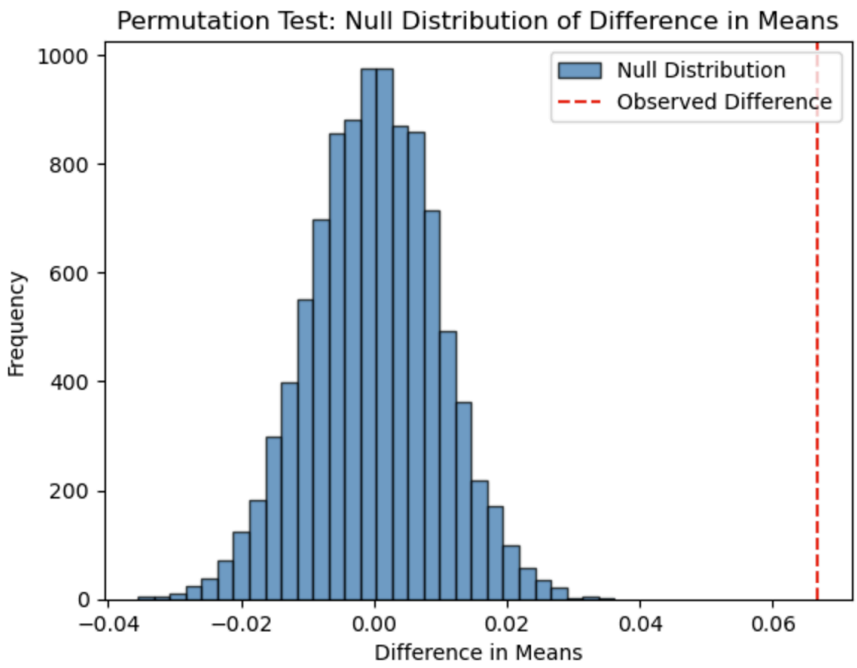
DSGA 1001- Introduction to Data Science Final Report

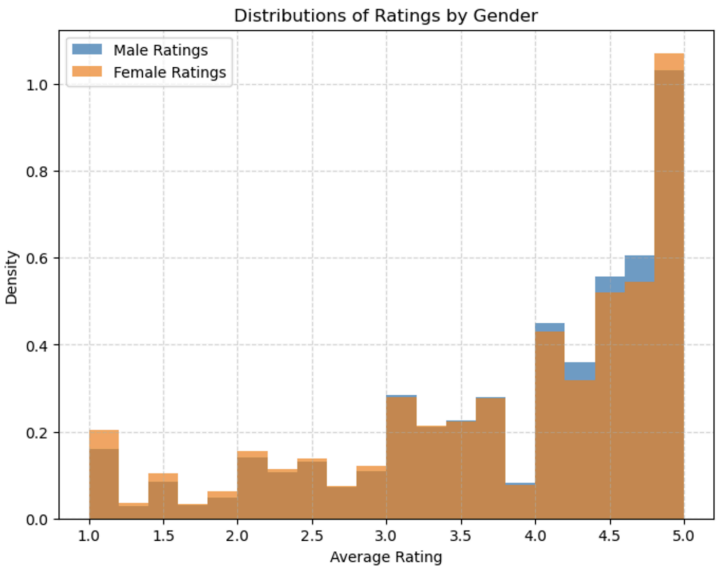
Name: Yueh-Han Chen, Pranav Thorat, Ahamed Foisal (IDS 41)

Q1



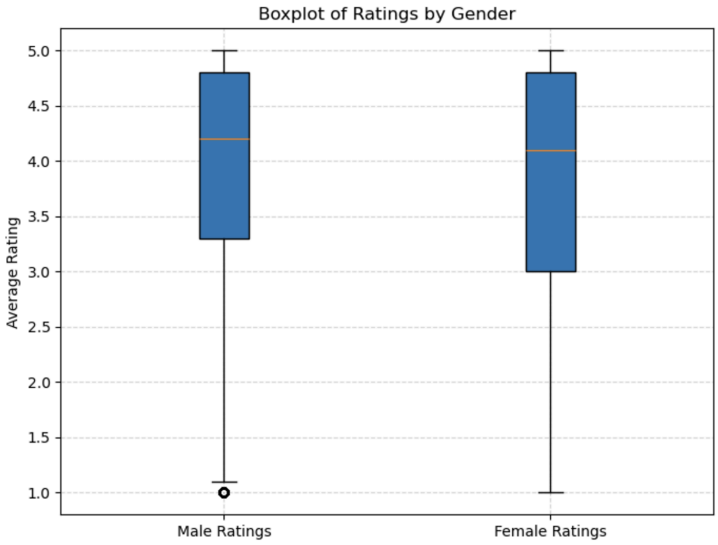
What we did: To investigate whether there is evidence of a pro-male gender bias in professor ratings, we performed a permutation test comparing the average ratings of male and female professors. A permutation test was chosen because the distribution of ratings was not normal, making it inappropriate to use parametric tests like the t-test. Null and Alternative Hypotheses: Null Hypothesis (H_0): There is no difference in average ratings between male and female professors (no gender bias). Alternative Hypothesis (H_a): Male professors have higher average ratings than female professors (pro-male gender bias). Findings: We calculated the observed mean difference between male and female professors' ratings as 0.067. By randomly shuffling the gender labels 10,000 times and computing the difference in means for each permutation, we constructed a null distribution. The resulting p-value was 0.0 (we use α=0.005 as the threshold), indicating that the observed difference is highly unlikely under the null hypothesis of no gender bias. This supports the alternative hypothesis that male professors tend to receive higher ratings than female professors.

Q2



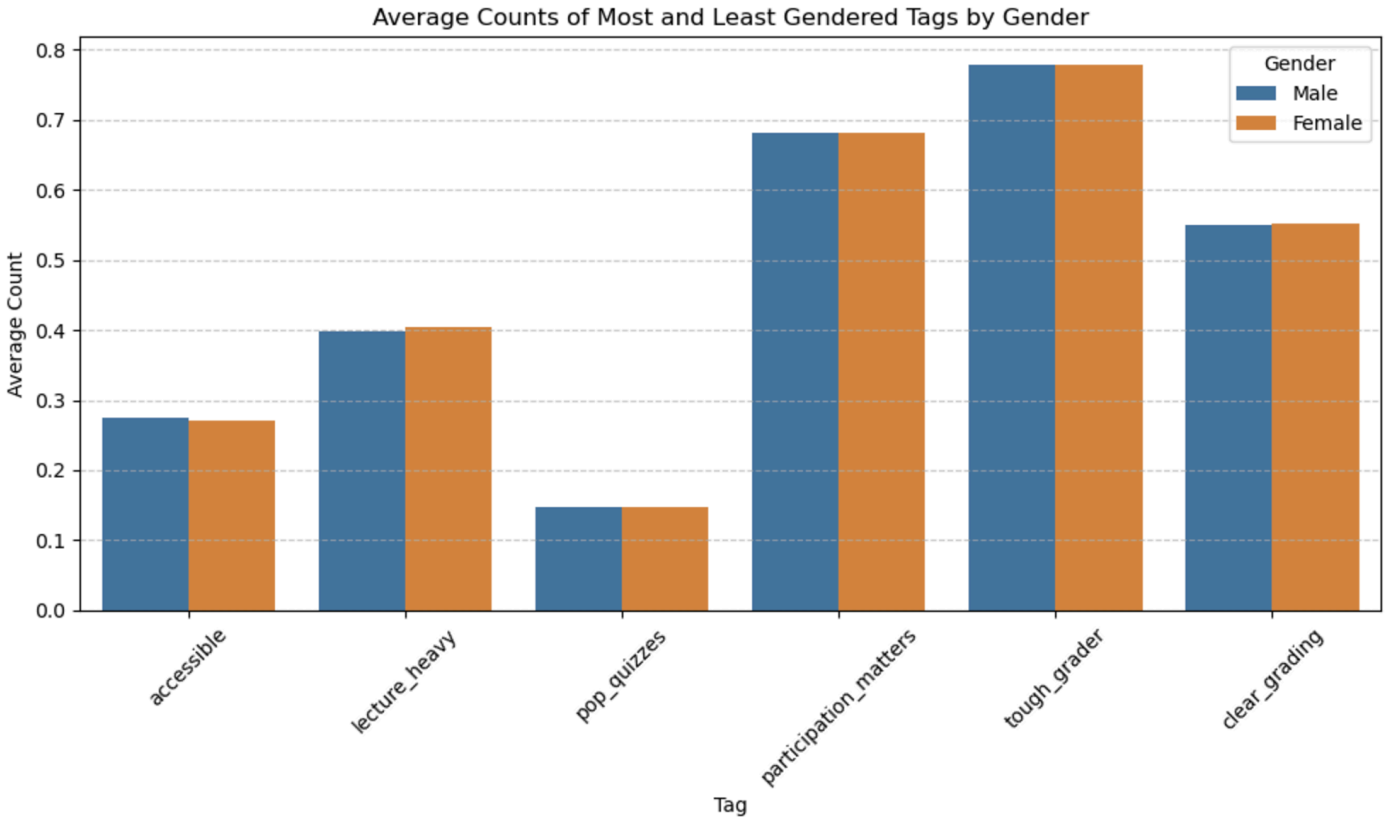
What we did: To determine whether there is a significant gender difference in the spread (variance) of the ratings distribution, we conducted a permutation test (which is appropriate in this case as it doesn't have assumptions about the distribution of the data). The null hypothesis posits that there is no difference in the variance of ratings between males and females, meaning any observed difference is due to random chance. Conversely, the alternative hypothesis asserts that there is a meaningful difference in variance between the two groups. Findings: The observed variance difference was -0.131, indicating that female ratings had a slightly greater spread than male ratings. However, with a p-value of 0.0 (based on 10,000 permutations)(we use α=0.005 as the threshold), the results are highly significant, leading us to reject the null hypothesis. This suggests that the variance in ratings is statistically different between genders.

Q3



After checking in with the professor during office hours, he pointed out that this question was meant to relate directly to Q1, so a one-sided test was conducted. The null hypothesis posited that there is no significant gender bias in student evaluations of professors, while the alternative hypothesis stated that there is a strong gender bias, with male professors enjoying a boost in ratings due to this bias. Cohen's d was used to measure effect sizes: for mean differences, it was computed as the difference in means normalized by the pooled standard deviation, and for variance differences, Prof. suggested using (var 1 −var 2)/2. Bootstrapping was employed to calculate 95% confidence intervals, a legitimate approach under these conditions as it relies on resampling to model the sampling distribution without parametric assumptions. The results showed a Cohen's d of 0.0599 for the mean difference with a CI of (-0.0170, 0.0171), and a Cohen's d of -0.0656 for the variance difference with a CI of (-0.0149, 0.0153). These findings suggest even though we found statistical significance in the above 2 questions, the effect size is actually very small as they include 0, which indicates not a significant practical significance.

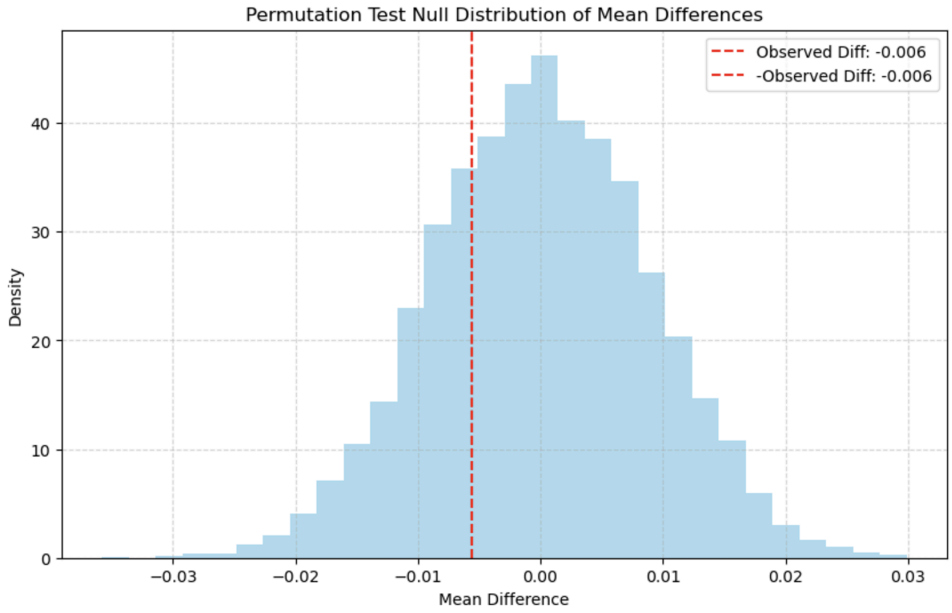
Q4



What We Did: To explore gender differences in the 20 tags awarded by students, we conducted a Mann-Whitney U Test for each tag. This test was chosen as the tags are count data and may not follow normal distributions. The data was first cleaned to remove rows with missing values in gender or tag counts. For each tag, the test compared the distributions of tag counts between male and female professors. We ranked the results based on p-values to identify the tags most

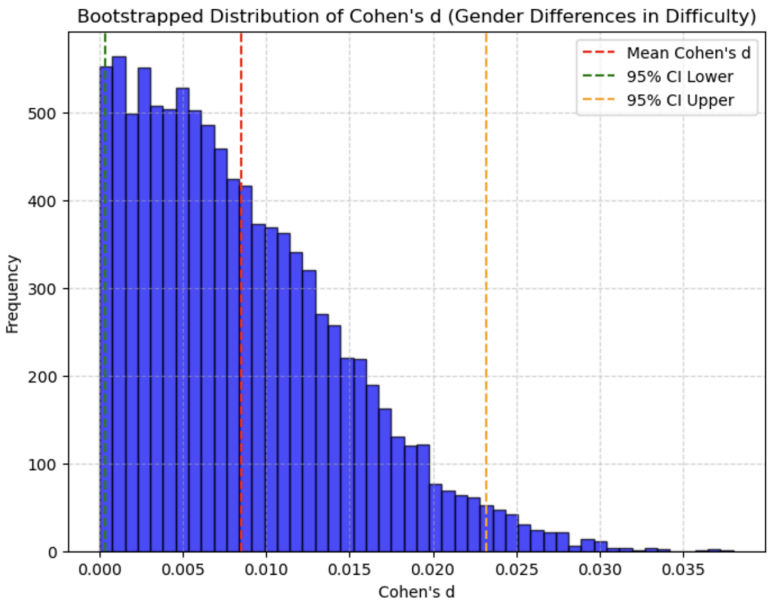
and least associated with gender differences. What We Found: The analysis revealed that the three most gendered tags (lowest p-values) were "accessible" (p=0.201), "lecture heavy" (p=0.216), and "pop quizzes" (p=0.224), though none showed strong statistical significance. Conversely, the least gendered tags (highest p-values) were "participation matters" (p=0.917), "tough grader" (p=0.929), and "clear grading" (p=0.969), indicating no meaningful gender differences for these tags.

Q5



To evaluate whether there is a significant gender difference in average difficulty ratings of professors, a two-sided permutation test was conducted as the question statement DID NOT mention the direction of the hypothesis, and the permutation test does not assume normality and is appropriate for comparing two independent groups. However, we ASSUME the data we have is representative of the population as the sample size is large enough (thousands). The null hypothesis stated no difference in difficulty ratings between male and female professors, while the alternative hypothesis posited a difference in either direction. The observed mean difference was -0.0056, with male professors having slightly lower ratings, and the permutation p-value was 0.5329 (we use α=0.005 as the threshold), far above the standard significance threshold (α=0.005). Descriptive statistics showed mean difficulty ratings of 2.84 (SD = 0.99) for male professors and 2.85 (SD = 0.99) for female professors. Since the p-value indicates that the observed difference is not statistically significant, we fail to reject the null hypothesis and conclude that there is no evidence of a meaningful gender difference in average difficulty ratings.

Q6

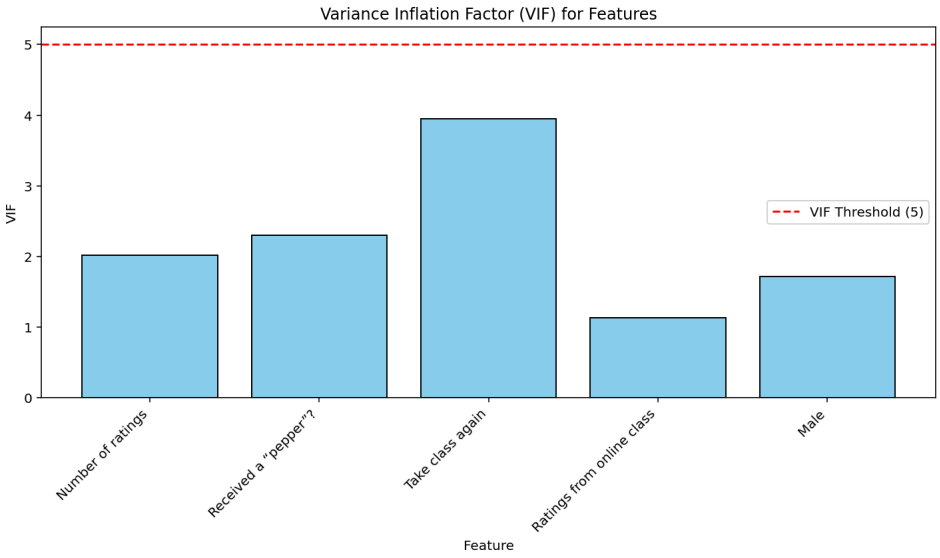


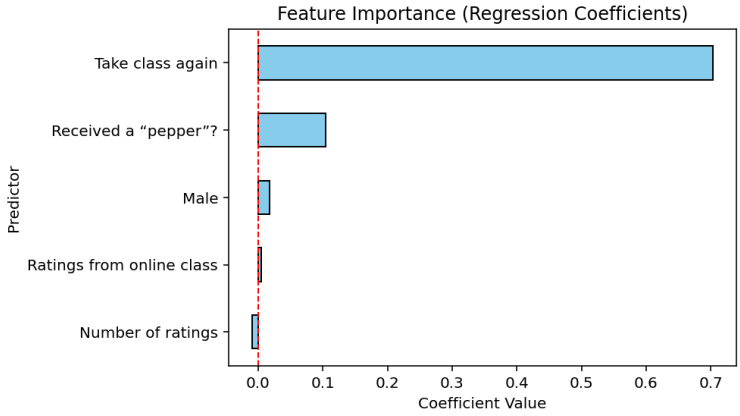
To quantify the likely size of the gender effect on average difficulty ratings, we calculated Cohen's d, a standardized measure of effect size, using bootstrapping to estimate the 95% confidence interval. Bootstrapping was chosen as it resamples the data repeatedly to generate a distribution of effect sizes without making strong parametric assumptions. However, we ASSUME the data we have is representative of the population as the sample size is large enough (thousands). Cohen's d was used because it provides a standardized way to measure the magnitude of the mean difference relative to pooled variability, allowing comparison across studies. Again, because the statements in questions 5 and 6 did not mention the direction of the hypothesis, we focused on measuring the effect size of the magnitude by using an absolute value of the mean diff. The bootstrap mean Cohen's d was 0.0085, with a 95% confidence interval

ranging from 0.0003 to 0.0232, indicating that the effect size is extremely small and likely negligible. This confirms that any gender difference in average difficulty ratings is both statistically and practically insignificant.

Q7

Addressing Multicollinearity, missing data and normalizing data





Dropped records with missing data and those with fewer than 5 ratings. Identified collinear features via a correlation heatmap and removed those highly correlated with other predictors (absolute correlation > 0.5) and less correlated with Average Rating (e.g., Average Difficulty and Female). Verified no multicollinearity using Variance Inflation Factor (VIF < 5). Standardized features to zero mean and unit variance using StandardScaler for regression.

A random number generator was used with N-number 11058720 for reproducibility.
A linear regression model was generated using the LinearRegression() function from scikit-learn.

RESULTS :
Regression Coefficients: *Take class again* (Coefficient: 0.7036) is the most influential predictor, strongly positively associated with ratings. *Received a "pepper"?* (Coefficient: 0.1046) also contributes positively but to a lesser extent. Other features, such as *Male* and *Ratings from online class*, have minimal impact, while *Number of ratings* has a slight negative effect (-0.0097).

Results from baseline model: This model was built using y_mean as the predicted value for the test instances which resulted in RMSE (Baseline) of 0.822400497935268

Model Performance:
RMSE: 0.3817, indicating low prediction error relative to the standard deviation of the data (0.846). The model effectively predicts ratings, with the predictors explaining a substantial portion of the variability while maintaining low multicollinearity. When compared to the baseline mode, this model performs better.
R^2 :0.7348 indicates that 73.48% of the variability in ratings is explained by the predictors.

Q8

Methodology : The tags dataset was normalized by dividing each tag's raw count by the corresponding number of ratings for each professor. This ensures that tag values are on a comparable scale, regardless of the number of ratings a professor received. To add substantiality and remove any bias from the results, rows with fewer than 5 ratings are removed. Address potential multicollinearity among the tag variables using variance inflation factor (VIF) analysis and other methods. Remove features with a Variance Inflation Factor (VIF) score of 5 or higher. A combination of features was tested, prioritizing the highest coefficients from the independent variables as determined by the regression model.

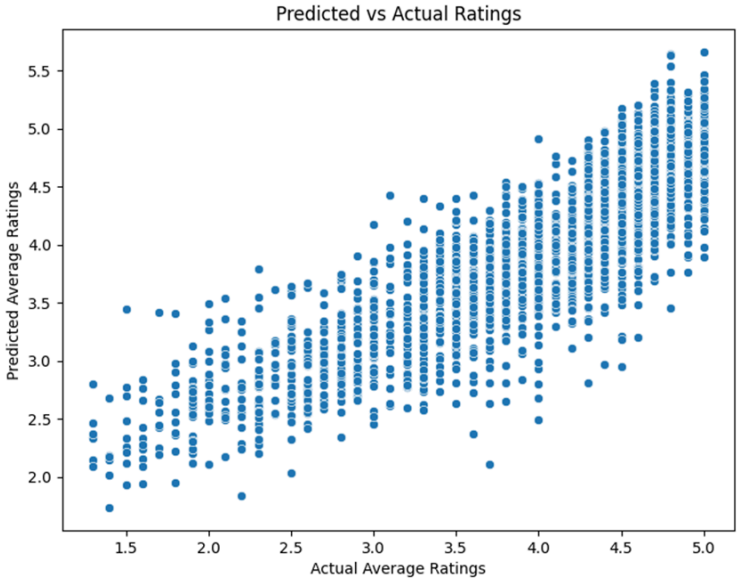
To address feature scaling, the code scales the features of the training and test datasets using the StandardScaler from sklearn. It first fits the scaler to the training data (X_train) to compute the mean and standard deviation, then transforms the training data by standardizing it (scaling to have zero mean and unit variance).

Linear Regression Results: R^2: 0.7509 ; RMSE: 0.4247

Tags like "Good_Feedback " (Regression Coefficient = 0.198539) and "Amazing_Lectures " (Regression Coefficient = 0.181090) emerged as strong predictors, aligning with intuitive expectations about factors contributing to perceived ratings.

Comparison: The R2 & RMSE values from the previous model are 0.7348 and 0.3817 : respectively, as compared to this model: R^2: 0.7509 ; RMSE: 0.4247. In this comparison, the second model with an R^2 of 0.7509 outperforms the first model's R^2 of 0.7348, suggesting it explains more of the variance in the data. Therefore we observe that the tags dataset captures variance amongst all features better than that in numeric data. One reason of why this happened could be that the number of features in tags data is higher than that in numerical data, leading to a high R^2.

However, the second model has a higher RMSE (0.4247 compared to 0.3817), indicating that its predictions are less accurate than the first model's. The reason might be that using numerical features helps the model learn better at the underlying patterns in the dataset, compared to the tags data we have, even with a fewer number of numerical features than the tag features.



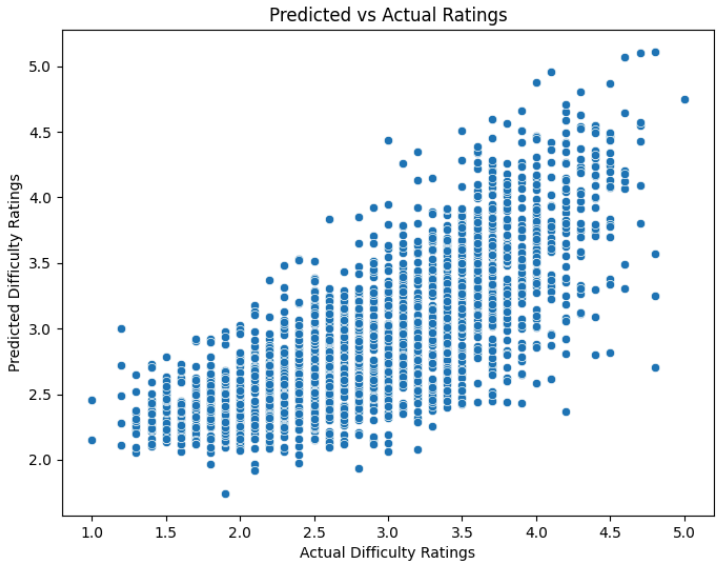
Q9

Methodology: Each tag's count is normalized by the number of ratings to ensure fairness, as professors with more ratings naturally accumulate more tags. This transformation helps remove bias associated with the volume of ratings. All tag features are standardized using StandardScaler, ensuring that all variables have a mean of 0 and standard deviation of 1. This step prevents large-scale features from dominating the regression.The analysis calculates the Variance Inflation Factor (VIF) for all tag features to detect multicollinearity: Features with a VIF greater than 5, indicating strong multicollinearity, are removed to enhance model interpretability and prevent redundancy. Ensuring that the regression coefficients represent unique contributions of each feature to the prediction.

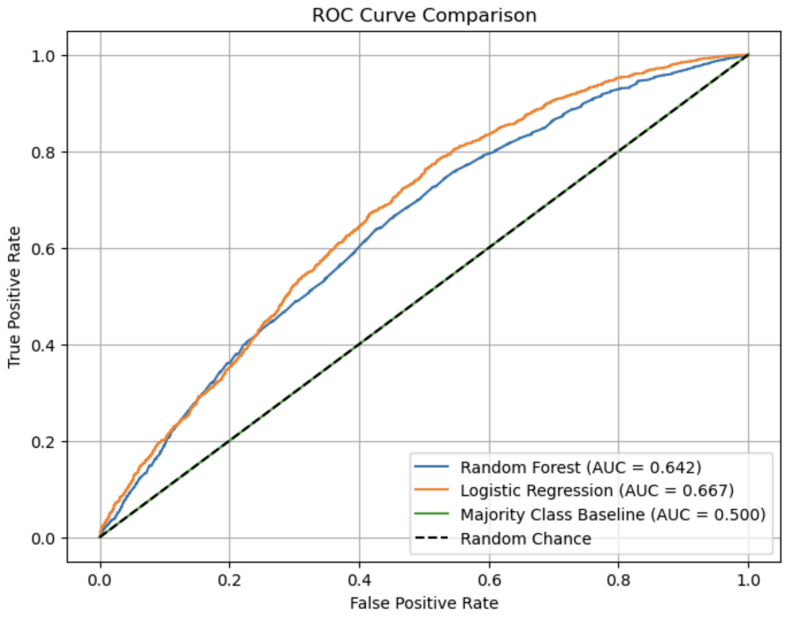
Results from baseline model: This model was built using y_mean as the predicted value for the test instances which resulted in RMSE (Baseline) of 0.7385071938406395

Linear Regression Model Metrics: R^2 = 0.5554; Root Mean Square Error (RMSE) = 0.452

Tags strongly predictive of Average Difficulty: Tags like "Tough grader" (Regression Coefficient = 1.672514) and "Test heavy" (Regression Coefficient = 1.325838) emerged as strong predictors, aligning with intuitive expectations about factors contributing to perceived difficulty. When compared to the baseline mode, this model performs better.



Q10



In this analysis, we aimed to predict whether a professor receives a "pepper" rating based on a variety of numerical, categorical, and tag-based features. We first performed preprocessing by removing rows where the number of ratings was less than 5 to ensure the reliability of the data. We then handled missing values by filling numerical columns with their means, categorical columns with their mode, and tag-based features with 0, assuming no tag was awarded. We removed multicollinearity by calculating the Variance Inflation Factor (VIF) for each feature and iteratively dropped features with VIF ≥ 5 to improve model stability; this process removed proportion_retaking and average_rating. To address the issue of class imbalance in the target variable (pepper), we applied SMOTE (Synthetic Minority Oversampling Technique), which generates synthetic samples of the minority class to balance the dataset, ensuring the models are not biased towards the majority class.

We trained three models: a random forest classifier, logistic regression, and a random guess baseline. The logistic regression model achieved the highest Area Under the Curve (AUC) score of 0.667, followed by the random forest model with an AUC of 0.642, and the baseline at 0.500. The AUC score measures the model's ability to distinguish between classes, with 0.5 indicating random guessing and 1.0 indicating perfect discrimination. The logistic regression model also achieved better overall accuracy (0.621) and recall (0.619), indicating its ability to correctly identify positive cases more frequently compared to the random forest model. These findings suggest that logistic regression is the most effective model for this task, providing a reasonable trade-off between sensitivity and precision in predicting "pepper" ratings.

Extra Credit

This analysis investigates whether professors with lower average ratings on RateMyProfessor.com tend to receive more student ratings, based on the assumption that dissatisfied students are more likely to leave online reviews.

Methodology:
The dataset was split into two groups based on the median average rating (calculated as 4.2, where actual value depends on the dataset). Professors with average ratings below the median were categorized as the "low rating" group, while those at or above the median formed the "high rating" group. The Mann-Whitney U test, a non-parametric statistical test, was used to compare the number of ratings between these groups due to the non-normal distribution of the data.

Null Hypothesis (H_0):
There is no difference in the number of student ratings between professors with lower average ratings and those with higher average ratings.

Alternative Hypothesis (H_1):
Professors with lower average ratings tend to receive more student ratings compared to those with higher average ratings. The number of ratings is associated with the professors' average ratings, and dissatisfied students (with lower ratings) are more likely to leave online reviews.

Results:
The Mann-Whitney U test yielded the following results:

- Median of Average Ratings: 4.2
- Test Statistic (U): 17624855.0
- P-value: 0.9999941190701485

Given the threshold for statistical significance (α=0.005), the test indicated no statistically significant difference in the number of ratings between low and high-rating groups.