Tanvi Patel
Principles of Data Science
Pascal Wallisch

Capstone Project

Preprocessing Steps:

I used my NYU N number (15956392) to seed the random number generator. The first row of each file appeared to be null or a leftover header row, so I dropped it and reset the index to keep things clean. To make both datasets easier to work with, I manually renamed the columns. I dropped rows with missing values under the “Average Ratings” and “Number of Ratings” columns. I printed out the median of “Number of Ratings” to understand how many ratings professors typically receive. The median was only 3, which I felt was too low to consider reliable, therefore I set a threshold of 5 or more ratings per professor for a meaningful analysis. I kept only those professors who had at least 5 ratings in the rmp_n dataset. I merged both datasets, aligning the rows across the two datasets by their index.

1. Is there evidence of a pro-male gender bias in this dataset?

I started by separating the dataset into two groups based on gender. I isolated the "Average Ratings" of male professors and female professors using the respective gender columns. Next I calculated the **median rating** for each group (male: **4.20;** female: **4.10**). To better understand the distribution of ratings, I plotted histograms for both male and female professors. Both distributions were right-skewed, which meant that the data were not normally distributed, so instead of a t-test, I used a non-parametric test (Mann–Whitney U test).

After running the Mann–Whitney U test, I found a statistically significant difference in ratings between male and female professors (**Mann–Whitney U statistic**: **50901762.5**; **p-value**: **0.0008**). This indicates that the difference in ratings is unlikely due to chance, meaning a potential pro-male bias in how professors are rated. I visualized this result using a boxplot, comparing average ratings by gender.

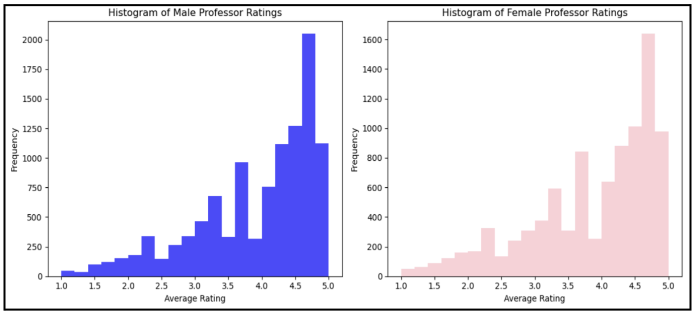


Figure 1. Histograms of Average Ratings by Gender

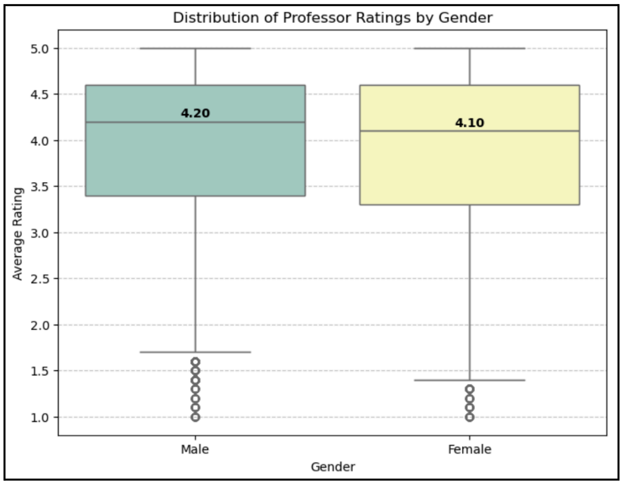


Figure 2. Boxplot Comparing Average Rating by Gender

2. Is there an effect of experience on the quality of teaching?

Since a few professors had extremely high numbers of ratings, I capped the data at the 99th percentile to prevent those outliers from skewing the results. Then, I computed the correlation between the number of ratings and average ratings and found a **correlation coefficient of 0.06**, showing a very weak positive relationship. To explore this further, I conducted a linear regression where the predator variable was "Number of Ratings" (experience) and outcome variable was "Average Ratings" (teaching quality). The results showed a statistically significant positive relationship (**p = 0.00**), a **small coefficient of 0.0048** (meaning for each additional rating, the predicted average rating only increases by 0.0048 points), and the **R^2 value** was **0.003** (meaning experience explains less than 1% of the variation in teaching ratings). I visualized this with a scatter plot and a regression line. While the regression line does slope slightly upwards, the data clearly show a wide spread, indicating that experience might matter a little, but there are many other factors that drive teacher ratings.

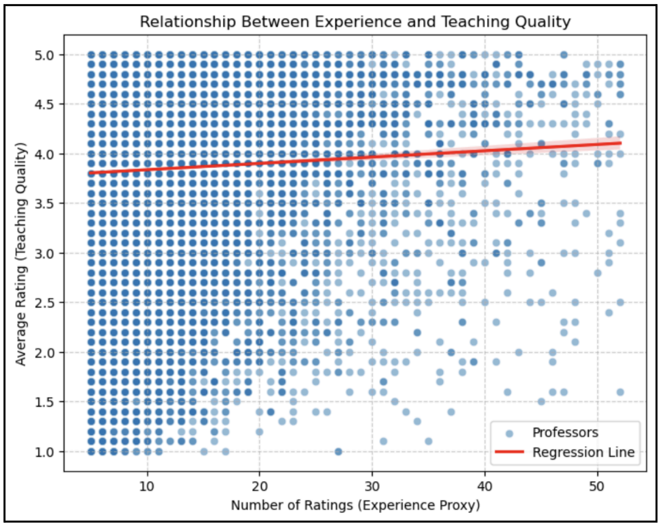


Figure 3. Scatterplot Showing Relationship Between Professor Experience and Teaching Quality

3. What is the relationship between average rating and average difficulty?

I conducted a Spearman correlation analysis between "Average Difficulty" and "Average Ratings". I chose Spearman correlation because it captures monotonic relationships and doesn't assume normality in the data. The results showed a **correlation coefficient of -0.6 with a p-value of 0.0**, which is both strong and statistically significant. This means that as the difficulty of a course increases, the professor's average rating tends to decrease. So, students generally give lower ratings to professors they find harder. To visualize this, I created a scatterplot with a regression line. The trend is downward-sloping, which reinforces the strong negative relationship between difficulty and ratings.

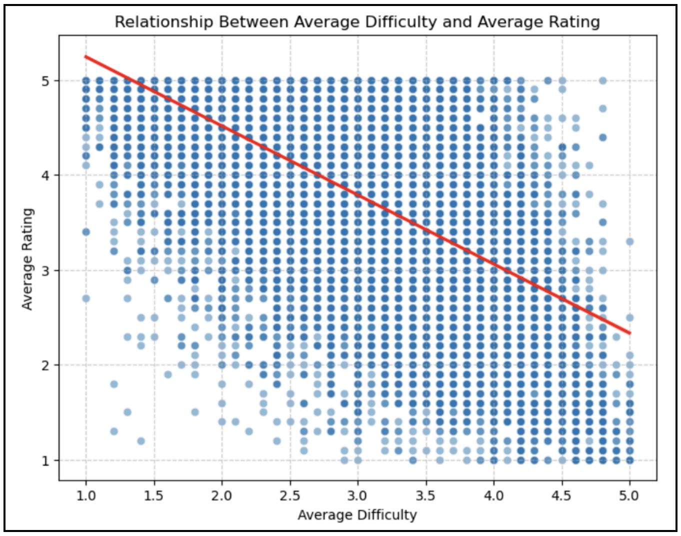


Figure 4. Scatter Plot Shows Relationship Between Average Difficulty and Average Rating

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

I used the proportion of online ratings (relative to total ratings) as a way to classify instructors. First, I cleaned the dataset by removing rows with missing values in key columns like "Online Ratings", "Number of Ratings", and "Average Ratings". I calculated a new column called Proportion Online, which represents the percentage of a professor's ratings that came from online courses. To split the data, I used a 50% threshold. Professors with more than 50% online ratings were classified as teaching "Many Online Classes". Those with 50% or less were grouped as "Few/No Online Classes". The histograms showed that the data distributions were not normal — both were skewed, with many ratings clustered near the upper end of the scale. Because the assumption of normality was violated, I chose to use the Mann–Whitney U test **(Mann-Whitney U Statistic: 9589124).** The test produced a statistically significant result **(p = 0.0)**, meaning that the difference in ratings between the two groups is unlikely due to chance. I visualized the result using a boxplot which clearly shows a higher median **(4.1)** for in-person-heavy professors and a wider spread and lower median **(3.8)** for online-heavy professors. This shows that students tend to rate professors who teach mostly in-person classes more favorably than those who teach mostly online.

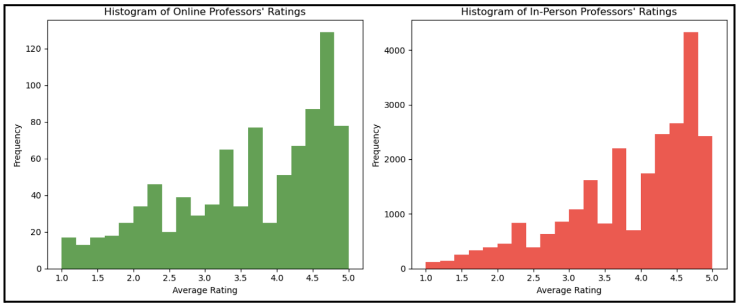


Figure 5. Histograms of Average Ratings by Teaching Modality

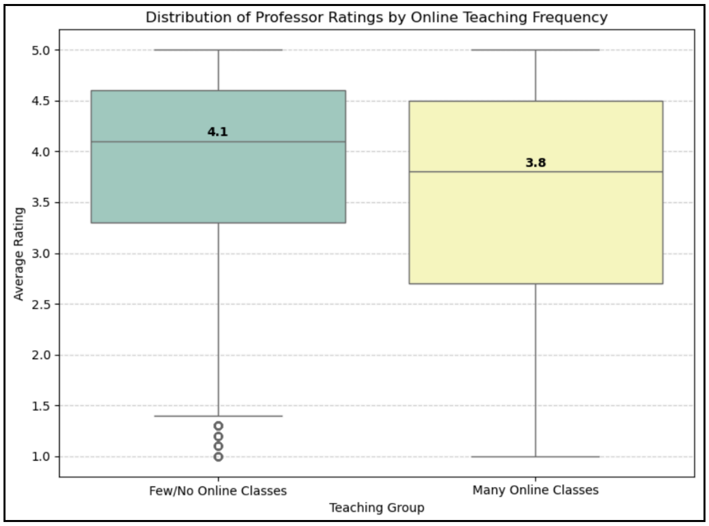


Figure 6. Boxplot of Average Ratings by Teaching Modality

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

I analyzed the relationship between Average Rating and Retake Percent. First, I cleaned the data by removing any rows with missing values in those two columns. I then calculated the Pearson correlation, which showed a very strong **positive correlation of 0.88**, meaning that, generally, the higher a professor's rating, the more likely students said they would take their class again. Next, I performed a linear regression, Retake Percent being the predictor and Average Rating being the outcome variable. The **coefficient was 0.0298**, indicating that for each 1% increase in retake rate, the average rating increases by about 0.03 points. The **R-squared was 0.775**, meaning ~77.5% of the variation in average ratings can be explained by the retake percent. The relationship was statistically significant **(p = 0.00).** I visualized the result with a scatterplot and regression line. The upward-sloping red line and clustering of points toward the upper-right corner reinforced the finding that professors with higher ratings are far more likely to be those students would take again.

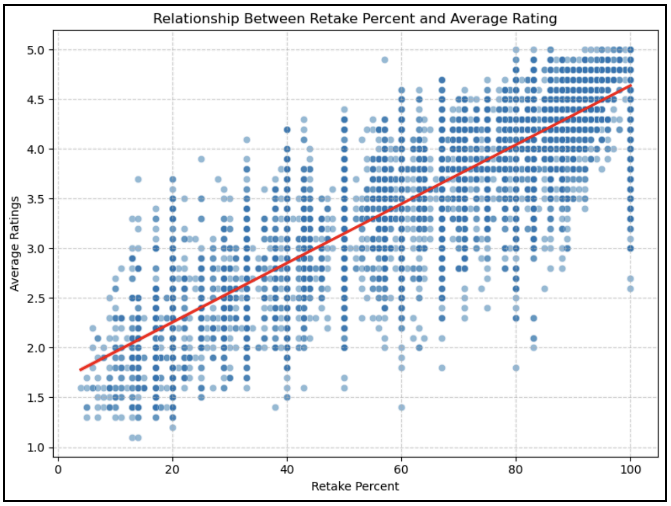


Figure 7. Scatterplot Showing Relationship Between Retake Percent and Average Rating

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

I started by splitting the dataset based on the “Pepper” column, where a value of 1 indicates a “hot” professor.I plotted separate histograms of average ratings for both hot and non-hot professors. Looking at the histograms showing the distribution of average ratings for hot/non-hot professors, the distributions were non-normal (skewed to the right) which led me to avoid using a t-test. I computed the **median rating** for each group (hot professors: **4.5**; non-hot professors: **3.6**). I conducted the Mann-Whitney U test (**statistic: 121910581.5; p-value: 0.0**), indicating that the difference in ratings between hot and not-hot professors is not due to chance. I created a boxplot comparing the two groups which clearly showed a higher/tighter rating distribution for hot professors and a lower/wider distribution for non-hot professors. Looking at the boxplot and difference in medians, it is clear that professors who are labeled as “hot” receive statistically significantly higher ratings than those who are not.

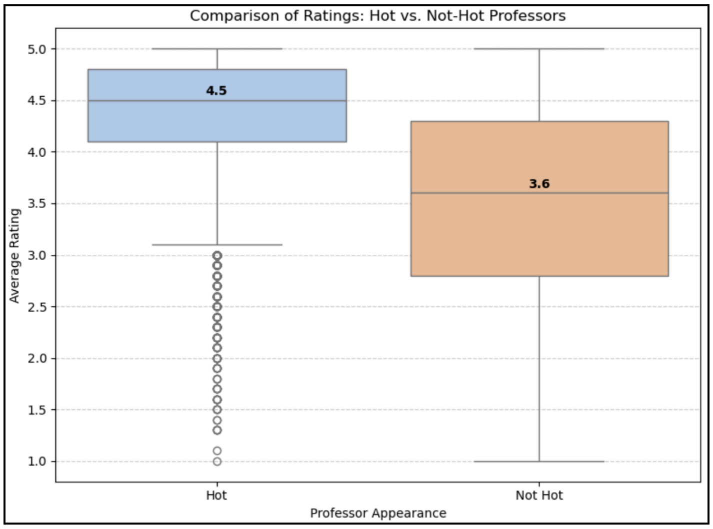


Figure 8. Boxplot Showing Average Ratings by Professor Appearance

7. Build a regression model predicting the average rating from difficulty

I built a simple linear regression model using one predictor: Average Difficulty. My outcome variable was Average Ratings. I started by cleaning the data, removing rows with missing values in either “Average Ratings” or “Average Difficulty”. I defined my predictor X as “Average Difficulty” and my target y as “Average Ratings”, then added a constant term for the intercept. I used OLS regression to fit the model. I computed R-squared and RMSE to evaluate the model performance. The model showed a statistically significant negative relationship between difficulty and ratings. The coefficient for **Average Difficulty** was **−0.73**, meaning that as difficulty increases by 1 point, the average rating tends to **drop by 0.73 points**. The **R-squared value was 0.38**, meaning **38% of the variation in ratings** can be explained by difficulty alone. That’s quite strong for a single predictor, and the negative beta indicates students tend to give lower ratings to professors perceived as harder. The **RMSE was 0.74**, indicating moderate prediction error. The model was statistically significant (**p=0.00; F-statistic:**

1.576e+04). The scatterplot and regression line visually confirmed that as difficulty rises, ratings decline.

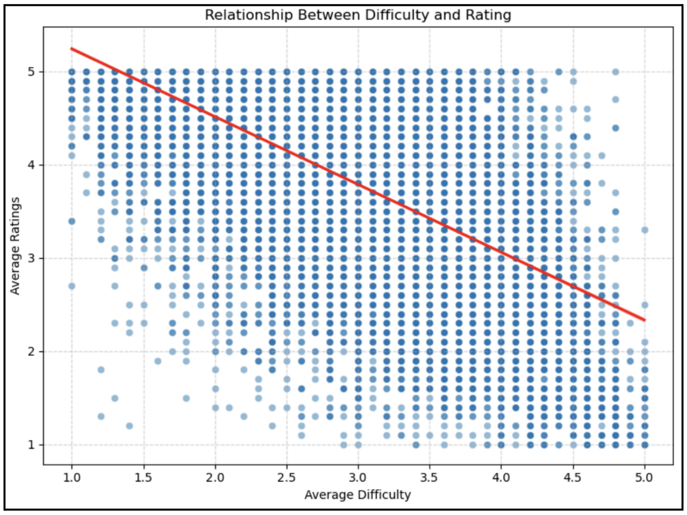


Figure 9. Regression Model of Average Ratings Predicted by Course Difficulty

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

I created a multiple linear regression model using all available factors. I first cleaned the data by dropping rows missing values for key predictors, including Average Difficulty, Number of Ratings, Retake Percent, Online Ratings, and gender indicators (Male, Female). The model achieved an **R-squared value** of **0.799**, meaning that approximately 80% of the variance in average ratings could be explained by these features, with a much lower **RMSE of 0.380** compared to the difficulty-only model's RMSE of 0.74. I also checked for multicollinearity by calculating Variance Inflation Factors (VIFs), and found that all predictors had VIFs between 1.01 and 1.38, suggesting no serious multicollinearity concerns. When comparing the full model to the difficulty-only model, the **R-squared improved by 0.415** (from 0.384 to 0.799). Examining the individual beta coefficients, I found that Average Difficulty had the strongest negative effect on ratings (**β = –0.2030, p < 0.001**), while Retake Percent had a strong positive influence (**β = 0.0266, p < 0.001**). Gender also showed small but statistically significant positive effects, with Male (**β = 0.0419, p < 0.001**) and Female (**β = 0.0293, p = 0.001**) both associated with slightly higher ratings after controlling for other variables. In contrast, the Number of Ratings (**β = 0.0004, p = 0.122**) and Online Ratings (**β = –0.0003, p = 0.872**) did not significantly predict ratings once other factors were considered.

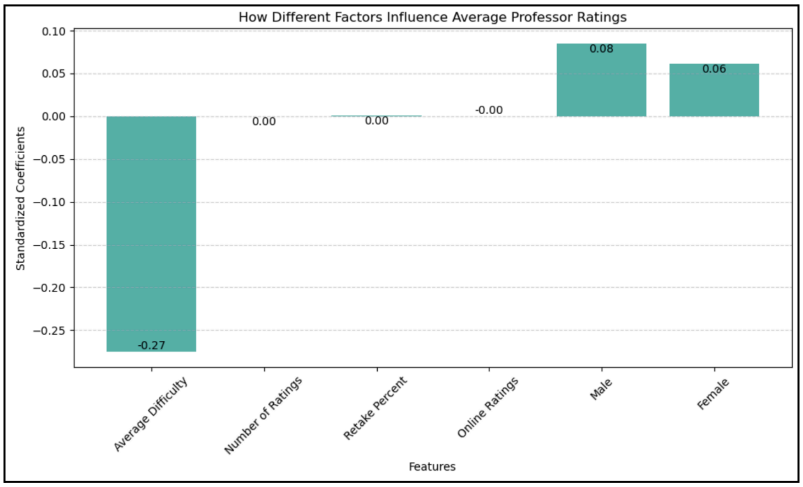
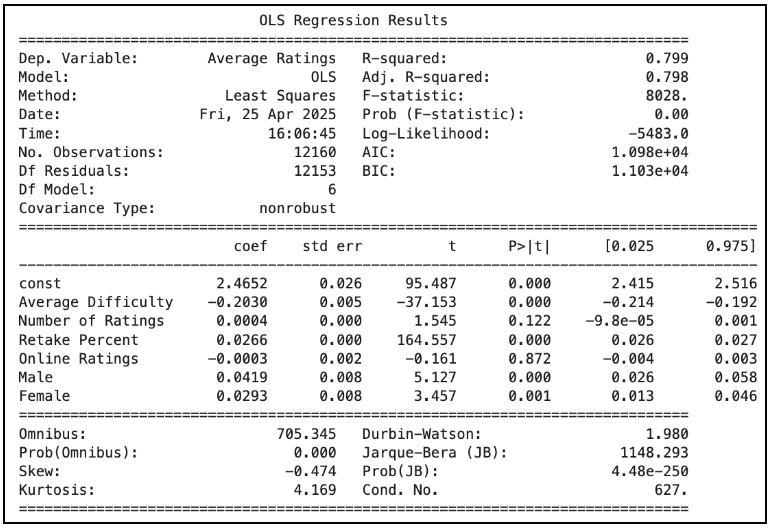


Figure 11. Barplot Showing Feature Importance Based on Standardized Coefficients



```
                            OLS Regression Results
==============================================================================
Dep. Variable:        Average Ratings   R-squared:                       0.799
Model:                            OLS   Adj. R-squared:                  0.798
Method:                 Least Squares   F-statistic:                     8028.
Date:                Fri, 25 Apr 2025   Prob (F-statistic):               0.00
Time:                        16:06:45   Log-Likelihood:                -5483.0
No. Observations:               12160   AIC:                         1.098e+04
Df Residuals:                   12153   BIC:                         1.103e+04
Df Model:                           6
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
const                 2.4652      0.026     95.487      0.000       2.415       2.516
Average Difficulty   -0.2030      0.005    -37.153      0.000      -0.214      -0.192
Number of Ratings     0.0004      0.000      1.545      0.122    -9.8e-05       0.001
Retake Percent        0.0266      0.000    164.557      0.000       0.026       0.027
Online Ratings       -0.0003      0.002     -0.161      0.872      -0.004       0.003
Male                  0.0419      0.008      5.127      0.000       0.026       0.058
Female                0.0293      0.008      3.457      0.001       0.013       0.046
==============================================================================
Omnibus:                      705.345   Durbin-Watson:                   1.980
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1148.293
Skew:                          -0.474   Prob(JB):                    4.48e-250
Kurtosis:                       4.169   Cond. No.                         627.
==============================================================================
```

Figure 12. OLS Regression Summary Table for All Predictors

9. Build a classification model that predicts whether a professor a “pepper” from average rating only

I built a binary classification model using logistic regression. I first preprocessed the data by dropping rows with missing values in either "Average Ratings" or "Pepper." Since the dataset was imbalanced (with far fewer hot professors than non-hot ones) I unsampled the minority class (professors with a pepper) so that both classes were equally represented. After balancing the data, I set up the model by using "Average Ratings" as the sole predictor and splitting the data into training and testing sets with an 80/20 stratified split.

I trained a logistic regression model on the training set and evaluated its performance on the test set. To assess how well the model distinguished between professors with and without peppers, I calculated the **AUROC**, which came out to **0.78**, indicating good discriminative ability. I also plotted the ROC curve, which showed the model performed much better than random guessing. In addition, I generated a classification report showing an overall **accuracy of 72%**. The model had a **precision** of **0.76** for predicting non-hot professors and **0.69** for predicting hot professors, while **recall** was **0.64** for non-hot and **0.80** for hot professors. Overall, I found that average rating alone is a fairly strong predictor of whether a professor receives a pepper, even though other factors might further improve prediction if added.

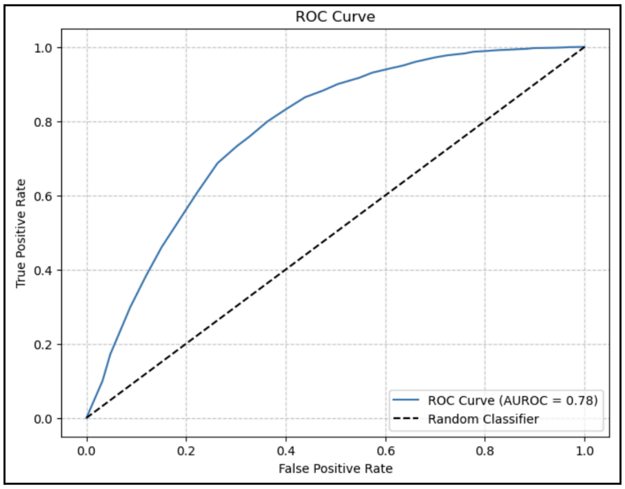


Figure 13. ROC Curve for Average Rating Model

10. Build a classification model that predicts whether a professor receives a "pepper" from all available factors.

I created a classification model using all available factors. I dropped rows missing values for key features including Average Ratings, Average Difficulty, Number of Ratings, Retake Percent, Online Ratings, Male, Female, and Pepper. Since the data was imbalanced, with fewer professors receiving a pepper, I upsampled the pepper group to create a balanced dataset. I then trained a logistic regression model using all the features and evaluated its performance. The model achieved an **AUROC of 0.79**, slightly better than the average rating-only model which had an AUROC of 0.78. The classification report showed an **overall accuracy of 73%**, with a **precision** of **0.76** for non-hot professors and **0.70** for hot professors, and **recall** values of **0.66** and **0.79**, respectively. Although the performance improved only modestly compared to using average rating alone (AUROC improvement of 0.017), the model using all factors was slightly better at identifying peppers, suggesting that while average rating carries most of the predictive power, additional features like difficulty, retake percent, and gender add some extra information.

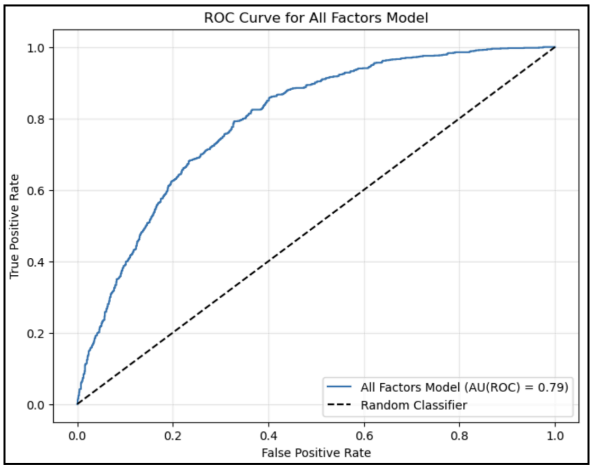


Figure 14. ROC Curve for All Factors Model

Extra Credit: Which Universities Have the Highest Pepper Rates?

I analyzed pepper rates across universities. First, I filtered the dataset to exclude professors with missing information on university name or pepper status. I then grouped the data by university and calculated each university's pepper rate as the number of "hot" professors divided by the total number of professors reviewed. To improve reliability, I restricted the analysis to universities with at least 10 professors reviewed. I then sorted the universities by pepper rate and plotted the top 10 universities with the highest rates. I found that Community College of Philadelphia had the highest pepper rate at 0.85, followed by Los Angeles Southwest College and College of the Sequoias, both at 0.80.

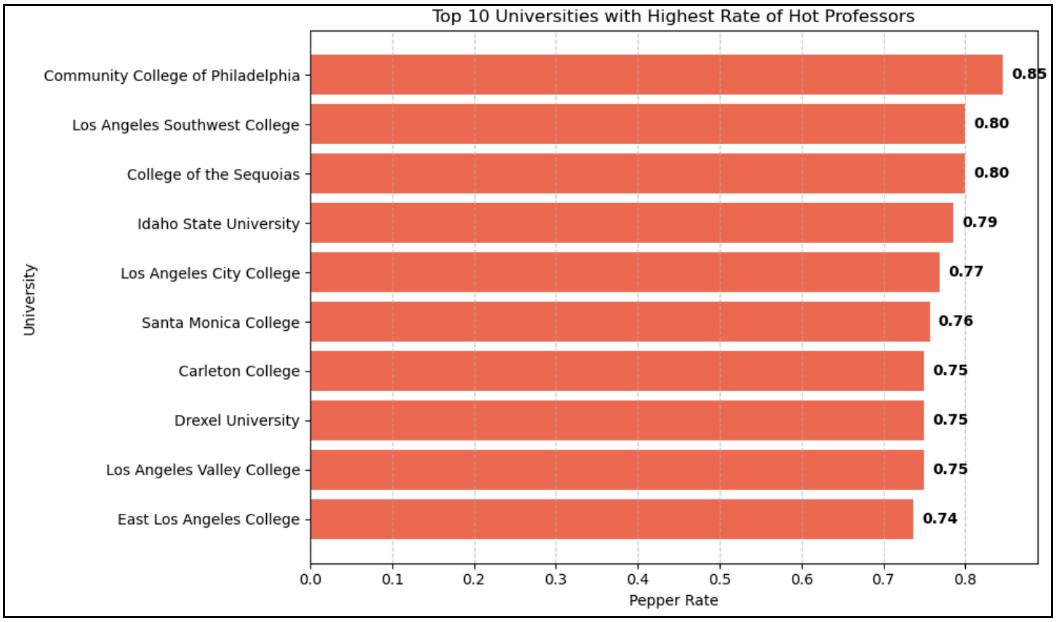


Figure 15. Bar Chart: Top 10 Universities with Highest Rate of Hot Professors