Capstone Project:
Assessing Professor Effectiveness (APE) with Data Science Methodologies

Authors: Bella Chang*, Kristina Fujimoto*, Emily Wang*

P. Wallisch (DS-GA 1001)

Dec. 9, 2024

**New York University, Center for Data Science*

Project Overview

This independent analysis uses RateMyProfessor.com data to explore potential gender bias in student evaluations of professors. The goal was to apply statistical and machine learning methods to assess how gender, perceived difficulty, and teaching tags influence overall professor ratings, and to explore whether appearance (pepper status) plays a role. All code can be found in the corresponding .py file in this repository.

Data Preprocessing

To preprocess the dataset, we began by merging all three datasets together, recognizing that each row corresponds to one professor. Since the column representing the proportion of students who would retake a class had approximately 86% of the values missing, we removed this column to preserve as much data as possible. We then removed null values row-wise because we found that all of the null values in the other columns were located in the same rows.

To clean the data, we also removed rows where “Male Prof?” and “Female Prof?” contained the same value (both 0 or 1), where the gender of the professor was unclear to avoid ambiguity in analysis. Next, to account for the discrepancy of “Average Rating” being more meaningful if a professor has more ratings, we used the median number of ratings (three) as a threshold (as opposed to mean, which is less robust to outliers), removing all rows where the professor's number of ratings is less than three. We decided to only accept data points where the professor’s number of ratings meets or exceeds the median to ensure that the average ratings are based on an adequate number of ratings, resulting in reliable, meaningful insights that are more representative of the data.

Lastly, to account for the imbalance of professors with more ratings having more tags, we normalized the tags data by dividing the number of tags for a given tag by the total number of tags that specific professor received, which scales all tag values between 0 and 1. While this results in professors with a small number of ratings receiving a much higher tag value for tags they are awarded, we found that this method offered an optimal tradeoff by effectively balancing the influence of the number of ratings for each professor. This ensures that the results of our analysis are not dominated by those with a higher number of ratings, while still preserving the relative frequency of tags for each professor. Lastly, we seeded our random number generators and models (using the random state parameter).

SECTION 1 – ANALYSIS

1(a): Evidence of Pro-Male Gender Bias in the Dataset

To evaluate whether there was evidence of a pro-male gender bias in the dataset, our group chose to conduct a null hypothesis significance test on the "average ratings" metric within our dataset, filtered by male and female professors.

We first observed the distribution of average ratings for male and female professors in Fig. 1. We also plotted the median for each distribution so we could see the middle of the ranges for each distribution without considering extreme values as another way of comparing their distributions. This gave us an initial intuition that the general shape of the distributions for each gender group was relatively similar, although it can be seen that the male professor average ratings gets considerably higher for larger average ratings.

For this question, our null hypothesis is that there is no pro-male gender bias in this dataset, and therefore any difference we see is due to chance. Our alternative hypothesis is that there is a pro-male gender bias in this dataset. As we are dealing with ratings data which is subjective in nature, we cannot confidently reduce our data to sample means, so we chose to conduct a one-tailed Mann-Whitney U test to observe the difference of central tendencies between the male and female groups. To conduct this test, we assumed sampling was conducted independently and randomly, our sample size was large, and the general shape of each distribution was relatively the same (as per our EDA).

Our test yielded a result of approximately 0.0000206, which is well below our p-value of 0.005. Thus, we have decided that there is a statistically significant difference between average ratings of male and female professors; in other words, our data supports the claim that there is a pro-male gender bias in this dataset.

1(b): Gender Difference in Ratings Distribution Spread

To evaluate if there is a gender difference in the spread of the ratings distribution, our group chose to run a null hypothesis significance test using the variances of the filtered male and female "average ratings" metric within our dataset.

Specifically, to target the key metric of the problem, we chose to run a permutation test with the absolute difference of the male and female variances on average rating as our test statistic. To run this test, we assumed that our actual data was drawn independently from a representative and large sample.

For this question, our null hypothesis is that there is no gender difference in the spread of the ratings distribution, and therefore any difference we see is due to chance. Our alternative hypothesis is that there is a gender difference in the spread of the ratings distribution.

We decided to find the absolute difference, rather than simply a difference, because we were interested in the general gender difference of the two distributions (no matter the direction). We conducted our permutation test with an arbitrary n=10,000, representing the number of times

we bootstrapped our data under the null hypothesis. We then found our p-value by counting the number of bootstrapped absolute variance differences that were greater than or equal to our observed value and dividing by 10,000: as per definition of a p-value, this is the literal implementation of the probability of observing the data we saw, or anything greater, under the null hypothesis.

Our test yielded a p-value of 0.0, with the observed absolute difference of variance being approximately 0.1, which can also be seen in Fig. 2. This p-value is well below our significance level of 0.005. Thus, we decided that there is a statistically significant gender difference in the spread of the ratings distribution in our dataset.

1(c): Gender Difference in Student-Allocated Tags

To evaluate if there is a gender difference in the tags awarded by students, our group chose a null hypothesis that there is no gender difference between male and female professors for any tag. The null hypothesis would test using the medians of the filtered male and female metrics for each tag, as tag data cannot be reduced to a means.

To start, we chose to run a Kolmogorov-Smirnov test on each of the 20 tags, divided by whether it was a female or male professor rating. This test was intended to determine whether the distributions of the tags between our two groups, male and female professors, was the same and we had a null hypothesis that they were. All the tags had p-values > 0.005 except for “Pop quizzes!” and their distributions were plotted in Figures 5-24, which means the majority were not distributed normally, but we decided to run the Mann Whitney U test on all 20 tags since the data cannot be reduced to a means.

Our null hypothesis was then tested with a Mann Whitney U test, which we ran individually on each tag for female and male professor data by comparing the medians. Based on the p-values, “Pop quizzes!” was the only one that had a p-value > 0.005 (0.0214) while all the other tags had significant p-values as seen in Fig 25 with all bars below the red dashed alpha of 0.005 line being significant. The three most gendered (lowest p-value) were “Hilarious”, “Amazing Lectures”, and “Lecture Heavy” while the three least gendered with a significant p-value (highest p-value) were “Tough Grader”, “Clear grading”, and “Don’t skip class or you will not pass”. From this, we can see that one gender receives more tags about the quality of the lectures than the other and that there isn’t a significant difference in gender about grading.

From the male and female professor data, we also calculated the Cohen’s d value for each tag. This revealed that although 19/20 of the tags do have a significant difference in gender, as indicated by p-values smaller than 0.005, the effect size of this difference is very small as indicated by most d values being smaller than 0.2. Thus, we decided that there is a significant difference in gender for the 19 tags with a p-value < 0.005 (except for “Pop quizzes!”), but this difference has a small effect.

1(d): Gender Difference in Average Difficulty

To evaluate if there is a gender difference in the average difficulty, our group chose a null hypothesis that there is no gender difference between male and female professors in terms of difficulty. The null hypothesis would test using the mean of the filtered male and female metrics for the average difficulty.

To clean the data, we had already removed all rows where “Male Prof?” and “Female Prof?” were equal to each other (both 0 or both 1), as to remove ratings where the gender of the professor was unclear. We started by testing the distributions of both groups (female and male professors' difficulty) by running a KS test, with a null hypothesis that they had the same distribution. The p-value was 0.99 which is much greater than our cutoff of 0.005, meaning that the p-value is not significant and the genders are normally distributed. You can also observe that their shapes are approximately normal in Fig. 26. Then we checked the variances of the genders by using bootstrapping.

Our null hypothesis for this test was that there is no gender difference in the spread of the difficulty distribution, and any difference observed is due to chance. Similar to question 2, we found the absolute difference because we were interested in the general gender difference for both distributions. The permutation test had an n of 10,000, representing the amount of times we bootstrapped under the null hypothesis. The p-values were calculated by the amount of absolute variance differences that were greater than or equal to the observed value, divided by 10,000. Our test yielded a p-value of 0.6885, with the observed absolute difference of variance being 0.0036, which can also be seen in Fig. 27. This p-value is much higher than our significance level of 0.005. Thus, we decided that there is not a statistically significant gender difference in the spread of the difficulty distribution in our data.

We then proceeded with running an independent sample t-test on the difficulty data, as is appropriate for normally distributed in both genders and similar variance data. The average difficulty data can also be reduced to a means, as done in independent sample t-test, because the difficulty ratings data has already been averaged. Our t-test provided a p-value of 0.887 and a t statistic of 0.142 which indicates that there isn’t a significant gender difference in average difficulty, as seen in Fig. 28.

After determining there is no gender difference in difficulty of professors, we determined the effect size by calculating Cohen’s d which was 0.001. This further indicates that there is no significance in the difference between male and female professors’ difficulty.

1(e): Comparing Student Tags in NY vs FL

For our own interest, we explored whether students in NY are more likely to award positive tags (e.g., “Caring,” “Inspirational”) than students in FL. We selected FL for comparison because it had a similar number of ratings to NY, making it a suitable benchmark. We addressed this question using a null hypothesis significance test and effect size analysis.

We began by categorizing tags as positive or non-positive within our dataset, then filtered for ratings from NY and FL. For each row, we calculated the mean of the positive tags, adding

this as a new column. It's important to note that this step further granularized the data, which could influence the hypothesis tests. However, given the dataset's imbalance and limitations, we deemed this approach appropriate. We also visualized the distributions and medians of positive tag means for both states (Fig. 39).

Our null hypothesis posits that NY and FL students are equally likely to award positive tags, with any observed differences due to chance. The alternative hypothesis suggests NY students are more likely to award positive tags. Since we compared two independent sample means with similar variances (something we noticed within our EDA), we conducted a one-tailed independent t-test. This choice assumed our data was randomly and independently sampled from a representative and sufficiently large population. Additionally, we used Cohen's d to measure the effect size of the difference between the two groups.

The t-test yielded a p-value of ~0.977, indicating no statistically significant difference between positive tags awarded in NY and FL. Thus, our data supports the null hypothesis. The effect size was approximately -0.064, suggesting that, practically, FL students might slightly favor positive tags more than NY students. However, this effect is small and therefore not practically meaningful.

Effect Sizes of 1(a) & 1(b)

To evaluate the effect size of gender bias in average ratings and the spread of average ratings, we calculated Cohen's d and an adapted metric for variance differences. These calculations used the "average rating" metric, filtered by male and female professors as two groups.

Our implementation of Cohen's d measured the difference between the means of male and female ratings, divided by their pooled standard deviation. For the spread effect size, we calculated the absolute difference between the variances of male and female ratings, divided by 2, adapting Cohen's d to fit our analysis without additional assumptions.

To estimate confidence intervals, we performed bootstrapping with 10,000 resamples, calculating both metrics in each round. The 2.5% and 97.5% percentiles of the bootstrapped values provided the 95% confidence intervals, while the means of the resampled metrics gave the final effect size estimates.

The results showed an effect size of ~0.064 for average ratings (95% CI: [0.041, 0.087]) and ~0.050 for variance differences (95% CI: [0.035, 0.066]), visualized in Figs. 3 and 4. Both represent small effect sizes, suggesting that while the differences are statistically significant (as shown in earlier questions), their practical impact is minimal.

Effect Size of 1(d)

To quantify the effect size of gender difference in average difficulty, we calculated Cohen's d which used the average difficulty divided by male and female professors. The function we used to calculate Cohen's d measured the difference between the means of male and female

difficulty, divided by the groups pooled standard deviation. To estimate the confidence intervals, we then did bootstrapping with 10,000 resamples and calculating Cohen's d in each resample. The 95% confidence interval was made of the 2.5% and 97.5% percentiles of the bootstrapped values and the means of each resample was the effect size estimates, Cohen's d.

Our results showed an average effect size of 0.00953 in the gender difference of difficulty for an average difficulty 95% confidence interval of [0.000383, 0.0265], as seen in Fig. 29. This represents a very small effect size, which further shows that the difference in difficulty between female and male professors is not significant, and the effect is small and not significant as well.

SECTION 2 – MODELING

2(a): Regression Model of Average Rating (only from numerical predictors)

To predict average rating from numerical predictors, we chose to use Ridge regression because it accounts for multicollinearity and improves generalization (i.e. prevents model from overfitting) by shrinking coefficients with a regularization term. Since the columns indicating "Female prof?" and "Male prof?" are highly correlated as depicted by the correlation heatmap in Fig. 30, we removed one of the two columns ("Male prof?") from the numerical predictors to prevent multicollinearity issues. In the initial preprocessing steps, we also removed the column indicating "proportion of students that said they would take the class again" from the numerical predictors because greater than 80% of the data was missing in that column and we wanted to keep as many data points as possible for more representative results.

Using RidgeCV (from the scikit-learn library), we iteratively tested multiple alpha values and used cross-validation with five folds, fitting a ridge regression model for each alpha on the training folds and testing on the validation fold to choose the best alpha (alpha=1), corresponding with the lowest average mean squared error across all fold splits. The model is refitted, with the best alpha value, on the training data, which we then used to predict values for the target variable ("Average Rating") and calculate corresponding R^2 and RMSE values.

The R^2 value of our model was 0.429 while the root mean squared error (RMSE) was 0.741. This suggests that the model explains 42.9% of the variance in the "Average Rating" target variable. We found that "Average Difficulty" is most strongly predictive of "Average Rating" with a beta coefficient of -0.514, and is negatively correlated, as visualized in Fig. 31. This suggests that classes perceived as more difficult are more likely to result in the professor having a lower average rating. Furthermore, "Received a Pepper?" is the second most strongly predictive factor with a coefficient of 0.293, suggesting that professors that are judged as "hot" by students are more likely to have a higher average rating. Lastly, the gender of the professor ("Female Prof?" column) is the third most strongly predictive factor but since it has a comparatively low beta of -0.03, it has a much weaker correlation with "Average Rating".

2(b): Regression Model of Average Rating (only from tags)

To predict average rating from all tags, we chose to use Ridge regression because it accounts for multicollinearity and prevents the model from overfitting by shrinking coefficients. We also made sure to check the correlation between features as seen in Fig. 32. We decided that all the features were reasonably uncorrelated for our purposes – most absolute correlations between features/tags were consistently low (within the range of [0, 0.2]), informing our decision to keep all features. To account for the imbalance of professors with more ratings having more tags, as described in the preprocessing step, we used the normalized tag values (found by dividing the number of a specific tag by the total number of tags the professor received for each tag).

We used the same process described in 2(a) with RidgeCV, using k-fold cross validation to find the most optimal alpha value of 10. Our Ridge regression model produced an R^2 value of 0.655 and a RMSE value of 0.576. Since the model explains 65.5% of the variance in the "Average Difficulty" variable (i.e. higher than the model from 2(a)) and the RMSE is lower compared to 2(a), this model using the tags as predictors performs better than the model using numeric predictors.

We found that the "Good feedback" tag is most strongly predictive of "Average Rating" with a beta coefficient of 0.285, as visualized in Fig. 33. Furthermore, "Respected" and "Amazing lectures" tags are the second and third most strongly predictive factors of "Average Rating" with beta coefficients of 0.230 and 0.223, respectively, suggesting with positive correlations, that professors who received more of these tags are more likely to have a higher rating. Lastly, tags such as "Lots of homework" and "Group projects" had the smallest beta coefficients in magnitude, with values of -0.0035 and -0.00231 respectively, suggesting that they have relatively weak correlations with "Average Rating".

2(c): Regression Model of Average Difficulty (only from tags)

To predict average rating from all tags, similarly to question 7 and 8, we chose to use Ridge regression because it accounts for multicollinearity and prevents the model from overfitting by shrinking coefficients. Following the same reasoning from question 8, to account for multicollinearity concerns, we checked the correlation between features/tags (Fig. 32) and concluded that all the features were reasonably uncorrelated for our purposes (within the range of [0, 0.2]), which informed our decision to keep all features. Similar to question 8, we used normalized tag values (found by dividing the number of a specific tag by the total number of tags the professor received for each tag).

We used the same process described in question 7 with RidgeCV, using k-fold cross validation to find the most optimal alpha value of 100. Our Ridge regression model produced an R^2 value of 0.473 and a RMSE value of 0.618. This indicates that our model explains 47.3% of the variance in the "Average Difficulty" variable.

As visualized in Fig. 34, we found that the "Tough grader" tag is most strongly predictive of "Average Difficulty" by far, with a beta coefficient of 0.338. This indicates that professors with a tough grader tag correlates with having a higher average difficulty rating. The "Clear grading" and "Hilarious" tags are the second and third most strongly predictive factors of "Average Difficulty" with beta coefficients of -0.113 and -0.093, respectively, indicating that professors with more clear grading and hilarious tags are likely to have a smaller average difficulty rating. Lastly, tags such as "Pop quizzes!" and "Lecture heavy" had the smallest beta coefficient values of 0.0097 and 0.0062 respectively, suggesting that they are weakly correlated with "Average Difficulty".

2(d): Classification Model of Receiving a "Pepper" (from all factors)

To predict whether a professor receives a pepper from all factors, tags and numerical data, we chose to use Logistic Regression because it provides a nonlinear model that connects the predictors and the outcomes, and is good for predicting binary values (which is the receives a pepper column). We used the same cleaned data as in previous questions which normalized tag values (dividing number of tags by total number of tags).

To start, we checked the class balance of the binary column “Received a pepper?” and saw that there was quite a bit of class imbalance: 62% of the values were no, while only 38% of the values were yes, as shown in Fig. 35. We then removed any non-numerical data such as “University” and then split the pepper data into 60% training and 40% testing which we then used to train and test our Logistic Regression model. The accuracy was 70.93% and our AU(RO)C was 0.774 as shown in Fig. 36. We used the AU(RO)C as a better metric for accuracy as there is class imbalance present, as well as metrics like sensitivity (0.576), specificity (0.792), precision (0.632), and NPV (0.751) which can be seen in the confusion matrix in Fig. 37. The AU(RO)C is better in our scenario because it analyzes more possibilities of classification thresholds which takes into account true positive and false positive rates (unlike in simple accuracy). This is also why the other metrics are better for measuring model performance than just accuracy.

In addition, we decided to find an optimal threshold for the model which balances sensitivity and specificity, to see whether this would improve the metrics and prediction accuracy. Our optimal threshold was 0.369 which yielded the same AU(RO)C of 0.774 as seen in Fig. 38, reduced the accuracy to 69% , but adjusted the other metrics to: sensitivity (0.776), specificity (0.637), precision (0.569), and NPV (0.821). While the sensitivity and specificity changed, the AU(RO)C stayed the same so the optimal threshold provides similar results to the original threshold. We then tried bootstrapping by resampling the no pepper (majority data), 10 times randomly, then training and testing the Logistic Regression model each time, to deal with the class imbalance. Each time returned the same AU(RO)C of 0.774. This shows that while our AU(RO)C of 0.774 isn’t ideal, the accuracy will not improve by dealing with class imbalance through bootstrapping or optimal thresholds.