Rebecca Yang
Pascal Wallish
Principles of Data Science
9 December 2024

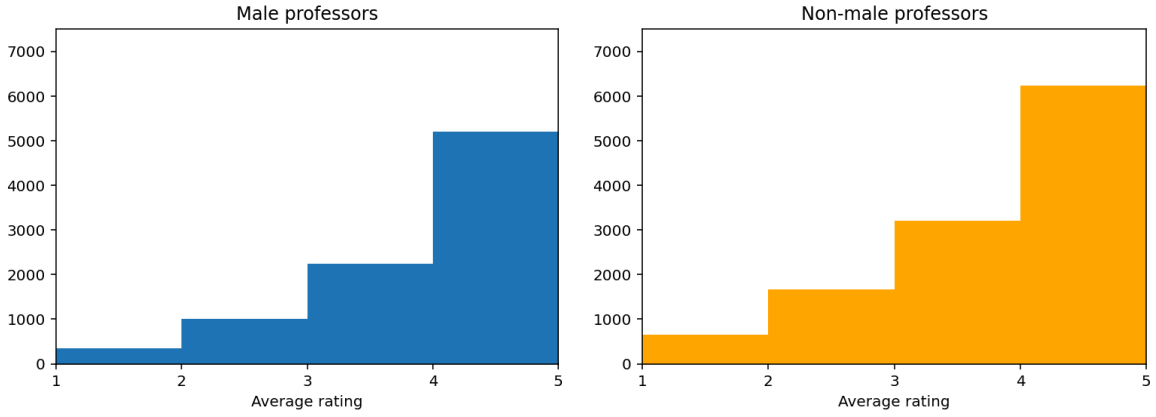
Capstone Project

0. Data cleaning

Because the average rating of a professor with a low number of ratings is likely to not be representative of their true average rating, all entries with a number of ratings less than 10 were removed from the dataset. NaN handling was conducted separately for each question in order to preserve as much data as possible. All rows containing a NaN in each relevant column were removed using NumPy.

1. Activists have asserted that there is a strong gender bias in student evaluations of professors, with male professors enjoying a boost in rating from this bias. While this has been celebrated by ideologues, skeptics have pointed out that this research is of technically poor quality, either due to a low sample size –as small as n = 1 (Mitchell & Martin, 2018), failure to control for confounders such as teaching experience (Centra & Gaubatz, 2000) or obvious p-hacking (MacNell et al., 2015). We would like you to answer the question whether there is evidence of a pro-male gender bias in this dataset. Hint: A significance test is probably required.

First, the distributions of average rating for male and non-male teachers was examined. This was done using the following histograms:



Although the distributions are similar in shape, both of them are heavily skewed, which suggests that a bootstrap confidence interval or permutation test should be conducted. A t-test would not be appropriate as the data is not normally distributed. I chose to take a bootstrap confidence

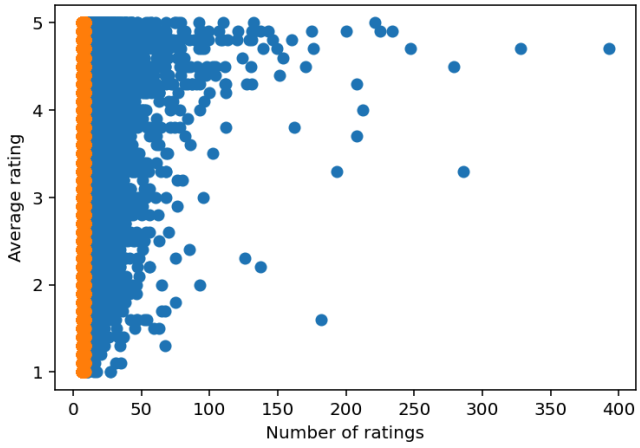
interval with confidence level 0.995. The null hypothesis is that a professor being male does not have a positive effect on their average rating. To control for teaching experience, which I used number of ratings as a proxy for, the data was separated into quartiles based on the number of ratings, and a bootstrap confidence interval was constructed separately for each quartile. This process produced this output:

```
q1 nonmale mean 99.5% CI: ( 3.68708 , 3.77927 )
q1 male mean: 3.87292
q2 nonmale mean 99.5% CI: ( 3.72127 , 3.82358 )
q2 male mean: 3.9193
q3 nonmale mean 99.5% CI: ( 3.75367 , 3.84509 )
q3 male mean: 3.92109
q4 nonmale mean 99.5% CI: ( 3.84129 , 3.92942 )
q4 male mean: 4.01073
```

The upper bound of the non-male mean average rating confidence interval is lower than the male mean average rating for all quartiles, so the null hypothesis is rejected at every level of experience. This shows that there is statistically significant evidence of pro-male bias in this dataset.

2. Is there an effect of experience on the quality of teaching? You can operationalize quality with average rating and use the number of ratings as an imperfect –but available –proxy for experience. Again, a significance test is probably a good idea.

Professors were categorized as “experienced” or “inexperienced” based on whether their number of ratings was higher or lower than the median number of ratings respectively. This was then graphed using a scatter plot:



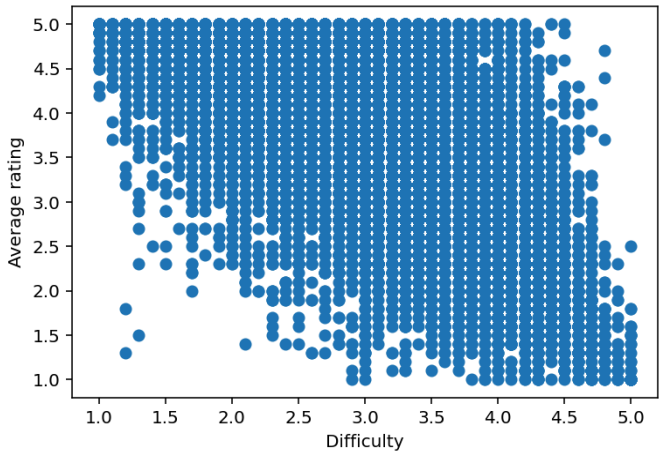
Due to the skewed distribution of average ratings, a bootstrap confidence interval on the mean average rating of inexperienced professors with a confidence level of 99.5% was conducted. The null hypothesis is that there is no difference in average rating based on number of ratings.

```
inexperienced mean 99.5% CI: ( 3.78486 , 3.83694 )
experienced mean: 3.89434
```

The mean rating of the experienced professors is not within this interval. The null hypothesis is rejected. Because the experienced professor average rating mean is above the confidence interval, it can be concluded that experience has a positive effect on the quality of teaching.

3. What is the relationship between average rating and average difficulty?

First, average difficulty was graphed against average rating on a scatterplot in order to determine what type relationship may exist.

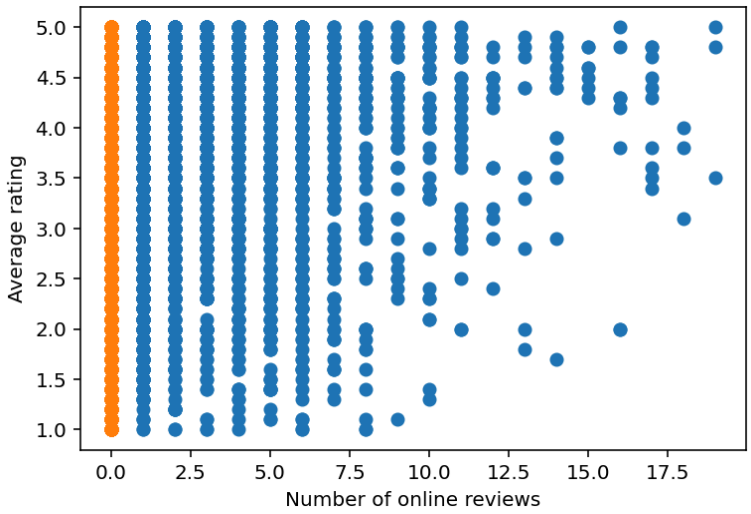


The relationship appears to be linear, so I calculated the correlation coefficient using NumPy, getting -0.62898 as a result (rounded to 5 decimal places). This is a moderate negative correlation, showing that average difficulty has a moderate negative linear relationship with average rating.

4. Do professors who teach a lot of classes in the online modality receive higher or lower ratings than those who don't? Hint: A significance test might be a good idea, but you need to think of a creative but suitable way to split the data.

I am interpreting "a lot" of classes as a large number of online classes rather than a large percentage of their classes being taught online. I am also using the number of online ratings as

a proxy for the amount of online classes taught. The data was split into “online” and “non-online” professors by comparing each professor’s number of ratings from online classes against the median number of ratings from online classes. The null hypothesis is that professors who have a large number of ratings from online classes do not differ in terms of average rating from professors who do not have a large number of ratings from online classes. Number of online reviews was graphed against average rating in a scatterplot:



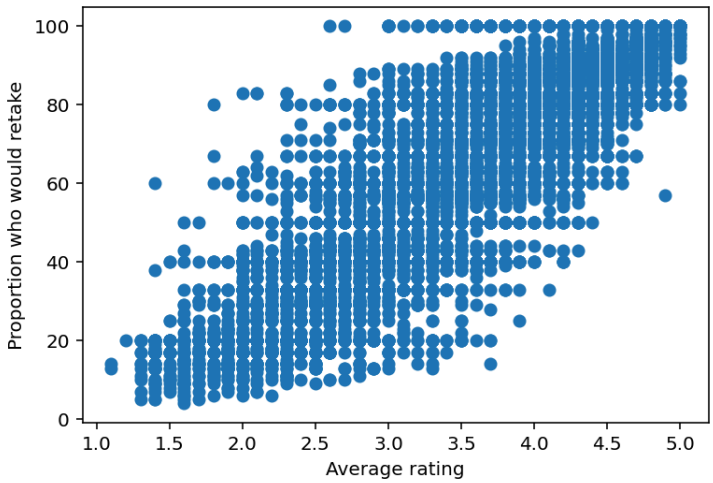
The distribution of average ratings among these groups is also skewed, so a bootstrap confidence interval on the mean average rating of online professors with confidence level 99.5% was calculated and compared against the mean average rating of non-online professors:

```
online mean 99.5% CI: ( 3.8042 , 3.88301 )
non-online mean: 3.85308
```

The non-online professor average rating mean is within the confidence interval, so the null hypothesis is not rejected, and it cannot be concluded that professors who teach a lot of classes online receive higher or lower ratings than those who do not. Given that teaching a large number of classes online has no effect on average ratings, the mean average rating of non-online professors is possible purely by chance alone.

5. What is the relationship between the average rating and the proportion of people who would take the class the professor teaches again?

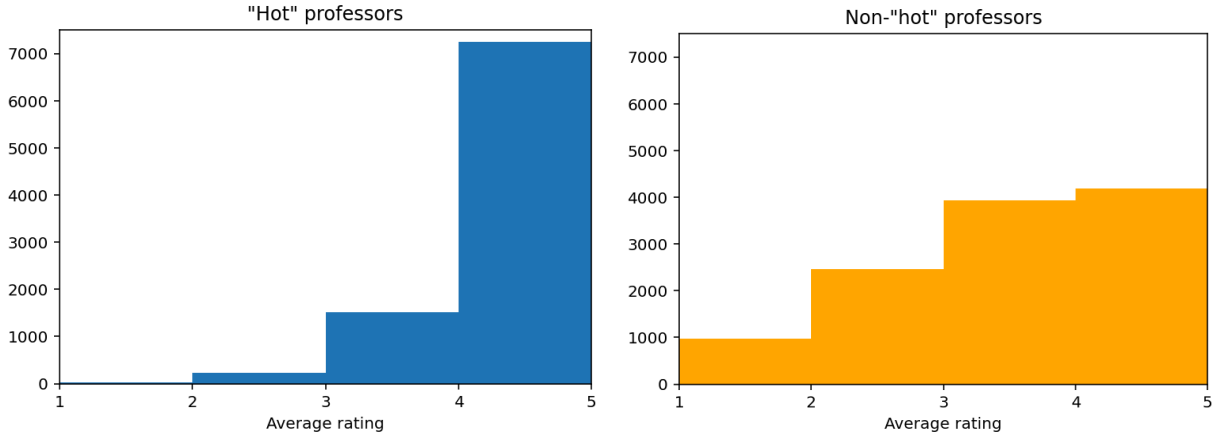
First, graph average rating was graphed against the proportion of students who would take the class again.



This relationship appears to be linear. The correlation coefficient was calculated using NumPy, getting 0.87930 as a result (rounded to 5 decimal places). This strong positive correlation means that average rating and the proportion of students who would take a class has a strong positive linear relationship.

6. Do professors who are “hot” receive higher ratings than those who are not? Again, a significance test is indicated.

The average rating distribution of professors who are “hot” and not “hot” were graphed using histograms.



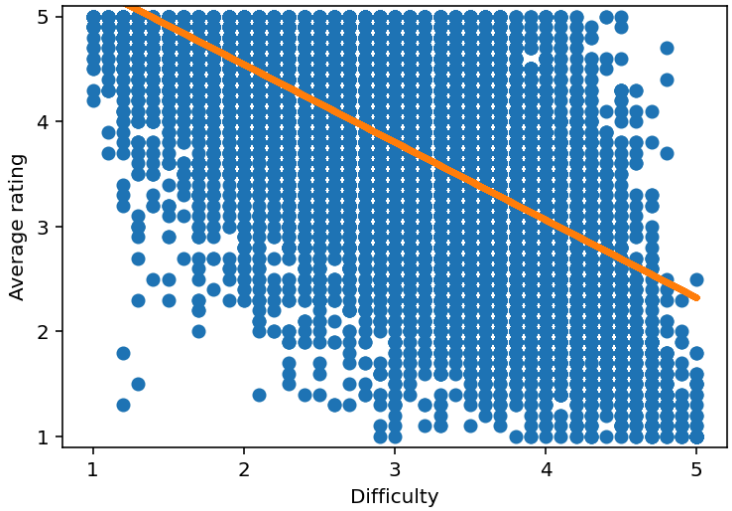
From these, it can be seen that the average rating distribution for both groups is heavily skewed, so a bootstrap confidence interval with a 99.5% confidence level was conducted on the mean average rating of non-”hot” professors. The null hypothesis is that attractiveness of the professor has no effect on their ratings.

```
non-hot mean 99.5% CI: ( 3.43329 , 3.48217 )
hot mean: 4.35438
```

The “hot” professor average rating mean is much lower than the lower bound of the confidence interval. I reject the null hypothesis. From these results, it can be concluded that attractiveness has a statistically significant effect on average ratings.

7. Build a regression model predicting average rating from difficulty (only). Make sure to include the R2 and RMSE of this model.

First, difficulty was graphed against average rating. The relationship between the two appears to be linear, so a linear regression is fit, using a sample of 60% of the data for training, and saving the remaining 40% of the data for testing.



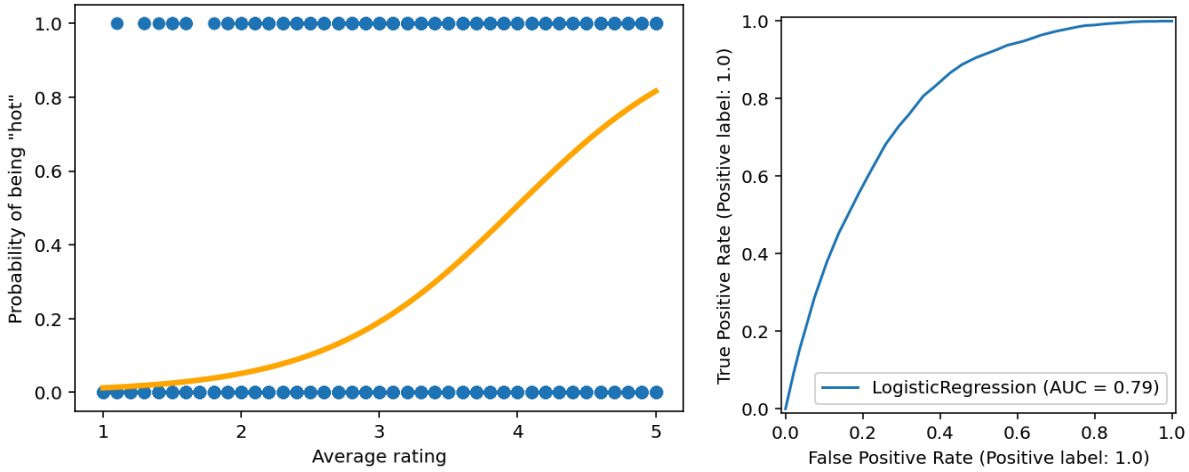
As difficulty increases by 1 unit, the average rating decreased by 0.74070 units. The R^2 of this model is 0.39211 and the RMSE is 0.72301 (both rounded to 5 decimal places). This means that 39.211% of the variation in the data is explainable by this model, which is not very high. The RMSE is also pretty high, showing that this model is not very accurate.

8. Build a regression model predicting average rating from all available factors. Make sure to include the R2and RMSE of this model. Comment on how this model compares to the “difficulty only” model and on individual betas. Hint: Make sure to address collinearity concerns.

Because some predictors are correlated (as seen in previous questions), a multiple linear Lasso regression was fit to the data using NumPy. Like the previous linear regression, 60% of the data was randomly sampled for use in training and the other 40% was used for testing. This yielded a R^2 of 0.77150. This means that 77.150% of the variance in data was explained by this model. It also has a RMSE of 0.40423, which is much lower than the previous model. The coefficients for the predictors were all 0 except for “hot”-ness, which has a beta of 0.02795. All numbers were rounded to the 5th decimal place.

9. Build a classification model that predicts whether a professor receives a “pepper” from average rating only. Make sure to include quality metrics such as AU(RO)C and also address class imbalances.

A logistic regression was fit to the data using the same cross-validation process as the linear regression models from questions 7 and 8. Class imbalances were addressed using the ‘balanced’ class weight setting for sklearn logistic regression. The mean accuracy of this model was 0.71510 and the AU(RO)C was 0.79, both of which are moderately strong.



10. Build a classification model that predicts whether a professor receives a “pepper” from all available factors. Comment on how this model compares to the “average rating only” model. Make sure to include quality metrics such as AU(RO)C and also address class imbalances.

A multinomial logistic regression was fit to the data using the process described in question 9. The mean accuracy of this model was 0.71629 which is slightly better than the single logistic

regression, but not much. The AU(RO)C was similarly slightly higher than the previous model. This shows that this model is marginally better at predicting whether a professor received a pepper based on their average rating.