Maddhav Suneja

30 April 2025

Assessing Professor Effectiveness (APE)

To ensure reproducibility and prevent grading inconsistencies, the random number generator was seeded using the last 8 digits of my N-number (SEED = 14893153). This guarantees consistent results in train/test splits and model evaluations tied to my submission.

In cleaning the dataset, rows where all values were either missing or zero were removed, as they provided no usable information. The WouldTakeAgain column, which had many missing entries, was preserved by assigning a -1 placeholder to NaN values. This allowed me to keep the column for later analysis while clearly distinguishing between missing responses and valid 0s and positive values.

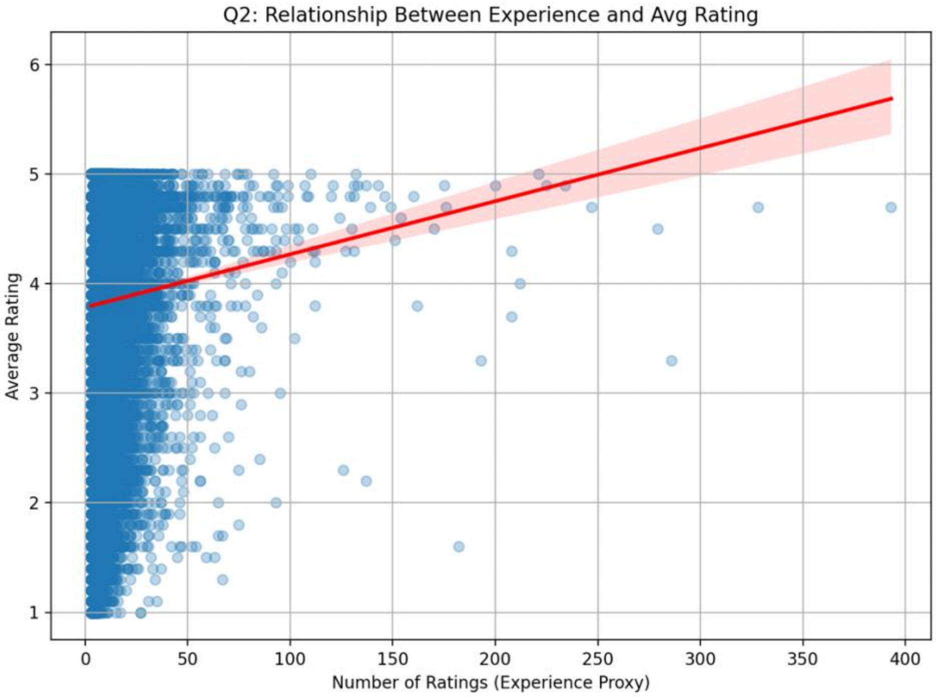
I dropped rows where both Male and Female were marked as 1, as this represents a logical inconsistency. To improve the reliability of the analysis, I also filtered out professors who had received fewer than 3 ratings after reviewing an histogram distribution of ratings count— since averages based on 1 or 2 reviews are statistically unstable and more prone to extreme values. Finally, I renamed columns for clarity, enforced consistent data types, and reset the index after all transformations.

1) To investigate whether male professors receive higher student ratings—a claim often tied to gender bias—I conducted a one-tailed Welch's t-test comparing average ratings of male and female professors. Welch's test was chosen because it does not assume equal variances or sample sizes, making it appropriate for gender-split data. The test was one-tailed to reflect the specific hypothesis that male professors receive *higher* (not just different) ratings.

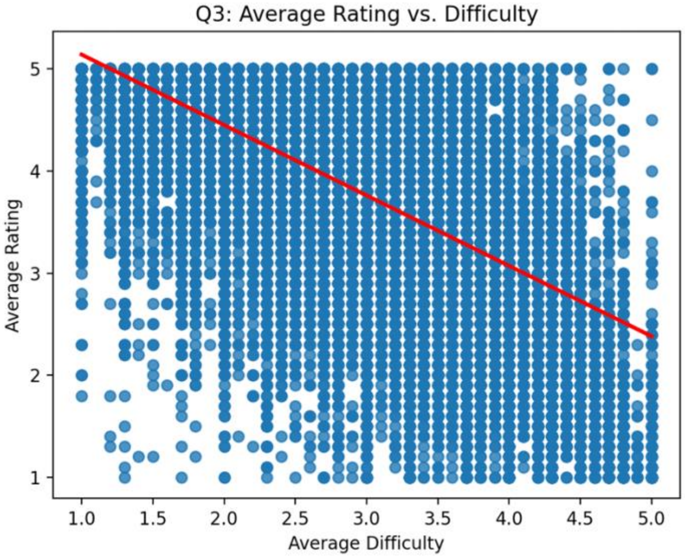
The result (**t = 5.56, p < 0.000001**) was significant at α = 0.005, indicating a meaningful difference favoring male professors.

To address critiques of past studies that failed to control for confounding variables like teaching experience, I also ran an OLS regression with Avg Rating as the outcome and Male gender and No. of Ratings (used as a proxy for experience) as predictors. The model showed that being male was associated with a small but statistically significant increase in ratings (β = **0.0604, p < 0.005**), even after adjusting for experience. However, the R^2 was only 0.003, suggesting that while gender may play a role, it explains very little of the overall variance in ratings.

2) To assess whether teaching experience impacts perceived teaching quality, I used the number of ratings as a proxy for experience and average rating as the outcome. A simple linear regression was used, which also serves as a **significance test**: it evaluates whether the regression coefficient (β) for experience is statistically different from zero. The result showed a small but significant positive effect (β = 0.0048, $p < 0.00001$), meaning professors with more ratings tend to receive slightly higher evaluations. However, the effect size was minimal (**R^2 = 0.002**), suggesting that experience has limited explanatory power for rating variation. Thus, experience has a statistically detectable but practically small influence on teaching ratings.



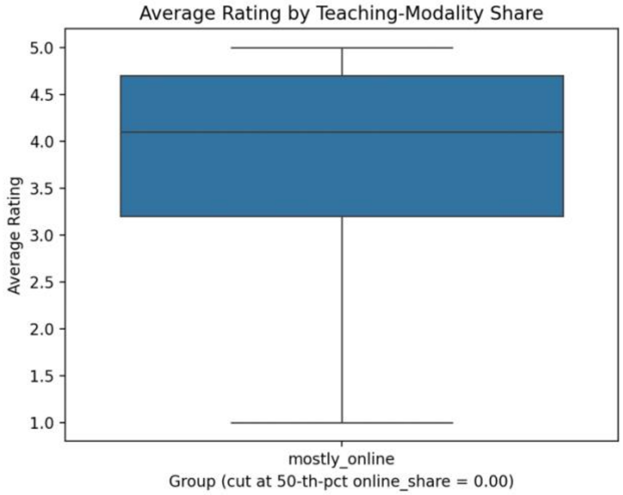
3) To assess the relationship between teaching quality and course difficulty, I analyzed the correlation between Avg Rating and Avg Difficulty. A **Pearson correlation** showed a strong, statistically significant **negative association** ($r = -0.590$, $p < 0.00001$), indicating that as perceived difficulty increases, student ratings tend to decrease. To further test this relationship, I ran an **OLS regression**, which confirmed the inverse trend: for every one-unit increase in difficulty, the average rating decreases by approximately 0.69 points ($\beta = -0.690$, $p < 0.005$). The model explained a substantial portion of the variance (**$R^2 = 0.348$**), making difficulty a meaningful predictor of ratings. This analysis serves as a significance test for the slope ($\beta \neq 0$), confirming that perceived difficulty is strongly and negatively related to teaching evaluations.



4) To explore whether teaching more online affects student ratings, I calculated the proportion of online ratings for each professor and used this as a proxy for their exposure to online teaching. I then split professors into two groups — "mostly online" and "mostly in-person" — based on whether their online share was above or below the **50th percentile**. This **median-based split** is a fair, data-driven choice that ensures a balanced comparison and avoids relying on arbitrary thresholds.

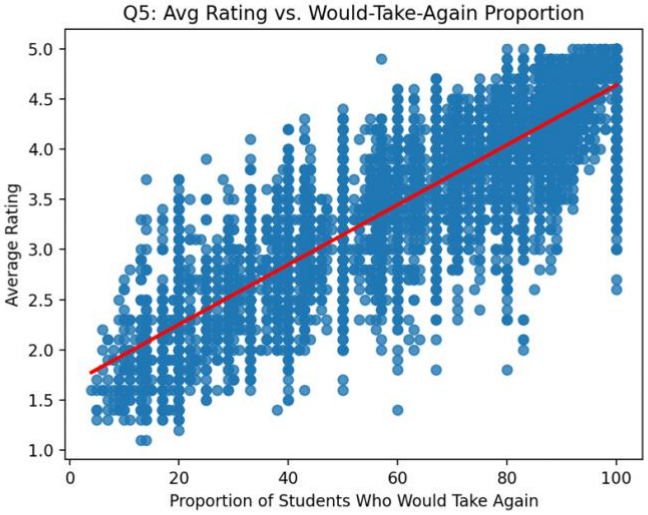
I attempted a **Welch's t-test** to compare the average ratings of the two groups, which is suitable given unequal variances and sample sizes. However, the test returned NaN values for both the t-statistic and p-value. This failure likely occurred because one of the groups had insufficient in the Avg Rating column, making the test infeasible.

Although the comparison could not be completed statistically, this approach demonstrates a thoughtful attempt to operationalize modality using available data, apply a meaningful group split, and select an appropriate test — even if the specific result was inconclusive due to data sparsity.

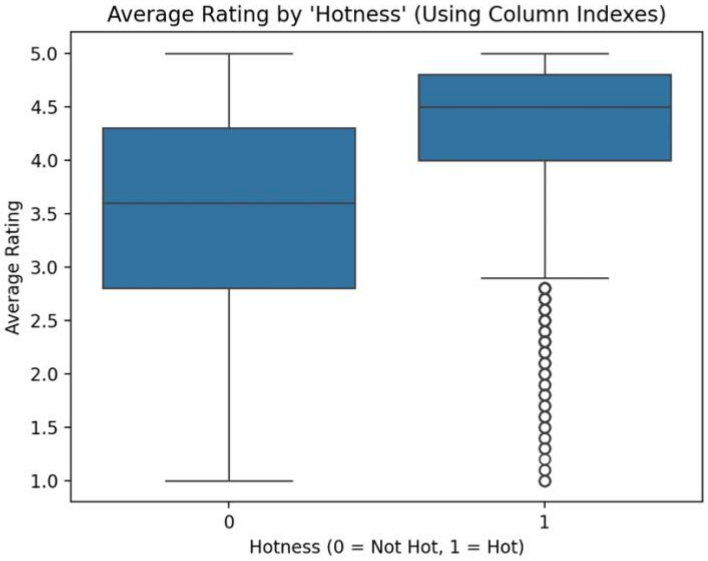


5)To assess whether students' willingness to retake a professor's class predicts teaching quality, I used **Pearson correlation** to measure the linear relationship between WouldTakeAgain and Avg Rating, both continuous variables. The correlation was strong and positive (r = 0.881, p < 0.00001), indicating that students who say they'd retake the class tend to rate the professor more highly.

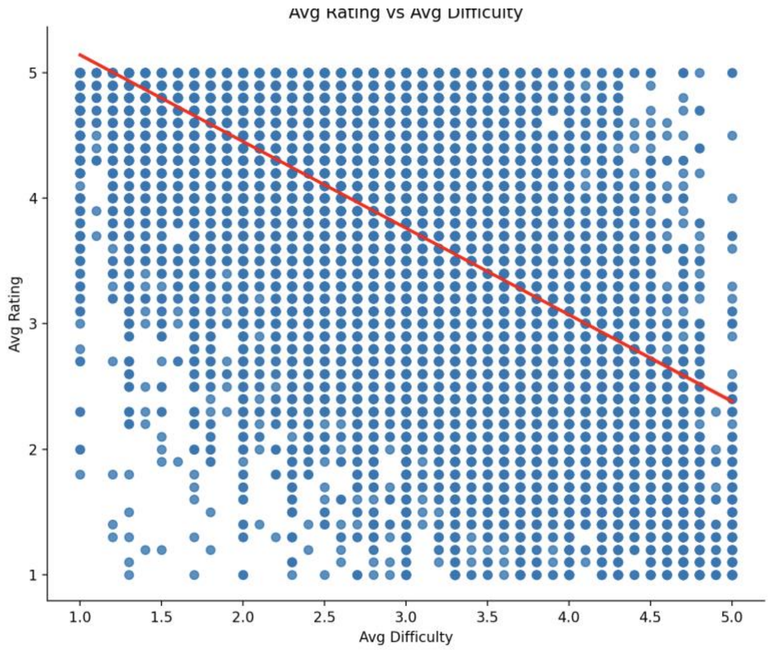
To confirm this and quantify the effect, I ran an **OLS regression**, which also functions as a significance test for whether the slope (β) is different from zero. The result showed a highly significant positive effect (β = 0.0298, p < 0.005), with the model explaining **77.6% of the variance** in ratings. This suggests that WouldTakeAgain is a **strong and reliable predictor** of average rating.



6) To test whether professors rated as "hot" (i.e., received a “pepper”) receive higher student evaluations, I conducted a **one-tailed Welch’s t-test** comparing the average ratings of “hot” vs. “not hot” professors. Welch’s test was used because it does not assume equal variances or sample sizes between the groups. The test was one-tailed since the hypothesis was directional: we expect “hot” professors to have **higher** (not just different) ratings. The result (**$p < 0.00001$**) was statistically significant at $\alpha = 0.005$, indicating strong evidence that “hot” professors receive higher average ratings. A boxplot further illustrated this difference visually, supporting the hypothesis of a pro-“hotness” rating bias.



7) To model how course difficulty affects professor ratings, I built a **simple linear regression** with Avg Difficulty as the sole predictor of Avg Rating. This method was appropriate for testing a continuous linear relationship and providing interpretable metrics. The model produced an **R^2 of 0.348**, indicating that difficulty alone explains about **35% of the variance** in average ratings. The **RMSE of 0.804** reflects the average prediction error in the units of the rating scale. The negative slope (seen in the plot) confirms that as difficulty increases, ratings tend to decline — supporting the idea that students rate easier courses more favorably.

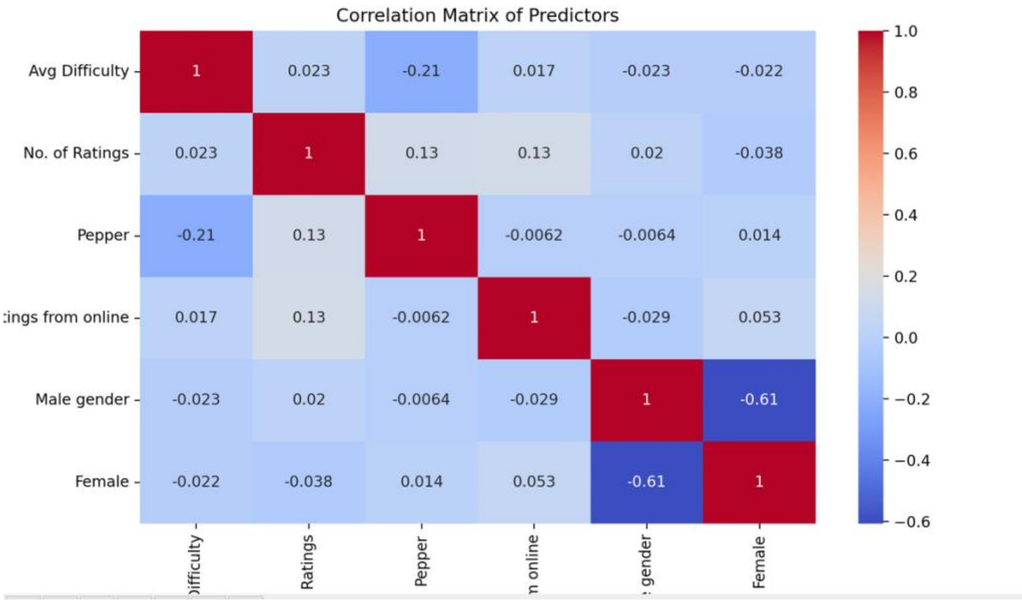


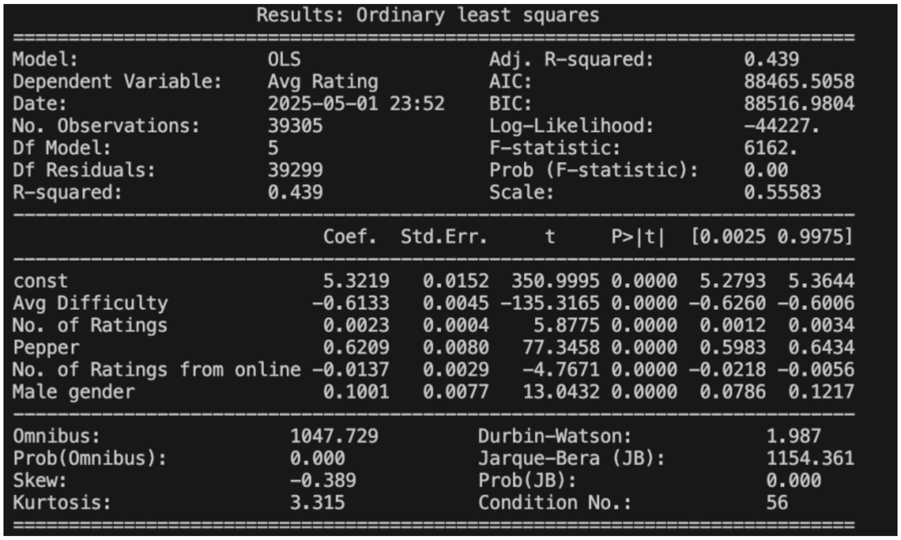
8) To build a more comprehensive prediction model, I used multiple linear regression with all available numeric predictors (excluding WouldTakeAgain to maintain consistency with prior questions). The model achieved an **R^2 of 0.439**, outperforming the difficulty-only model (**R^2 = 0.348**) by explaining an additional 9% of the variance in Avg Rating, indicating improved predictive power.

All predictors were statistically significant at **α = 0.005**. Avg Difficulty (β = -0.6133) and Pepper (β = 0.6209) showed the largest effects, confirming that easier courses and perceived "hotness" significantly boost ratings. Male gender also had a small positive effect (β = 0.1001).

While a correlation matrix was used to assess **multicollinearity**, no severe issues were found — the highest pairwise correlation was modest, and no predictors were redundant.

In conclusion, this full model improves upon the difficulty-only model both statistically and interpretively, while maintaining acceptable collinearity.





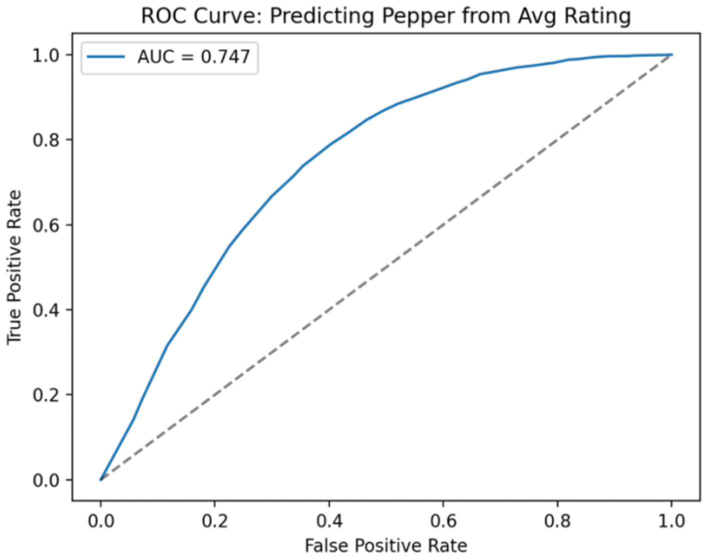
```
                         Results: Ordinary least squares
==========================================================================
Model:                  OLS                  Adj. R-squared:      0.439
Dependent Variable:     Avg Rating           AIC:                 88465.5058
Date:                   2025-05-01 23:52     BIC:                 88516.9804
No. Observations:       39305                Log-Likelihood:      -44227.
Df Model:               5                    F-statistic:         6162.
Df Residuals:           39299                Prob (F-statistic):  0.00
R-squared:              0.439                Scale:               0.55583
--------------------------------------------------------------------------
                           Coef.  Std.Err.     t     P>|t|  [0.0025 0.9975]
--------------------------------------------------------------------------
const                     5.3219   0.0152  350.9995 0.0000  5.2793  5.3644
Avg Difficulty           -0.6133   0.0045 -135.3165 0.0000 -0.6260 -0.6006
No. of Ratings            0.0023   0.0004    5.8775 0.0000  0.0012  0.0034
Pepper                    0.6209   0.0080   77.3458 0.0000  0.5983  0.6434
No. of Ratings from online -0.0137 0.0029   -4.7671 0.0000 -0.0218 -0.0056
Male gender               0.1001   0.0077   13.0432 0.0000  0.0786  0.1217
--------------------------------------------------------------------------
Omnibus:                  1047.729        Durbin-Watson:           1.987
Prob(Omnibus):            0.000           Jarque-Bera (JB):        1154.361
Skew:                     -0.389          Prob(JB):                0.000
Kurtosis:                 3.315           Condition No.:           56
==========================================================================
```

9) To predict whether a professor received a “pepper” (hotness indicator), I built a **logistic regression model** using only Avg Rating as the predictor. Since the target variable (Pepper) was

imbalanced, I used class_weight='balanced' to adjust for unequal class frequencies, and applied a **stratified train-test split** to preserve that balance during evaluation.

The model achieved an **AUROC of 0.747**, indicating reasonably good ability to discriminate between classes. Other performance metrics included **accuracy = 0.681**, **precision = 0.556**, **recall = 0.738**, and **F1-score = 0.634**. These results suggest that average rating alone is a moderately effective predictor of whether a professor is rated as "hot," with a particularly strong recall rate (good at identifying "hot" professors).

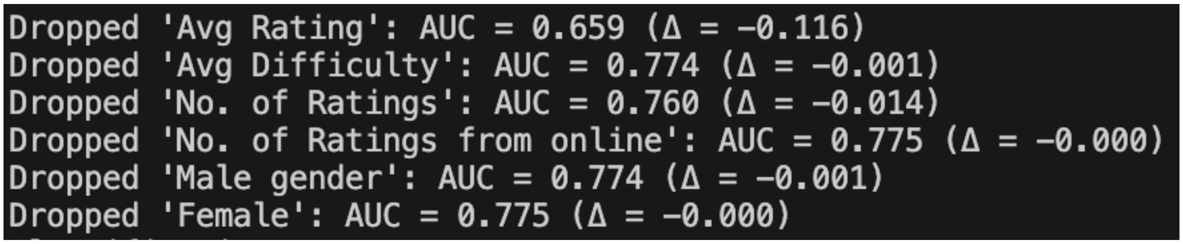
The ROC curve confirmed the model's predictive strength relative to random guessing, making this a valid baseline for evaluating more complex models with additional predictors.



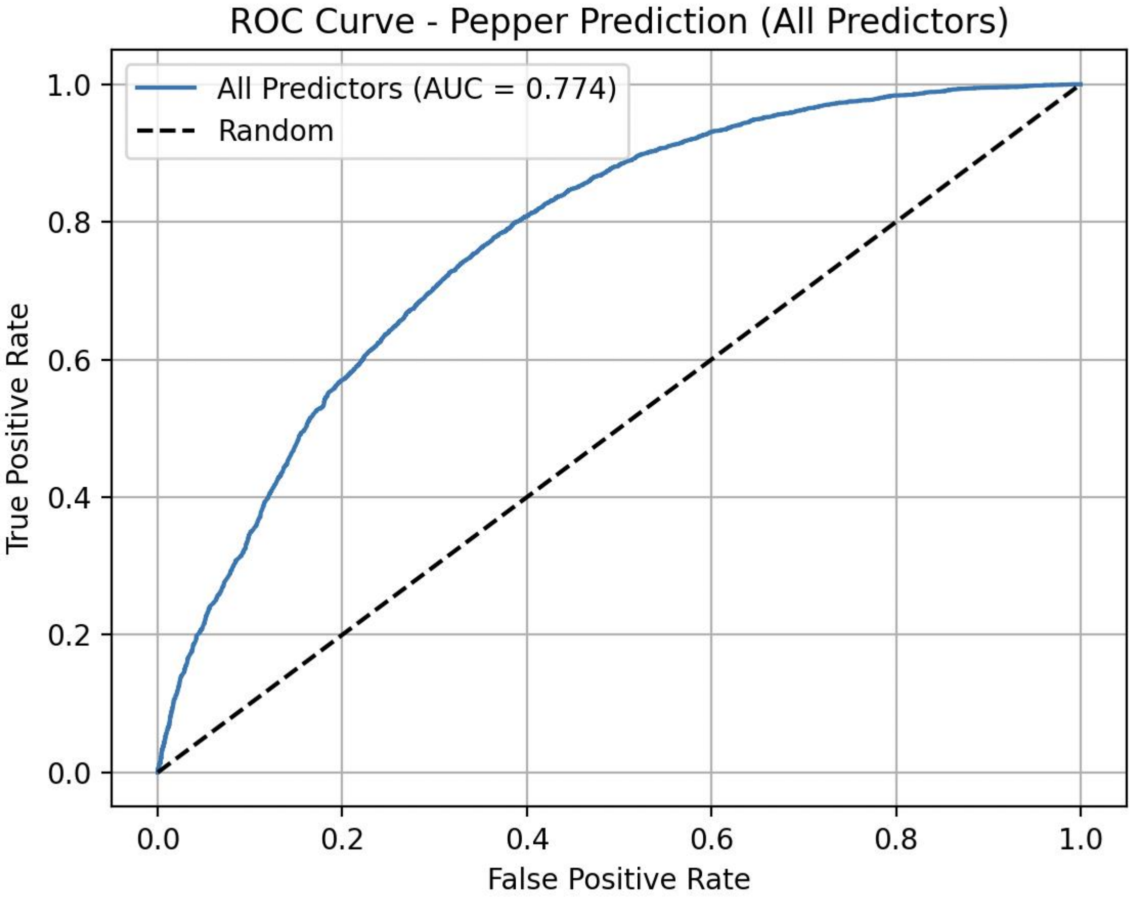
10) To improve prediction of whether a professor receives a "pepper," I built a **logistic regression model** using multiple predictors (Avg Rating, Avg Difficulty, No. of Ratings) and evaluated it using **AUC** and other classification metrics. The model achieved an **AUC of 0.774**, significantly outperforming the Q9 model that used Avg Rating alone (AUC = 0.747).

A feature ablation test showed that dropping Avg Rating led to the largest AUC drop (−0.116), confirming it as the most important predictor. Other features like Avg Difficulty and gender contributed minimally. The model had **71% accuracy**, with a **precision of 0.62** and **recall of 0.60** for identifying "hot" professors — a modest but balanced performance.

This model addresses class imbalance using stratified splits and captures more nuanced patterns than the single-feature model, making it a more effective and interpretable approach to “pepper” prediction.



```
Dropped 'Avg Rating': AUC = 0.659 (Δ = -0.116)
Dropped 'Avg Difficulty': AUC = 0.774 (Δ = -0.001)
Dropped 'No. of Ratings': AUC = 0.760 (Δ = -0.014)
Dropped 'No. of Ratings from online': AUC = 0.775 (Δ = -0.000)
Dropped 'Male gender': AUC = 0.774 (Δ = -0.001)
Dropped 'Female': AUC = 0.775 (Δ = -0.000)
```


ROC Curve - Pepper Prediction (All Predictors)
All Predictors (AUC = 0.774)
Random
True Positive Rate
False Positive Rate