Capstone Project

Nirvan Mahesh Shukla

January 7, 2025

1 Preamble

My name is Nirvan Shukla and I worked on the Capstone Project as a solo group. I used my N-Number for the random seed, which I generated by running `rng = np.random.default_rng(seed = N_number)`. When it was used, it was often used as `rng` in place of `random`. For the data pre-processing, I decided it would be easier to use one succinct data-frame, which I would be able to parse through as needed for the problem set. I loaded in each CSV using the Pandas library, and took a look at each dataset, when I ran into a problem: the column names for each dataset were represented by the first row of data. So, in order to clean this up, I created a quick function `fix_df` to take in each data-frame, and shift the column row, and every subsequent row down by one unit. I verified the integrity of the data-frame by checking to make sure the number of rows matched the spec sheet, and then I made sure to rename the columns with their appropriate monikers. I also made sure to convert every column with numerical data into a float data type, so that I would be able to run various hypothesis tests, and regression models on them. Since each teacher has a different amount of ratings, simply looking at the average rating would be an unfair way to judge a professor. In order to deal with the average rating, I looked at a few ways of dealing with the data. I thought about setting an arbitrary threshold for the number of ratings to look at, but I saw that the 25th percentile for ratings was 1 rating, which would have been the minimum to consider anyway. So, I decided to try a weighted average instead of just eliminating data. The weighted average calculation was used to adjust the average rating for professors to account for variability in the number of ratings. Without adjustment, professors with very few ratings might have extreme averages (e.g., very high or very low) that are less reliable due to insufficient data. This approach ensures that professors with a larger number of ratings contribute more weight to their average than those with fewer ratings, which stabilizes the estimate and reduces the influence of outliers. The formula combines the professor's observed average rating, weighted by the number of ratings they received, with the global average rating, weighted by a fixed constant k. The constant k acts as a regularization parameter, balancing the influence of the global average against the professor's individual average. For professors with very few ratings, the adjusted average leans more toward the global mean, reflecting lower confidence in their observed average. For

professors with many ratings, the adjusted average closely aligns with their observed average, reflecting higher confidence in the reliability of the data. This method was used to create a fairer and more robust comparison of ratings across professors, ensuring that the analysis was not disproportionately influenced by ratings with low sample sizes. We used a k value of 3 because that represented the median of the number of ratings. This same methodology was applied to the average difficulty. Therefore, for the problem set, an adjusted rating and adjusted difficulty rating was used to answer the questions.

2 Problem 1

To assess whether there is a statistically significant difference in adjusted ratings between male and female professors, I conducted an independent t-test. The null hypothesis for this analysis was that there is no difference in the mean adjusted ratings between male and female professors. Since I am using adjusted ratings data, I considered it to be interval level data and not purely ordinal, therefore, a T test would be valid. The test revealed a T-statistic of 6.91 and a p-value of 5.05×10^{-12}. This p-value is far below the significance threshold of 0.005, leading us to reject the null hypothesis. The results indicate that there is a statistically significant difference in the adjusted ratings between male and female professors. The high T-statistic signifies that the difference in mean adjusted ratings between genders is substantial, with male professors receiving significantly higher adjusted ratings compared to female professors. These findings suggest strong evidence against the null hypothesis, supporting the conclusion that gender is associated with differences in adjusted ratings in this dataset.

3 Problem 2

To investigate whether there is a statistically significant difference in the distribution of adjusted ratings between male and female professors, I conducted a Kolmogorov-Smirnov (KS) test. The null hypothesis for this analysis was that the distributions of adjusted ratings for male and female professors are identical. The KS test was chosen because it is a non-parametric test that compares the spread of two datasets. The test produced a KS Statistic of 0.0240 and a p-value of 1.74×10^{-7}. This p-value is below the alpha threshold of 0.005, leading to the rejection of the null hypothesis. This leads us to say there is a statistically significant difference in the distribution of adjusted ratings between male and female professors. The KS Statistic of 0.0240 suggests that the maximum difference between the cumulative distribution functions of the two groups is small but statistically significant due to the large sample size. Overall, the findings reveal that male and female professors have statistically different distributions of adjusted ratings.

4 Problem 3

To evaluate the likely size of the effects observed in the analyses of gender differences in adjusted ratings (Question 1) and the distribution of adjusted ratings (Question 2), I calculated effect sizes along with their 95% confidence intervals. For Question 1, Cohen's d effect size was calculated to measure the magnitude of the difference in mean adjusted ratings between male and female professors. The result was Cohen's $d = 0.0582$ with a 95% confidence interval of $(0.0347, 0.0826)$. This indicates a very small effect size, suggesting that while the difference in adjusted ratings is statistically significant, its actual impact is quite small. The confidence interval confirms the effect, since it does not include zero. For Question 2, I used Cohen's f^2 effect size to quantify the significance of the difference in the distribution of adjusted ratings for the KS test. The Cohen's f^2 was calculated as -0.0441, which also suggests a negligible effect size. Additionally, the 95% confidence interval for the KS Statistic was $(0.0158, 0.0362)$. This small KS Statistic, coupled with the small f^2, supports the conclusion that while the distributional differences between male and female professors are statistically significant due to the large sample size, the practical significance of these differences is minimal. Overall, while both analyses suggest statistically significant gender differences in adjusted ratings and their distributions, the effect sizes indicate that the practical implications of these differences are limited.

5 Problem 4

To evaluate whether there are statistically significant gender differences in the tags awarded to professors, I conducted permutation tests for each of the 20 tags in the dataset. The null hypothesis for each test was that the distribution of a given tag is identical for male and female professors. A permutation test was chosen because it is a non-parametric method that does not rely on assumptions about the distribution of the data, making it ideal for binary data like these tags. The results revealed several tags with statistically significant differences between genders, with p-values below the significance threshold of 0.005. The top three most gendered tags, as measured by the smallest p-values and largest observed differences, were 'Hilarious' Observed Diff = 0.4136, $p = 0.0000$), 'Amazing lectures' Observed Diff = 0.2280, $p = 0.0000$, and 'Respected' Observed Diff = 0.2414, $p = 0.0000$. Male professors were significantly more likely to be tagged with these attributes than female professors. In contrast, the least gendered tags, based on their larger p-values, were 'Tough grader' Observed Diff = 0.0212, $p = 0.1363$, 'Lots of homework' Observed Diff = -0.0219, $p = 0.0865$, and 'Lots to read' =Observed Diff = 0.0232, $p = 0.0447$. For these tags, the differences between genders were small and not statistically significant in most cases. Overall, the results highlight that tags such as 'Hilarious,' 'Amazing lectures,' and 'Respected' show pronounced gender differences, favoring male professors, while other tags demonstrate negligible differences. These findings

suggest potential biases in how students perceive and tag professors, with some tags being clearly more influenced by gender than others.

6 Problem 5

To determine whether there is a statistically significant difference in average difficulty ratings between male and female professors, I conducted an independent t-test. The null hypothesis for this analysis was that there is no difference in the mean difficulty ratings between male and female professors. The independent t-test was chosen because the data has been adjusted into interval like data, rather than ordinal. The test produced a T-statistic of -3.03 and a p-value of 0.0024. This p-value is below the significance threshold of 0.005, leading to the rejection of the null hypothesis. These results indicate that there is a statistically significant difference in the mean difficulty ratings between male and female professors. The negative T-statistic suggests that male professors have, on average, slightly lower difficulty ratings compared to female professors.

7 Problem 6

To evaluate the size of the effect in the gender differences in average difficulty ratings, I calculated Cohen's d as a measure of effect size along with its 95% confidence interval. The calculated Cohen's d was -0.0255, with a 95% confidence interval of $(-0.0483, -0.0029)$. This very small effect size suggests that while the difference in average difficulty ratings between male and female professors is statistically significant, it is negligible in practicality. The negative direction of the effect size indicates that male professors tend to have slightly lower average difficulty ratings compared to female professors. The confidence interval supports this finding, since it does not include zero. Overall, the results show that while there is a statistically significant gender difference in average difficulty ratings, the magnitude of this difference is very small.

8 Problem 7

To determine which factors are most predictive of average ratings for professors, I built a regression model using all of the available numerical predictors in the dataset. The predictors included variables such as adjusted difficulty, proportion of students who would take the class again, number of ratings, and others from the rmpCapstoneNum dataset. The model was evaluated using R^2 and RMSE to assess its quality. The model achieved an R^2 of 0.807, indicating that approximately 80.7% of the variance in average ratings can be explained by the predictors included in the model. The RMSE was 0.297, showing a reasonably small average error in predictions of the average ratings. Among the predictors, the most strongly associated factor was the proportion of students who would

take the class again, with a coefficient of 0.486, indicating a strong positive relationship with average ratings. Adjusted difficulty was also a notable predictor, but it had a negative coefficient of -0.127, suggesting that higher perceived difficulty is associated with slightly lower average ratings. Other predictors, such as receiving a pepper (0.085), being male (0.020), and being female (0.010), had much smaller coefficients, indicating weaker contributions to the prediction of average ratings. These results suggest that students' willingness to take the class again is the strongest predictor of average ratings, reflecting a potential link between overall satisfaction and perceived teaching quality, which makes sense. Meanwhile, difficulty and gender-related variables play smaller roles. The high R^2 value indicates that the model effectively captures the factors influencing average ratings. As for co-linearity, I used a correlation matrix to see how each variable was correlated with each other to see if any cross-validation would be necessary for the data. I filtered the dataset to see if any pairs of variables had a correlation over $|0.5|$, and no results were returned. Since 0.5 itself is not a very high correlation, I decided the variables were sufficiently unrelated to each other, and left the model as is.

9 Problem 8

To assess the extent to which various student-evaluated tags predict adjusted ratings, I built a regression model using the 20 tags as predictors. The model achieved an R^2 of 0.345, indicating that approximately 34.5% of the variance in adjusted ratings can be explained by the tags provided by students. The RMSE was 0.452, reflecting the average error in predicting adjusted ratings. While the R^2 value is moderate, it suggests that the tags contribute somewhat meaningfully to the prediction of adjusted ratings but leave a lot of the variability unexplained. The most predictive factors were "Tough grader" (Coefficient $= -0.195$), which showed a strong negative association with adjusted ratings, and "Good feedback" (Coefficient $= 0.104$), which had a strong positive relationship. Other tags with positive coefficients included "Caring" (0.081) and "Amazing lectures" (0.061), suggesting that these characteristics are highly valued by students. On the other hand, tags such as "Lecture heavy" (-0.071) and "Group projects" (-0.047) were negatively associated with adjusted ratings, indicating potential areas of student dissatisfaction. These findings demonstrate that specific tags play a significant role in shaping students' perceptions of professors, particularly traits like fairness, feedback quality, and emotional connection. Now, in order to deal with co-linearity, I created a correlation matrix to see which pairs of coefficients were correlated to each other. This time, I set a threshold of $|0.6|$ to ensure that I was only handling pairs of tags that were very correlated. I found that "Respected" and "Caring" ($r = 0.731$), "Respected" and "Inspirational" ($r = 0.721$), "Respected" and "Amazing lectures" ($r = 0.695$), "Caring" and "Good feedback" ($r = 0.693$), and "Inspirational" and "Amazing lectures" ($r = 0.686$). These high correlations suggest that there is redundant information from some of these tags, which

can inflate the importance of related predictors in the model. To address multicollinearity, I decided to employ lasso and ridge regression methods to see which performed better. After applying Ridge regression, the R^2 dropped to 0.246, with an RMSE of 0.485, indicating that regularization reduced overfitting but at the cost of some explanatory power. Conversely, after applying Lasso regression, the R^2 improved to 0.346, with an RMSE of 0.440, slightly outperforming the initial model. Lasso regression also reduced the number of features with non-zero coefficients, simplifying the model by focusing on the most impactful predictors. These findings demonstrate that specific tags play a significant role in shaping students' perceptions of professors, particularly traits like fairness, feedback quality, and emotional connection. However, the moderate R^2 suggests that while these tags are important, other unmeasured factors also significantly influence adjusted ratings.

10 Problem 9

To predict average difficulty from all available tags in the dataset, I built a regression model using the tags as predictors. The regression model achieved an R^2 of 0.265, indicating that approximately 26.5% of the variance in average difficulty can be explained by the tags given by students while the RMSE was 0.434. The R^2 value suggests that the tags have moderate predictive power, but a large amount of the variability in average difficulty remains unexplained. The most predictive tag was "Tough grader" (Coefficient = 0.182), which showed a strong positive association with average difficulty, indicating that professors who are considered to be tough graders are generally rated as more difficult. Conversely, "Caring" (Coefficient = -0.105) had the strongest negative relationship with average difficulty, suggesting that professors perceived as caring are rated as less difficult. Other notable predictors included "Clear grading" (Coefficient = -0.086) and "Accessible" (Coefficient = 0.061). In order to deal with colinearity, I created a correlation matrix to see which pairs of coefficients were correlated to each other. I set a threshold of $|0.6|$ to ensure that I was only handling pairs of tags that were very correlated. I found that "Respected" and "Caring" ($r = 0.731$), "Respected" and "Inspirational" ($r = 0.721$), "Respected" and "Amazing lectures" ($r = 0.695$), "Caring" and "Good feedback" ($r = 0.693$), and "Inspirational" and "Amazing lectures" ($r = 0.686$). This makes sense because it is the same set of tags as the last problem, so naturally their correlations would remain the same to each other. To address multicollinearity, I decided to use both of the lasso and ridge regression methods to see which performed better. After applying Ridge regression, the R^2 dropped to 0.182, with an RMSE of 0.458, indicating that regularization reduced overfitting but at the cost of some explanatory power. Conversely, after applying Lasso regression, the R^2 improved to 0.272, with an RMSE of 0.430, very slightly outperforming the first model. Lasso regression also reduced the number of features with non-zero coefficients, simplifying the model by focusing on the most impactful predictors. These findings demonstrate that specific tags play a signif-

icant role in shaping students' perceptions of professors' difficulty, particularly being a tough grader and having many papers.

11 Problem 10

To predict whether a professor receives a "pepper" based on all available factors (both numerical and tag data), I built a classification model using Logistic Regression. The Logistic Regression model achieved an overall accuracy of 72%, with a macro average F1-score of 0.72. For professors who did not receive a pepper, the model achieved a precision of 0.76 and a recall value of 0.67, indicating it is more precise but slightly less sensitive in predicting this class. For professors who received a pepper, the model achieved a precision of 0.69 and a recall of 0.78, meaning it is more sensitive to identifying professors who received a pepper but has slightly lower precision, leading to a few false positives. The AU(RO)C for Logistic Regression was 0.806, indicating the model has good discriminatory ability between the two classes. Class imbalance concerns were addressed by using the `class_weight="balanced"` parameter, ensuring that the performance was not skewed by the differing proportions of professors who received/did not receive a pepper. These results suggest that the model performs well overall, with reasonably good precision and recall for both classes, and is capable of effectively distinguishing between professors who are and are not likely to receive a "pepper".

12 Extra Credit

I decided to try and determine whether there is a statistically significant difference in the number of ratings received by male and female professors. The null hypothesis for this analysis was that there is no difference in the distribution of the number of ratings between male and female professors. A Mann-Whitney U-test was chosen because the number of ratings is a discrete count variable that is not normally distributed, making this non-parametric test for 2 groups correct. The test produced a U-statistic of $412,610,851.0$ and a p-value of 2.55×10^{-13}. This p-value is far below the significance threshold of 0.005, leading to the rejection of the null hypothesis. The results indicate that there is a statistically significant difference in the distribution of the number of ratings received by male and female professors. Given the extremely low p-value, the results strongly suggest that gender is associated with differences in how many ratings professors receive.