Assessing Professor Effectiveness (APE)

Capstone Project

Introduction to Data Science (DS-GA 1001)

12/19/2025

Authors:

- Shreya Vijay Rathi (N-number: N14048293)
- Adarsh Tiwari (N-number: N17883578)

Preprocessing Statement:
All preprocessing steps (handling missing values, scaling, imputation strategies and feature selection) were applied consistently across analyses unless explicitly stated otherwise. The Minimum ratings threshold was set to 10. Missing values were handled either by row-wise deletion or median imputation depending on the question-specific modeling goal.Missing numerical values were imputed with the median and missing tag values were filled with 0.
All analyses involving randomness were seeded using the N-number of Adarsh Tiwari to ensure reproducibility and compliance with academic integrity guidelines.

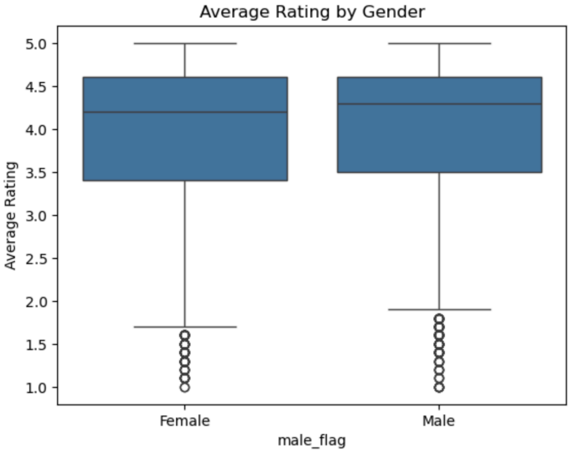
Question 1: Difference in Average Ratings

Approach: Compared group means using Welch's t-test due to unequal variances and large-sample robustness.

Findings:

- The coefficient for male instructors (male_flag) was 0.0075.
- Male mean rating: 3.964, Female mean rating: 3.895
- t-statistic: 3.42, p-value: 0.00063

Conclusion: Male professors receive, on average, 0.075 higher rating points than non-male professors. The effect is statistically significant (p < 0.005). However, the magnitude is practically negligible on a 1–5 rating scale



Question 2: Gender Difference in Rating Spread

Approach: F-test for equality of variances (comparing male vs. female rating variances)

Findings:

- Male variance: 0.7354 and Female variance: 0.8017
- F-statistic: 1.09 and P-value: 0.00763

Conclusion There is no statistically significant difference in rating dispersion between male and female professors at this stricter significance level.

Question 3: Likely Size of Gender Effects on Ratings

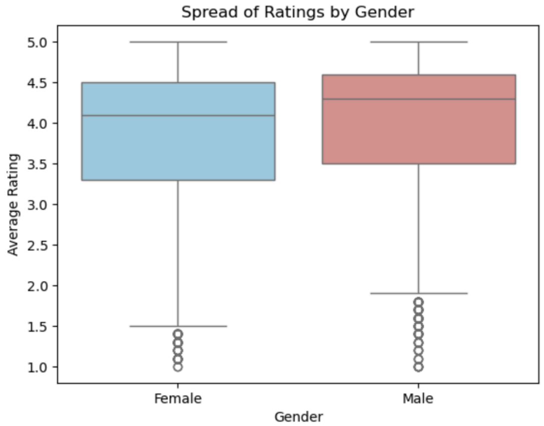
Approach: Estimation of effect sizes with 95% confidence intervals.

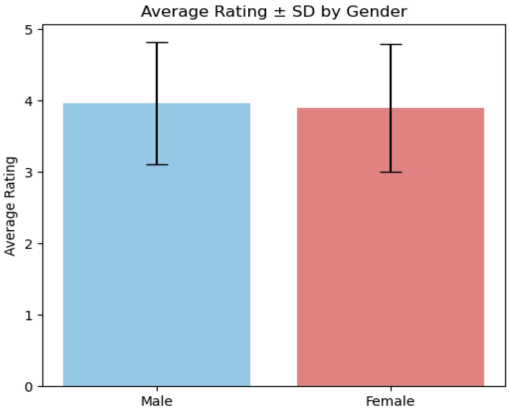
Findings:

- Mean difference (male - female): 0.0689
- 95% CI for mean difference: (0.0294, 0.1084)
- Variance ratio (male/female): 0.9173
- 95% CI for variance ratio: (0.8607, 0.9774)

- Difference in standard deviation (male - female): -0.0378

Conclusion: Male professors have slightly higher average ratings, but the difference is small. The spread of ratings is nearly equal between genders, indicating no meaningful difference in dispersion.





Question 4: Numerical Features → Average Rating

Approach: Effect sizes for each of the 20 tags were estimated using normalized tag proportions and 95% confidence intervals to quantify the likely magnitude of gender differences.

Findings:

- **Most gendered tags:** Hilarious, Caring and Amazing Lectures show the largest differences.
- **Least gendered tags:** Accessible, Tough Grader and Pop Quizzes exhibit minimal differences.

Conclusion: A few tags show measurable gender bias but the effect sizes are generally small and the majority of tags show little to no difference between male and female professors.

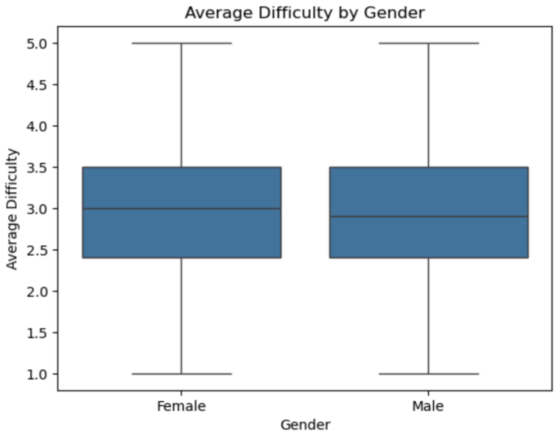
Question 5: Gender Difference in terms of Average Difficulty

Approach: Compared average difficulty ratings between male and female professors using Welch's t-test to account for unequal variances. Calculated mean difference and 95% confidence interval to quantify effect size.

Findings:

- Male variance: 0.5913, Female variance: 0.5862
- T-statistic: -0.2749, P-value: 0.7834 → not statistically significant
- Mean difference (male - female): -0.0048, 95% CI: (-0.0393, 0.0296)

Conclusion: There is no meaningful gender difference in perceived course difficulty. Both the mean difference and spread of difficulty ratings are very similar between male and female professors.



Question 6: Likely Size of Gender Effect on Average Difficulty

Approach: Estimated the mean difference in average difficulty between male and female professors and computed a 95% confidence interval to assess the likely magnitude of the effect.

Findings:

- Mean difference (male - female): -0.0048
- 95% CI: (-0.0393, 0.0296)

Conclusion: There is no meaningful difference in average difficulty between male and female professors; the confidence interval includes zero, indicating the effect is negligible.

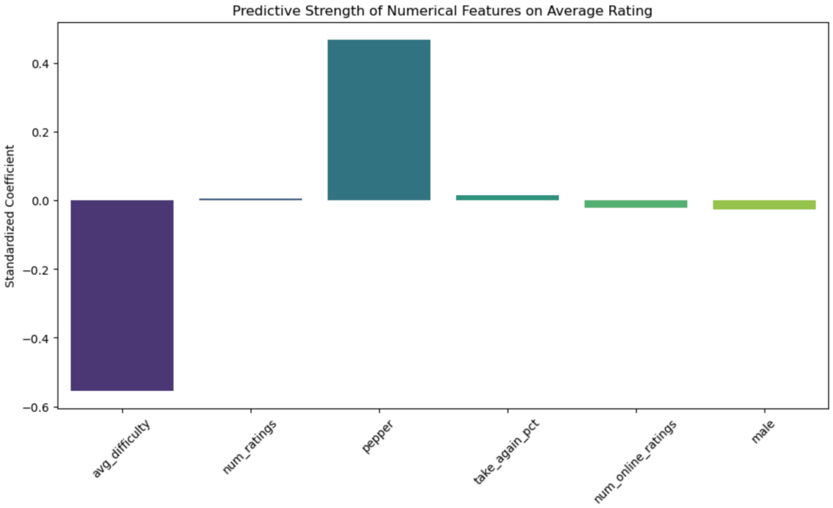
Question 7: Regression to predict average ratings.

Approach: OLS regression predicting average rating from numerical variables (avg_difficulty, num_ratings, pepper, take_again_pct, num_online_ratings, male, female) after handling missing values and checking multicollinearity.

Findings:

- $R^2 = 0.688$ → model explains ~69% of variance.
- Strongest predictors: avg_difficulty: -0.427, pepper: 0.383, take_again_pct: 0.018

Conclusion: Ratings are mainly driven by course difficulty, “pepper” status, and take-again percentage.



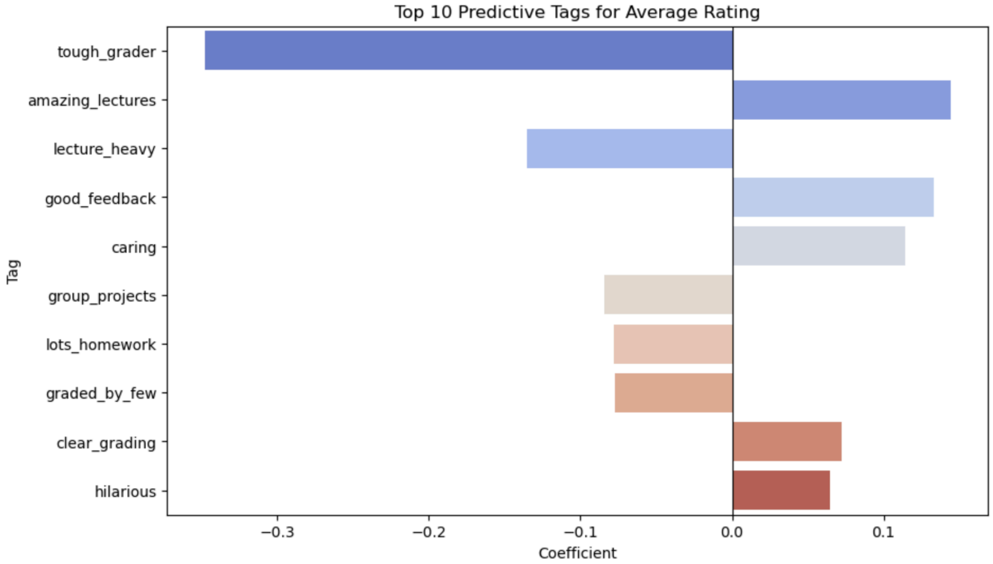
Question 8: Regression with Tags to predict Average Rating

Approach: OLS regression using all 20 tags; features standardized and multicollinearity checked via VIF. Missing ratings dropped.

Findings:

- $R^2 = 0.179$, RMSE = 1.021
- Top predictive tags: tough_grader (-0.348), amazing_lectures (0.144), lecture_heavy (-0.135), good_feedback (0.133), caring (0.114)

Conclusion: Tags provide interpretable insights but are weaker predictors than structured course metrics. Tags explain less variance than numerical predictors (R^2 ~0.37) but highlight qualitative influences on ratings



.

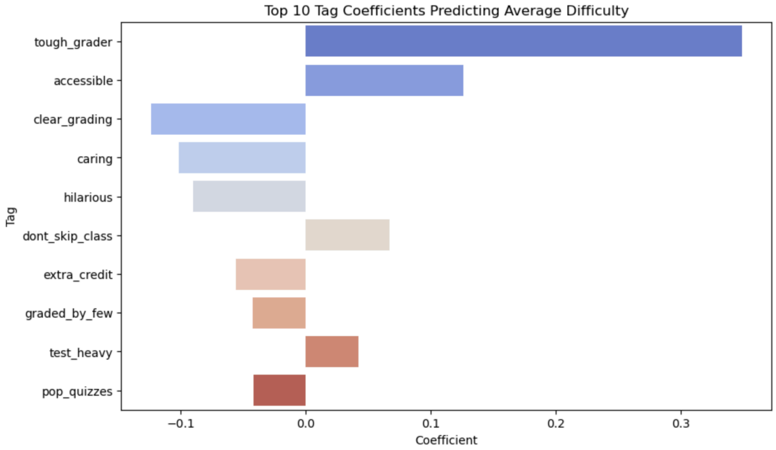
Question 9: Regression Predicting Average Difficulty from Tags

Approach: OLS regression using all 20 tags; multicollinearity checked via VIF; standardized features.

Findings:

- $R^2 = 0.143$, RMSE = 0.811
- Strongest positive predictor - tough_grader (0.348) → courses perceived as more difficult
- Strongest negative predictors - caring (-0.101), clear_grading (-0.124) → reduce perceived difficulty.

Conclusion: Tags partially explain perceived course difficulty, with tough_grader having the most impact. Results are consistent with Model A but slightly smoothed by imputation.



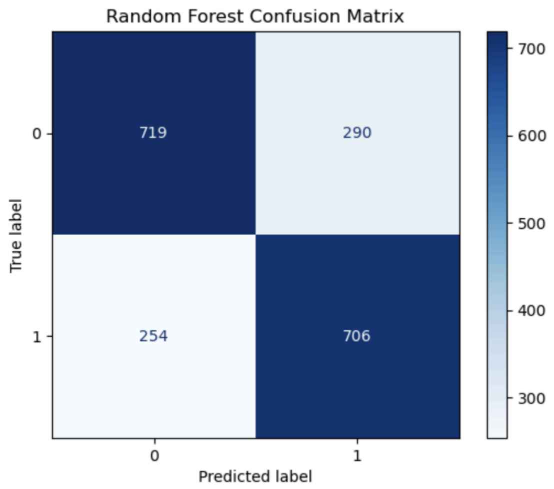
Question 10: Classification — Predicting "Pepper"

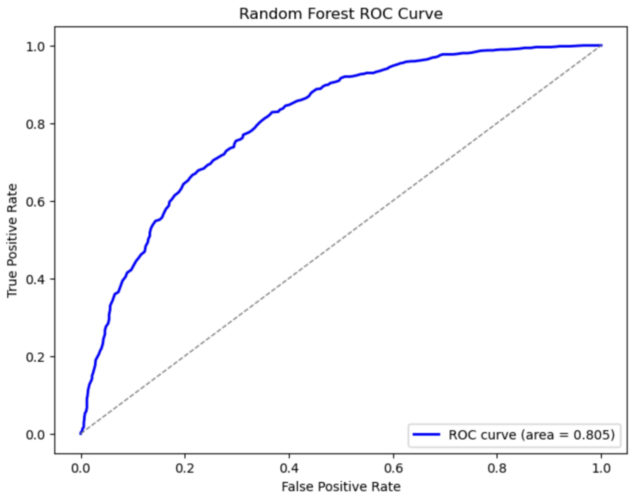
Approach: We combined numerical features (e.g., avg_rating, avg_difficulty) and tag indicators to predict whether a professor receives a "pepper." To reduce multicollinearity and dimensionality, PCA was applied before training the models. Two classification models were used: Random Forest and Logistic Regression with class imbalance addressed using balanced weights.

Findings:

- Random Forest: AUROC = 0.805, Accuracy = 72%. Class-wise performance: Class 0 → Precision 0.74, Recall 0.71, F1-score 0.73; Class 1 → Precision 0.71, Recall 0.74, F1-score 0.72
- Logistic Regression: AUROC = 0.814, Accuracy = 73%. Class-wise performance: Class 0 → Precision 0.76, Recall 0.70, F1-score 0.73; Class 1 → Precision 0.71, Recall 0.77, F1-score 0.74

Conclusion: Using both numerical and tag-based features, with dimensionality reduction and proper handling of missing data, allows reliable prediction of "pepper" awards. Logistic Regression provides slightly better discrimination, but both models are suitable for this classification task.





Extra Credit Analysis: Teaching Styles and Course Insights

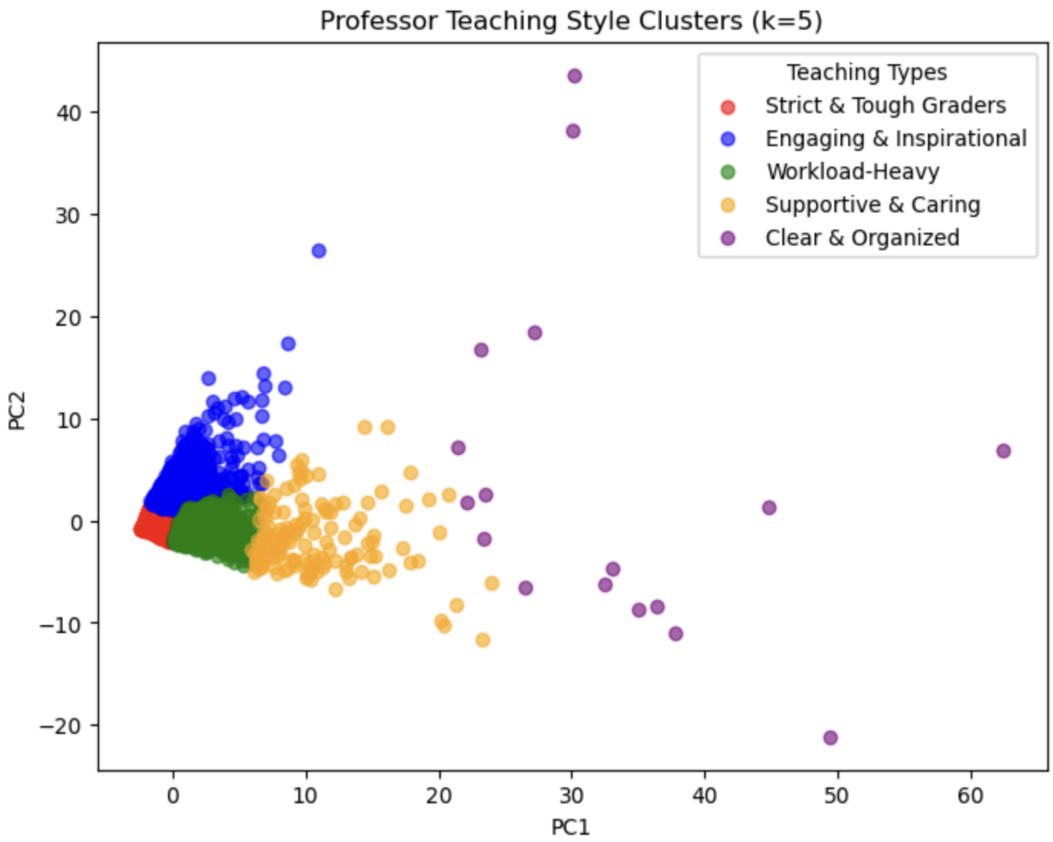
1. Professor Teaching Archetypes

Approach: Used K-Means clustering (k=5) on normalized teaching tags to identify common teaching styles.

Findings:

- **Teaching Archetypes**: Professors cluster into five distinct types. Some types are characterized by caring and good-feedback traits with high ratings and moderate difficulty, while others, marked by “tough grader” traits, show lower ratings and higher difficulty. A few rare types are defined by unique traits like hilarity or extreme rigor.
- **State-Wise “Hotness”:** Regional differences exist in the fraction of professors receiving a “pepper,” indicating variation in perceived popularity across U.S. states.
- **Stop Courses**: The most difficult courses are associated with tough grading professors, while the most popular courses tend to be taught by professors with caring or good-feedback traits.

Conclusion:
Teaching styles clearly cluster into meaningful archetypes, which influence course difficulty, ratings, and popularity, with notable regional patterns.


Professor Teaching Style Clusters (k=5)
Teaching Types
Strict & Tough Graders
Engaging & Inspirational
Workload-Heavy
Supportive & Caring
Clear & Organized
PC2
PC1
40
30
20
10
0
−10
−20
0
10
20
30
40
50
60

-