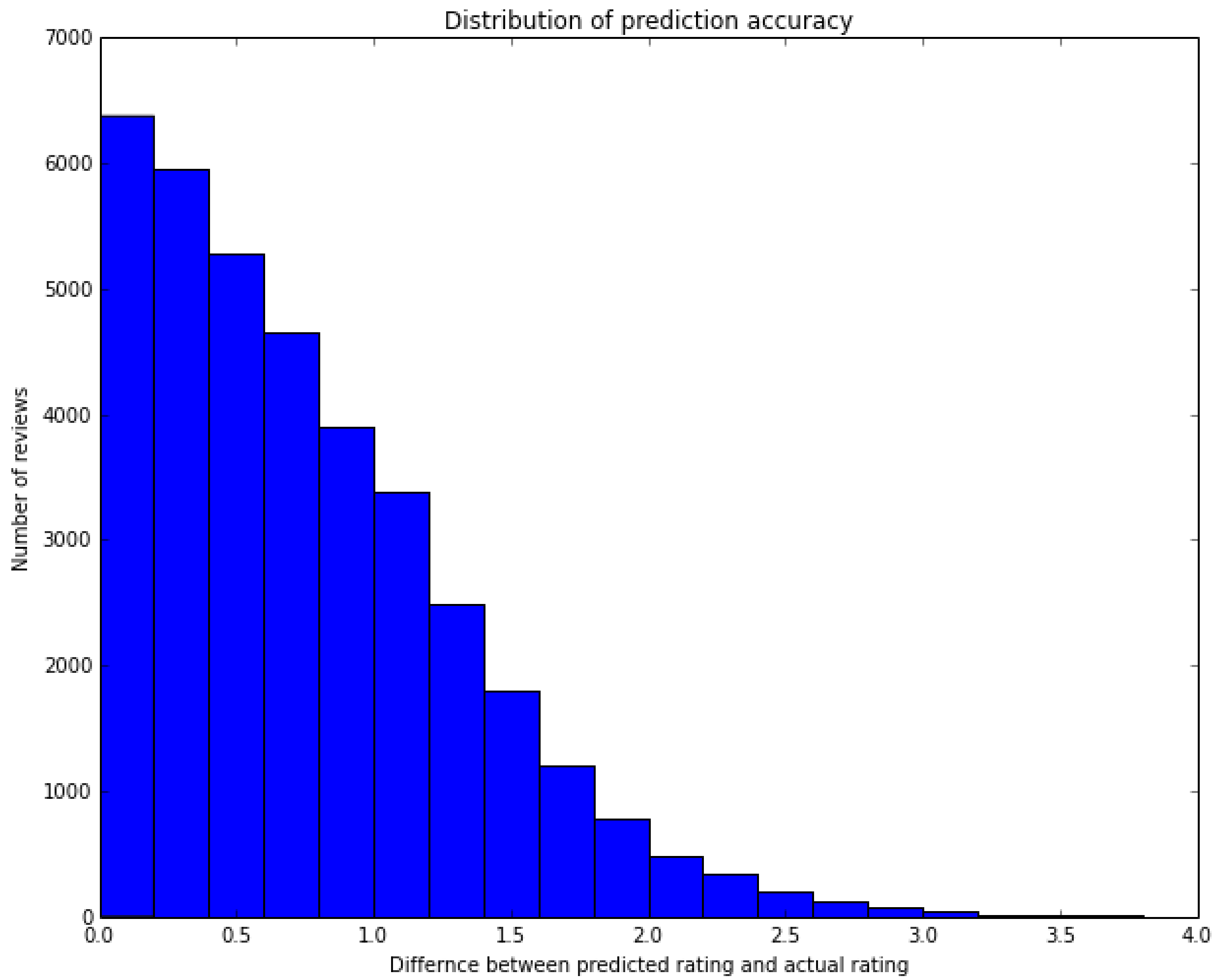
The premise here was to test whether an average sentiment polarity (around 0) corresponds to an average rating (around 3). This plot compares normalized sentiment polarity and normalized average rating. The difference between the 2 is represented on the x-axis, and users were put into bins based on this difference. As expected, the resultant plot took the form of a Gaussian, but the offset in this case is the most critical part.



In order to find the exact offset of this Gaussian from 0, we plotted a Gaussian curve for this data, and found the offset to be -0.07. Thus we concluded that average sentiment does correspond to average rating, for the most part.

Prediction of Review Ratings

Using these first 30 ratings, we proceeded to test how good our predictions of a randomly selected set of future ratings for each given user are, based on the sentiment of current review text and average sentiment and rating of the first 30 reviews.

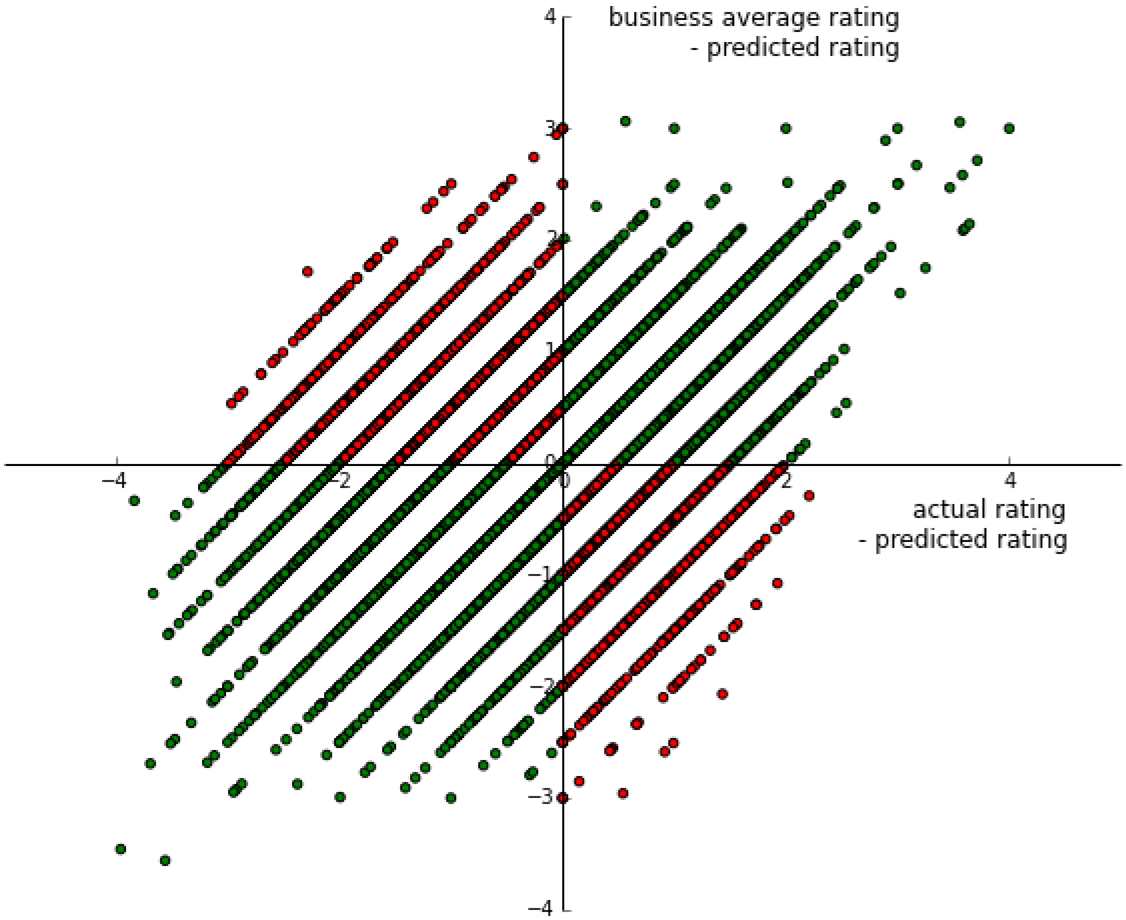


One of the more critical parts of our analysis is presented in this plot. Using the difference between the user's sentiment of a randomly selected future review and the average sentiment for the first 30 reviews, and the average rating doled out by the user for the first 30 reviews, we tried to predict (linearly) the rating given to the current review, bounding it between 1 and 5. The absolute error in this prediction is on the x-axis, and the number of users falling into the corresponding bin on the y-axis.

Just over 70% of the reviews fall into bins where the absolute difference between predicted and actual rating is less than 1.

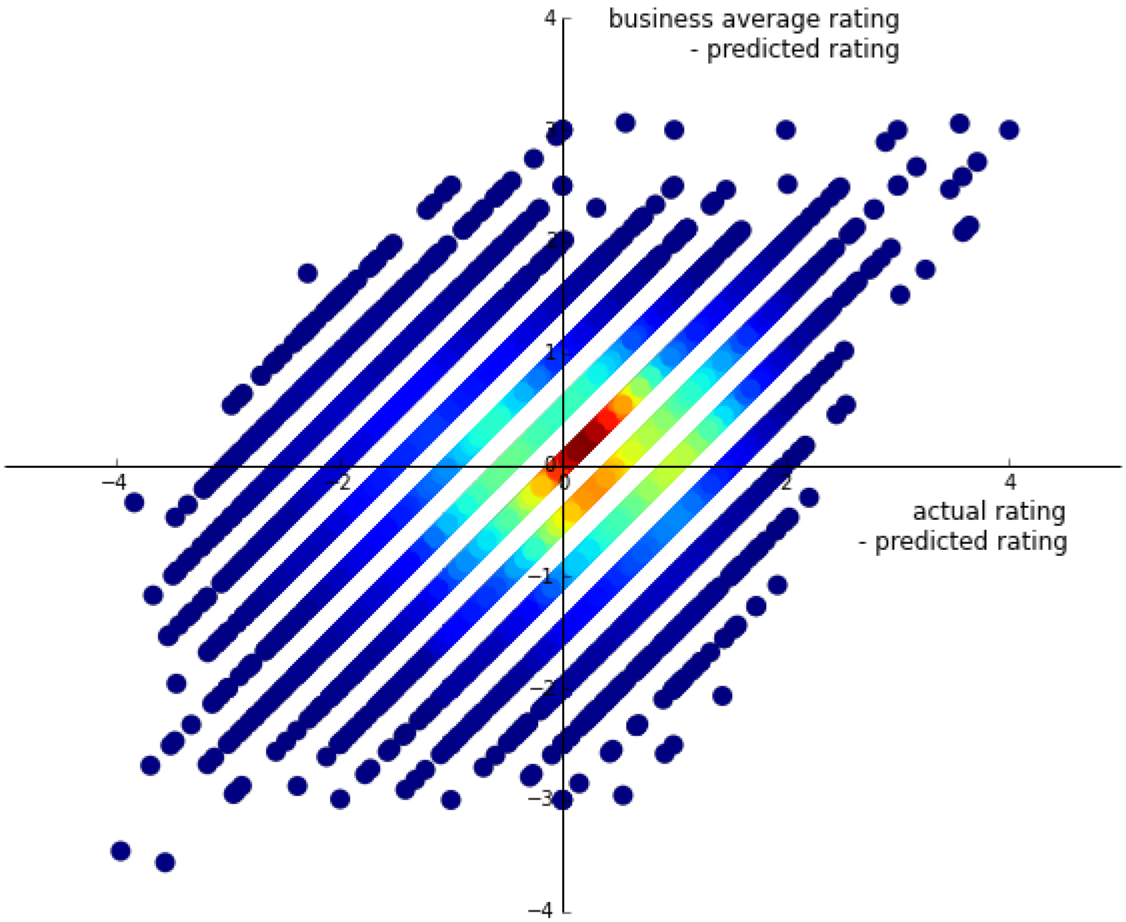
Bias in Ratings

After prediction, we decided to test whether there was an bias involved in a user's rating of a business. Namely, when our predicted rating was inaccurate, was this difference due to a bias caused by the existent average business rating?



According to our hypothesis, if the predicted rating for a review exceeds the actual rating, this is due to a positive bias caused by a higher business average rating (exceeding the predicted rating as well). Thus we created a scatter plot comparing the difference between actual and predicted rating (on x-axis) and difference between business average rating and predicted rating (on y-axis). For this hypothesis to be correct, all plotted points should be in the first or third quadrants (i.e. if actual rating is higher than predicted rating then business rating should be higher than predicted rating as well, and vice versa).

The results in this case were not very conclusive. The green points are the ones that satisfy the above hypothesis, and the red ones do not. All in all, just over 63% of the points fell into the first and third quadrants. We calculated the correlation factor between these 2 features to be 0.34.

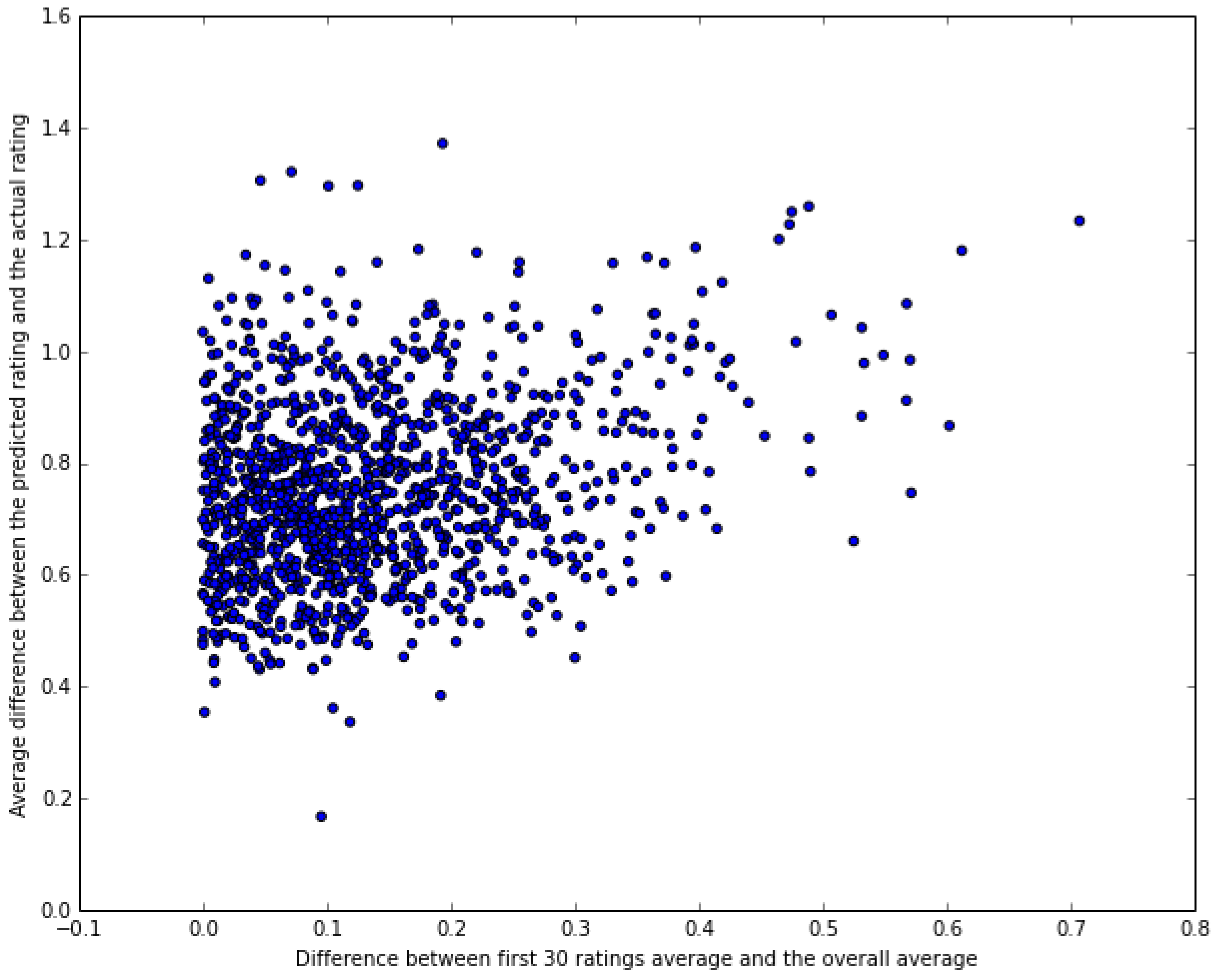


This is the same graph but gives a better idea about density of points. The highest density lies near 0, where actual rating, predicted rating, and business rating converge to the same value.

Our conclusion from this graph is that there must be multiple other factors affecting and biasing a user's review rating away from the predicted rating, not just the business average rating. We expect a model incorporating more of these factors to explain these biases a lot better. Although we see some correlation here, it is relatively weak, and thus the business average rating is a factor but not a dominant one.

Analyzing another possible cause of inaccuracy

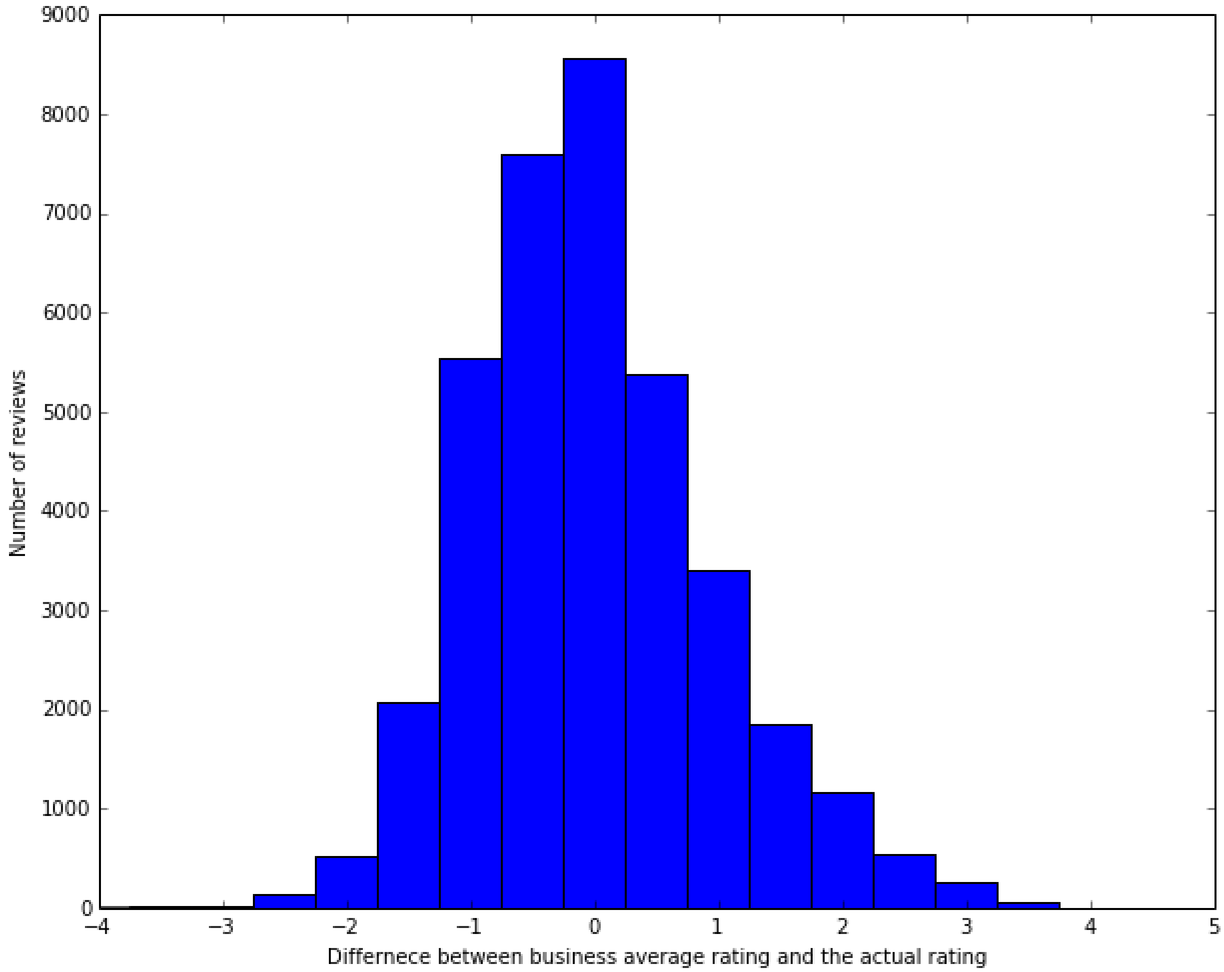
After the previous graph not yielding expected results, we considered another possible cause for inaccuracy in predictions, and tried to correlate users for whom the first 30 reviews do not indicate well future ratings with users for whom we had inaccurate predictions.



Our thought behind this plot was that perhaps it is the users for which the first 30 reviews are not indicative of their future reviews for whom the predictions are inaccurate. However, this was not the case. With a correlation factor of 0.27, this was not a significant factor in prediction inaccuracy.

When users are most likely to review

This was another interesting (albeit unrelated) question we thought to ask: Are users more likely to review when they feel similarly about a business's average rating, or when they think their review would make more of a difference by virtue of the fact that they disagree with the existent average rating?



This graph seeks to give an idea of how closely users rate to an existing business rating. On the x-axis is the difference between the business average rating and a user's rating, and on the y-axis the number of reviews that correspond to a particular difference (these differences are in multiples of 0.5).

Interestingly, the most number of reviews fall in bins where the actual rating by the user ends up being within 0.5 of the business average rating (58% of the total, to be precise). This means that users do tend to rate similarly to the existing business average. Whether this is something the user considers before he decides to write a review, or an influencing factor in his review already is unknown.

Looking Ahead

- Segregate users susceptible to these biases and not
- Develop a more intricate model for either rating prediction or biasing factors

Assumptions/Limitations

- User may adjust for cognitive bias caused by anchoring before they actually write the review, and so the bias is already present in the review text
- Business average is only the latest one, and therefore not the average at the time that the user writes the review. Here we assume that a business average rating will not fluctuate too much after a given point of time, much like our discovery for individual users.
- Sentiment analysis is a flawed technique, and doesn't take into account human expression such as sarcasm
- User is relatively consistent in language and coverage of content between reviews
- Anchoring bias in this case may also depend on the personality of the user, and may not bias all users in the same direction
- We didn't have all reviews for each user in the given dataset. Limited number probably truncated by date.