DS-GA 1001: Capstone Project

Group Name: CAP 67
Group Members: Jee Ho Tae, Shuai Yang, Lucas Yao

Author Contributions

Jee Ho Tae: Data preprocessing and cleaning steps, Statistical analysis (Questions 1–4, Extra Credit), significance testing, effect size estimation, confidence interval construction, and visualization

Shuai Yang: Regression modeling (Questions 7–9), interpretation of model results, collinearity diagnostics, plotting and figure styling.

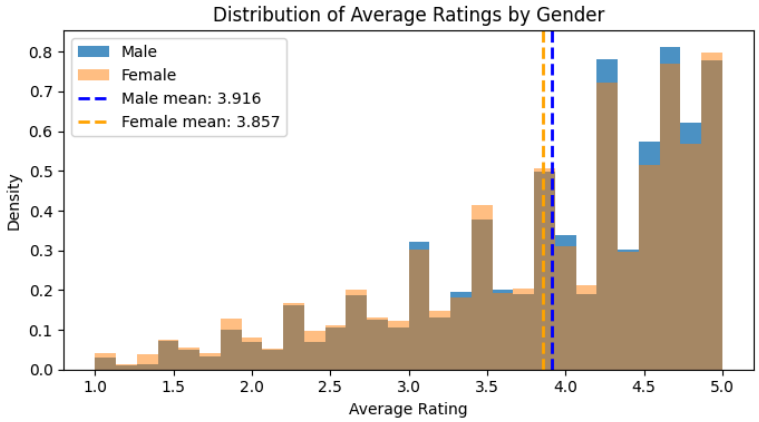
Lucas Yao: Classification modeling (Question 5,6,10), managed the comparative analysis of gender-based difficulty ratings and developed the logistic regression classification model for "pepper" status, utilizing AUROC metrics and balanced class weights to handle dataset imbalance.

Data Preprocessing and Cleaning

- We seeded the RNG with Lucas's N-Number (18049010). Used it in questions 3, 7, 8, 9, 10.
- All preprocessing steps were applied consistently across all analyses. The three provided datasets (rmpCapstoneNum.csv, rmpCapstoneQual.csv, and rmpCapstoneTags.csv) were first loaded and concatenated column-wise, as each row corresponds to the same professor across files.
- To improve the reliability of summary statistics, we set a threshold of at least 5 ratings for each professor, since average ratings based on very few observations are highly unstable and the mean number of ratings is 5.374.
- Gender information was handled by including only professors with exactly one of the gender indicators (male, female) equal to 1, excluding ambiguous cases. A single binary gender variable was then constructed (gender (male=1,female=0)).
- For the tag data, raw tag counts were normalized by dividing each tag by the total number of ratings for that professor, yielding average tag rates. This normalization ensures that tag comparisons are not confounded by differences in the number of ratings received, as professors with more ratings would have more tags.
- The resulting cleaned dataset was used for all subsequent statistical tests, visualizations, and models reported in this project.

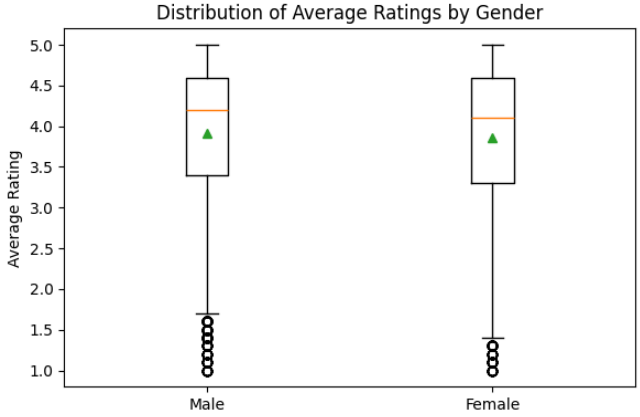
1. We would like you to answer the question whether there is evidence of a pro-male gender bias in this dataset.

To assess whether there is evidence of a pro-male gender bias in student evaluations, we divided the dataset into male and female professors and compared their average ratings. Male professors (n = 10015) have a mean average rating of 3.916, while female professors (n = 8407) have a mean average rating of 3.857. The histogram of average ratings by gender shows substantial overlap between the two distributions, but the mean rating for male professors is slightly shifted to the right of that for female professors, consistent with a small pro-male difference in average rating. To formally test whether this observed difference is statistically significant, we used Welch's two-sample t-test (H0: there is no gender difference in average ratings), which does not assume equal variances and accommodates unequal sample sizes. The test yielded t = 4.273 with $p = 1.94 \times 10^{-5}$. Since p < 0.005, we reject the null hypothesis and conclude that there is statistically significant evidence of a difference in average ratings by gender in this dataset, with male professors receiving slightly higher ratings on average. However, the effect size is small, where Cohen's d ≈ 0.063 indicates a very small standardized difference despite the strong statistical significance.



2. Is there a gender difference in the spread (variance/dispersion) of the ratings distribution?

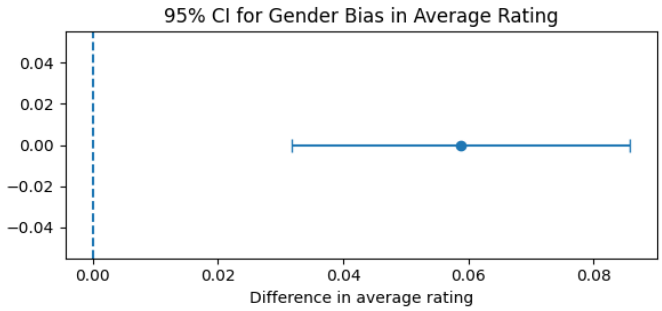
To assess if there is a gender difference in the spread of the ratings distributions, we divided the dataset into male and female professors and computed the sample variances (0.824 and 0.902 respectively). Female professors exhibit a slightly larger variance, which indicates that there is greater dispersion in their average ratings. This is also visible in the box plot, where the female distribution shows a slightly wider interquartile range and longer whiskers. Consequently, the magnitude of this difference appears to be modest. We used the two-sided F-test for equality of variances by placing the larger sample variance (female) in the numerator. The F-statistic of 1.094 implies that the sample variance for female professors is about 9.4% greater than that for male professors. Since the p-value (1.688×10^{-5}) < 0.005, we reject the null hypothesis of equal variances and conclude that the spread of average ratings differs significantly by gender.



3. What is the likely size of both of these effects (gender bias in average rating, gender bias in spread of average rating), as estimated from this dataset?

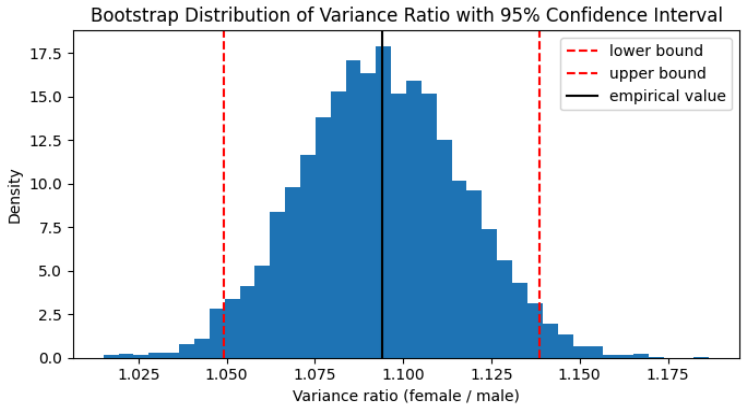
Gender bias in average rating:

In our sample, the mean average rating is 3.916 for male professors and 3.857 for female professors, so the mean difference is 0.059 points. Since both gender groups are very large (male = 10015, female = 8407), the sampling distribution of the difference in the sample means is well approximated by a normal distribution through the central limit theorem. Thus, we calculated the standard error of the difference in means and combined it with the 95% critical z value (1.96) to form a 95% confidence interval of [0.032, 0.086]. The interval lies above zero, which indicates that male professors tend to receive slightly higher average ratings than female professors by 0.032 ~ 0.086 points. The corresponding plot supports this interpretation, where the horizontal error bar is entirely above zero.



Gender bias in the spread of average ratings:

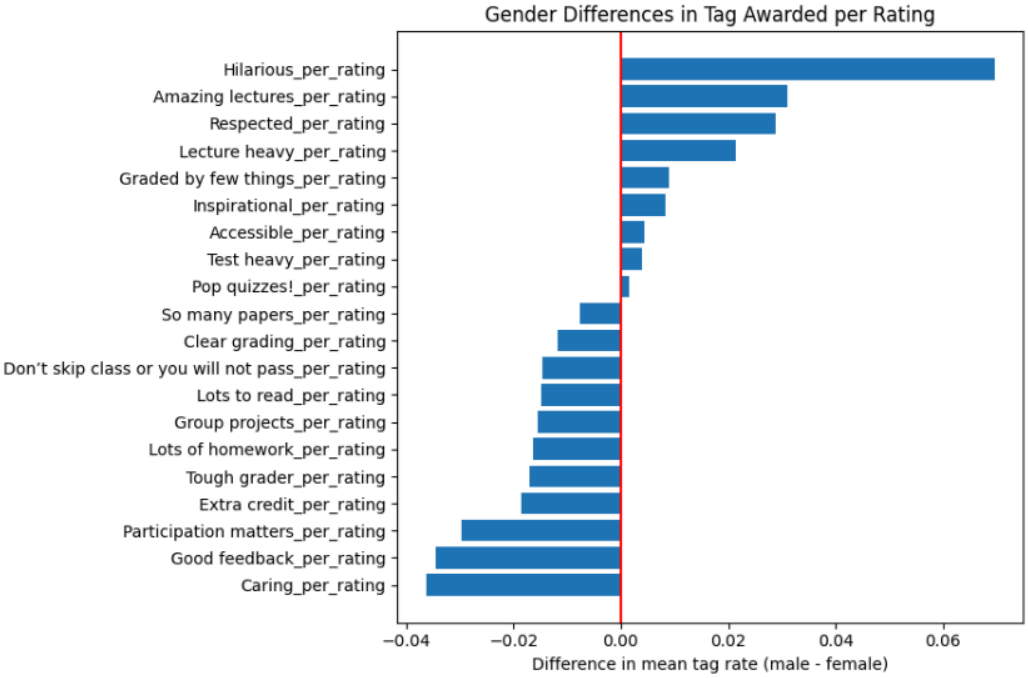
For the difference in spread, we focused on the variance ratio between female and male professors' average ratings, which is the natural effect measure underlying the F-test for equality of variances. Therefore, we used bootstrapping to construct a 95% confidence interval. Specifically, we repeatedly resampled male and female professors with replacement, maintaining the original group sizes (10015 and 8407), recomputed the variance ratio for each resample, and took the 2.5th and 97.5th percentiles of the bootstrap distribution. Across 5,000 bootstrap resamples, the 95% confidence interval was [1.049, 1.139]. The interval lies above 1, so it suggests that female professors' average ratings are about 4.9–13.9% more variable than those of male professors. The bootstrap histogram reflects this conclusion, where the confidence bounds lie entirely to the right of 1.



4. Is there a gender difference in the tags awarded by students?

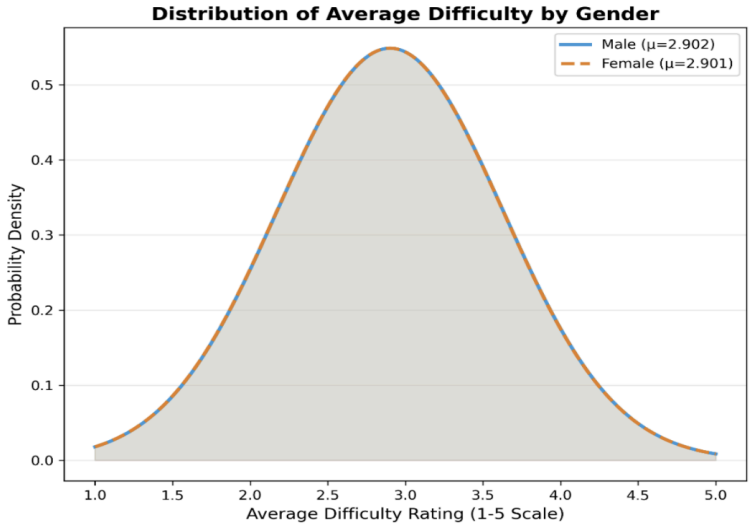
For each of the 20 normalized tag rates, we compared male and female professors using Welch's t-test (H0: no gender difference in the average tag rate). Using $\alpha = 0.005$, 19 out of 20 tags showed a statistically significant gender difference. The only tag that was not statistically significant was *Pop quizzes!* ($p = 0.312$). The horizontal bar graph shows the difference in mean tag rate (male - female) for all tags, where bars to the right of zero indicate a higher average tag rate for male professors, and bars to the left indicate a higher average tag rate for female professors. Consistent with the test results, most bars are separated from zero, but the absolute magnitudes of these differences are generally small. Among these, the top 3 most gendered (lowest p-value) are *Hilarious* ($p = 4.422 \times 10^{-153}$, difference = +0.0696), *Amazing lectures* ($p = 7.307 \times 10^{-42}$, difference = +0.0310), and *Caring* ($p = 3.309 \times 10^{-36}$, difference = −0.0363). The first two favor male professors (positive differences), while *Caring* favors female professors (negative difference). On the other hand, the bottom 3 least gendered (highest p-value) are *Test heavy* ($p = 0.00028$, difference = +0.0040), *Accessible* ($p = 0.004266$,

difference = +0.0045), *Pop quizzes!* (p = 0.312293, difference = +0.0015). Since *Pop quizzes!* is not statistically significant at α = 0.005, there is insufficient evidence of a gender difference for that tag. Overall, these results indicate that gender differences in tag usage are statistically detectable across most tags, but the absolute differences in average tag rates are generally small.



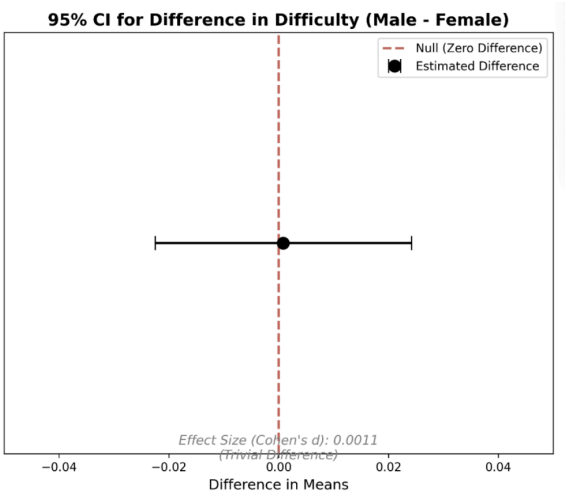
5. Is there a gender difference in terms of average difficulty?

In our restricted sample of 18,422 professors, the mean average difficulty rating for male professors is 2.902 compared to 2.901 for female professors, resulting in a negligible mean difference of only 0.001 points. Given the very large sizes of both gender groups (male = 10015, female = 8407), we applied a Welch's two-sample *t-test* to evaluate the significance of this difference. The test produced a T-statistic of 0.0712 and a corresponding P-value of 0.943, which is substantially higher than the standard 0.05 significance threshold. Because the P-value is so large, we fail to reject the null hypothesis. This indicates that there is no statistically significant gender difference in average difficulty ratings. The microscopic gap observed is well within the range of what would be expected by random sampling error alone.

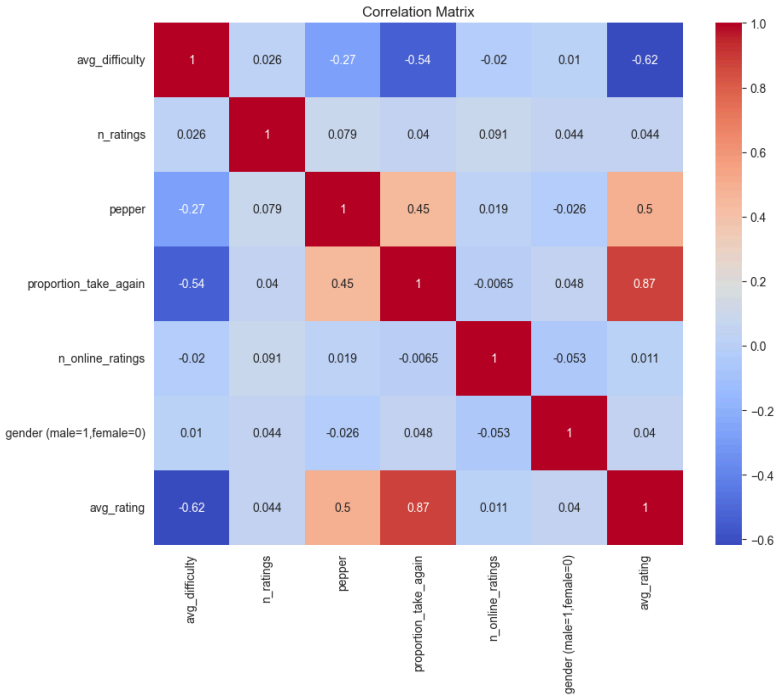


6. Please quantify the likely size of this effect at 95% confidence.

To quantify the magnitude of this effect, we calculated Cohen's d, which resulted in an extremely small value of 0.0011. This indicates that the difference between the two group means is less than one-thousandth of a standard deviation, confirming that any practical difference is trivial. Furthermore, because our large sample sizes (male = 10015, female = 8407) allow for a normal approximation of the sampling distribution, we constructed a 95% confidence interval for the mean difference. The resulting interval is [-0.0225, 0.0242], which contains zero. This inclusion of zero indicates that we cannot rule out the possibility that there is no true difference in the population, a finding entirely consistent with our non-significant t-test result.



7. Build a regression model predicting average rating from all numerical predictors.

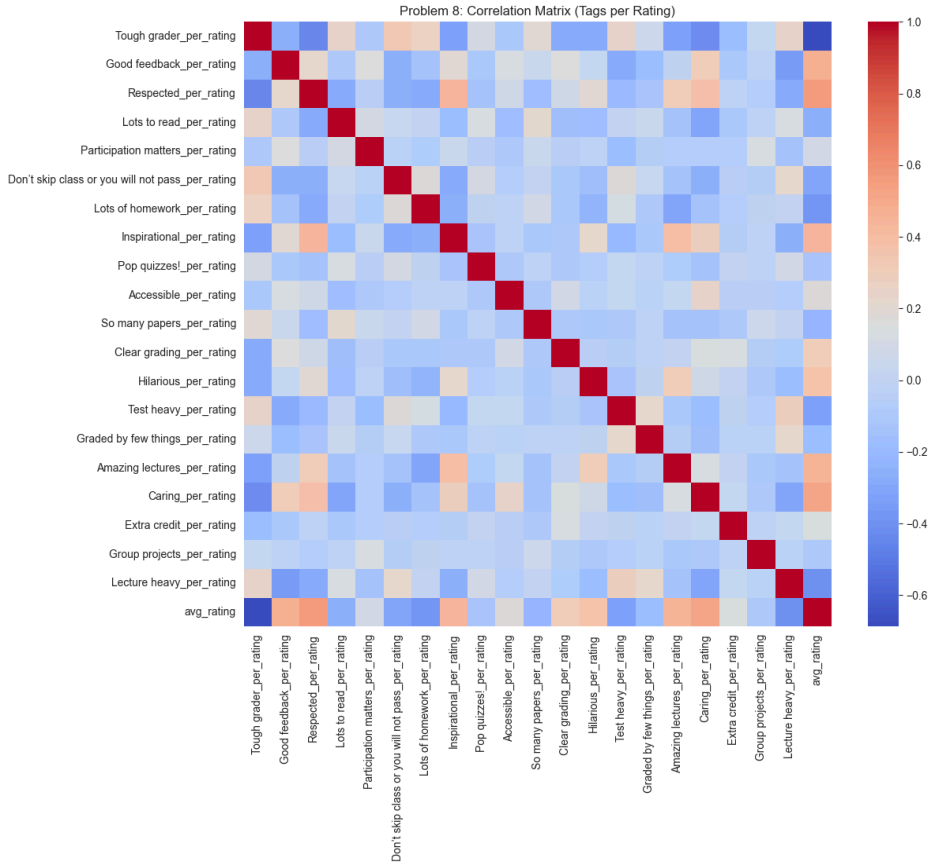


Average ratings computed from small numbers of ratings are statistically unstable and introduce noise into our model. Therefore, to obtain a more reliable estimate of the target variable, we remove rows with n_rating lower than 10. After performing the train-test split, we fit an initial model with all numerical predictors and found that the R^2 is 0.81708 and RMSE is 0.35440. Then, we performed feature selection using correlation heatmap. We can observe that *average difficulty*, *the proportion of students who would take again*, and "*pepper*" (a binary

indicator of perceived instructor attractiveness) have the strongest correlation with average rating of the professor. However, before fitting these in our final model, we also need to address the possible multicollinearity among the three predictors by using Variance Inflation Factor (VIF). The VIF of each predictor is below 2, indicating low correlation among them. Therefore, all three predictors can be included in our model. After standardizing, we fitted our final model, yielding R^2 of 0.81762 and RMSE of 0.35387. The strongest predictor is *the proportion of students who would take the class again*, with the largest absolute coefficient of 0.578718.

8. Build a regression model predicting average ratings from all tags.

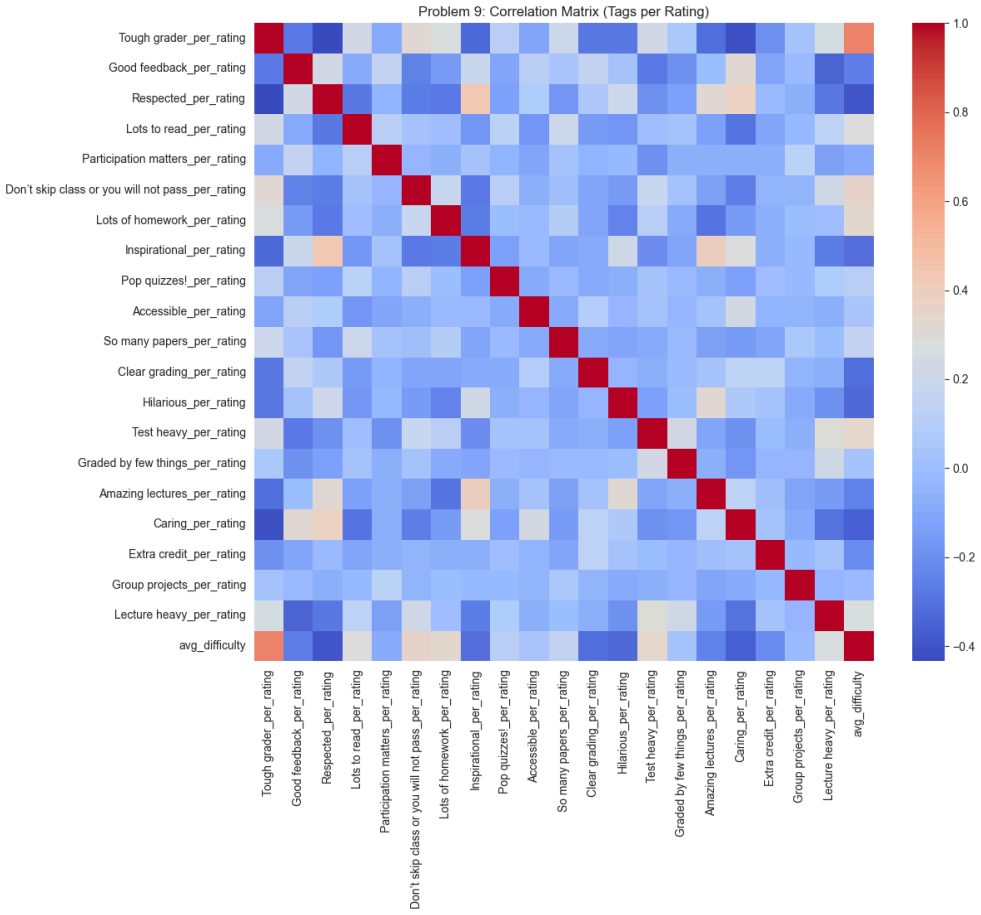
After performing the train-test split, we fitted our initial model using all the normalized tags as predictors. The model has a R^2 of 0.77826 and RMSE of 0.41384. After creating the correlation heatmap for feature selection, we selected the top 12 predictors by their absolute value of the correlation with the average rating larger than 0.3. The VIF verification process found that all predictors had a VIF value lower than 2, which means that there are no multicollinearity concerns for this model. The final model has a R^2 of 0.76235 and RMSE of 0.42842. And we also found that the strongest predictor is Tough grader. The numerical-predictor model clearly outperformed the tag-based model with a higher R^2 and a lower RMSE. From the perspective of model efficiency, the numerical-predictor model achieves better performance with far fewer predictors.



9. Build a regression model predicting average difficulty from all tags

After creating the train-test split, we first fit an initial model using all normalized tags as predictors and *average difficulty* as target variable. The full tag model achieves an R^2 of 0.64001 and RMSE of 0.46322. This tells us that the tags explain a meaningful portion of the variation in the average rating. We then performed feature

selection using the correlation matrix and selected the top 10 tags as our predictors. As usual we addressed the potential multicollinearity using the VIF, where all selected predictors have VIF smaller than 2. Therefore, there are no serious collinearity concerns. Using these selected predictors after standardization, the reduced model achieves an R^2 of 0.60830 and RMSE of 0.48320. The reduced model performs slightly worse than the full model (lower R^2 and higher RMSE), which is expected after removing predictors. Therefore, if the primary goal is predictive accuracy, we would retain the full tag model. However, the reduced model provides a more interpretable summary, and it confirms that *Tough grader* remains the strongest predictor of average difficulty.

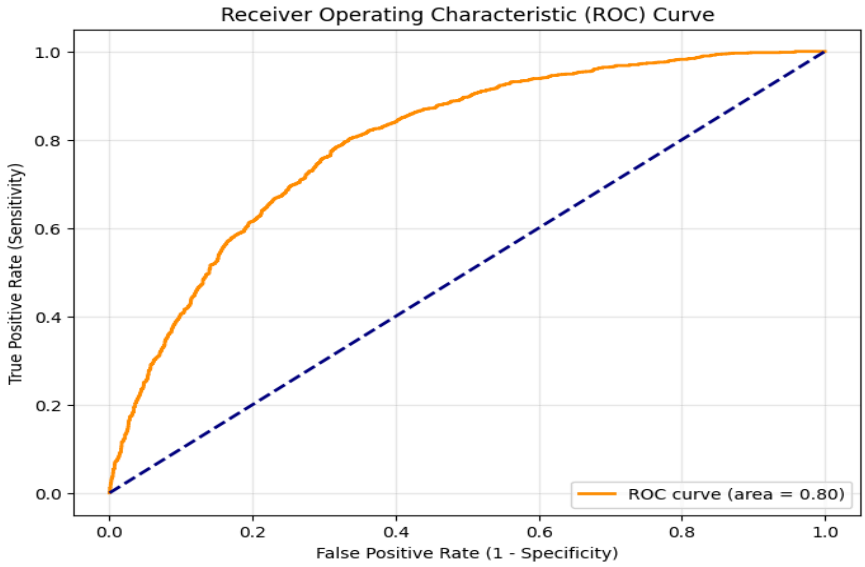


10. Build a classification model that predicts whether a professor receives a "pepper" from all available factors (both tags and numerical).

Our objective was to predict "*pepper*" status (*pepper=1*) using a logistic regression model trained on all numerical features and tag counts. We filtered for *n_ratings* > 5 to ensure data quality, resulting in a sample of 25,368 professors. To address class imbalance without over-relying on accuracy, which the lecture notes identify as a flawed metric for imbalanced data, we employed class_weight='balanced'. This approach follows a soft margin philosophy, allowing for slack to ensure robust performance across both categories. We evaluated the model using the Area Under the Receiver Operating Characteristic Curve (AUROC):.

- AU(RO)C (0.7962): Indicates a strong ability to distinguish between "*pepper*" and "*non-pepper*" professors.
- Recall / Sensitivity (0.7875): Demonstrates high "power," capturing ~78.8% of true "*pepper*" cases.
- Specificity (0.6769): Measures "selectivity," or the ability to avoid false positives.
- Precision (0.6420): Shows that when the model predicts a pepper, it is correct ~64.2% of the time.

Based on the coefficients, *Average rating* is the strongest predictor (1.2542), followed by performance tags like *Inspirational* (1.0362), *Amazing lectures* (0.9191), *Hilarious* (0.7512), and *Tough grader* (0.3203) . Interestingly, we discovered that being a *Tough grader* (0.3203) is influential in receiving a *pepper*, which is contradictory with the common belief that they are viewed negatively.



Extra credit: What is the proportion of states with a statistically significant gender difference in professor's average ratings?
To assess whether there is gender bias in professors' average ratings at the state level, we conducted separate Welch two-sample t-tests within each U.S. state, comparing male and female professors' average ratings. We restricted the analysis to states with at least 30 male and 30 female professors, resulting in 42 states being tested. Using $\alpha = 0.005$, only one state (Tennessee) exhibited a statistically significant gender difference, with male professors receiving higher average ratings (difference = 0.287, p = 0.0046). The scatterplot of state-level mean rating differences shows that while multiple states have positive or negative mean differences (including some comparable in magnitude to Tennessee), only Tennessee meets the statistical significance threshold. Thus, the proportion of states exhibiting evidence of gender bias in average ratings is 0.024.