Intro to Data Science Capstone

Paola Calle, Mario Tan

December 2025

1 Introduction

Student evaluations of teaching are a controversial metric often criticized for potential bias. This study provides a rigorous statistical examination of RateMyProfessor data to quantify the extent of gender bias and determine the structural drivers of average ratings as given by university students.. We aim to utilizing signifiance tests for inference and normalized linear regression for prediction. By combining statistical hypothesis testing, natural language processing, and machine learning, we move beyond anecdotal evidence to quantitatively deconstruct the relationships between gender, difficulty, and student ratings.

Code Availability

The underlying code and all figures for our analytics is available in the project repository: Every question is associated with a single .ipynb in analysis/. For example, Q1.ipynb is all code for question 1.

https://github.com/paolacalle/ape-capstone/tree/main

2 General Data Cleaning

2.1 Gender

After an initial EDA, we notices that gender had faulty data in the sense that the gender of some observations were unclear. Given that many of our observations requires comparative analysis between gender, we decided to exclude cases where an observation were marked for both female and male or neither.

2.2 Students who would take the class again

The original dataset included a 5th Column for a categorical variable (The % of students who report they would take the class again). Due to the low response rate (85% N/A responses) and its limited utility for numerical regression compared to tags and difficulty scores, this feature was removed during preprocessing to streamline the dimensionality of the model.

2.3 Tags Normalization

Because professors receive very different numbers of ratings, raw tag counts are not directly comparable. We therefore normalize each tag by the total number of ratings for the professor. For professor i and tag t, the tag intensity is defined as

$$\text{Intensity}_{i,t} = \frac{\text{TagCount}_{i,t}}{\text{NumRatings}_i}.$$

This normalization yields a per-rating measure that reflects how frequently students award each tag while controlling for differences in exposure across professors. Comparing the correlation matrices before and after normalization (Figure 10), we observe a reduction in strong pairwise correlations, indicating that normalization mitigates the influence of raw tag frequency and reduces apparent collinearity among tag features.

Summary statistics for the normalized tag features indicate that z-score standardization was applied correctly, with means approximately equal to zero across all tags. The median value for each tag equals its minimum, reflecting strong right-skewness: most professors receive low tag intensity scores, while a smaller subset receives substantially higher values. Several tags (e.g., `extra_credit` and `group_projects`) exhibit large maximum values, indicating the presence of rare but highly intense observations. Despite this skewness, all features retain nontrivial variability and are therefore retained for subsequent modeling, with coefficient magnitudes interpreted in relative rather than unit-based terms.

Table 4 shows that all normalized tag features exhibit substantial spread and strong right-skewness, with several tags displaying rare but highly intense values.

3 Questions

3.1 Pro-male gender bias

Average ratings are strongly left-skewed and ordinal (Figure 1), so we use a Mann–Whitney U test rather than a t-test. Testing the one-sided hypothesis that male professors receive higher ratings yields a statistically significant result (U = 346,126,732.5, p = 4.14×10^{-6}).

Exploratory analysis shows no clear relationship between average rating and number of reviews, nor any gender-specific trend (Figure 2). As a robustness check, we stratify the data by the median number of reviews and repeat the Mann–Whitney U test within each stratum. In both strata, ratings for male professors remain statistically higher (Table 1).

3.2 Gender Difference in the Spread

To assess gender differences in the dispersion of professor ratings, we conducted non-parametric permutation tests using both the variance and the interquartile range (IQR) as measures of spread. For each test, gender labels were randomly permuted 10,000 times under the null hypothesis that ratings are exchangeable with respect to gender, and the absolute difference in dispersion between groups was recomputed for each permutation.

Figures 3 and 4 show the resulting permutation null distributions for the variance and IQR tests, respectively. The red dashed line indicates the observed dispersion difference. While the variance-based test indicated a significant difference in dispersion, the IQR-based test (p = 0.005) provides robust evidence that the central spread of ratings differs between male and female professors. We see that female professors have a significantly higher variance (and wider IQR) than male professors. This confirms the result is highly significant because a difference this large cannot be explained by chance. Female professors are apparently more "polarizing" than male professors or rather students have less consensus on female professors (a wider mix of very high and very low grades) compared to males, whose ratings cluster more tightly around the higher end of the ratings distribution.

3.3 Gender Difference Effects

The estimated gender difference in average ratings is small. The standardized mean difference corresponds to a Cohen's $d = 0.060$ with a 95% confidence interval of $[0.043, 0.077]$, indicating that male professors' mean ratings are higher by approximately 0.06 standard deviations. Because ratings are bounded and left-skewed, we additionally report a robust, unstandardized effect. On the original 1–5 rating scale, the median rating for male professors exceeds that for female professors by 0.067 points, with a 95% bootstrap confidence interval of $[0.048, 0.086]$, suggesting a small but precisely estimated shift in typical ratings.

Gender differences are more pronounced in the dispersion of ratings. The variance ratio is $\mathrm{Var}_{\mathrm{male}}/\mathrm{Var}_{\mathrm{female}} = 0.90$ with a 95% confidence interval of $[0.88, 0.92]$, while the IQR ratio is $\mathrm{IQR}_{\mathrm{male}}/\mathrm{IQR}_{\mathrm{female}} = 0.83$ with a 95% confidence interval of $[0.83, 0.89]$. Ratios are used to compare dispersion because variance and IQR are scale-dependent quantities, and ratios provide a unit-free measure of relative spread with a natural null value of one. These results indicate that ratings for male professors are consistently less variable than those for female professors, even though differences in central tendency are small in magnitude.

Note. All confidence intervals are estimated via bootstrap resampling with 2,000 iterations.

3.4 Gender Differences in Student-Awarded Tags

To evaluate whether students award qualitative tags differently by professor gender, we employ a two-case analysis designed for the strongly zero-inflated nature of the tag data. Because students may award at most three tags per rating, most professor–tag combinations take a value of zero.

Case 1 (Tag Awarding Probability). We first test for gender differences in whether a tag is awarded at all by comparing the proportion of zero tag intensities between male and female professors using permutation tests. Among the most gendered tags, we observe statistically significant differences for *accessible*, *good_feedback*, and *respected* (all $p = 0.0002$). The positive differences for *accessible* and *respected* indicate that female professors are more likely to receive zero instances of these tags, meaning students are less likely to award them at all. In contrast, the negative difference for *good_feedback* indicates that female professors are more likely to receive this tag.

In contrast, *tough_grader*, *no_skip*, and *clear_grading* exhibit weak evidence of gender differences, with non-significant p-values and small effect sizes, indicating similar probabilities of being awarded to male and female professors. These results indicate that gender differences in student-awarded tags primarily manifest in whether a tag is applied at all.

Case 2 (Non-Zero Tag Intensity). We next examine gender differences in tag intensity conditional on a tag being awarded at least once (as plotted in Figure 6). This conditional analysis asks whether gender differences persist in tag intensity once a tag has been awarded. We compare non-zero tag intensities using permutation tests on the median.

Three tags exhibit statistically significant differences in non-zero intensity. For *tough_grader* and *good_feedback*, female professors receive higher median tag intensity among those who receive the tag, whereas *hilarious* exhibits the opposite pattern, with male professors receiving higher intensity. Tags such as *test_heavy*, *papers*, and *extra_credit* show no statistically significant differences in non-zero intensity. Moreover, many tags (including *caring*, *amazing_lectures*, *accessible*, and *respected*) yield p-values of 1.0, which occurs because the median non-zero intensity is identical across genders for these tags.

Summary: Taken together, the two-case analysis indicates that gender differences in student-awarded tags are driven primarily by differences in the likelihood of receiving a tag rather than by differences in intensity once awarded. Tags associated with perceptions and interpersonal attributes exhibit the strongest gender differences, while tags describing grading structure or course policies are largely gender-neutral.

Likelihood (Case 1): The biggest differences are in who gets the tag at all. Male Professors: Are far more likely to be tagged as Hilarious (+13%), Amazing Lectures (+7%), and Respected (+6%). Female Professors: Are far more likely to be tagged as Participation Matters (-4%), Caring (-3%), and Good Feedback (-3%).

Intensity (Case 2): Once a tag is awarded, the gender gap disappears for most tags, with a few key exceptions: ”Hilarious” is the only tag where men get it more often AND more intensely, while ”Tough Grader” and ”Good Feedback” are tags where, if a woman gets them, she gets them more intensely than a man would.

3.5 Class Difficulty Gendered?

We find no statistically significant difference in average difficulty ratings between male and female professors (Mann–Whitney U test, $p = 0.64$). Consistent with this result, the estimated effect size is negligible (Cliff's $\delta = 0.0024$), indicating virtually no practical difference in perceived course difficulty across genders.

Note. Cliff's δ is reported as a nonparametric effect size appropriate for ordinal data.

3.6 Class Difficulty Gendered Effect

The estimated mean difference in average difficulty ratings (female − male) is 0.0057 (95% bootstrap CI $[-0.0044, 0.0230]$), ruling out any substantively meaningful gender difference in perceived course difficulty. This conclusion is consistent with the negligible effect size estimated using Cliff ($\delta = 0.0024$).

3.7 Regression model for average rating from number columns.

We fit a linear regression model to predict average professor rating using all numerical predictors from `rmpCapstoneNum.csv`. The variable `would_take_again_prop` was removed due to substantial missingness

(43,239 cases). To reduce redundancy, num_online_reviews was converted into a binary indicator (has_online_ratings), and the separate male and female columns were combined into a single gender_code variable (0 = male, 1 = female). All predictors were standardized prior to analysis.

A correlation analysis showed a positive association between average rating and pepper, and a negative association with avg_difficulty. Moderate correlation was observed between pepper and num_ratings, while other predictors exhibited relatively weak pairwise correlations, suggesting limited collinearity (Figure 9). To further assess variance structure, we conducted a PCA, which showed that four principal components were required to explain approximately 85% of the variance (Figure 11). Comparing a regression model using these four PCs to a model using the original five predictors via 5-fold cross-validation yielded nearly identical performance (mean RMSE = 0.81, mean R^2 = 0.35), indicating that multicollinearity does not meaningfully affect model performance. R^2 = 0.35 means that our model explains 35% of the variation in professor ratings. This is a solid result for behavioral data. RMSE = 0.81 means that on average, the model's prediction is off by 0.81 stars. Since the standard deviation of ratings is 1.11, our regression model is significantly better than just guessing the average.

Examining the average absolute coefficients across folds from the model, avg_difficulty was the most strongly predictive factor (highest aboslute value for all Beta-values), followed by pepper, while gender_code and has_online_ratings had comparatively smaller effects (Table 2). We can reasonably predict that students give worse ratings to professors that they perceived as being difficult or teaching difficult classes.

3.8 Regression model for average rating from tags.

As discussed in Section 3.4, normalizing tag counts with respect to the number of reviews substantially mitigates strong correlations among tag variables. As an additional diagnostic, we conduct a PCA to assess whether the variance in the normalized tag features is concentrated in a small number of latent components.

The first principal component explains approximately 11% of the total variance, with no clear elbow observed in the scree plot . Approximately 17 principal components are required to explain about 85% of the cumulative variance, indicating that variance is broadly distributed across features rather than concentrated in a low-dimensional structure. Using the first 17 components in a regression model yields an average RMSE of 0.79 and an average R^2 of 0.50, with low variance across cross-validation folds. Overall, PCA offers limited dimensionality reduction benefits and primarily serves as a diagnostic confirming reduced but non-negligible collinearity among normalized tag features.

Given these results, we retain all original features and fit a linear regression model. To assess feature importance while accounting for variability across folds, we examine the mean absolute value of each coefficient across the five cross-validation folds.

Across five cross-validation folds, the features with the largest mean absolute coefficient magnitudes are consistently related to qualitative teaching attributes. In particular, tough_grader and good_feedback exhibit the strongest effects, appearing among the top coefficients in all five folds. Additional stable predictors include respected, caring, and amazing_lectures, indicating that student perceptions of instructional quality and engagement are the primary drivers of average ratings (Table 3).

Model Comparison. The tag-based regression model outperforms the numerical predictor model, achieving a higher average R^2 (0.50 vs. 0.35) and a slightly lower RMSE (0.79 vs. 0.81). While the numerical predictors exhibit limited collinearity and a low-dimensional variance structure—requiring only four principal components to explain approximately 85% of the variance—the tag features display broadly distributed variance, with approximately 17 components needed to reach a similar threshold. PCA-based regression yields performance comparable to using the original features in both cases, indicating that multicollinearity does not materially affect predictive accuracy. Overall, the stronger performance of the tag-based model suggests that qualitative student perceptions captured by tags are more directly predictive of average ratings than structural or contextual numerical attributes.

3.9 Predicting Average Difficulty from Tag Features

We build a regression model to predict average difficulty using all tag features from rmpCapstoneTags.csv. As discussed in Section 3.4, normalizing tag counts with respect to the number of reviews substantially mitigates strong correlations among tag variables. Compared to average rating, the distribution of average difficulty is more symmetric (Figure 15), supporting stable linear modeling.

To assess potential multicollinearity among the normalized tag features, we first conduct a PCA as a diagnostic. The scree plot and cumulative variance explained (Figures 13 and 14) show that approximately 16 components are required to explain 85% of the total variance, indicating that variance is broadly distributed across features rather than concentrated in a low-dimensional structure. A regression model fit using these components yields comparable but slightly worse predictive performance, using five-fold cross-validation (mean RMSE $= 0.81$, mean $R^2 = 0.48$), suggesting that PCA does not meaningfully improve prediction.

Given these results, we retain the original normalized tag features and fit a standard linear regression model, evaluated using five-fold cross-validation. This model achieves a mean RMSE of 0.79 (variance 4.8×10^{-5}, SD 0.007) and a mean R^2 of 0.50 (variance 2.2×10^{-5}, SD 0.005).

To assess feature importance while accounting for fold-to-fold variability, we examine the mean absolute value of each coefficient across the five cross-validation folds, and inspect the minimum and maximum coefficient values to verify directional consistency. The most strongly predictive tag is `tough_grader` (mean coefficient $= -0.28$), followed by `good_feedback` (0.24), `respected` (0.17), `caring` (0.18), and `amazing_lectures` (0.17). These features appear among the top coefficients in all five folds, indicating stable and consistent effects. Overall, perceived grading strictness and instructional quality are the primary drivers of average difficulty ratings.

3.10 Classification Model for Professor "Pepper"

Despite right-skewness in several tag features, logistic regression remains stable (as described in Section 2, and these features are retained due to their potential discriminative value for predicting `pepper`.

3.10.1 Predicting `pepper`

We build a binary classification model to predict whether a professor receives a `pepper` using all available predictors, including both numerical variables and normalized tag features. Exploratory analysis of the target variable reveals moderate class imbalance: approximately 72% of professors do not receive a pepper, while 28% do. Consequently, a naive classifier predicting only the majority class would achieve roughly 72% accuracy, rendering accuracy an unreliable evaluation metric.

To address this imbalance, we employ stratified five-fold cross-validation and fit a logistic regression classifier with standardized predictors. Model quality is assessed using threshold-independent metrics that are robust to class imbalance. Across folds, the model achieves a mean AUROC of 0.776 (see 17 and a mean precision–recall AUC of 0.548, indicating meaningful discrimination between professors with and without a `pepper` beyond the majority-class baseline.

Using the default probability threshold of 0.5 yields high accuracy (0.75) but poor recall for the minority class: only 29.7% of professors who receive a `pepper` are correctly identified. To improve minority-class detection, we apply threshold moving based on the precision–recall curve and select a decision threshold of 0.278 that maximizes the F1 score. At this threshold, recall for the positive class increases substantially to 75.2%, while precision decreases to 45.8%. Balanced accuracy improves from 0.61 to 0.71, and the F1 score increases from 0.40 to 0.57. As expected under class imbalance, overall accuracy decreases to 0.68 due to an increase in false positives (Table 5).

Overall, threshold moving substantially improves detection of professors who receive a `pepper` and provides a more appropriate operating point than the default cutoff. Because accuracy is highly sensitive to both class imbalance and threshold choice, we emphasize AUROC, precision–recall performance, balanced accuracy, and recall when evaluating model quality.

Lowering the decision threshold increases recall for the minority class from 29.7% to 75.2%, demonstrating why accuracy alone is misleading under class imbalance.

3.11 Extra Credit: Does the joint distribution of (major, pepper) differ between female and male professors?

To assess whether the joint distribution of academic major and `pepper` status differs between female and male professors, we conduct a chi-square test of independence using a combined (major, pepper) categorical variable. The test indicates a statistically significant association between gender and the joint major–pepper distribution ($\chi^2 = 1399.28$, $df = 19$, $p < 10^{-280}$).

To further examine whether this association is driven primarily by pepper status, we perform stratified chi-square tests within each pepper group. Significant associations between gender and major are observed both among professors without a pepper ($\chi^2 = 1031.16$, $p < 10^{-200}$) and among those with a pepper ($\chi^2 = 367.92$, $p < 10^{-70}$). These results indicate that gender differences in major distribution persist regardless of pepper status.

Gender is associated with differences in major distribution both overall and within pepper strata, but the magnitude of this association is small despite strong statistical significance.

4 Conclusion

Ultimately, we can conclude from our given dataset that RateMyProfessor ratings are less a measure of objective pedagogical quality and more a reflection of transactional ease and entertainment value. While gender bias evidently persists in the language students use, the numerical scores are fundamentally driven by the desire for a high grade and to a lesser degree, an engaging classroom experience. For university administrators, we recommend that this implies that raw student evaluations be normalized by department and difficulty level to ensure a fair assessment of faculty performance as opposed to ratings given by students who may not be representative of the wider population.

A Appendix

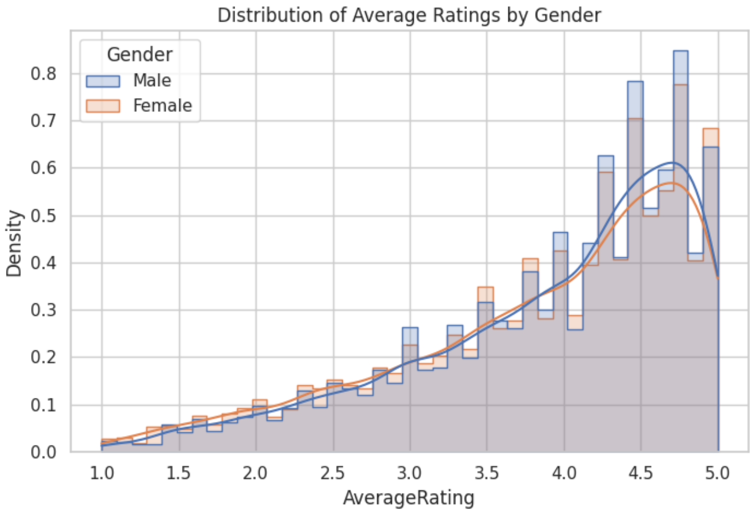


Figure 1: Distribution of average professor ratings according to gender. The distribution is left-skewed with a heavy concentration near the upper bound. Male Professors are granted slightly higher ratings than female professors.

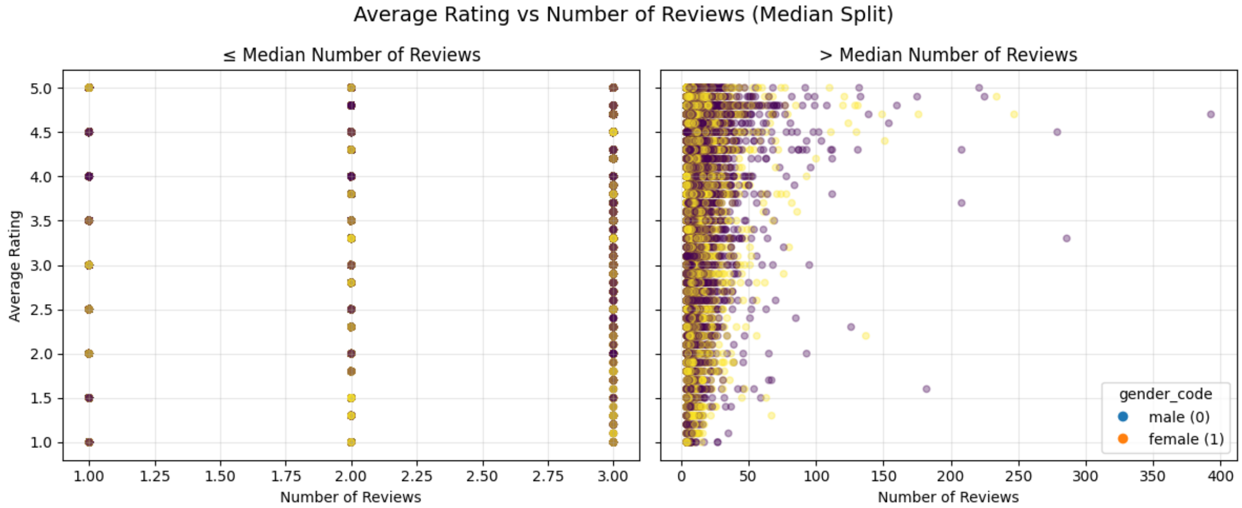


Figure 2: Average rating versus number of reviews, stratified by median review count and colored by gender. No clear relationship between review volume and rating is apparent.

Table 1: Mann–Whitney U tests stratified by review volume

Review Volume	n (M)	n (F)	Mean (M)	Mean (F)	Median (M)	Median (F)	p-value
Below median	14,700	14,107	3.855	3.787	4.3	4.0	0.000964
Above median	12,463	10,818	3.905	3.843	4.2	4.1	0.000039

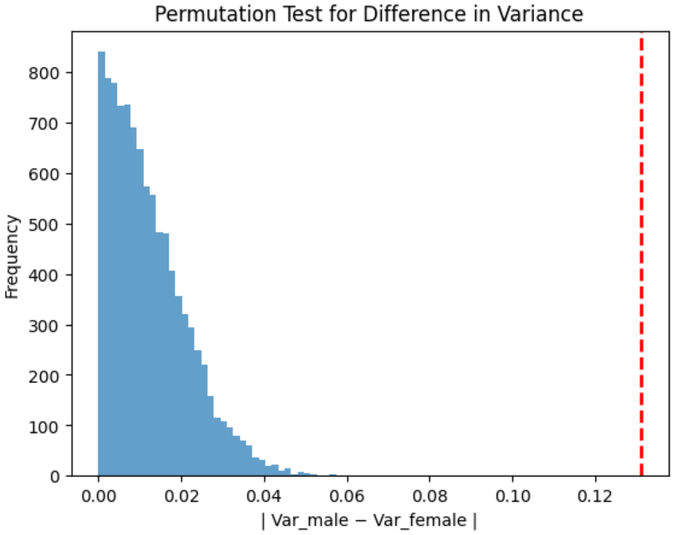


Figure 3: Permutation test null distribution for the absolute difference in variances between male and female professors' ratings. The dashed line denotes the observed statistic of about 0.13.

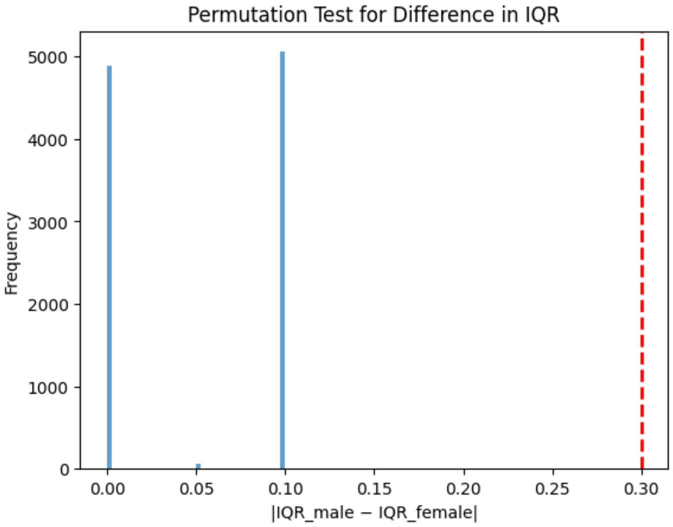


Figure 4: Permutation test null distribution for the absolute difference in interquartile ranges (IQR) between male and female professors' ratings. The dashed line denotes the observed statistic of about 0.30 ($p = 0.005$).

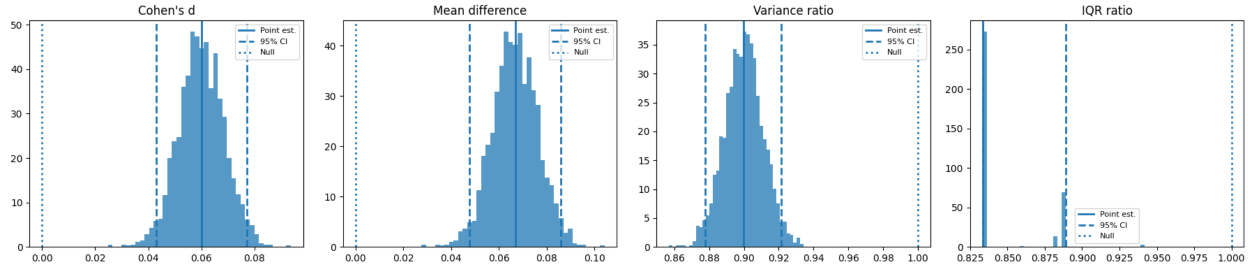


Figure 5: Bootstrap distributions (2,000 resamples) for four effect size estimates: Cohen's d, mean difference, variance ratio, and IQR ratio.

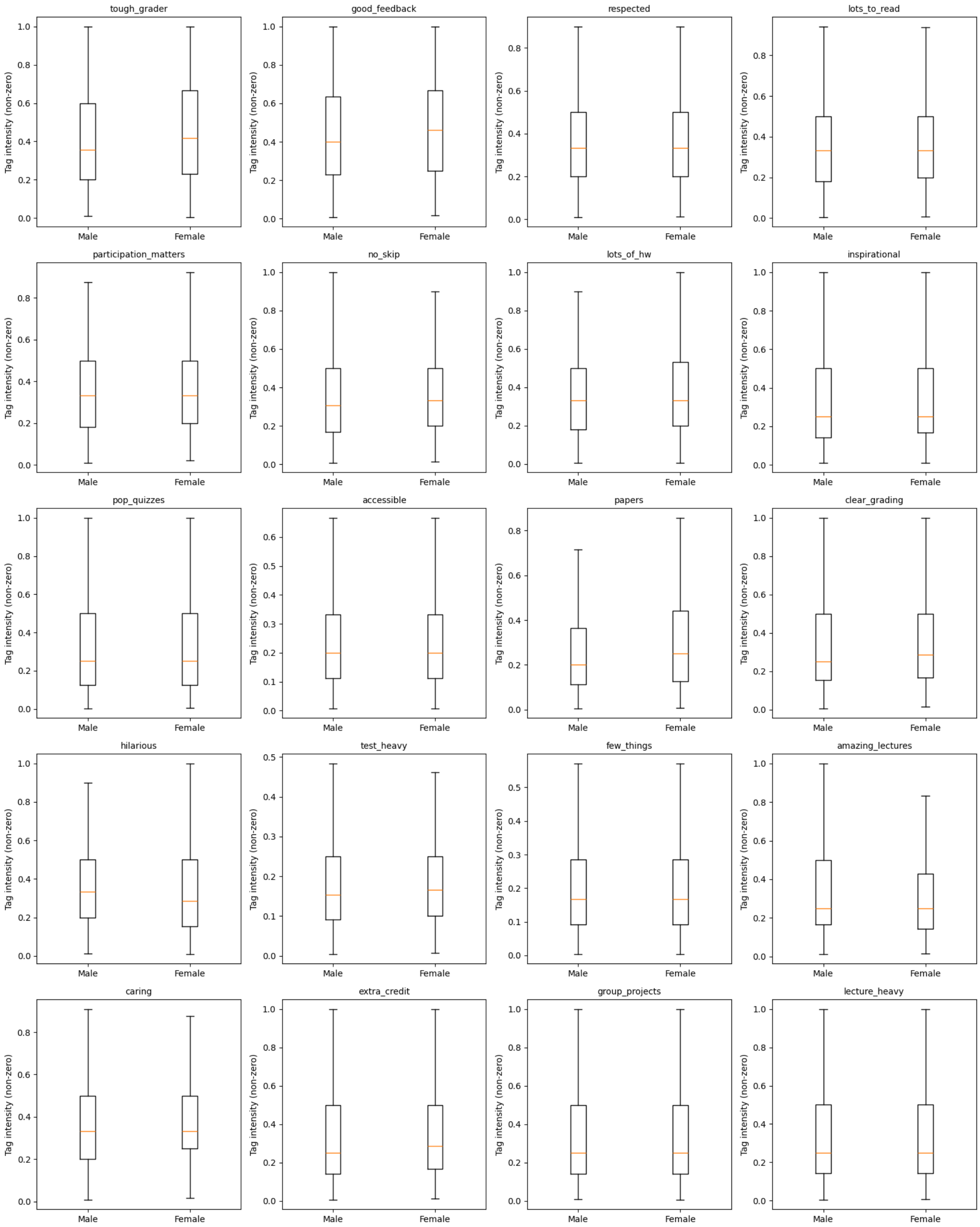


Figure 6: Boxplots of non-zero tag intensity by professor gender.

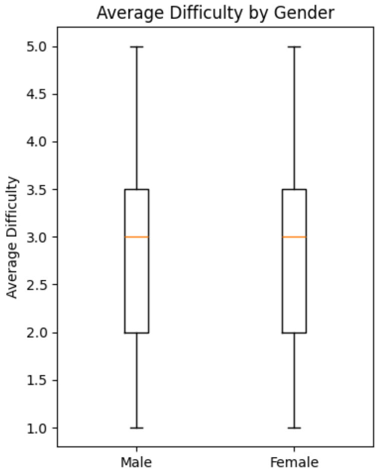


Figure 7: Distribution of average difficulty ratings by professor gender.

Table 2: Average absolute coefficients across 5-fold cross-validation

Feature	Mean $|\beta|$	Times in Top 5
num_ratings	0.387	5
avg_difficulty	0.299	5
pepper	0.203	5
gender_code	0.181	5
has_online_ratings	0.032	5

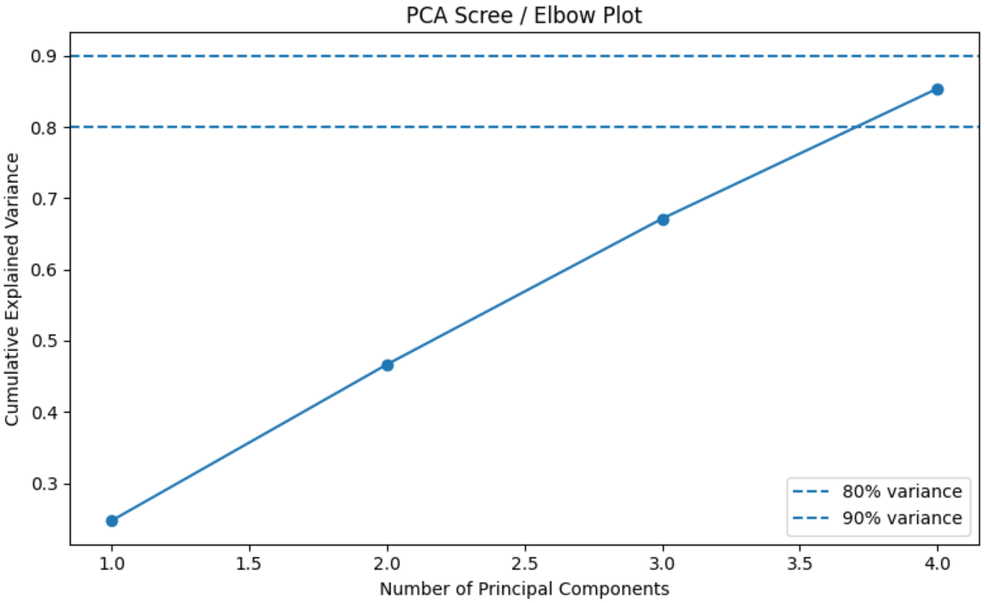


Figure 8: PCA scree (elbow) plot showing cumulative explained variance.

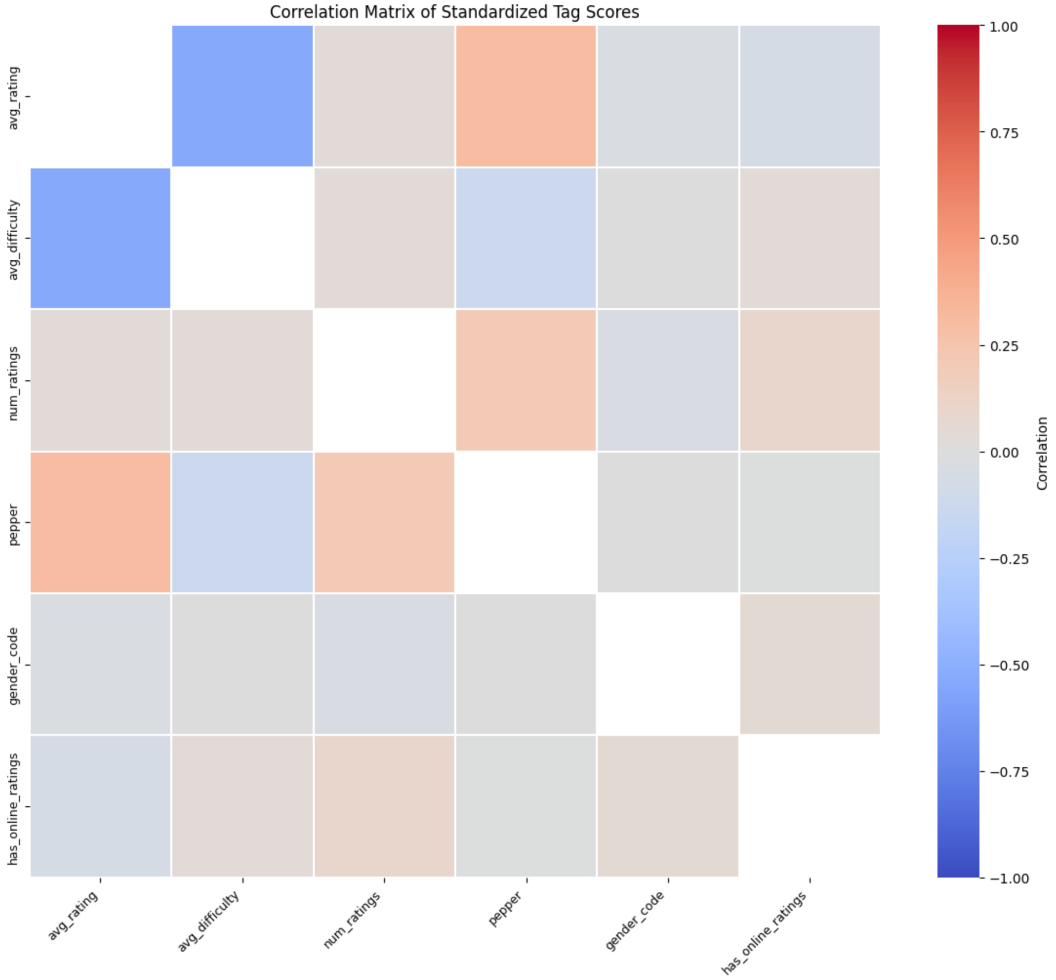
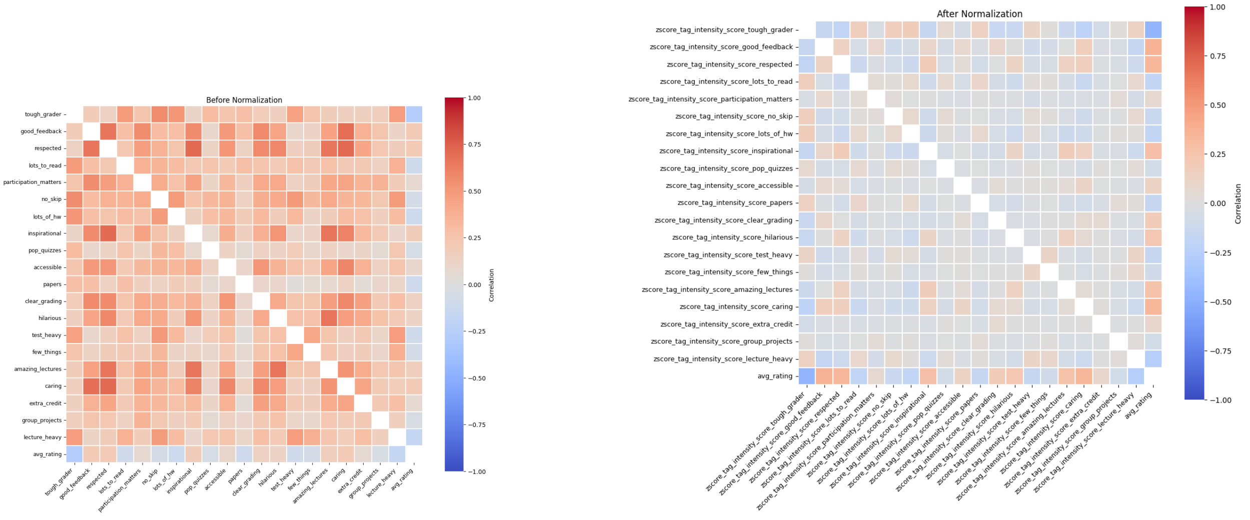


Figure 9: Correlation matrix of standardized predictors and average rating.



(a) Before normalization.

(b) After normalization (z-score tag intensity).

Figure 10: Correlation matrices of tag features and average rating before vs. after normalization. Normalization reduces the dominance of raw tag frequency and makes correlations more comparable across tags.

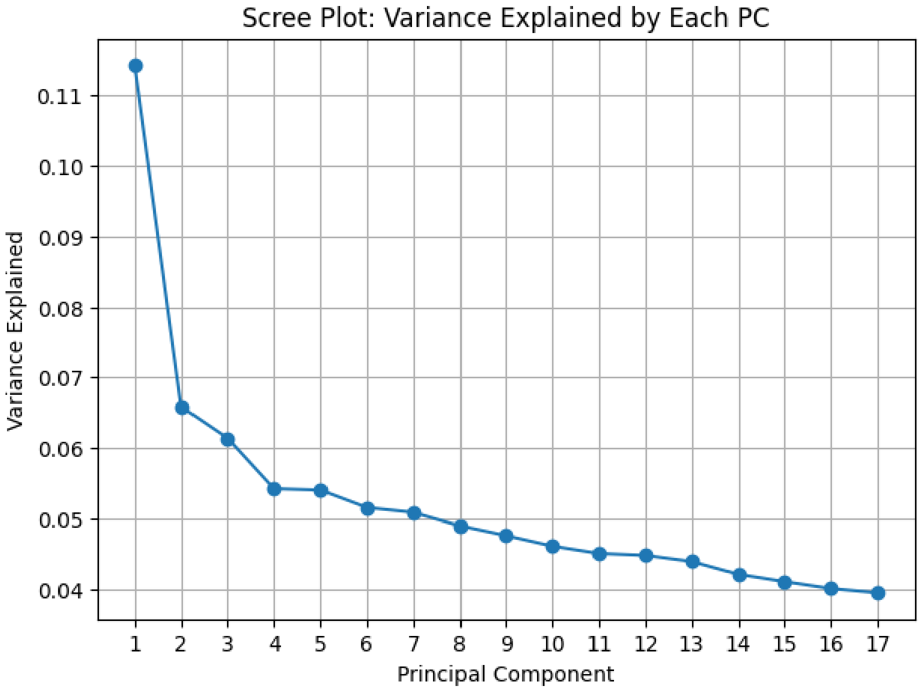


Figure 11: Scree plot showing the variance explained by each principal component.

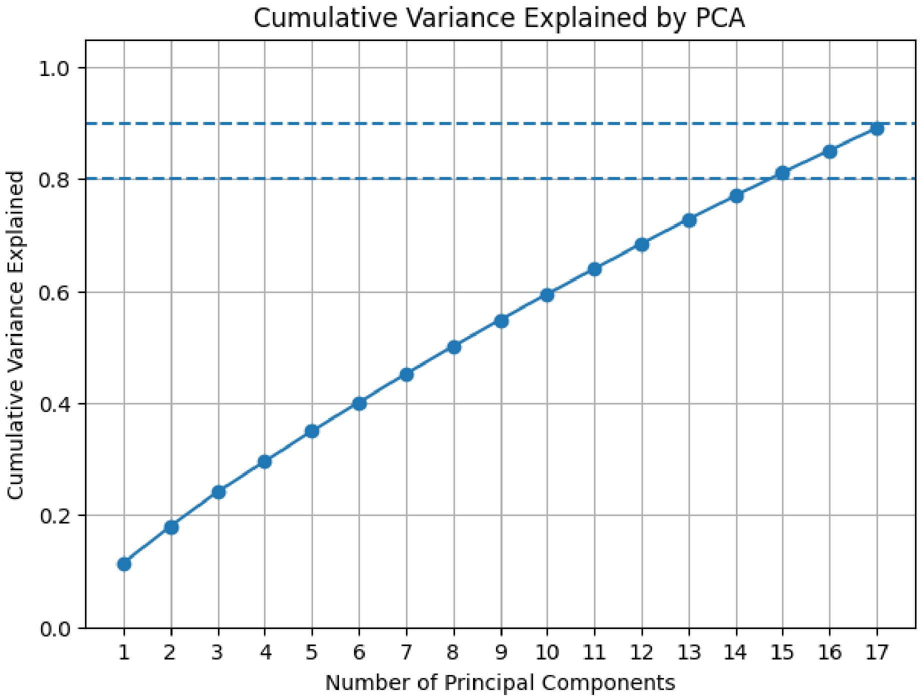


Figure 12: Cumulative variance explained by PCA as a function of the number of components.

Table 3: Top tag features by mean absolute coefficient magnitude across five cross-validation folds.

Feature	**Mean $\lvert\beta\rvert$**	**Folds in Top 5**
Tough grader	0.275	5
Good feedback	0.247	5
Respected	0.178	5
Caring	0.178	5
Amazing lectures	0.168	5

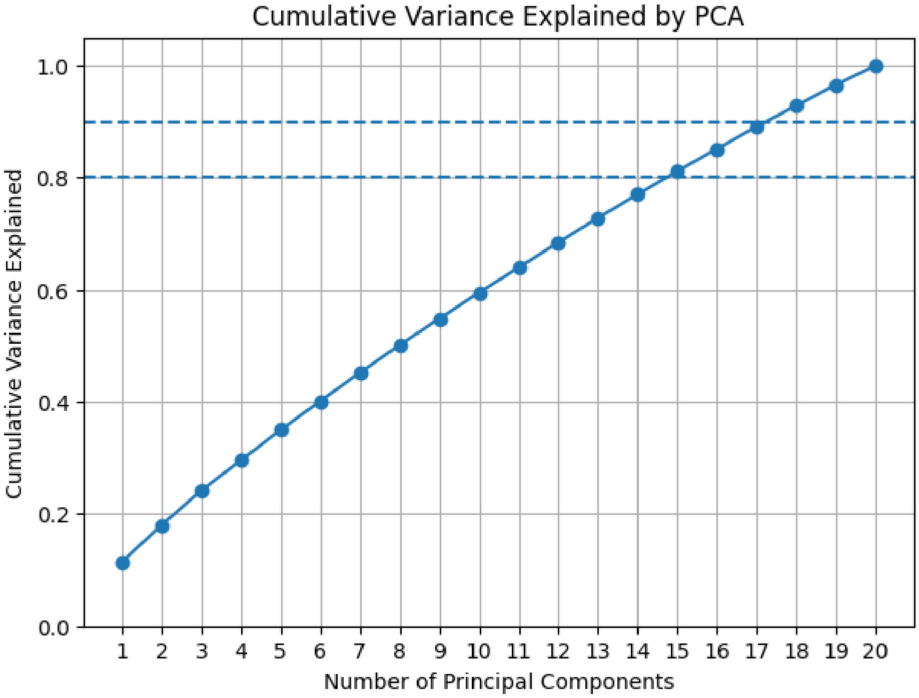


Figure 13: Cumulative variance explained by PCA on normalized tag features for average difficulty. Approximately 16 principal components are required to explain 85% of the total variance.

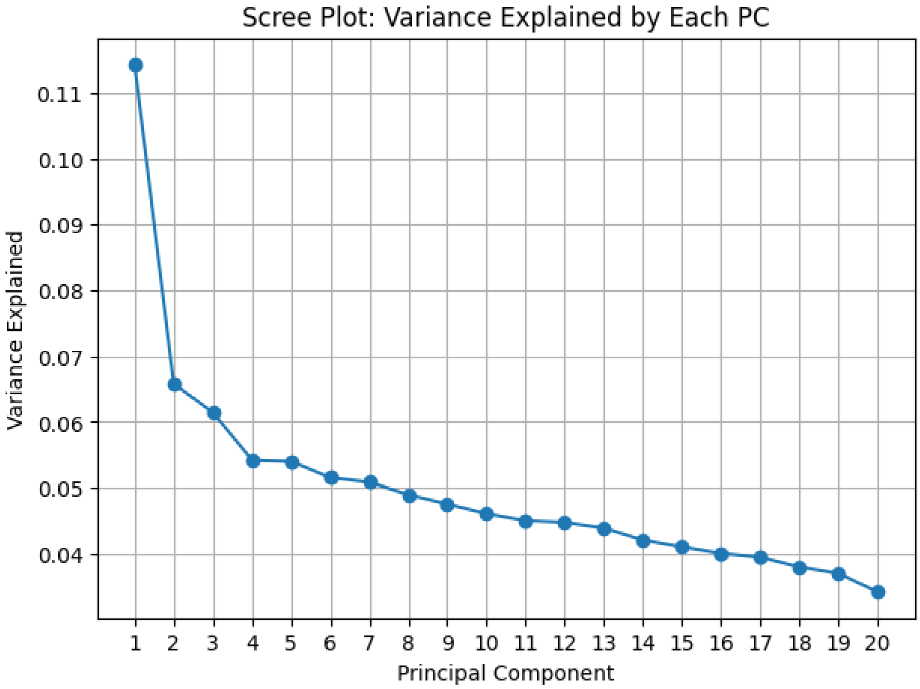


Figure 14: Cumulative variance explained by PCA on normalized tag features for average rating. Variance is broadly distributed across components, with approximately 17–18 components required to reach 85%.

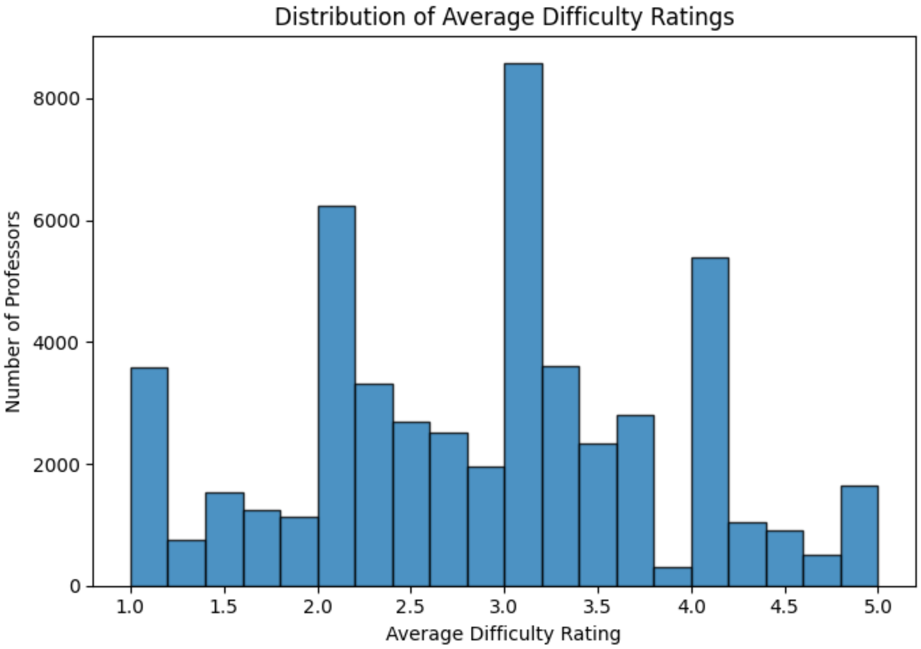


Figure 15: Distribution of average difficulty ratings across professors. Although derived from an ordinal scale, averaging across many ratings yields a smooth empirical distribution.

Table 4: Summary statistics for normalized tag features (z-scored).

Tag Feature	**Min**	**Mean**	**Median**	**Max**
Tough grader	-0.59	0.00	-0.59	6.29
Good feedback	-0.79	0.00	-0.44	5.46
Respected	-0.60	0.00	-0.60	7.15
Lots to read	-0.53	0.00	-0.53	7.33
Participation matters	-0.62	0.00	-0.62	6.62
No skip	-0.56	0.00	-0.56	11.66
Lots of homework	-0.53	0.00	-0.53	7.21
Inspirational	-0.46	0.00	-0.46	9.06
Pop quizzes	-0.24	0.00	-0.24	16.04
Accessible	-0.36	0.00	-0.36	7.27
Papers	-0.25	0.00	-0.25	15.72
Clear grading	-0.55	0.00	-0.55	8.23
Hilarious	-0.47	0.00	-0.47	8.35
Test heavy	-0.23	0.00	-0.23	11.47
Few things	-0.25	0.00	-0.25	10.07
Amazing lectures	-0.45	0.00	-0.45	9.70
Caring	-0.68	0.00	-0.68	6.60
Extra credit	-0.38	0.00	-0.38	70.73
Group projects	-0.25	0.00	-0.25	129.50
Lecture heavy	-0.42	0.00	-0.42	9.58

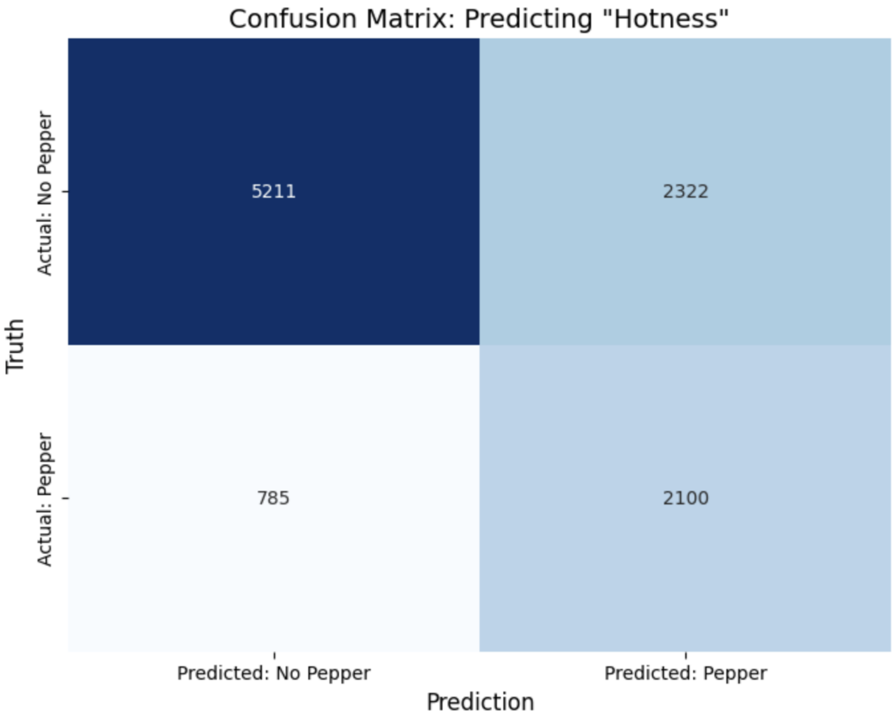


Figure 16: Confusion Matrix shows that while the model is ”trigger-happy” (prone to False Positives), it is highly effective at capturing the signal of attractiveness. The results confirm that on RateMyProfessor, the line between being a ”Good Teacher” and a ”Hot Teacher” is heavily blurred by student perceptions of quality.

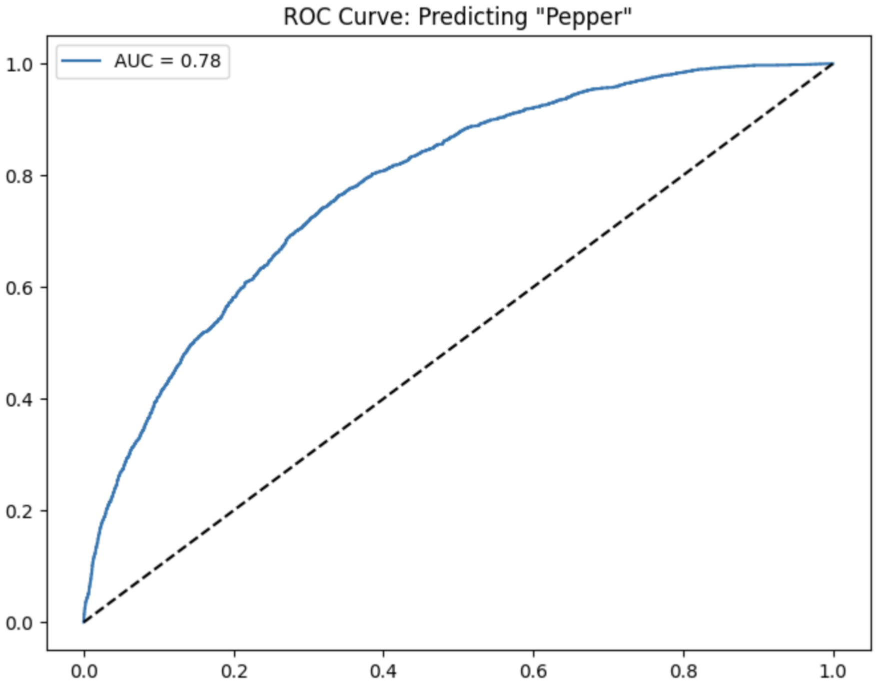


Figure 17: The model achieved an AU(RO)C score of 0.78, indicating strong predictive capability. This score suggests that the model is effective at distinguishing between professors who receive a ”pepper” and those who do not, performing significantly better than random chance (0.50).

Table 5: Classification performance for predicting `pepper` at different decision thresholds. Lowering the threshold improves minority-class detection at the cost of additional false positives, illustrating why accuracy is misleading under class imbalance.

Metric	**Threshold = 0.50**	**Threshold = 0.278**
True Negatives (TN)	34,911	24,830
False Positives (FP)	2,751	12,832
False Negatives (FN)	10,142	3,577
True Positives (TP)	4,284	10,849
Accuracy	0.752	0.685
Balanced Accuracy	0.612	0.706
Precision (Pepper)	0.609	0.458
Recall (Pepper)	0.297	0.752
F1 Score (Pepper)	0.399	0.569
Specificity (No Pepper)	0.927	0.659
False Positive Rate (FPR)	0.073	0.341
False Negative Rate (FNR)	0.703	0.248

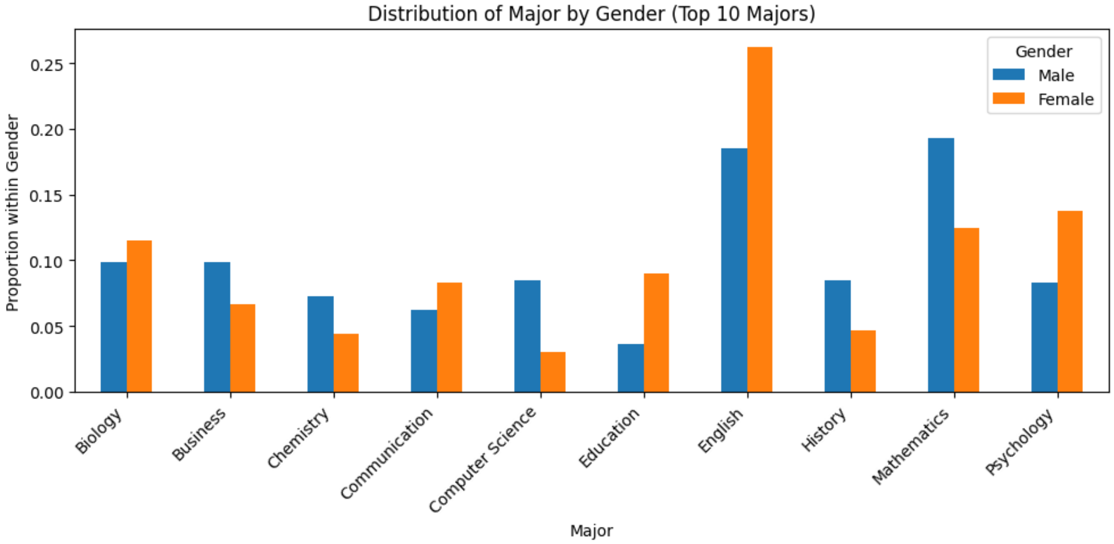


Figure 18: Distribution of academic majors by gender for the ten most common majors. Bars show the proportion of professors within each gender belonging to a given major. Differences reflect variation in representation across fields rather than differences in evaluation or labeling.