Rushil Johal
NYU Center for Data Science
December 2024

Principles of Data Science Capstone: Assessing Professor Effectiveness Using RateMyProfessor

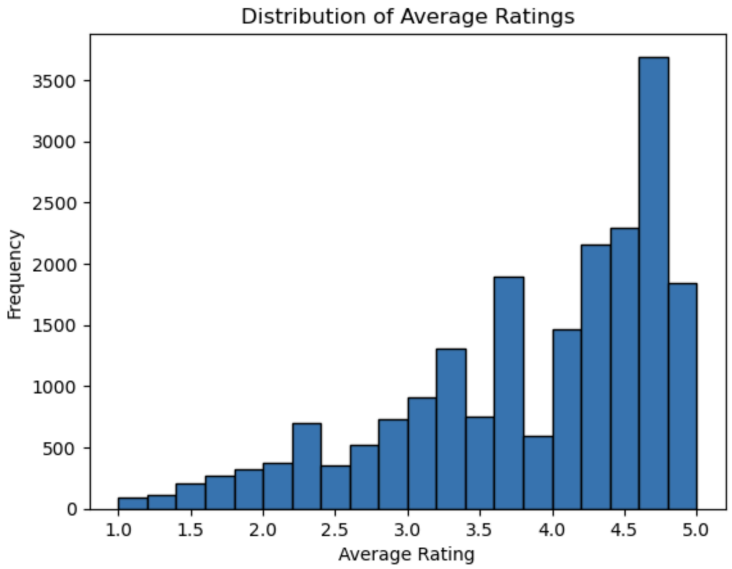
Data Preprocessing and Cleaning

To begin, I seeded the random number generator using my eight digit N-number to ensure the replicability of any random sampling or data splits used in subsequent analyses.

To ensure the dataset is suitable for analysis, I loaded the 'rmpCapstoneNum.csv' and 'rmpCapstoneQual.csv' datasets. I first renamed all columns for clarity to match their descriptions (e.g. Average_Difficulty). I removed all missing values in critical columns, such as Average_Rating and Number_of_Ratings. Only professors with more than 5 ratings were retained to ensure meaningful average ratings and avoid yielding extreme average ratings, leaving 20,577 valid data records out of the original 89,892. The datasets were then merged on their indices, combining numerical metrics with qualitative fields in a new cleaned dataset, 'cleaned_rmp_data.csv,' which I will use going forward for analysis. Lastly, Boolean fields (e.g. Received_Pepper and the genders) were normalized to binary values (0 or 1).

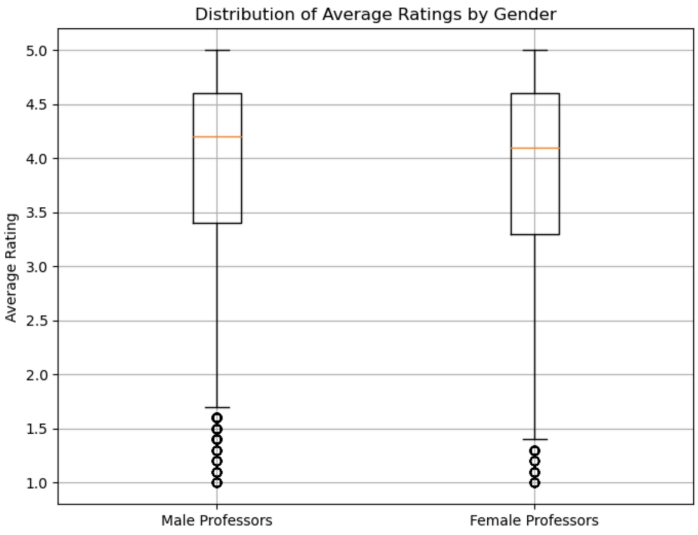
A histogram below shows the distribution of Average_Ratings, highlighting its skew toward higher ratings. Also below is an overview of 'cleaned_rmp_data.csv.'

```
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Average_Rating        20577 non-null  float64
 1   Average_Difficulty    20577 non-null  float64
 2   Number_of_Ratings     20577 non-null  float64
 3   Received_Pepper       20577 non-null  int64
 4   Proportion_Take_Again 11938 non-null  float64
 5   Online_Ratings        20577 non-null  float64
 6   Male                  20577 non-null  int64
 7   Female                20577 non-null  int64
 8   Major_Field           20577 non-null  object
 9   University            20577 non-null  object
 10  State                 20577 non-null  object
dtypes: float64(5), int64(3), object(3)
```



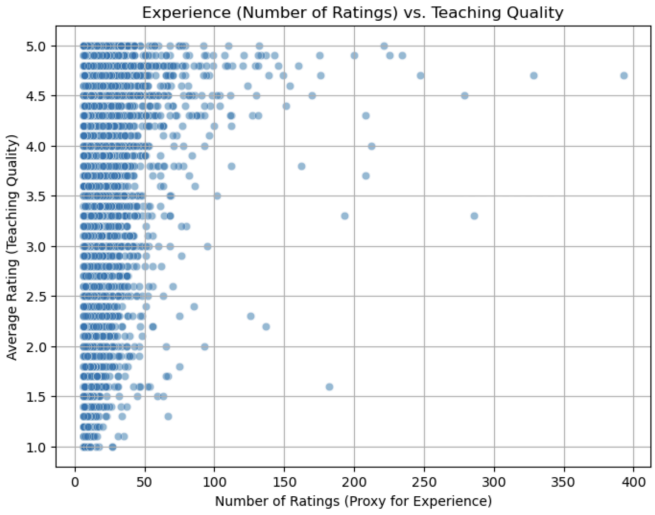
1) **Activists have asserted that there is a strong gender bias in student evaluations of professors, with male professors enjoying a boost in rating from this bias. While this has been celebrated by ideologues, skeptics have pointed out that this research is of technically poor quality, either due to 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. Hint: A significance test is probably required.**

I performed a two-sample t-test comparing the average ratings of male and female professors. I first extracted ratings for male (Male = 1) and female (Female = 1) professors. A boxplot visualizes the distributions of ratings by gender. The t-test, assuming unequal variances, yielded a t-statistic of 3.74 and a p-value of 0.00018, below the alpha threshold (0.005). Male professors had a slightly higher mean rating (3.93) compared to female professors (3.87), though the effect size appears small given the overlapping distributions and similar standard deviations (male: 0.90, female: 0.93). Overall, this suggests a statistically significant but very modest pro-male bias in ratings.



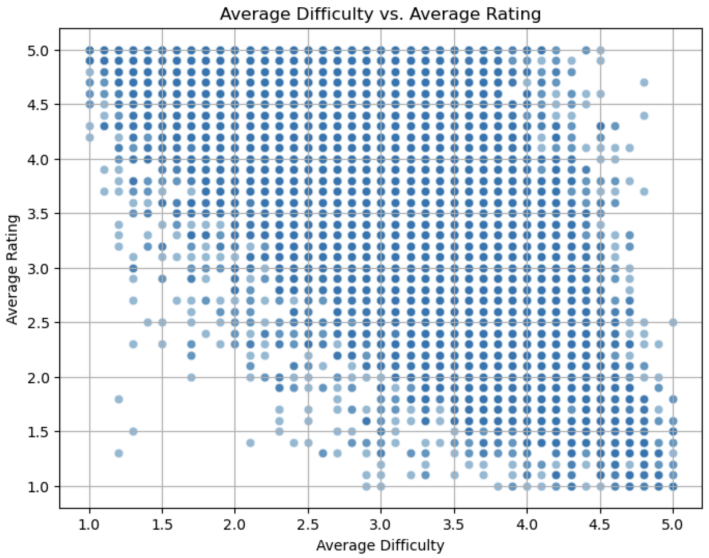
2) **Is there an effect of experience on the quality of teaching? You can operationalize quality with average rating and use the number of ratings as an imperfect – but available – proxy for experience. Again, a significance test is probably a good idea.**

I used Number_of_Ratings as a proxy for experience and Average_Rating for teaching quality. I computed the Spearman correlation coefficient to capture any non-linear and/or monotonic relationships while being it more robust to outliers. The correlation was weak (ρ ≈ 0.037), but statistically significant (p-value ≈ 8.91e-8), likely due to the large sample size. A scatterplot illustrated no substantial trend between experience and quality, as data points were widely dispersed. The median number of ratings was 9, indicating that most professors had a moderate amount of experience. Overall, while the significant p-value suggests a relationship exists, the weak correlation implies minimal practical effect of experience on the quality of teaching.



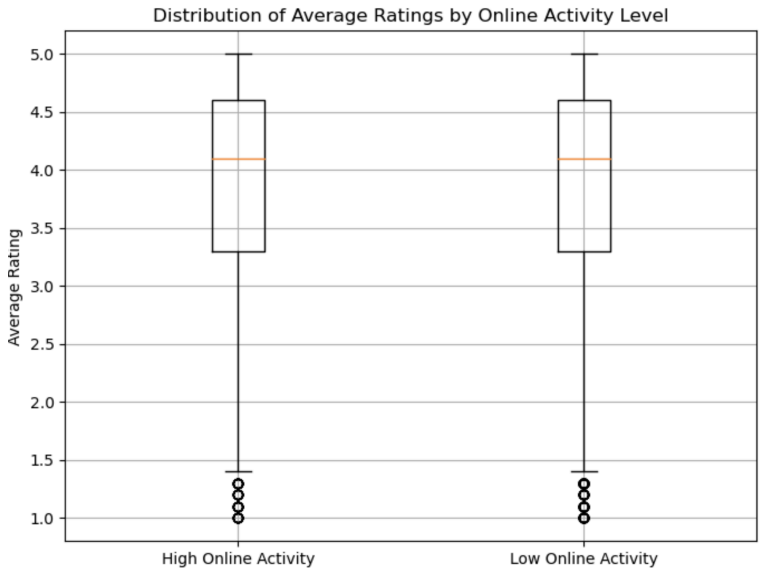
3) What is the relationship between average rating and average difficulty?

I calculated the Spearman correlation coefficient due to the same case for Question #2. The correlation coefficient was -0.612, indicating a strong, statistically significant negative relationship (p-value < 0.001). This suggests that professors with higher difficulty scores tend to receive lower ratings. The scatterplot confirms this inverse relationship, where as average difficulty increases, average rating decreases. On average, the mean difficulty was 2.94, and the mean rating was 3.85, highlighting that most professors are rated above average while being perceived as moderately difficult.



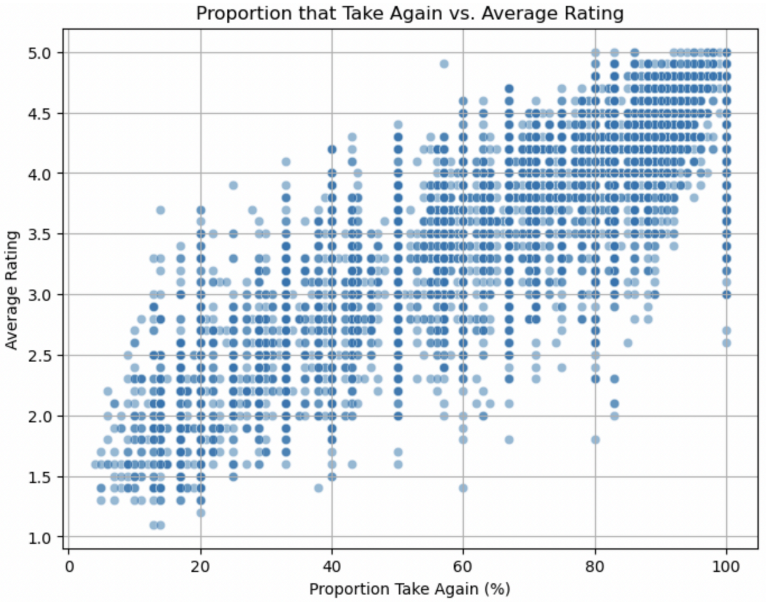
4) Do professors who teach a lot of classes in the online modality receive higher or lower ratings than those who don't? Hint: A significance test might be a good idea, but you need to think of a creative but suitable way to split the data.

I used the median number of online ratings (proxy for activity) to split the dataset into two groups. Professors with online ratings above the median were classified as "High Online Activity," while others were in the "Low Online Activity" group. A two-sample t-test revealed no significant difference in ratings (t-statistic ≈ -0.718, p-value ≈ 0.473). The mean ratings were very similar: 3.84 for high online activity and 3.85 for low. The boxplot confirms the similarity, as distributions largely overlap. Overall, online teaching activity appears unrelated to average ratings.



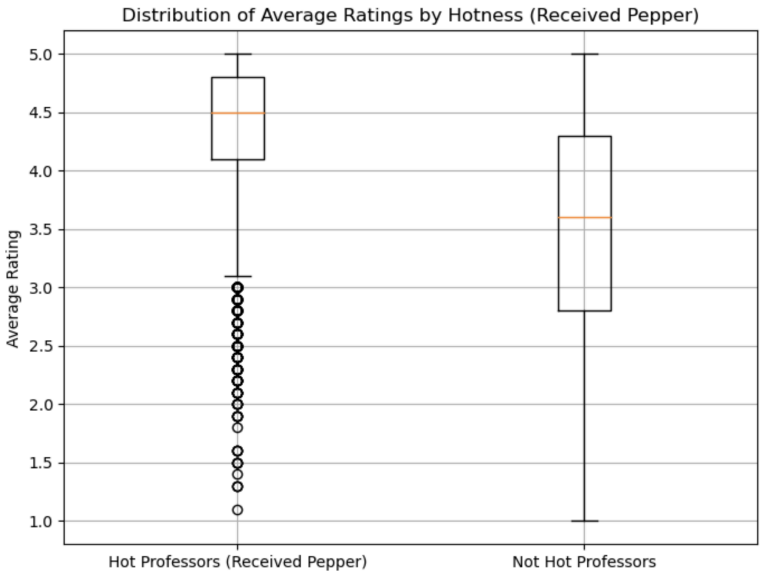
5) **What is the relationship between the average rating and the proportion of people who would take the class the professor teaches again?**

I first filtered the dataset to remove missing values in the Proportion_Take_Again column. Then, I calculated the Spearman correlation coefficient due to the same reasons like for Questions #2 and #3, which showed a strong positive relationship ($\rho \approx 0.850$, p-value < 0.001). This indicates that professors with higher ratings are more likely to have a higher proportion of students willing to retake their class. The scatterplot also confirms this direct relationship, where as Proportion_Take_Again increases, so does the Average_Rating. On average, 76.46% of students would retake a professor's class, and the mean rating was 3.94. Overall, this suggests that student satisfaction is highly linked to their willingness to retake a course.



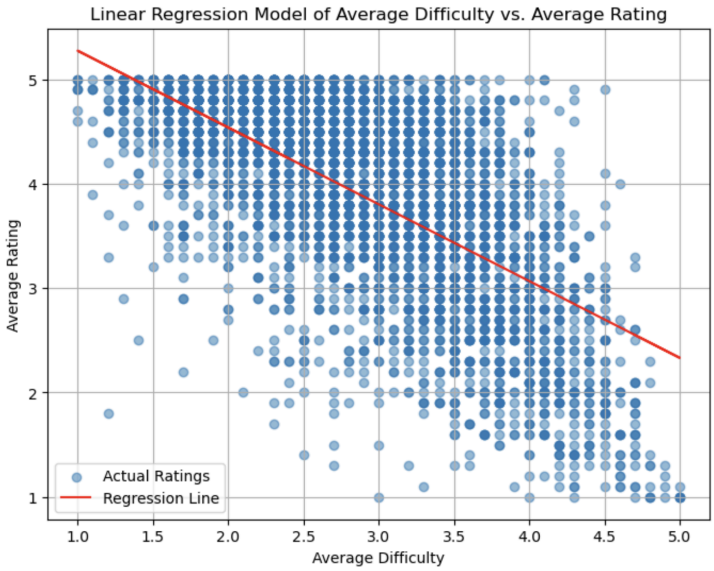
6) **Do professors who are "hot" receive higher ratings than those who are not? Again, a significance test is indicated.**

I split the dataset into two groups: professors with a pepper (Received_Pepper = 1), labeled as "hot", and those without (Received_Pepper = 0). A two-sample t-test revealed a highly significant difference (t-statistic ≈ 83.03, p-value < 0.001). Hot professors had a much higher mean rating of 4.35 compared to 3.46 for non-hot professors. The boxplot also illustrates the clear disparity between the two groups. These findings suggest a strong association between perceived attractiveness (hotness) and number of ratings. Thus, professors who are "hot" tend to receive higher ratings.



7) **Build a regression model predicting average rating from difficulty (only). Make sure to include the R2 and RMSE of this model.**

I built a linear regression model with Average_Difficulty as the independent variable and Average_Rating as the dependent variable. After splitting the data into training (80%) and testing (20%) sets, I fit the model and evaluated its performance. The model achieved an R2 of 0.408, indicating that about 40.8% of the variability in ratings can be explained by difficulty. The root mean squared error (RMSE) was 0.714, suggesting moderate prediction accuracy. The regression line's slope of -0.735 confirms a negative relationship between difficulty and ratings, with an intercept of 6.008. The plot also shows a clear downward trend, confirming that as difficulty increases, ratings tend to decrease.

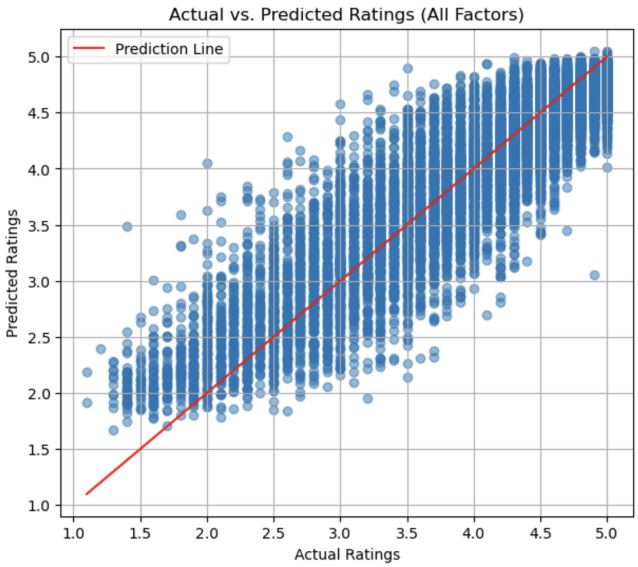


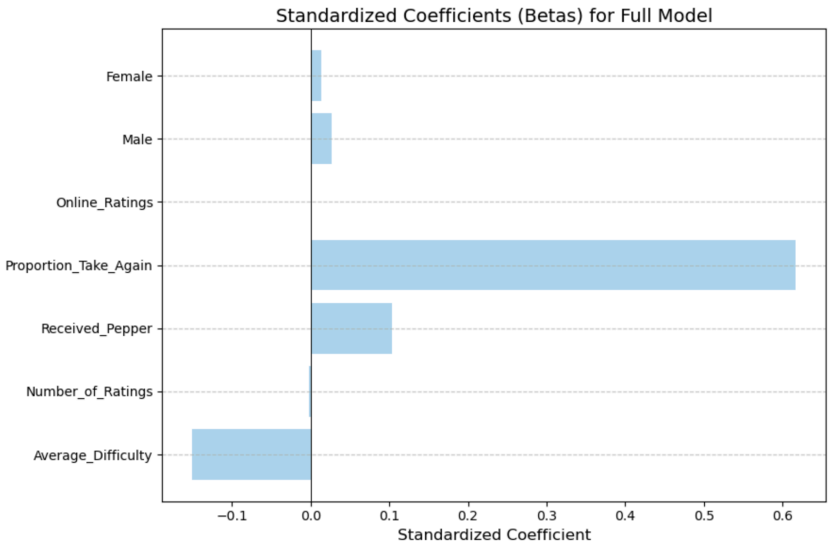
8) **Build a regression model predicting average rating from all available factors. Make sure to include the R2 and RMSE of this model. Comment on how this model compares to the "difficulty only" model and on individual betas. Hint: Make sure to address collinearity concerns.**

I built an ordinary least squares (OLS) regression model predicting average rating from all available numerical factors, ensuring missing data was removed. I also standardized all predictors before fitting the regression model to address multicollinearity and reviewed p-values and coefficients for redundancy, which were very minimal. The model achieved an R2 of 0.810, indicating that 81.0% of the variability in ratings is explained by the predictors, significantly outperforming the "difficulty-only" model (R2 ≈ 0.408). The RMSE was 0.368, reflecting better predictive accuracy than the simpler model (RMSE = 0.714). The regression coefficients revealed that Proportion_Take_Again (β ≈ 0.616) was the most impactful predictor, Received_Pepper ($\beta \approx$ 0.103) was also impactful but less, while Average_Difficulty (β ≈ –0.150) still showed a negative effect. Overall, the full model substantially outperforms the difficulty-only and all factors significantly improved predictions.

The plot comparing predicted and actual ratings shows tight clustering along the prediction line, confirming the model's accuracy. The bar chart also compares the different standardized coefficients (betas) with each other.

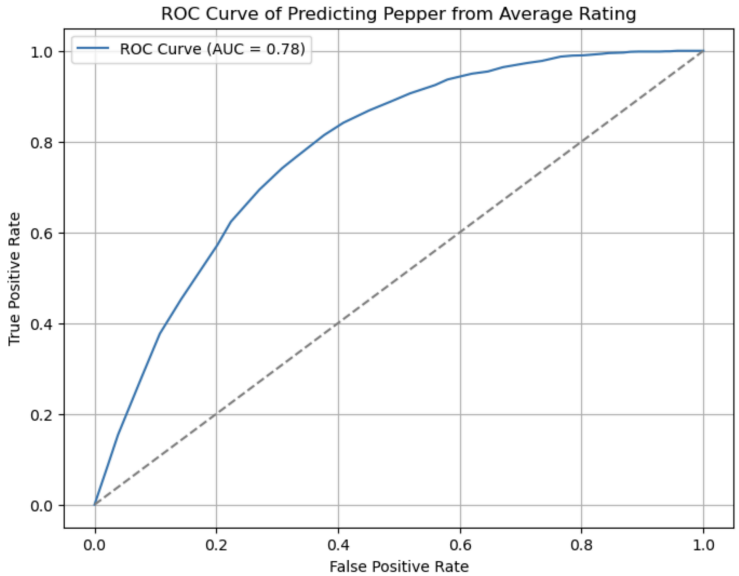
```
Coefficients: const    3.936706
x1       -0.150336
x2       -0.002003
x3        0.103063
x4        0.616491
x5       -0.000353
x6        0.026007
x7        0.013476
```





9) Build a classification model that predicts whether a professor receives a "pepper" from average rating only. Make sure to include quality metrics such as AU(RO)C and also address class imbalances.

I built a logistic regression model using Average_Rating as the sole predictor. The dataset was split into training (80%) and testing (20%) sets using stratification to address the class imbalance, as only 43.8% of professors received a pepper. After training, the model achieved an AUROC of 0.782, indicating good discriminatory power. The confusion matrix revealed 1,600 true negatives, 1,337 true positives, 713 false positives, and 466 false negatives, with an overall accuracy of 71.4%. The precision for non-pepper is 77% while for pepper is 65%. The ROC curve further demonstrated the model's ability to distinguish between classes. While the model performs well overall, the class imbalance may still influence recall (69% for non-pepper, 74% for pepper).



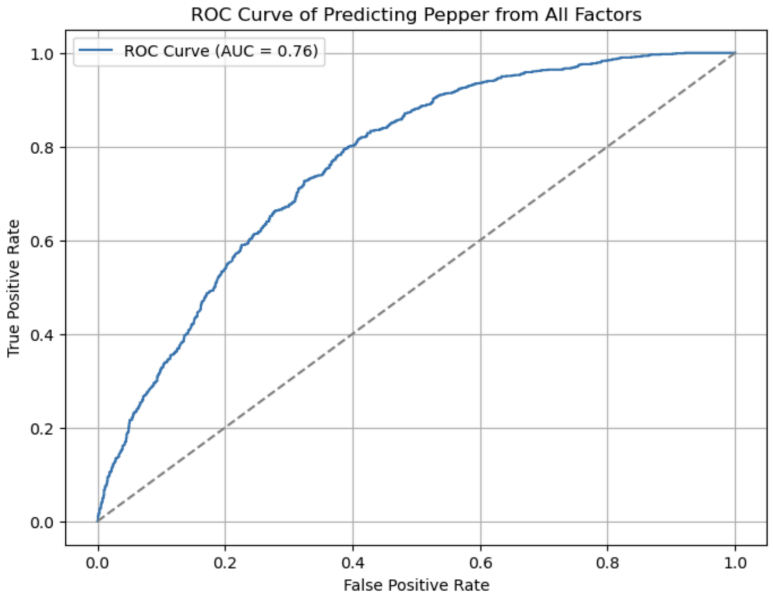

```
AUROC: 0.781737767601147
Professors Receiving Pepper: 0.43815910968557126
Overall Accuracy: 0.7135568513119533
Confusion Matrix:
[[1600  713]
 [ 466 1337]]
Classification Report:
              precision    recall  f1-score   support

           0       0.77      0.69      0.73      2313
           1       0.65      0.74      0.69      1803

    accuracy                           0.71      4116
   macro avg       0.71      0.72      0.71      4116
weighted avg       0.72      0.71      0.71      4116
```

10) Build a classification model that predicts whether a professor receives a "pepper" from all available factors. Comment on how this model compares to the "average rating only" model. Make sure to include quality metrics such as AU(RO)C and also address class imbalances.

I also built a logistic regression model but with six predictors: Average_Difficulty, Number_of_Ratings, Proportion_Take_Again, Online_Ratings, Male, and Female. I also applied a stratified train-test split to address class imbalances (like I also did for Question #9), ensuring that the proportion of professors receiving a "pepper" (positive class) was preserved in both training and testing sets, avoiding biases in the model's performance. The model achieved an AUROC of 0.765 and an accuracy of 70.0%, slightly below the single-factor model (AUROC ≈ 0.782, accuracy = 71.4%). The confusion matrix shows 859 true negatives, 808 true positives, 412 false positives, and 309 false negatives. This indicates similar performance but no significant improvement over the single-factor model. The ROC curve also highlights the model's moderate discriminatory power. Overall, adding more predictors increased model complexity but did not provide substantial gains in classification performance, suggesting potential overfitting.



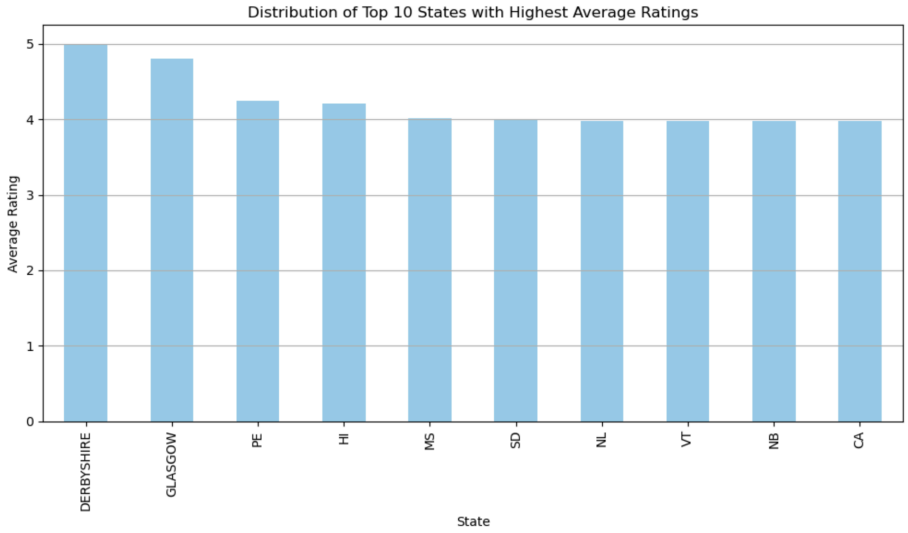

```
AUROC (All Factors): 0.7645802972021694
Overall Accuracy: 0.698073701842546
Confusion Matrix:
[[859 412]
 [309 808]]
Classification Report:
              precision    recall  f1-score   support

           0       0.74      0.68      0.70      1271
           1       0.66      0.72      0.69      1117

    accuracy                           0.70      2388
   macro avg       0.70      0.70      0.70      2388
weighted avg       0.70      0.70      0.70      2388
```

Extra credit: Tell us something interesting about this dataset that is not trivial and not already part of an answer (implied or explicitly) to these enumerated questions [Suggestion: Do something with the qualitative data, e.g. major, university or state by linking the two data files]

I decided to analyze how average professor ratings vary by state. I grouped the data by State and computed the mean Average_Rating for each state. The results revealed notable variation, with Derbyshire having the highest average rating of 5.0 and Edinburgh the lowest at 2.8, which both are in the UK and not the US. On average, states had a mean rating of 3.81, with a standard deviation of 0.29, indicating moderate consistency in ratings across states. The bar chart shows the top 10 states with the highest average ratings, reflecting regional differences in student evaluations. Overall, these findings suggest that cultural, institutional, and/or even demographic factors may influence professor ratings, having the potential for further exploration.




```
count    65.000000
mean      3.814908
std       0.294110
min       2.800000
25%       3.726030
50%       3.816667
75%       3.906557
max       5.000000
Name: Average_Rating, dtype: float64
Highest-rated state: DERBYSHIRE (5.0)
Lowest-rated state: EDINBURGH (2.8)
```