Beckett Newton
Professor Pascal Wallisch
Principles of Data Science
Fall 2024
Capstone Project

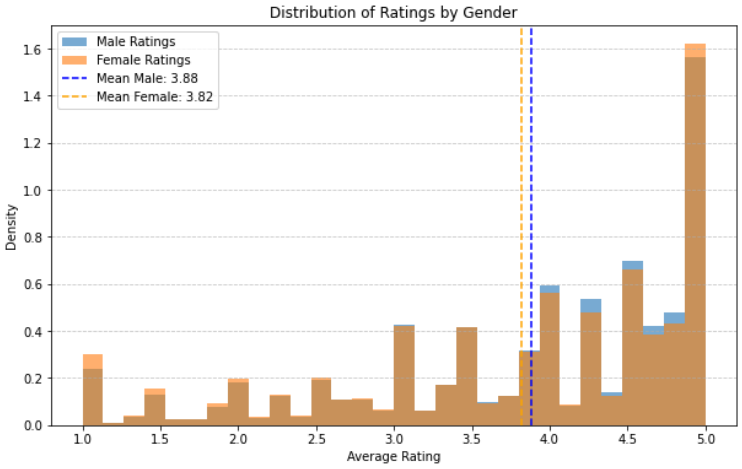
My Approach to Data Cleaning/Pre-Processing:

As mentioned in the spec sheet for this project, there are two major components to consider when cleaning this dataset: handling NaNs and determining how to handle data based on the number of ratings. To handle NaNs, I took a targeted approach due to the high prevalence of missing data (86% of rows). Most missing values were concentrated in "the proportion of students that said they would take the class again". Removing all rows with any NaNs would have unnecessarily eliminated much of the dataset. Instead, I removed rows with more than one NaN (22% of the dataset), ensuring that missing values were confined to this single column. For the acceptance of data based on the number of ratings, I carefully considered three options: accepting all data, setting a threshold, or implementing a weighting system. I decided to accept all data for several reasons. This decision avoids introducing biases that could arise from excluding professors with fewer ratings, ensuring a diverse and representative dataset. Excluding professors with five or fewer ratings would eliminate 54% of the data, disproportionately biasing the analysis toward more experienced or popular professors. Similarly, setting a threshold would be inherently subjective, and weighting based on the number of ratings could unintentionally amplify confounds, such as over-representing experienced professors. While ratings from fewer responses are noisier, the dataset's size (89,000+ professors) ensures sufficient statistical power to detect meaningful patterns. The Central Limit Theorem further supports this choice, as the mean rating distribution approximates normality, allowing statistical methods like z-tests to naturally adjust for variability. By avoiding exclusions or complex weighting systems, this approach balances inclusivity and simplicity while leveraging robust statistical frameworks to handle uncertainty. The RNG was seeded using my N-number (15179480).

Question 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.

Analysis of Gender Differences in Professor Ratings

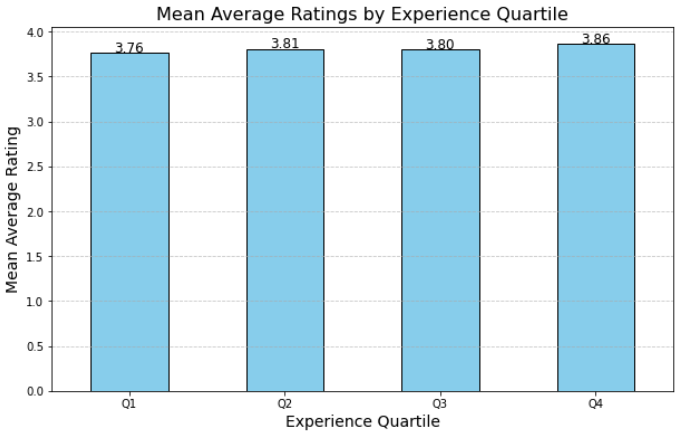
***Mean Rating (Male Professors)**: 3.88, **Mean Rating (Female Professors)**: 3.82, **Z-Statistic**: 6.51, **P-Value**: 7.29e-11*



To investigate whether there is evidence of a pro-male gender bias in this dataset, I conducted a two-tailed z-test to compare the average ratings of male and female professors. The z-test was chosen because the sample size is large, allowing the Central Limit Theorem to ensure that the sampling distribution of the mean is approximately normal. Additionally, the z-test is well-suited for comparing two population means when the population standard deviations are unknown but can be estimated from the sample, as is the case here. Using a pooled standard deviation, the z-statistic was calculated to test the null hypothesis that there is no difference in average ratings between male and female professors. The alternative hypothesis was that male professors receive higher ratings. The z-test results (Z-Statistic: 6.51, P-Value: 7.29e-11) show a statistically significant difference between male and female professors' average ratings, suggesting that the difference in the mean average rating between male and female professors is unlikely due to chance alone. As such, there is statistical evidence of a pro-male gender bias in professor ratings in this dataset. Despite the large sample size providing a level of comfort in the results, the small magnitude of the difference underscores the need for caution when interpreting the practical implications, as other confounds that cannot be accounted for given the data - such as course type (elective or requirement) or subject matter - could impact ratings.

Question 2: Is there an effect of experience on the quality of teaching? You can operationalize quality with average rating and use the number of ratings as an imperfect –but available –proxy for experience.

Analysis of the Effect of Experience on Teaching Quality



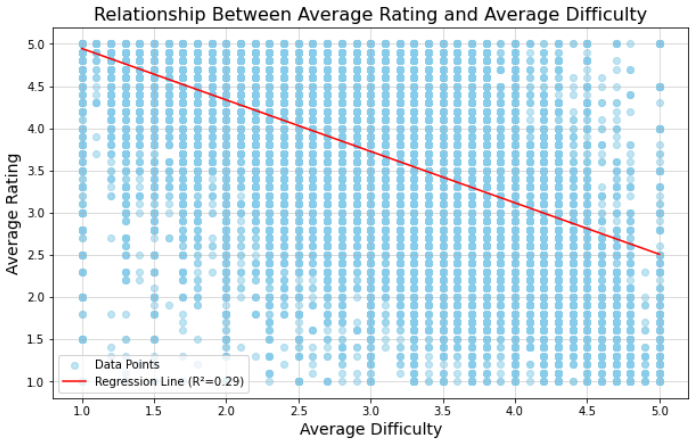
***Mean Ratings by Quartile**: **Q1** (Lowest 25% of Experience): 3.76, **Q2**: 3.81, **Q3**: 3.80, **Q4** (Highest 25% of Experience): 3.86, **ANOVA Test Results**: **F-Statistic**: 22.14, **P-Value**: 2.55e-14*

To assess whether teaching experience impacts teaching quality, I divided professors into quartiles based on the number of ratings: Q1 (least experienced) to Q4 (most experienced), each representing 25% of the data. I used quartiles instead of regression to avoid assuming a linear relationship between experience and teaching quality and to allow for clearer group comparisons, which make it easier to observe trends and interpret results, especially if the relationship is non-linear or influenced by outliers. For each quartile, I calculated the mean average rating and performed a one-way ANOVA to test for significant differences in ratings across the quartiles. An ANOVA test was chosen because it is specifically designed to compare the means of multiple groups (3+) - in this case, the four experience quartiles - to determine if any observed differences are statistically significant. This makes it well-suited for testing whether teaching quality varies systematically across levels of experience. Professors in Q4 had the highest mean rating (3.86), while Q1 had the lowest (3.76). Ratings in Q2 (3.81) and Q3 (3.80) were similar. While the magnitude of the results were small, the ANOVA results ($F = 22.14$, $p = 2.55e\text{-}14$) showed highly statistically significant differences between the quartiles and suggest that the general increase in average rating as the number of ratings (experience) increased was not likely due to chance alone.

Question 3: What is the relationship between average rating and average difficulty?

Analysis of the Relationship Between Average Rating and Average Difficulty

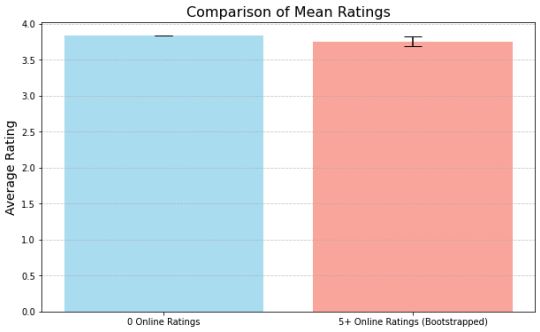
***Correlation Coefficient**: −0.54, **P-Value (Correlation)**: 0.00 (highly significant), **Linear Regression Results**: Slope: −0.61, Intercept: 5.56, R^2: 0.29*



To investigate the relationship between average rating (a measure of teaching quality) and average difficulty (a measure of course difficulty), I calculated the Pearson correlation coefficient to quantify the strength and direction of the linear relationship between average rating and average difficulty. Pearson's method is well-suited here because both variables are continuous, and the scatter plot suggested a linear trend. The correlation coefficient was −0.54, indicating a moderate negative relationship, and the p-value was $p<0.005$, demonstrating that this relationship is statistically significant. The analysis reveals a statistically significant negative relationship between average rating and average difficulty: as courses are perceived to be more difficult, their average ratings tend to decrease. While the results confirm a significant relationship, the R^2 value of 0.29 suggests that factors other than difficulty (e.g., subject matter, teaching style, or grading leniency) also play a substantial role in determining ratings. These findings align with the expectation that students may rate challenging courses more harshly, reflecting their perceptions of difficulty rather than solely the quality of teaching.

Question 4: Do professors who teach a lot of classes in the online modality receive higher or lower ratings than those who don't?

Analysis of Online Teaching and Ratings



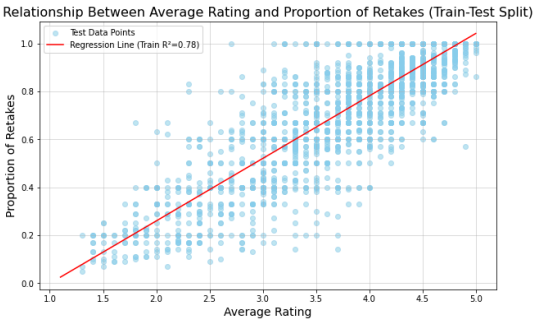
***Bootstrapped mean rating for professors with 5+ online ratings**: 3.75, **95% Confidence Interval (Group 5+)**: [3.69,3.82], **Mean rating for professors with 0 online ratings**: 3.83, **T-Statistic**: −2.296, **P-Value**: 0.02*

To determine whether professors who teach a lot of classes in the online modality receive higher or lower ratings than those who don't, professors were divided into two groups: those with exactly 0 online ratings (59,077 professors) and those with at least 5 online ratings (880 professors). These groupings were chosen to distinguish professors with no engagement in online teaching from those with meaningful online activity. Given the significant disparity in group sizes (59,077 professors with 0 online ratings vs 880 professors with 5+ online ratings), I applied bootstrapping to the smaller group (professors with 5+ online ratings). Bootstrapping allowed me to create a reliable distribution of the mean rating for this group without assuming a specific underlying distribution. For this analysis, I generated 10,000 bootstrap samples, each of size 880, and computed the mean average rating for each sample. The confidence interval for the mean was derived from the 2.5th and 97.5th percentiles of the bootstrapped distribution. To compare the two groups, I conducted a Welch's t-test, which is ideal for datasets with unequal variances and sample sizes. This test compares the mean ratings of the two groups to assess whether the observed difference is statistically significant. While the p-value (0.0219) indicates some evidence of a difference, it does not meet the observed significance threshold (α=0.005). As a result, we cannot conclude that the difference is statistically significant under the criteria set for this project.

Question 5: What is the relationship between the average rating and the proportion of people who would take the class the professor teaches again?

Analysis of the Relationship Between Average Rating and Proportion of Retakes

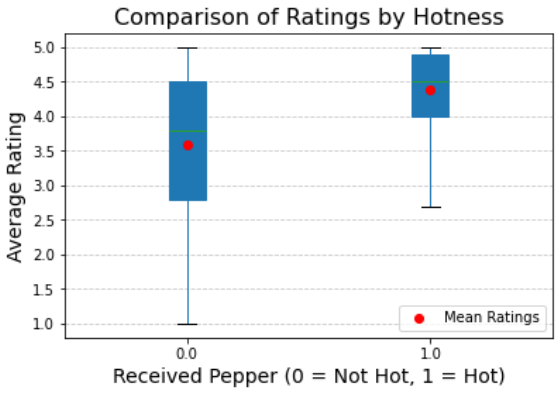
***Pearson's Correlation Coefficient**: 0.88, **P-Value (Correlation)**: 0.00 (highly significant), **Linear Regression Results**: Slope: 0.0026, Intercept: −0.0026, R^2 (training): 0.78, R^2 (test): 0.77*



To explore the relationship between average rating (a measure of teaching quality) and the proportion of students willing to retake a professor's class, I calculated the proportion of retakes as the number of students willing to retake divided by the total number of reviews. This transformation standardized the data to a 0–1 scale, ensuring consistency and comparability across observations. To validate the findings, I split the dataset into training and testing subsets (80% train, 20% test) using a random seed derived from my N-number. This ensured the analysis was reproducible and that the model's performance could be assessed on unseen data. The relationship was quantified using the Pearson correlation coefficient, which measures the strength and direction of a linear relationship. A linear regression was then performed on the training data to model the mathematical relationship and evaluate how well average rating predicts the proportion of retakes. The test data was used to compute R^2, providing a measure of the model's generalizability. These techniques were chosen because they are well-suited for examining linear relationships, with the correlation coefficient providing an overall strength metric and the regression offering a predictive framework. The Pearson correlation coefficient on the training set was 0.88, indicating a strong positive relationship. The p-value ($p<0.005$) confirmed that the relationship is statistically significant. The linear regression model trained on the data yielded an R^2 value of 0.78, showing that 78% of the variability in the proportion of retakes could be explained by the average rating. On the test data, the model maintained a strong R^2 value of 0.77, underscoring its robustness. This analysis demonstrates a significant, consistent relationship between teaching quality (as measured by average rating) and students' willingness to retake a professor's class. While the slope of the regression is relatively small, it reflects the tight clustering of data points, which strengthens the observed trend.

Question 6: Do professors who are “hot” receive higher ratings than those who are not?

Analysis of "Hotness" and Ratings



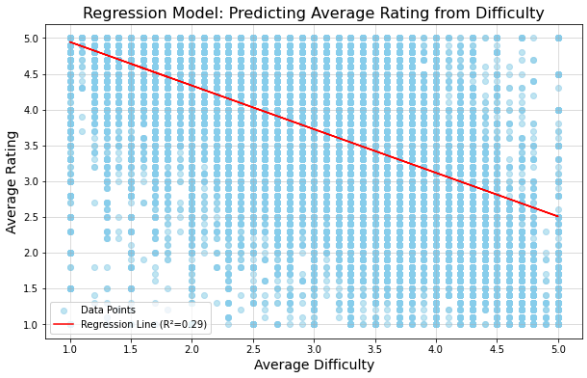
***Not Hot Professors** (ReceivedPepper=0): Mean rating = 3.58, **Hot Professors** (ReceivedPepper=1): Mean rating = 4.38, **T-Statistic**: 113.11, **P-Value**: 0.00 (highly significant)*

To determine whether professors deemed "hot" receive higher ratings, I used the Received Pepper column as a binary indicator (1 for "hot" and 0 for "not hot"). A two-sample Welch's t-test was conducted to compare the mean ratings of the two groups. This test was chosen because it accounts for potential differences in variances and is well-suited for the large sample size in this analysis, providing robust and reliable results. The analysis revealed that professors marked as "hot" had an average rating of 4.38, significantly higher than the 3.58 average for professors not marked as "hot." The t-statistic was 113.11, an extraordinarily large value that reflects the substantial difference between the groups. The p-value was <0.005, indicating that the observed result is far beyond what would be expected under the null hypothesis.These results provide strong evidence that professors judged as "hot" tend to receive significantly higher ratings. This disparity may stem from biases in student evaluations, where physical attractiveness influences perceptions of teaching quality. Alternatively, it could reflect other correlated traits, such as charisma or enhanced communication skills, that also contribute to higher ratings. However, the large effect size suggests that subjective factors related to appearance likely play a prominent role. This analysis highlights a statistically and practically significant difference in ratings based on perceived "hotness." While the results are robust, they underscore the potential influence of subjective biases in teaching evaluations.

Question 7: Build a regression model predicting average rating from difficulty.

Analysis of Regression Model Predicting Average Rating from Difficulty

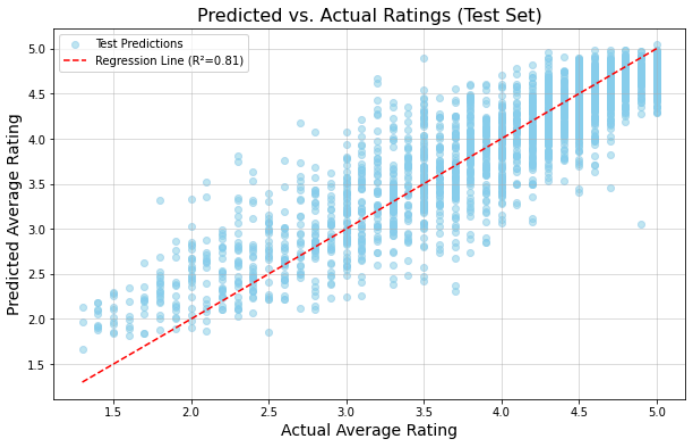
***R^2 (training):** 0.29, **RMSE (training)**: 0.95, **R^2 (test):** 0.28, **RMSE (test)**: 0.96*



To examine the relationship between average difficulty and average rating, I built a simple linear regression model using non-missing data for these variables. The model predicts a professor’s average rating based solely on the reported difficulty of their class. To ensure reproducibility and evaluate the model’s generalization ability, I implemented a train/test split using 80% of the data for training and 20% for testing, seeded to my unique N-number. On the training set, the model achieved an R^2=0.29 and RMSE = 0.95, indicating that approximately 29% of the variability in average rating is explained by average difficulty. While statistically significant, this result suggests that other factors beyond difficulty contribute substantially to ratings. The RMSE reflects the average error in predicting ratings, which is moderate given the typical rating range of 1 to 5. On the test set, the model produced an $R^2 = 0.28$ and RMSE = 0.96, demonstrating consistent performance on unseen data. These results confirm that the model generalizes well and that its predictions remain stable across different subsets of the data. The model demonstrates a significant negative relationship between difficulty and ratings: as perceived difficulty increases, ratings decrease (Slope = −0.61). However, the relatively low R^2 value suggests that difficulty alone is not sufficient to fully explain variations in ratings. Other factors, such as teaching quality, class size, or student biases, likely play a significant role in shaping evaluations. While difficulty is a statistically significant predictor of ratings, its explanatory power is limited. The moderate RMSE on both the training and test sets suggests room for improvement in predictive accuracy.

Question 8: Build a regression model predicting average rating from all available factors.

Analysis of Full Regression Model Predicting Average Rating



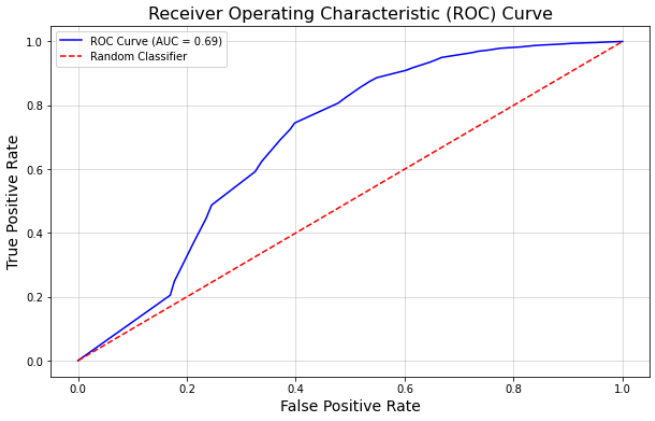
R² (training): *0.81,* ***RMSE (training)****: 0.37,* ***R² (test):*** *0.80,* ***RMSE (test)****: 0.38*

To predict average rating using all available factors, I built a multiple linear regression model with the following predictors: Average Difficulty, Number of Ratings, Received Pepper (hotness indicator), Proportion Retake, Number of Online Ratings, Male (binary indicator), and Female (binary indicator). The dataset was split into 80% training and 20% testing using a random seed tied to my N-number to ensure reproducibility. Predictors were standardized to mitigate collinearity and enable direct comparison of coefficients. The model generalizes well, maintaining high explanatory power and low prediction error across both datasets. Proportion retake (β=0.62, p<0.005) was the strongest predictor, showing that students' willingness to retake a class has the most substantial positive impact on ratings. Received pepper (β=0.10, p<0.005) was also a significant predictor, as professors judged as "hot" received higher ratings, highlighting the influence of subjective biases. Average difficulty (β=−0.15, p<0.005) retained a significant negative relationship with ratings but is less impactful compared to the difficulty-only model. Number of ratings and number of online ratings were not statistically significant and Male (β=0.03) and Female (β=0.01) were small but statistically significant positive effects, potentially reflecting subtle biases in evaluations. Predictors were standardized to mitigate collinearity. The condition number of 2.08 confirms no severe collinearity concerns, ensuring stable and interpretable coefficient estimates. The full model ($R^2 = 0.81$) explains significantly more variance than the difficulty-only model ($R^2 = 0.29$). Its RMSE (0.37) is much lower than that of the difficulty-only model (0.95), confirming superior predictive accuracy. The inclusion of Proportion retake and received pepper as key predictors highlights the importance of additional factors in explaining variations in ratings. The full model significantly outperforms the difficulty-only model, demonstrating that factors like proportion retake and received pepper play a critical role in predicting ratings. Proportion Retake emerges as the strongest determinant, emphasizing the centrality of students' willingness to retake a class. The influence of Received Pepper underscores the role of subjective biases, such as perceived "hotness," in shaping evaluations. While the model performs well, incorporating interactions between predictors or adding course-specific features could further enhance predictive power and provide deeper insights into the factors influencing teaching evaluations.

Question 9: Build a regression model predicting average rating from all available factors.

Analysis of the Classification Model Predicting "Hotness" (Pepper)

Accuracy*: 64%,* ***AUROC*** *(Area Under the ROC Curve): 0.69,* ***Not Hot****: precision of 0.86 (high confidence in predictions for the majority class), recall: 0.60 (captures most "not hot" professors), F1-Score: 0.71,* ***Hot****: precision of 0.42 (some false positives for the minority class), recall: 0.74 (captures most "hot" professors due to oversampling), F1-Score: 0.54, weighted average F1-Score: 0.66*

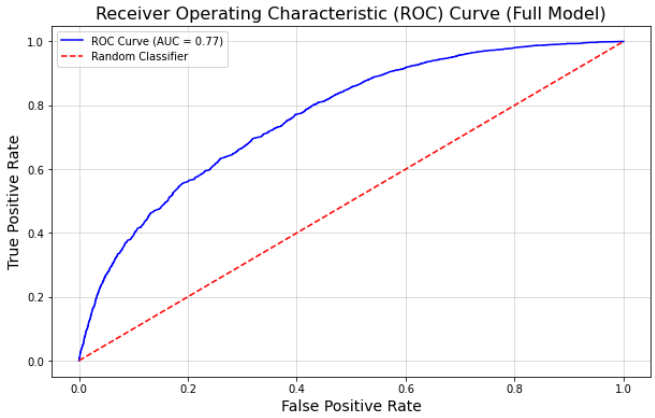


To predict whether a professor receives a "pepper" (hotness indicator) based solely on their average rating, I developed a logistic regression model. The data was split into 80% training and 20% testing, using a random seed for reproducibility. The minority class ("hot") was underrepresented, so I applied oversampling during training to improve the model's recall for this class. This approach ensures better recognition of minority instances, albeit at the cost of increased false positives. The model achieved an AUROC of 0.69, indicating moderate

predictive power and outperforming a random classifier. The high recall for the "hot" class (0.74) reflects the effectiveness of oversampling in improving sensitivity, though the reduced precision indicates more false positives. The "not hot" class showed balanced performance, with higher precision but moderate recall. The model demonstrates moderate success in predicting "hotness" using average rating alone. Oversampling effectively addressed class imbalance, enhancing recall for the minority class. However, relying solely on average rating limits explanatory power, as other factors (e.g., subject matter, class size) likely contribute to "pepper" designation. This classification model underscores the potential of using average ratings to predict "hotness," achieving a balanced tradeoff between precision and recall. Further improvements could incorporate additional predictors to enhance both performance and interpretability.

Question 10: Build a classification model that predicts whether a professor receives a “pepper” from all available factors.

Analysis of the Full Classification Model Predicting "Hotness" (Pepper)

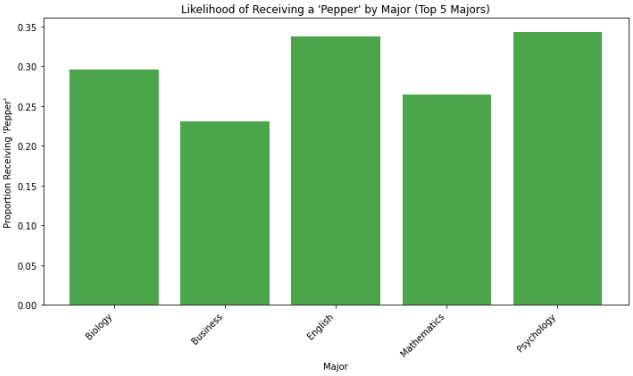


***Accuracy**: 68%, **AUROC** (Area Under the ROC Curve): 0.76, **Not Hot**: precision: 0.88, recall: 0.70, F1-Score: 0.78, **Hot**: precision: 0.44, recall: 0.73, F1-Score: 0.55, weighted average F1-score: 0.69*

To predict whether a professor receives a "pepper" (hotness indicator) using all available factors, I built a logistic regression model with predictors including Average Rating, Average Difficulty, Number of Ratings, Proportion Retake, Number of Online Ratings, Male, and Female. The dataset was split into 80% training and 20% testing, with missing values imputed using the median. Class imbalance was addressed by oversampling the minority class ("hot") during training.
The full model achieved an AUROC of 0.76, outperforming the "average rating only" model (AUROC = 0.69). Accuracy improved from 64% to 68%, reflecting better classification overall. Oversampling effectively increased the model’s sensitivity to the minority class, maintaining strong recall (0.73) while balancing precision and recall across classes. The inclusion of additional predictors, such as Average Difficulty and Proportion Retake, allowed the full model to capture more nuanced patterns, leading to improved classification metrics. While the "hot" class precision remains lower due to oversampling, the overall balance between precision and recall improved. The full model's ROC curve further highlights its superior ability to distinguish between classes compared to the simpler model. The full model significantly outperforms the "average rating only" model, demonstrating the value of incorporating multiple predictors. While the model effectively handles class imbalance and achieves better classification metrics, further enhancements, such as feature engineering or ensemble methods, could improve precision for the "hot" class and overall performance.

EXTRA CREDIT: “Hotness” of Professors by Major

***F-Statistic:** 41.32, **P-value:** 1.46e-34, **Biology:** 0.30, **Business:** 0.23, **English:** 0.34, **Math:** 0.26, **Psychology:** 0.34*



I wanted to try and see if there was a relation between a certain major and the likelihood of receiving a pepper. In other words, are professors in certain fields hotter than others? Two datasets were provided: one with numerical data (e.g., ratings and gender) and another with qualitative attributes (e.g., major, university, and state). The datasets were merged on their shared row indices to create a unified dataset containing all variables. Rows with missing values in the qualitative fields were removed to ensure alignment between datasets. The top 5 most common majors were identified based on frequency in the dataset. This step ensured sufficient representation for each major in the analysis. For each of the top 5 majors, the proportion of professors who received a "pepper" was calculated by taking the mean of the received pepper variable within each group. The count of professors in each major was also calculated to confirm robust

sample sizes. A bar chart was created to display the proportion of professors receiving a "pepper" for each of the top 5 majors. This visualization highlights differences in perceived attractiveness across academic disciplines. Handling missing values ensured that the analysis relied only on complete and meaningful records. Removing duplicates in columns avoided confusion during aggregation. Analyzing by major allowed for targeted insights into academic disciplines, aligning with the goal of exploring field-specific trends. Using the mean of the received pepper variable provided an intuitive measure of likelihood, as this binary variable directly reflects the proportion of "pepper" recipients. Focusing on the most common majors ensured a balanced trade-off between granularity and interpretability, avoiding noise from underrepresented groups. The bar chart effectively communicated differences between majors, making the results accessible and engaging. The analysis revealed notable differences in the likelihood of receiving a "pepper" across the top 5 majors. These findings suggest that students' perceptions of attractiveness may vary significantly by field, potentially reflecting underlying biases or cultural associations tied to specific disciplines. I conducted a an ANOVA test to determine the significance of the findings and determined that the results were indeed statistically significant (F-statistic: 41.32, P-value: 1.46e-34) and the variation in proportion of professors receiving a pepper determined by major is unlikely due to chance alone.