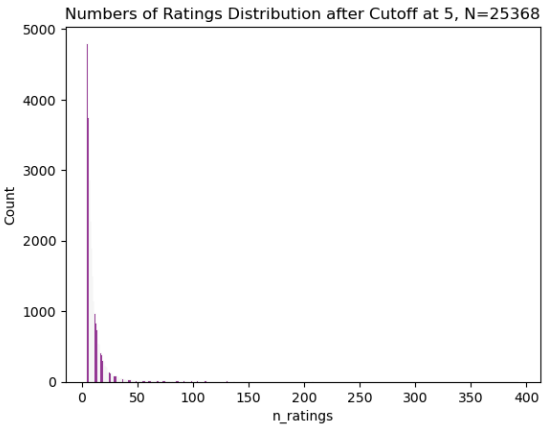
Krysten Nguyen - IDS Capstone

Preprocessing



My seed for RNG is 15113718. Throughout the analysis, alpha level of 0.005 is the threshold for statistical significance.

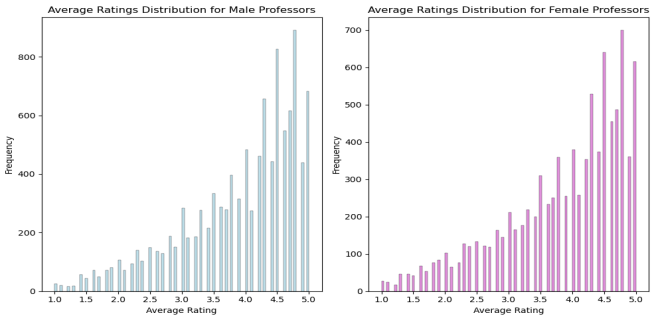
To clean the data, I concatenate the 3 types of data into 1 dataframe by rows (by professor), then remove empty rows where qualitative data is null (N=70004). Because professor rating and difficulty is reported as an average, I filter out professors with at least 5 ratings so that the students' extreme sentiments are neutralized and average ratings are more representative and preserve 36.2% of the sample (N=25368). Plotting the distribution of rating number, we observe an extreme skew, with the majority of rating numbers below 10.

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.

There are 10791 male and 9183 female professors by gender classification data provided. This adds up to less than N=25368 possibly due to classification false negatives.

The normality condition doesn't hold for our samples: there are more above-average ratings in both genders. However, comparing the means is still meaningful because the sample consists of the rating average for each professor. I perform an upper-tailed Welch t-test to compare the mean of rating distribution, not assuming equal variance.
H0: The average rating for male professors is similar to that for female professors.
Result: TtestResult(statistic=4.101, pvalue=2.062e-05, df=19183.438)

The result is statistically significant. We have reason to conclude male professors are rated more highly than female professors (observed differences in the mean of these average ratings are not plausibly just due to sampling error).



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.

To test the gender difference in the variance of nonnormal rating distribution, I perform a permutation test with test statistics as the variance difference between the 2 samples. I then double-check my result with Levene test which tests the null hypothesis that all input samples are from populations with equal variances when normality is not met.
H0: The rating distribution for both genders has similar spread
Result:

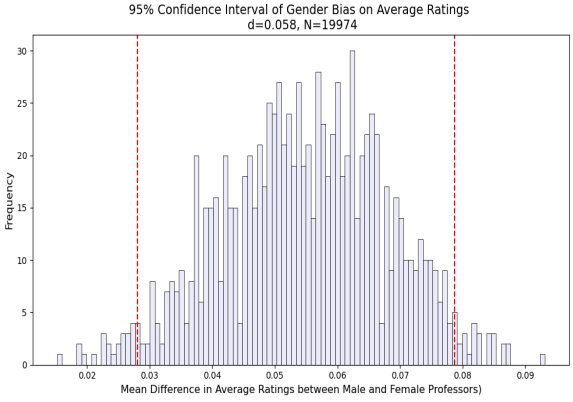
- PermutationTestResult(statistic=-0.071, pvalue=0.0004)
- LeveneResult(statistic=19.156, pvalue=1.211e-05)

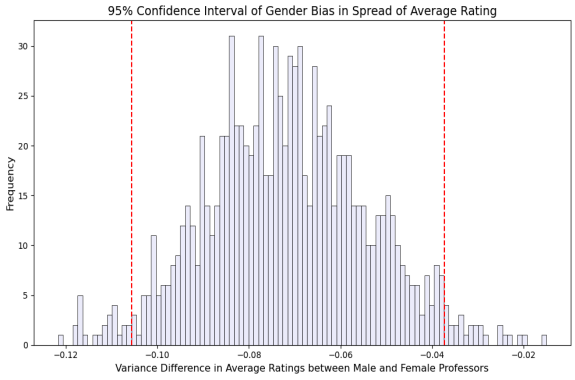
Both of the test results are statistically significant. We have reason to conclude a gender difference between the variance of rating distribution (the observed gender differences in this spread are unlikely to be due to mere sampling error)

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.

I perform 1000 iterations of Bootstrap to compute 95% CIs for differences in mean and variance of ratings between genders through resampling with replacement. I use bootstrapping to avoid normality assumptions about data distribution. I also Cohen's d to standardize the mean difference as an effect size for gender disparity.
Result:
Cohen's d: 0.058
Average ratings:
95% CI for mean difference: (0.028,0.079)
95% CI for variance difference: (−0.106,−0.037),

Gender bias exists in average ratings (males rated slightly higher), but the effect size (d=0.058) is minimal. A negative 95% CI for variance difference indicates that female professors have slightly higher variability in ratings.





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.

I conduct 20 Mann-Whitney U tests, one for each tag, to assess gender differences in the frequency of tags awarded by students. I choose an upper-tailed Mann-Whitney U as a non-parametric test to compare distributions for gender bias without assuming the normality of tag frequencies.
H0: There is no difference in the number of tag awarded to male and female professors
Result:
Top 3 most gendered (lowest p-values):

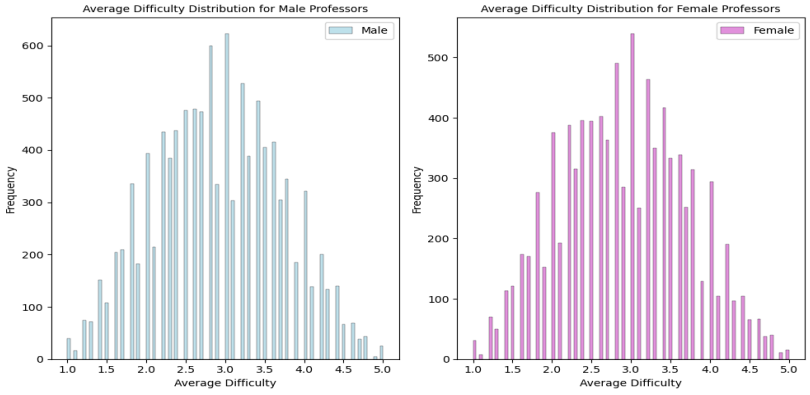
- Hilarious: 1.012e-150
- Amazing lectures: 6.594e-39
- Respected: 5.7e-29

Top 3 least gendered (highest p-values):

- Caring: 0.999
- Participation matters: 1.0
- Group projects: 1.0

The tags *Hilarious, Amazing lectures,* and *Respected* show strong statistically significant gender differences, suggesting students perceive male professors higher on these attributes. Conversely, the tags *Caring, Participation matters,* and *Group projects* show no gender bias, implying that these qualities are attributed similarly regardless of gender. This seems logical as *Participation matters* and *Group projects* depend less on personal judgment and more on the course structure.

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



The average difficulty for both genders is distributed normally. I perform an upper-tailed Welch t-test to account for possible differences in variance.
H0: Average difficulty is similar between male and female professors.
Result:
TtestResult(statistic=0.0852, pvalue=0.466, df=19450.748)
The result is not statistically significant. There is no evidence for gender difference in average difficulty.

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

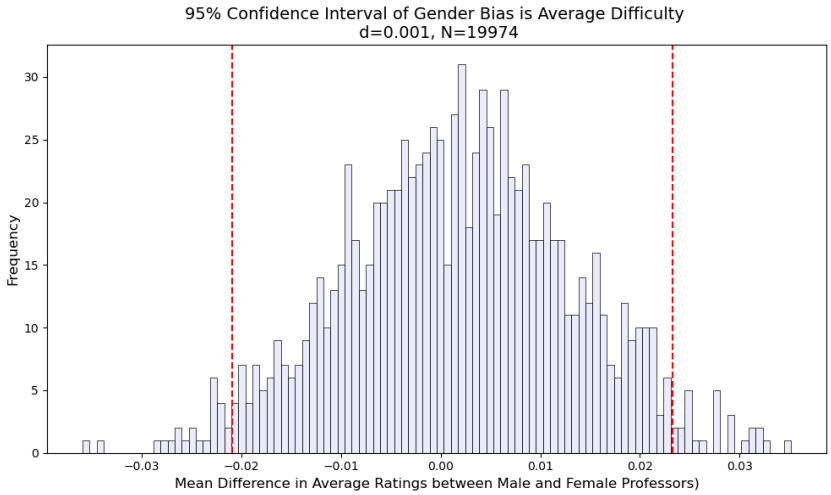
I perform 1000 iterations of Bootstrap, to compute 95% CIs for differences in mean of average difficulty between genders through resampling with replacement. I chose bootstrapping to avoid normality assumptions about data distribution. I also Cohen's d to standardize the mean difference as an effect size for gender disparity.

Result:

Cohen's d: 0.001

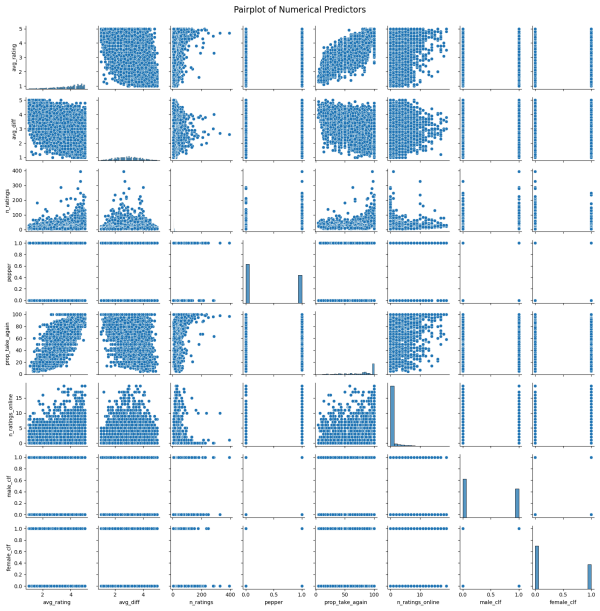
Average ratings: 95% CI for mean difference:(-0.021, 0.023)

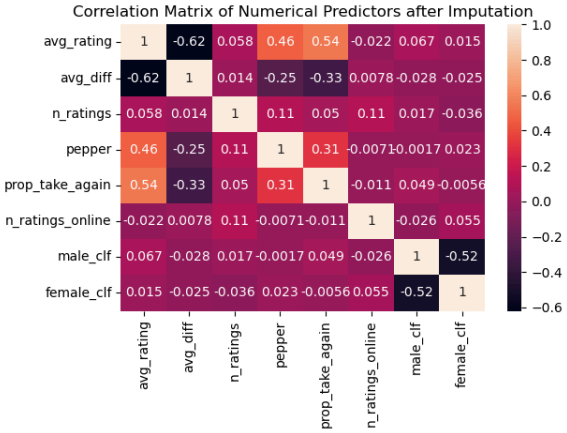
The 95% CI interval includes 0, indicating no statistically significant gender difference in average difficulty ratings at the 95% confidence level. The small range of the CI suggests that any potential difference if it exists, is possibly due to sampling error.



7. Build a regression model predicting average rating from all numerical predictors (the ones in the rmpCapstoneNum.csv) file. Make sure to include the R^2 and RMSE of this model. Which of these factors is most strongly predictive of average rating? Hint: Make sure to address collinearity concerns.

I first graph pair plot and correlation matrix for all numerical predictors to observe correlation and address potential multicollinearity. The column "probability would take again" contains N/A value, so I impute it by its mean. I drop "female classification" from the regression model since it is not linearly independent from "male classification".

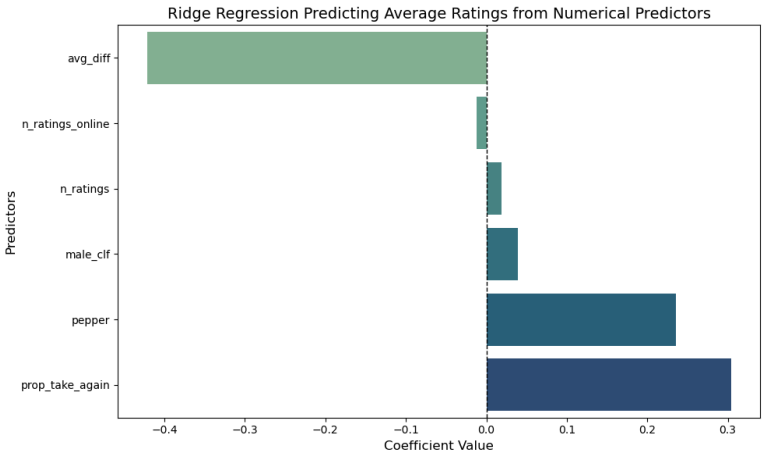




I split my imputed data into 80-20 train-test set, then within the training set, set aside 20% for validation test, with random_state as my N-number seed. I fit and transform training predictors with Standard Scaler and transform validation and test predictors so that independent numerical variables are centered and scaled to have the same variance.

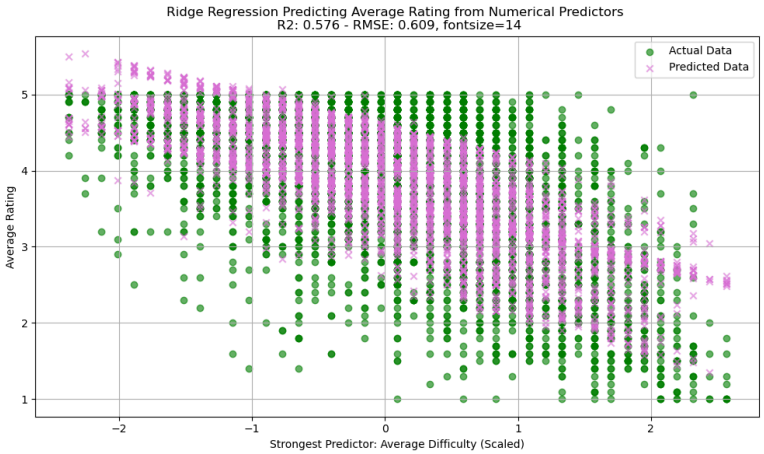
I choose to fit the data into Ridge regression to prevent overfitting and mitigate multicollinearity by adding the L1 penalty term. I tune hyperparameter alpha, and pick the alpha whose model prediction on the validation set has the lowest RMSE. I record R^2 and RMSE for the best model prediction on training, validation and test set.

Result: Ridge Regression best alpha:
0.01
-- R^2
Training set: 0.569
Validation set: 0.576
Test set: 0.576
-- RMSE
Training set: 0.623
Validation set: 0.620
Test set: 0.609

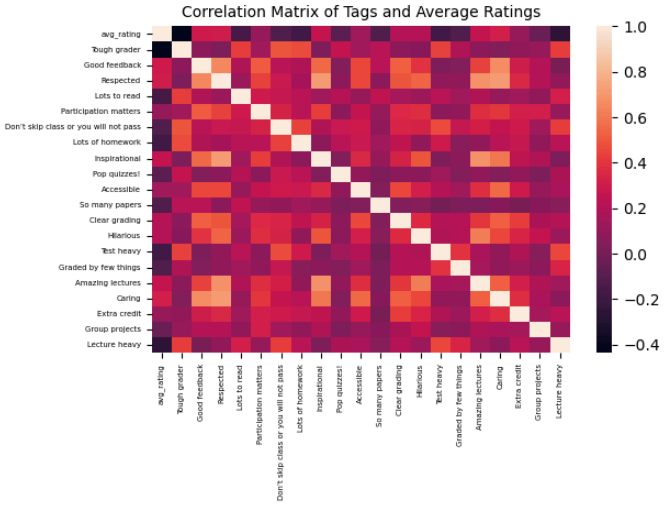


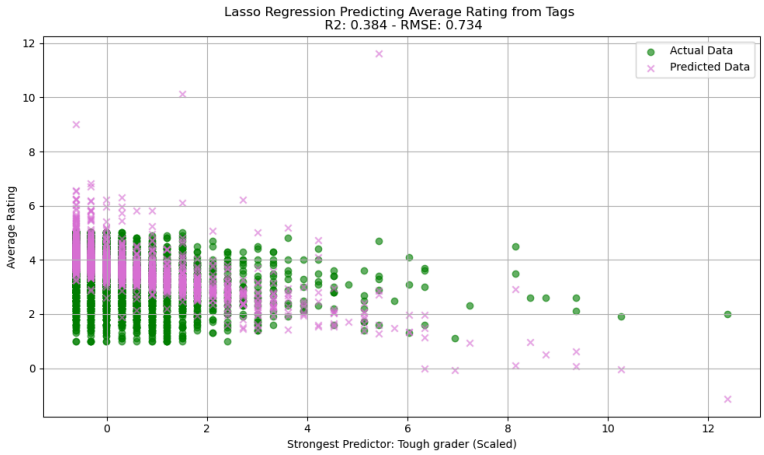
Ridge regression model performs consistently across all datasets, with a moderate COD ($R^2 \approx 0.57$), suggesting that about 57% of the variance in the target variable is explained by the predictors. Lower RMSE on the test set suggests the model performs slightly better on unseen data, demonstrating the ability to generalize.

Observing the betas for the weight of each predictor, *Average difficulty* has the strongest negative coefficient at -0,4 and has the most influence on the model, indicating that *Average ratings* are negatively associated with *Average difficulty*. *Probability would take again* and *Pepper* has moderately positive coefficients (0.3, 0.2), suggesting they also positively impact the predicted ratings but to a lesser extent than *Average difficulty*.



8. Build a regression model predicting average ratings from all tags (the ones in the rmpCapstoneTags.csv) file. Make sure to include the R^2 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.





Plotting the correlation matrix of average ratings and tag counts, multicollinearity seems to be a concern. I would use Lasso regression to allow for feature selection and reduce collinear features' weights to 0.

I split my imputed data into 80-20 train-test set, then within the training set, set aside 20% for validation test, with random_state as my N-number seed; standardize the predictors to have equal scale. I tune hyperparameter alpha for L2 penalty, and pick the alpha producing model prediction on the validation set has the lowest RMSE. I record R^2 and RMSE for the best model prediction on training, validation and test set.

Result:

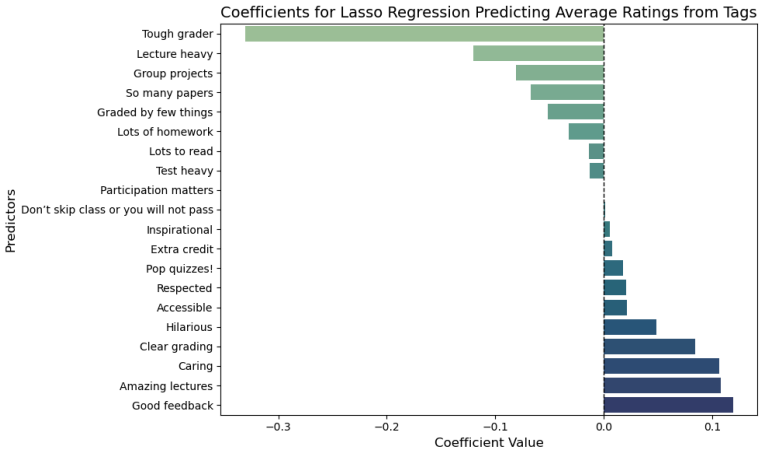
Lasso Regression best alpha: 0.01

-- R^2

Training set: 0.372
Validation set: 0.351
Test set: 0.384

-- RMSE

Training set: 0.752
Validation set: 0.768
Test set: 0.734

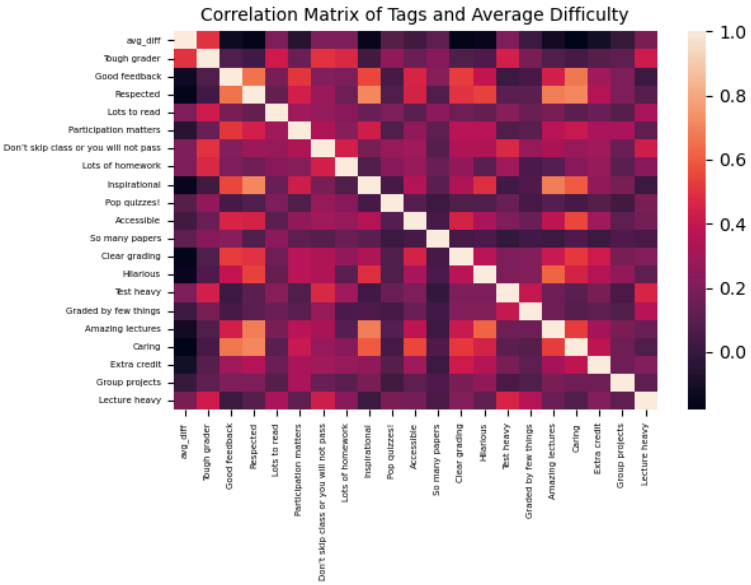


Lasso regression model results in an R^2 of 0.384, suggesting that about 38.4% of the variance in the average ratings is explained by the tag predictors. Higher R^2 and lower RMSE on the test

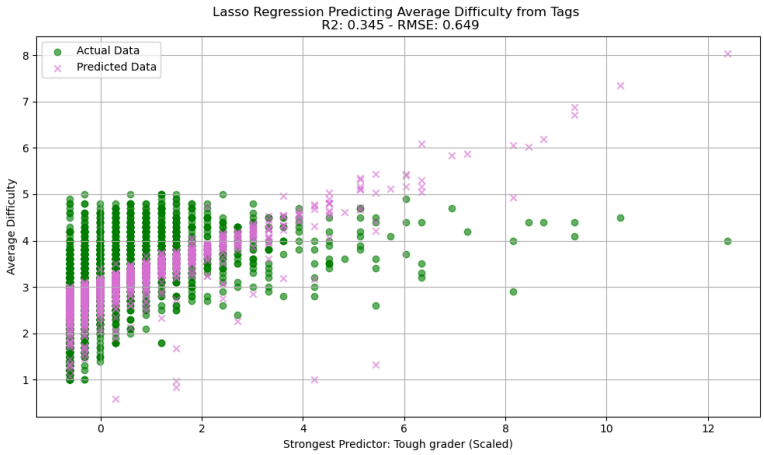
set suggest the model performs slightly better on unseen data, although underfitted to predict average ratings, compared to the Ridge regression model with numerical predictors. Linear relationship between average ratings and numerical features is stronger than with tag counts. Multiple linear regression is likely an inappropriate model to fit tag count data.

Amongst the coefficients for tag predictors, *Tough grader* has the strongest negative coefficient around -0.32 and has the most influence on the model, indicating that average ratings is negatively associated with the number of *Tough grader* tags. Other negative weight tags includes *Lecture heavy, Group Projects, So many paper*, showing that students give professors they perceive as 'tough' and demand more participation and assignments lower ratings. Positive teaching qualities like *Good feedback, Amazing lectures,* and *Caring* are weighted positively in students' rating of the professor.

9. Build a regression model predicting average difficulty from all tags (the ones in the rmpCapstoneTags.csv) file. Make sure to include the R^2 and RMSE of this model. Which of these tags is most strongly predictive of average difficulty? Hint: Make sure to address collinearity concerns.

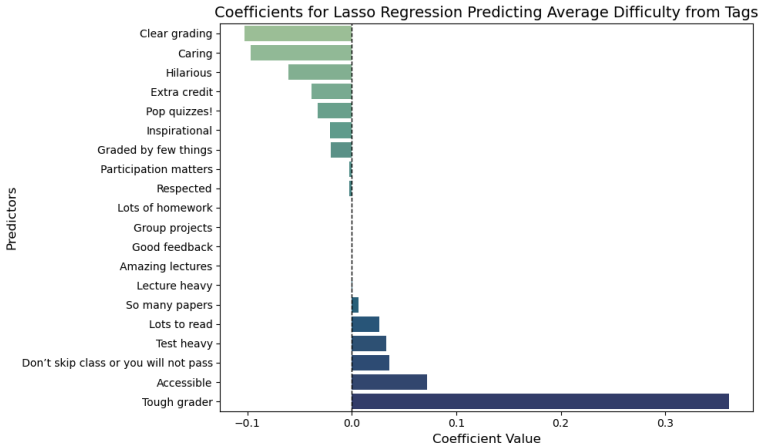


Plotting the correlation matrix of average difficulty and tags count, similarly to Question 8, multicollinearity in tags is a concern as many of these can be grouped into negative or positive sentiments. I use Lasso regression to allow for feature selection and reduce collinear features' weights to 0. Similarly, I split data into train-validation-test sets and standardize the predictors to have equal scale. I tune hyperparameter alpha for L2 penalty, and pick the alpha whose model prediction on the validation set has the lowest RMSE. I record R^2 and RMSE for the best model prediction on training, validation and test set.



Lasso Regression best alpha: 0.01
-- R^2
Training set: 0.321
Validation set: 0.313
Test set: 0.3475
-- RMSE
Training set: 0.666

Validation set: 0.665
Test set: 0.649



Lasso regression model results in an R^2 of 0.345, suggesting that about 34.5% of the variance in the average ratings is explained by the tag predictors. Higher R^2 and lower RMSE on the test set suggest the model performs slightly better on unseen data. However, the model is underfitted to average difficulty. Multiple linear regression is likely an inappropriate model to fit tag count data.

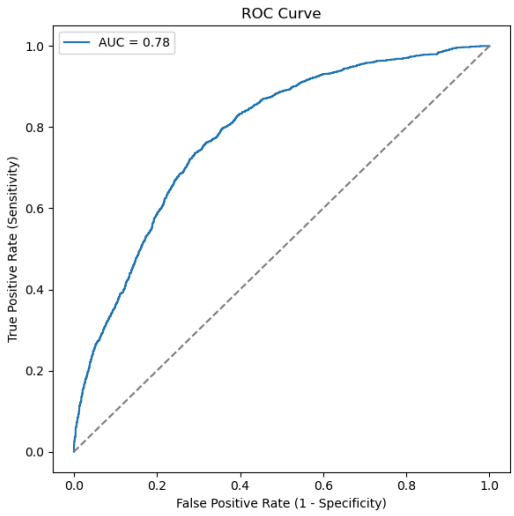
Tough grader is again the strongest coefficient at around 0.35, indicating that average difficulty is positively associated with the number of *Tough grader* tags - which is a logical correlation because amongst other factors, how lenient or tough the professors are with grades determines students' perception of class difficulty. Conversely, *Clear grading* and *Caring* tags are negatively associated with professor difficulty.

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.

There is a notable class imbalance with 57.99% of professors not receiving a "pepper" (class 0) and 42.01% receiving one (class 1), which will result in better classification for "no pepper" class.

I split 80-20 train and test sets, scale predictors with Standard Scaler and fit the training data in a multinomial logistic regression. The metrics assessing the prediction on test sets.

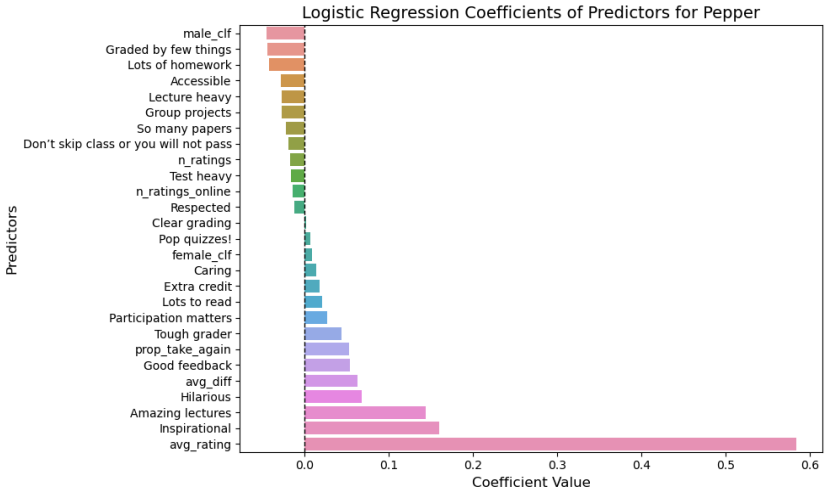
Result: AUROC = 0.78



	precision	recall	f1-score	support
0.0	0.78	0.71	0.74	2913
1.0	0.65	0.74	0.69	2161
accuracy			0.72	5074
macro avg	0.72	0.72	0.72	5074
weighted avg	0.73	0.72	0.72	5074

Optimal threshold (maximizing TPR - FPR): 0.999998

AUC = 0.78 suggests that the model is moderately good at distinguishing between the two classes. The optimal decision threshold for classification is extremely close to 1 showing the model is heavily biased toward classifying observations as class 0 (no pepper). This threshold setting reduces false positives but also results in fewer predicted pepper cases (class 1), therefore a lower precision but higher recall for class 1.



Average Rating is the strongest positive coefficient (0.583), meaning higher-rated professors are more likely to receive a "pepper." *Inspirational* and *Amazing lectures* are also positive predictors for a professor receiving a "pepper" - there's a correlation between teaching style and student admiration. Negative tags include *Male classification, Graded by Few Things, Lots of homework* discourage students from awarding a "pepper."

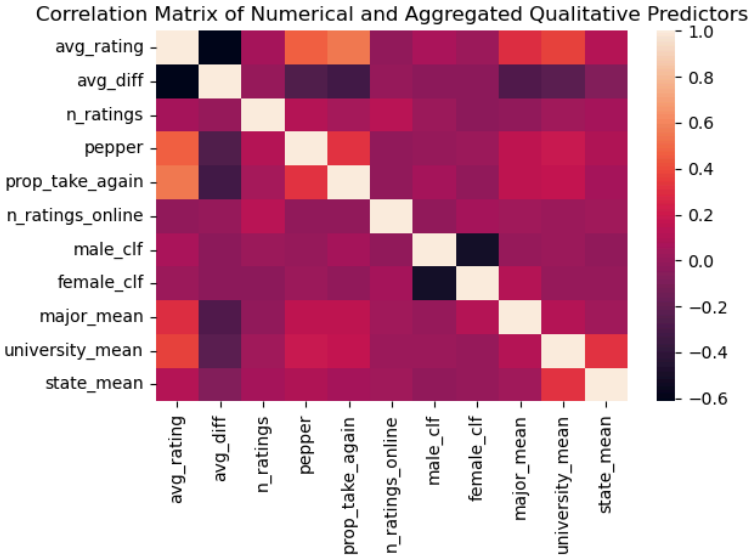
The coefficients suggest that students are heavily influenced by teaching qualities and enjoyment factors like inspiration, humor)when deciding whether to award a "pepper," while factors associated with academic demand tend to have a negative impact.

Extra credit: Tell us something interesting about this dataset that is not trivial and not already part of an answer (implied or explicitly) to these enumerated questions[Suggestion: Do something with the qualitative data, e.g. major, university or state by linking the qualitative data to the two other data files (tags and numerical)].

Because regression with numerical data results in the best model predicting average ratings, I want to explore whether additional dimensions with qualitative data improve the performance of the regression model.
Qualitative predictors (major, university, state) were aggregated by computing the mean average rating for each category in the training dataset. This creates numerical features that summarize the target variable based on qualitative groupings:

- major_mean: Mean average rating per major.
- university_mean: Mean average rating per university.
- state_mean: Mean average rating per state.

The aggregation transformed qualitative predictors into meaningful numerical features while avoiding one-hot encoding too many dimensions.
These aggregated features were mapped to both the training and test datasets. For unseen categories in the test set, the global mean from the training set was used to avoid data leakage. The test set is further split into training (60%) and test (40%) sets, followed by an additional split of the test set into validation (50%) and final test (50%) sets. StandardScaler was applied to normalize all numerical features, ensuring consistent scaling across predictors.



I use Lasso regression model to identify important predictors by shrinking less significant, collinear, coefficients toward zero.
I tune the model with a a range of alpha values, and the best alpha was chosen based on the lowest validation RMSE.

Result:

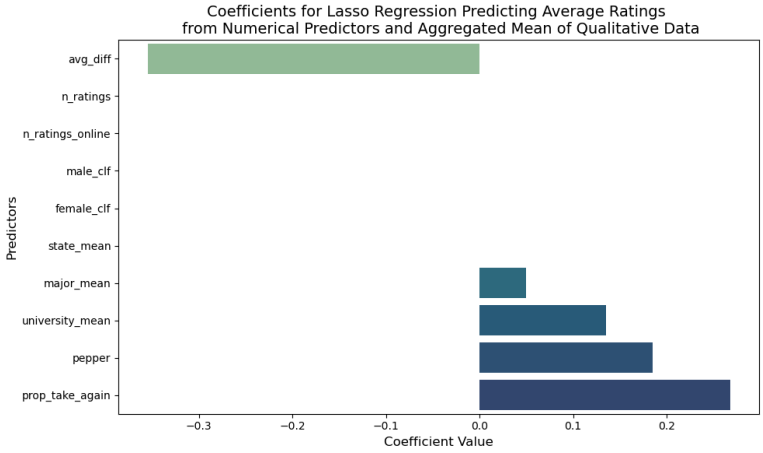
Lasso Regression best alpha: 0.05

-- R^2

Training set: 0.596
Validation set: 0.538
Testing set: 0.549

-- RMSE

Training set: 0.604
Validation set: 0.637
Testing set: 0.634



The R^2 indicate that the model explains about 55% of the variance in average ratings on the test set, RMSE is stable around 0.63 for both validation and test set, indicating generalizability. However, these metrics are slightly worse than the Ridge regression using only numerical values. In this case, adding more dimensions with mean average rating for each qualitative category (major, university, state) does not improve model performance.

The strongest predictors for this model are numerical data: *Average difficulty* (negative coefficient), *Probability would take again, Pepper. Aggregated ratings by university* also contribute positively to average ratings prediction. Predictors that Lasso shrinks towards 0 include: *Aggregated state state mean rating*, *female and male classification* and *number of ratings* - which does not add any independent dimensions to our regularized regression model.