Data Capstone Project - Group #52

Contributors:

Nikolas Prasinos (np3106)
Kyeongmo Kang (kk5739)
Alexander Pegot-Ogier (ap9283)

Authors' contributions:

We split the 10 questions between us, while constantly consulting each other, whenever an obstacle arose.

Nikolas Prasinos: Questions 5, 6, 7, 10

Kyeongmo Kang: Questions 2, 4, 8

Alexander Pegot-Ogier: Questions 1, 3, 9

All of the authors completed the extra credit task, verified code cohesion and replicability, and wrote the report.

Preprocessing Summary:

General Data cleaning

For preprocessing the data, we decided first to drop the rows for all columns that had no number of ratings associated with a professor. This makes sense because realistically if there are no ratings, the professor should not be accounted for. Moreover, we also noticed that the feature "The proportion of students that said they would take the class again" has a very large number of NaN values and initially thought of dropping it. However, we saw that it actually encapsulates a lot of information, since it is a significantly predictive factor of some of the features we want to predict in certain questions and therefore decided against dropping it. Subsequently, we decided to impose a threshold where we only account for professors who have ratings higher than the mean number of ratings. This is done, in order to increase the credibility and reliability of "average" estimates in our dataset, such as ratings and difficulty.

General Data Transformations and Normalizations

When using the tags dataset, we normalized the data by dividing the number of tags awarded for each specific tag, with the number of ratings that each professor has. We did that because a raw number of tags received by a professor is not meaningful if not placed in context, with the number of ratings they have received from their students.

General assumptions:

1) We are assuming that if the professor has male and female not selected, then is considered as not identified and we don't include them when trying to find insight about gendered features.
2) Professors are independent among them, the same as students ranking the professors.
3) Ratings (such as average ratings and average difficulty in our dataset) are considered to be ordinal data (ranked), which has an order but not necessarily equal intervals. Given its robustness to skewness, often observed in rating distributions, using the median as a comparison metric is the most appropriate approach.
4) Our alpha significance level is set at 0.005 to account for concerns regarding alpha inflation.

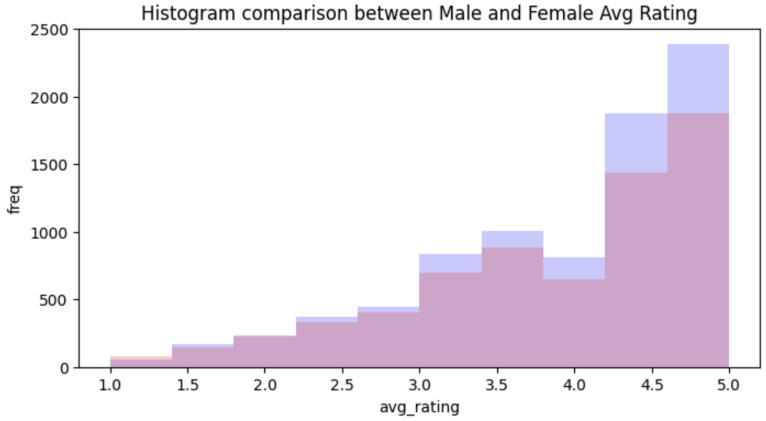
1. Activists have asserted that there is a strong gender bias in student evaluations of professors, with male professors enjoying a boost in rating from this bias. While this has been celebrated by ideologues, skeptics have pointed out that this research is of technically poor quality, either due to 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.

D: We compared the average ratings of male and female professors, explicitly including only those professors whose gender was with high confidence identified. Given the skewness of the data (skewed to the left), we chose a Mann-Whitney U Test to compare the distributions. We visualized the data using a histogram provided below.

H0: There is no gender bias in students' evaluations - males and females are rated the same.

H1: There is a difference in gender bias in students' evaluations - male professors are rated higher average than female professors

Y: We performed Mann-Whitney U Test because it's a non-parametric test that does not assume normality, making it suitable for this scenario.



F: The test statistic returned a p-value of **0.00039**. The U test statistic is **283.5.** The histogram shows that the average rating for male professors is visually higher than the average rating for female professors.

A: Given the extremely low p-value (**0.00039**), we **drop the null hypothesis** and conclude that there is a significant difference between male and female professors, indicating that male professors are rated higher than female professors.

2. Is there a gender difference in the spread (variance/dispersion) of the ratings distribution? Again, it is advisable to consider the statistical significance of any observed gender differences in this spread.

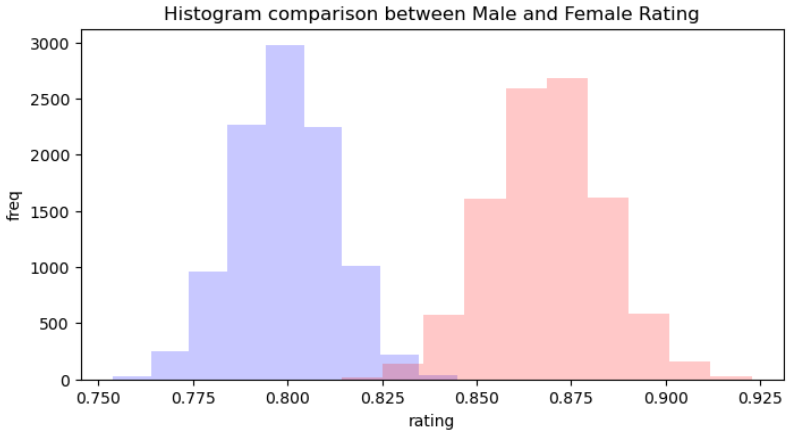
D: We compared the average rating distributions of male and female professors to assess gender differences in spread (variance/dispersion). We used bootstrapping to get the distribution of the variances between males and females and then did a two-sided KS test to compare them.

H0: Variances of male and female ratings have the same distribution.

H1: Variances of male and female ratings have significantly different distributions.

Y: Bootstrapping was used to analyze variance as it is robust for non-parametric data. The two-sided KS test was applied to validate whether variance distributions of male and female ratings differ significantly.

F: The KS test produced a p-value close to **0**, showing a highly significant result. Bootstrapping also revealed noticeable variance differences between genders(in the histogram below red if for females and blue is for males).



A: With a p-value near **0.0** and visual evidence from the attached figure, we **drop the null hypothesis** and conclude that male and female ratings follow different distributions.

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? Please use 95% confidence and make sure to report each/both.

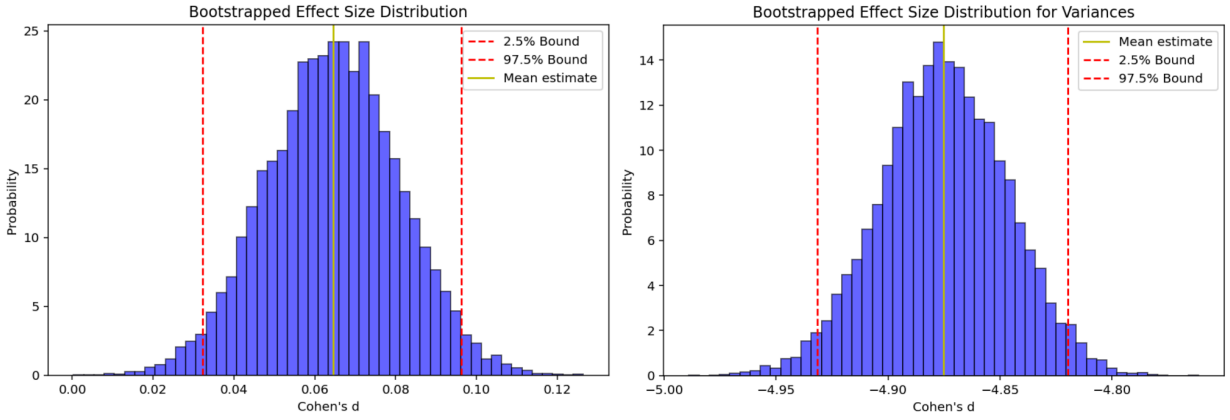
D: To estimate the average ratings effect sizes, we used the bootstrapping method to estimate Cohen's d formula. For the effect size of the spread (variances), we applied a second bootstrapping process, building on the variances obtained in Question 2. This allowed us to compute a standardized effect size for the variance differences using a modified Cohen's d formula, specifically adapted for comparing the variability between groups.

Y: We used the Cohen's d effect size to quantify the magnitude of the differences in average ratings while using the bootstrap method allows us to estimate our confidence intervals without the data being normally distributed. For the variance effect, we performed a second bootstrapping over the previously calculated variances from question 2. We performed a second Cohen's d to quantify differences in spread.

F: Average Ratings: The mean Cohen's d was **0.0647, having a 95% CI between [0.0328, 0.0979].**

Spread of Ratings: The mean Cohen's d of that effect of the variance difference was **-4.875**, and the 95% CI was **[-4.93169026 -4.81967219].**

A: These results suggest that male professors receive higher ratings on average compared to female professors, based on Cohen's effect size of approximately **0.0647**. Additionally, the large negative value for the Cohen's d for the variance differences **(-4.875)** indicates that female ratings exhibit significantly greater variability compared to males.



4. Is there a gender difference in the tags awarded by students? Make sure to teach each of the 20 tags for a potential gender difference and report which of them exhibit a statistically significant different. Comment on the 3 most gendered (lowest p-value) and least gendered (highest p-value) tags.

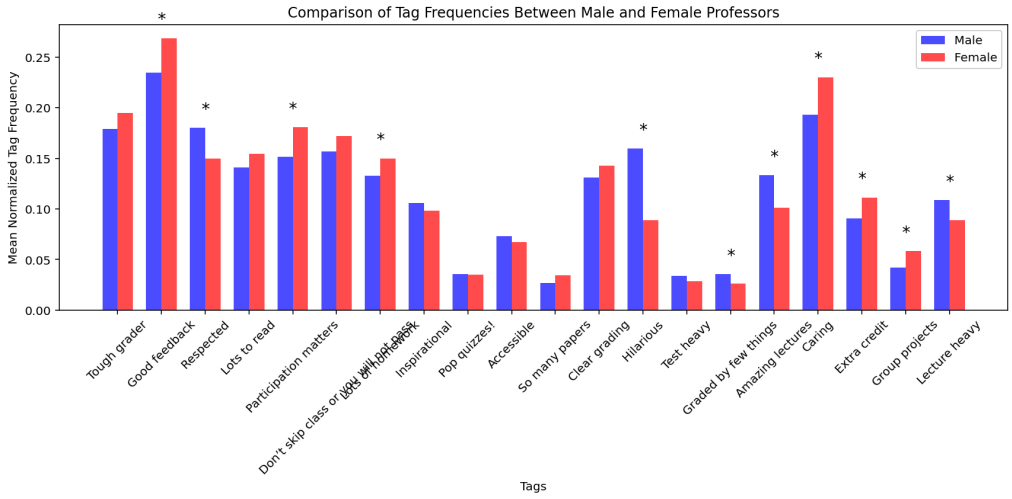
D: We compared the tags awarded to professors by gender using two-sided Chi-squared test tests to identify statistically significant differences.

H0: There is no association between gender and the tags awarded (tags are distributed equally across genders).

H1: There is an association between gender and the tags awarded (tags are distributed differently across genders).

Y: We use the two-sided Chi-squared test, because we have categorical data. More specifically, we have counts of tags awarded to professors and therefore Chi-squared is the appropriate significance test. Also our significance level is 0.005, because we want to reduce the effect of alpha inflation.

F: The three most gendered tags (lowest p-values) are **"Hilarious"** (p-value: **6.680e-38**), **"Amazing Lectures"** (p-value: **2.519e-09**), and **"Caring"** (p-value: **3.529e-08**). The three least gendered tags are **"Inspirational"** (p-value: **0.1238**), **"Accessible"** (p-value: **0.1852**), and **"Pop Quizzes"** (p-value: **0.872**). Our analysis shows that 11 out of 20 tags have statistically significant gender differences.



A: Since **11 tags have p-values below 0.005**, we **drop the null hypotheses**, and conclude that there is a significant gender difference in 11 out of 20 tags awarded by students.

5. Is there a gender difference in terms of average difficulty? Again, a significance test is indicated.

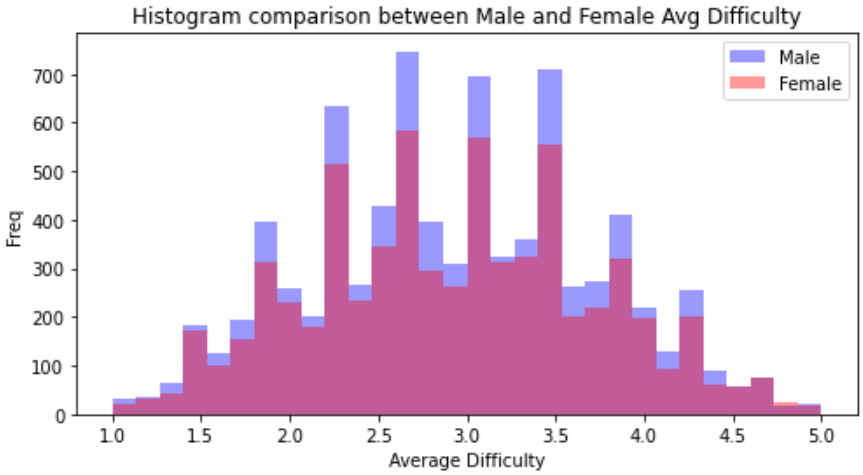
D: We compared the average difficulty attributed to each professor, by separating them into two groups, one for exclusively males and one for exclusively females. We used a two-sided Mann-Whitney U test to identify statistically significant differences.

H0: There is no gender difference in the average difficulty.

H1: There is a gender difference in the average difficulty.

Y: We used the Mann-Whitney U test because we want our test to best take advantage of the ordinality of our data.

F: The two-sided Mann-Whitney U test yielded a p-value of **0.968**

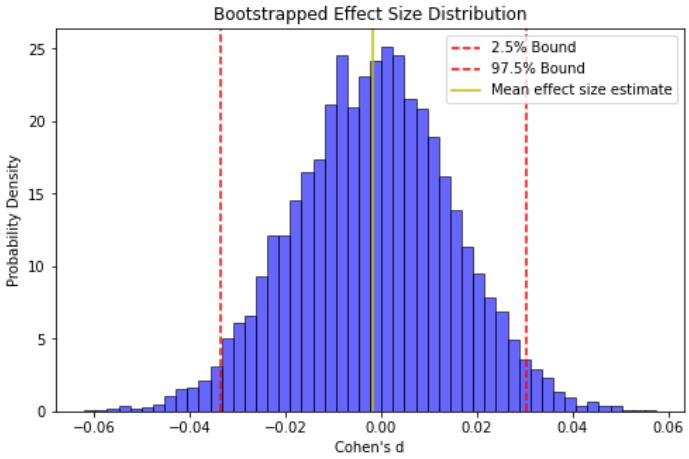


A: Given that the p-value is **0.968**, we conclude that we **can not drop the null hypothesis**, meaning that we can't say whether there is a statistically significant difference between male professors' average difficulty and female professors' average difficulty.

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

D: To estimate the effect size of the gender difference in average difficulty, we used the bootstrapping method to create the distribution of Cohen's d. We quantify that using a 95% confidence interval.

Y: We used the bootstrap method in order to create a distribution of our statistic, in this case, the effect size (Cohen's d.). That way we can get the mean of the effect size, the confidence interval, as well as a proper visualization of its distribution.



F: The mean Cohen's d is **-0.0016, having a 95% CI between [-0.0336, 0.0304].**

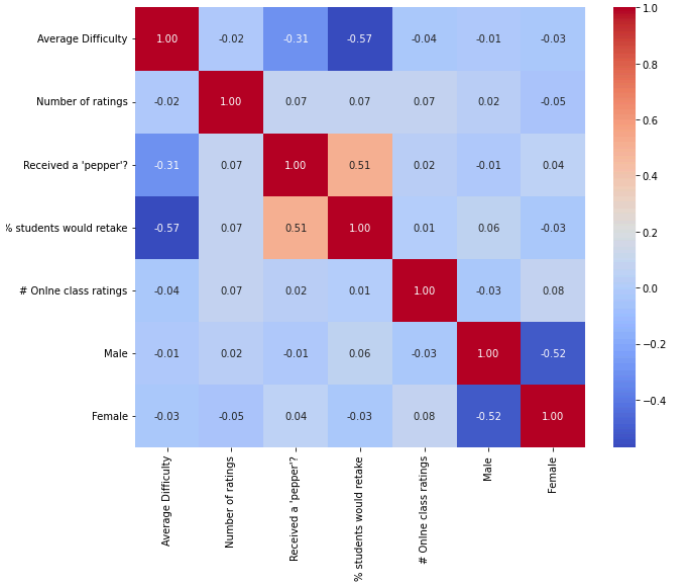
A: A Cohen's d of **-0.0016** is extremely close to zero, indicating almost no practical effect size. The 95% confidence interval, which spans from approximately **-0.034 to 0.030**, further suggests that any observed difference could be attributed to random variation rather than a true effect.

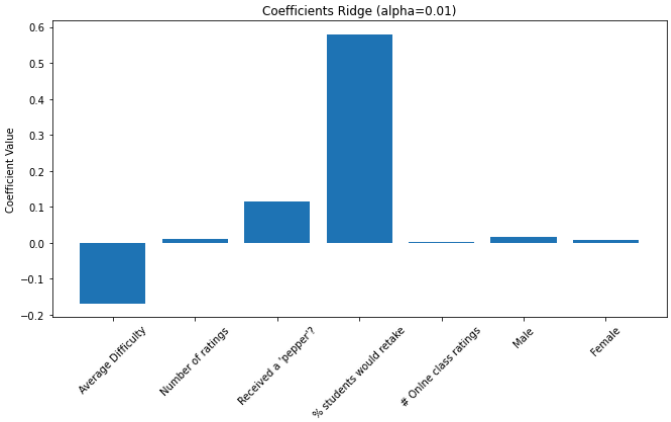
7. Build a regression model predicting average rating from all numerical predictors (the ones in the rmpCapstoneNum.csv) file. Make sure to include the R2 and RMSE of this model. Which of these factors is most strongly predictive of average rating? Hint: Make sure to address collinearity concerns.

D: We cleaned the data by performing row-wise NaN removal. Afterwards we standardized the data so that all the features have zero mean and unit variance. Then we split the data into 80% training set and 20% testing set and performed multiple linear regression, ridge regression and lasso regression.

Y: Cleaning the data is essential, so there won't be any row-column pairs without a valid value. Standardization was performed, because our data were not on the same value range and it's generally a good practice. Importantly, we visualized the correlation matrix, in a heatmap, in order to check multicollinearity. Then we performed multiple linear regression, ridge regression and lasso regression, to test which one would be better for our data and to address multicollinearity issues.

F: Given the correlation matrix, there are no strong correlations between the predictors. In all regression models, we arrive at almost identical results, with an extremely small improvement by the ridge regression model. **R^2 = 0.8410 and RMSE = 0.3306**. We plot the values of the coefficients and see the impact each has on predicting the Average Rating.





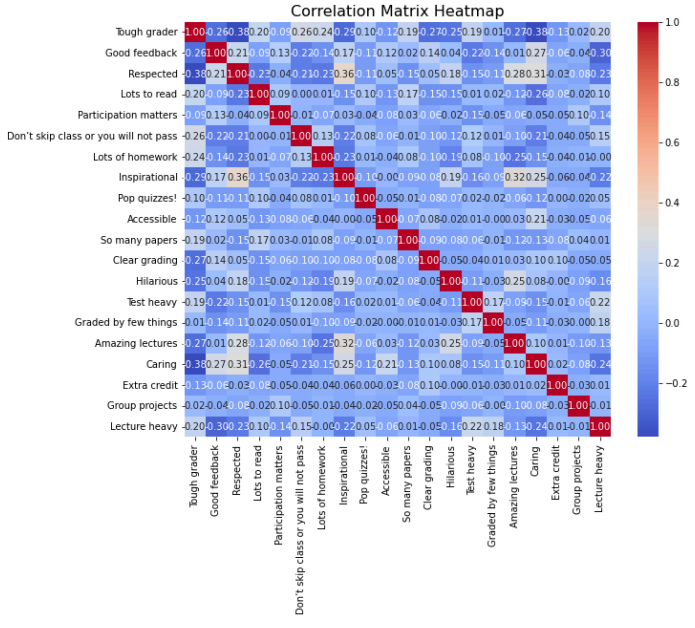
A: Our model performs well enough yielding: **R^2 = 0.8410 and RMSE = 0.3306**. There may exist slight multicollinearity between the features "Received a pepper" and "% students would retake" and "Male" and "Female", however, we address these issues using ridge and lasso regression. It is easy to see that the most strongly predictive factor is the percentage of students who would retake the class, which also makes intuitive sense.

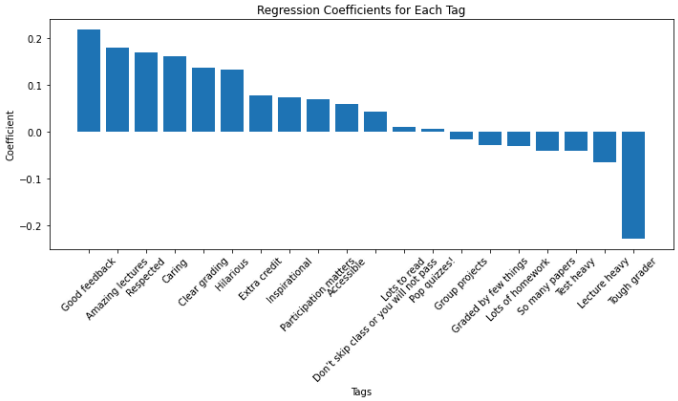
8. Build a regression model predicting average ratings from all tags (the ones in the rmpCapstoneTags.csv) file. Make sure to include the R2 and RMSE of this model. Which of these tags is most strongly predictive of average rating? Hint: Make sure to address collinearity concerns. Also comment on how this model compares to the previous one.

D: We normalized tag data by dividing each tag count by the number of ratings. To address multicollinearity, we visualized a correlation matrix of the tags and calculated variance inflation factors (VIF). The dataset was split into training (80%) and test (20%) sets for model evaluation.

Y: Normalizing the tag counts ensured comparability across professors, regardless of the number of ratings, and avoided bias toward professors with more ratings. This allowed us to build a linear regression model to predict average ratings from tags while controlling for multicollinearity.

F: The model achieved an **R^2 = 0.7339** and an **RMSE = 0.4842**. The top predictor was "**Tough Grader"**, with a coefficient of **-0.2271**, indicating a strong negative association with average ratings. The highest VIF was also "**Tough Grader"** at **1.65**, while all other tags had VIF values close to 1, suggesting minimal multicollinearity between the predictors.





1	Tough Grader	1.654798
3	Respected	1.439228
17	Caring	1.414295
8	Inspirational	1.391394
2	Good Feedback	1.313172
16	Amazing Lectures	1.286716
20	Lecture Heavy	1.275913
7	Lots of Homework	1.253289
4	Lots to Read	1.224156
13	Hilarious	1.212977
12	Clear Grading	1.200905
6	Don't Skip Class	1.179783
14	Test Heavy	1.163934
5	Participation Matters	1.129465
15	Graded by Few Things	1.110317
11	So Many Papers	1.102484
10	Accessible	1.086536
19	Group Projects	1.068989
18	Extra Credit	1.068303
9	Pop Quizzes	1.052375

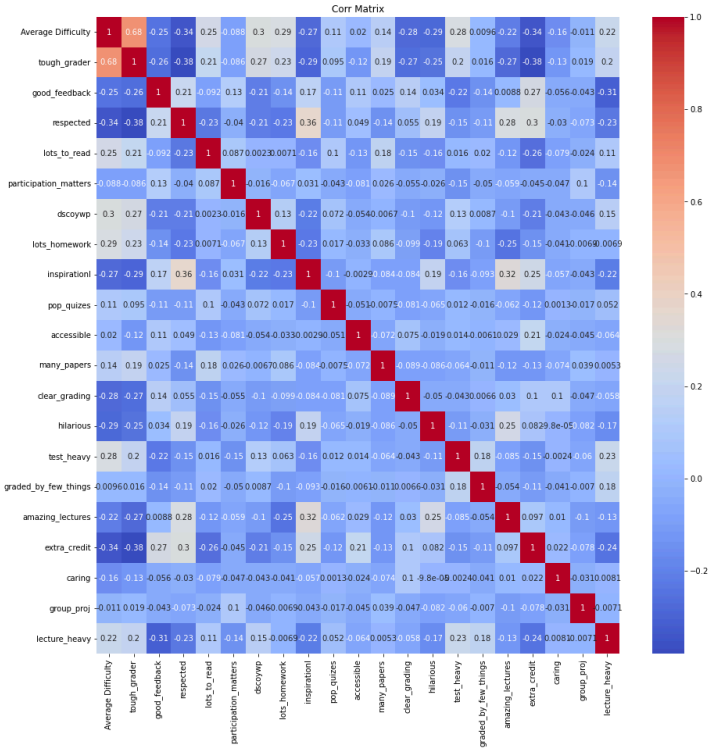
A: The model explains **73.39%** of the variance in average ratings, with an average prediction error of **0.4842** units. "**Tough Grader"** was the most influential feature, suggesting professors with this tag are rated negatively correlated. As all VIF values are close to 1, multicollinearity is negligible. However, the model may still be influenced by unmeasured confounders or nonlinear relationships. It is also worth mentioning that this model performs slightly worse than the previous one, indicating that the numerical data from ratemyprofessor.com, are more strongly predictive of the average rating of professors, than the tags data.

9. Build a regression model predicting average difficulty from all tags (the ones in the rmpCapstoneTags.csv) file. Make sure to include the R2 and RMSE of this model. Which of these tags is most strongly predictive of average difficulty? Hint: Make sure to address collinearity concerns.

D: We merged the Average Difficulty column from the main dataframe with the tags dataframe to build a linear regression model predicting Average Difficulty from the tags. The tags were normalized by dividing their counts by the number of ratings for each professor to ensure proportionality. To address multicollinearity concerns, we visualized a correlation matrix of the tags. The data was split into training (80%) and test (20%) sets for model evaluation. Again a multiple linear regression model was used.

Y: Again normalizing the tag counts ensured comparability across professors, regardless of the number of ratings, and avoided bias toward professors with more ratings.

F: The model achieved an **R^2 = 0.579**, with an RMSE of **0.518**. The top predictor was Tough Grader, with a coefficient of **1.736**, indicating a strong positive association with predicted difficulty.



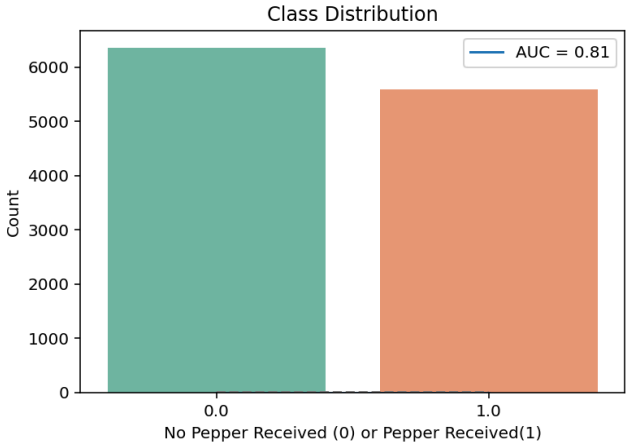
A: The model explains **57.9%** of the variance in Average Difficulty, with predictions deviating by **0.518** units on average. "**Tough Grader"** was the most influential feature, suggesting that professors labeled as tough graders significantly increased predicted difficulty ratings by **1.736** units. The second most influential feature was **"Test Heavy"** with a coefficient value of **1.283** and the third was **"Accessible"** with a coefficient value of **0.948**. Despite its reasonable fit, the model may be affected by unmeasured confounders or residual multicollinearity.

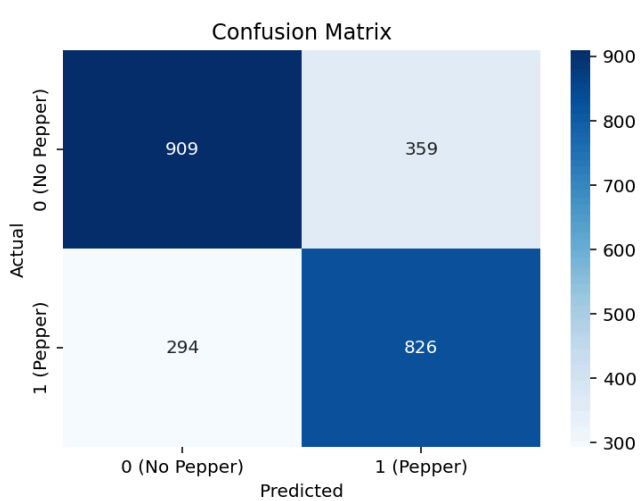
10. Build a classification model that predicts whether a professor receives a "pepper" from all available factors (both tags and numerical). Make sure to include model quality metrics such as AU(RO)C and also address class imbalance concerns.

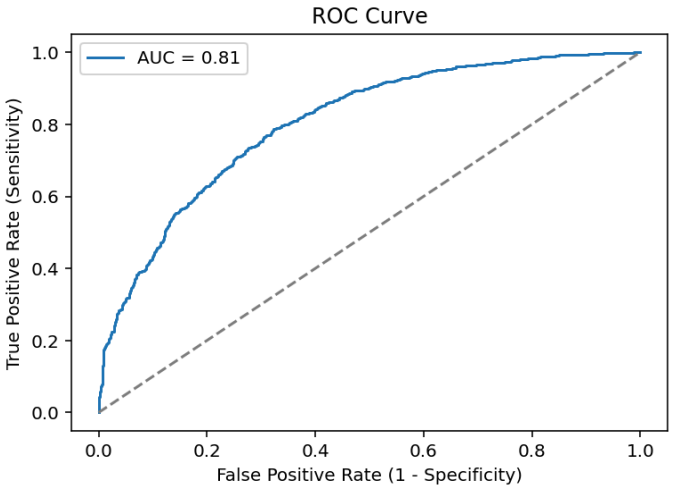
D: We cleaned the data by performing row-wise NaN removal. Then we normalized the tag features by dividing them by the number of ratings that each professor has. Afterwards, we used a MinMax scaler for the numerical features, so that they all had the same value range from 0 to 1. We checked for class imbalance by plotting the class distribution, we split the data into 80% training set and 20% testing set and then we used multiple logistic regression to do the classification.

Y: Again cleaning the data and normalizing the features is paramount for the correct classification process. We used logistic regression, which is suitable for binary classification, due to its sigmoid nature. Moreover, it handles imbalanced data with weights, which is one way to address the class imbalance concerns. Not to mention that logistic regression (similar to linear regression but for classification), offers a straightforward and interpretable model. Coefficients can be directly linked to feature importance and odds ratios, which is useful for understanding relationships in the data.

F: We plotted the class distribution to check for class imbalance. Also, we plotted both the heatmap of the confusion matrix, as well as the ROC curve. The accuracy of the model is **0.76** and the AU(RO)C is **0.83**.







A: Judging from the class distribution plot, we see that the two classes have similar counts, indicating that there isn't a severe class imbalance. Furthermore, given our accuracy and AU(RO)C values, **0.76** and **0.83**

respectively, we conclude that the model performs well. More specifically, the high number of true positives and true negatives indicates strong performance. The slight class imbalance may affect the number of false positives and false negatives, but not in a drastic way, given our use case.

Extra Credit

D: We wanted to find which of the 3 most popular majors has the most difficult professors on average. To do that we cleaned the data and found the most popular majors. Then we plotted their average difficulty histograms, along with their KDEs using Gaussian Kernels. Then we used Levene's test to check for homogeneity of variances.

H0: No statistical difference between the variances of the 3 majors.

H1: There is a significant difference between the variances of the 3 majors.

Afterwards we performed a Welch's one-way ANOVA significance test, to see if there was a significant difference between the average difficulty between the majors.

H0: No difference between the 3 majors.

H1: Significant difference between the 3 majors.

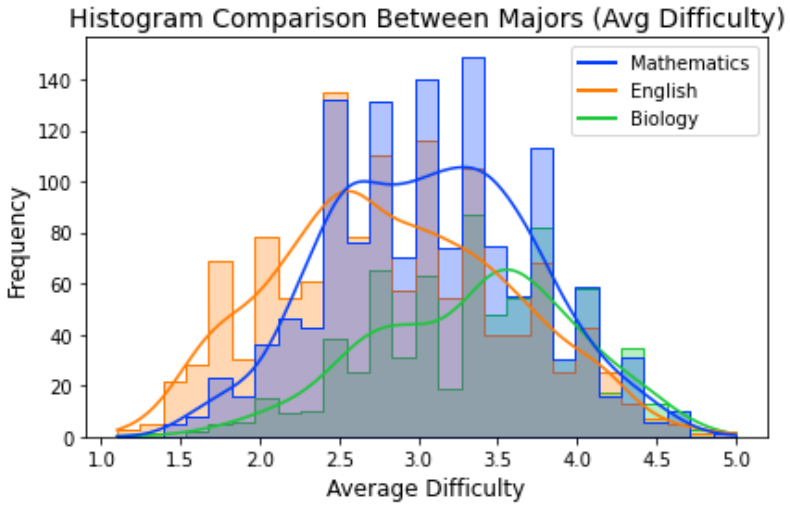
If the null hypothesis is dropped and we find a statistically significant difference between the 3 majors, we performed a one-sided Mann-Whitney U test, to test which of the two most likely majors, judging from the aforementioned histogram plot, has a higher average difficulty stats and if it is statistically significant.

H0: Mathematics average difficulty is not less than the Biology one

H1: Mathematics average difficulty is significantly less than the Biology one

Y: We plotted the average difficulty histograms between the 3 majors and their KDEs, to visualize and get an idea of the distributions. We performed Levene's test to check for homogeneity of variances, because the one-way ANOVA test assumes homogeneity of variances between the data. We then performed Welch's one-way ANOVA, because of the unequal variances and because we wanted to compare more than two groups (in our case 3). Then we selected Mathematics and Biology because, by looking at the histogram we can see that they are less skewed to lesser average difficulty values than English. We then performed the Mann-Whitney U test again because we want to take advantage of the ordinality of the data, which in our case are ratings of average difficulty by students.

F: The histogram can be seen below and we see how different the distributions are. The p-values that are yielded by the significance tests are, for the Levene's test the p-value is **8.331e-07**, for the Welch's one-way ANOVA the p-value is **4.228e-56** and for the one-sided Mann Whitney U test the p-value is **6.038e-21**.



A: For the first test, given the p-value of the Levene's test, which is **8.331e-07**, we **drop the null** and conclude that there is a statistically significant difference between the variances. That's why we have to use Welch's one-way ANOVA. For the second test, given the p-value of Welch's one-way ANOVA, which is **4.228e-56**, we **drop the null** and conclude that there is a statistically significant difference between the average difficulty between Mathematics, English, and Biology. For the third test, given the p-value of the one-sided Mann-Whitney U test, which is **6.038e-21**, we **drop the null** and conclude that the average difficulty of the professors in the Mathematics major is significantly less than the one in the Biology major.