Navaneeth Unnikrishnan
Principles of Data Science II
Professor Wallisch
24 April 2025

Project Report - Assessing Professor Effectiveness

Preprocessing Summary

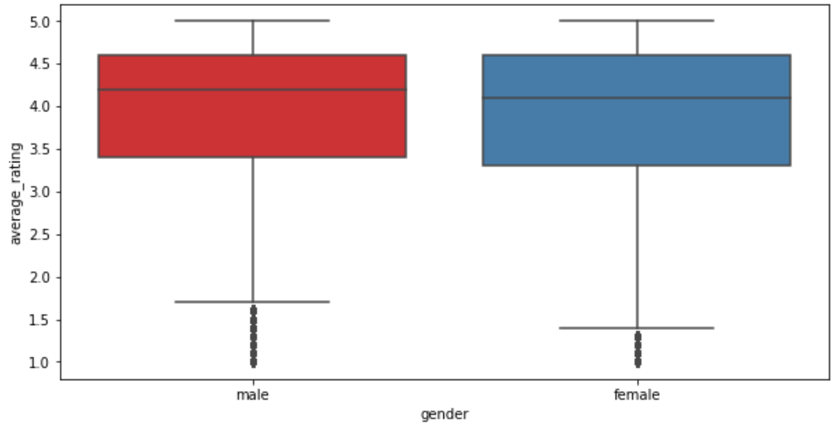
To prepare the data for analysis, I cleaned my `rmpCapstoneNum.csv` file containing qualitative data on professor evaluations. I first removed all rows with missing values in important columns, including `average_rating`, `average_difficulty`, and `num_ratings`. As per instructions, to ensure reliability in evaluating teaching quality, I filtered the dataset to include professors with only five or more ratings (to remove outliers). I also created a `gender` column, combining the data from the original dataset into a single categorical variable, having the options `male`, `female`, and `not known`.

1. Is there evidence of a pro-male gender bias in the data?

To analyze gender bias skewing towards males in the data, I cleaned the dataset to include only professors whose gender was identified as either "male" or "female", removing any unknowns. Then, I compared the average ratings of male and female professors using Welch's t-test, assuming that both groups follow a normal distribution but *don't have equal variances*. These were the following numbers I got:

- **Sample size (male): 10791**
- **Sample size (female): 8407**
- **Mean rating of male professors: 3.91**
- **Mean rating of female professors: 3.86**
- **t-statistic: 4.19**
- **p-value: 2.85 x 10^(-5)**

Since the p-value is *below the alpha threshold of 0.05,* we reject the null hypothesis. This gives us **statistically significant evidence of a pro-male bias in student evaluations of teaching,** as male teachers receive slightly higher ratings than female teachers on average. Below is the boxplot I made, showing the distribution of average ratings by gender.

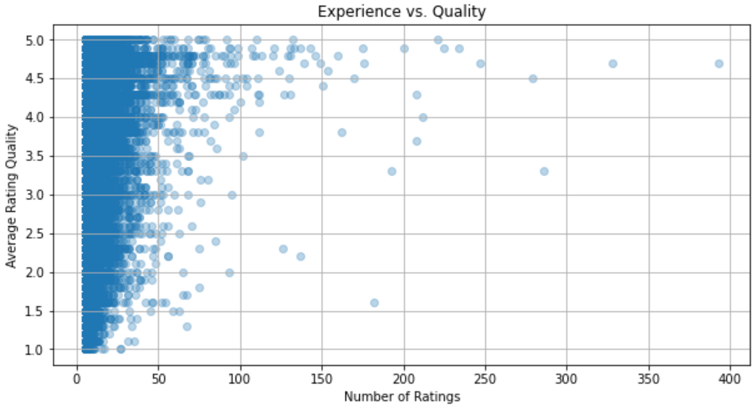


2. Does teaching experience impact professor ratings?

To assess whether teaching experience affects professors' ratings, I looked at the `num_ratings` a professor received to represent experience and the `average_rating` a professor received to represent quality. Here, I conducted a Pearson correlation test as both variables, mentioned before, are *continuous* and *numeric*. Additionally, both variables are measured within a range, as `num_ratings` counts how many students submitted reviews, and `average_rating` is the mean of rating on a one-to-five scale. Pearson's method directly answers whether more experience is linearly associated with higher quality. Here are the numbers I got:

- **Pearson correlation coefficient: 0.0578**
- **p-value: 3.00 x 10^(-20).**

This correlation is **statistically significant at the 0.05 level**, but the **correlation coefficient is close to zero**. Hence, while professors with more ratings/experience tend to have higher ratings, **the correlation between both variables is weak.** Below is the scatter plot I made showing the relationship between experience and teaching quality.

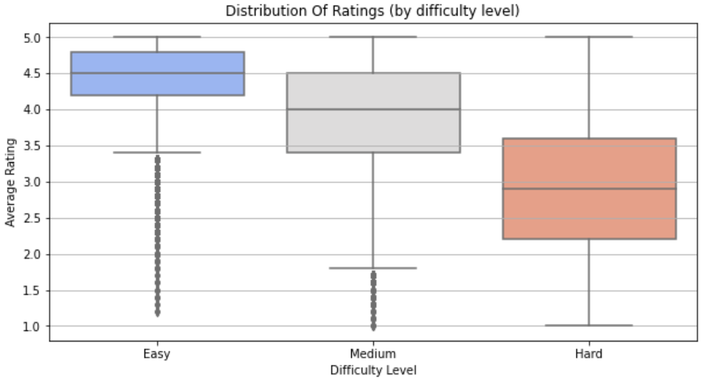


3. **What is the relationship between average rating and average difficulty?**

To answer this question, I utilized `average_difficulty` as the predictor and `average_rating` as the outcome. I conducted a Pearson correlation test to see whether difficulty and rating have a linear relationship. These were the numbers that I got:

- **Pearson correlation coefficient: -0.619**
- **p-value: 0.0**

This shows a **strong, statistically significant negative relationship between both variables**. As **courses become more complex**, students tend to give professors lower ratings. The result is important at the significance level of 0.05, and **since the correlation is strong**, perceived difficulty is meaningfully related to perceived teaching quality. For my visualization, I created a box plot that categorized **difficulty scores into three categories: easy (1 - 2.5), medium (2.5 - 3.5), and hard (3.5 - 5.0).**

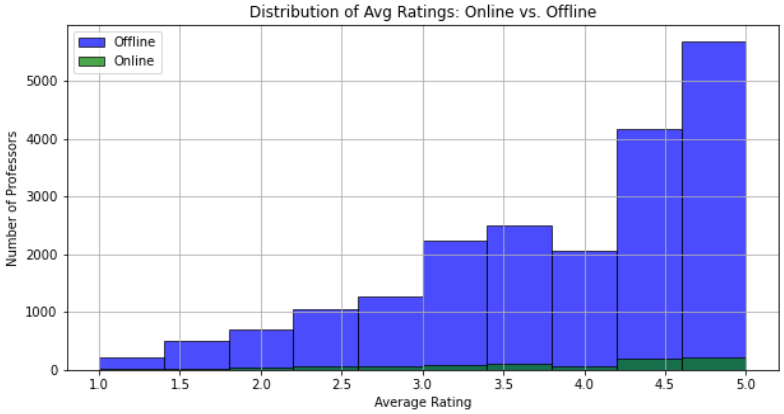


4. **Do professors who teach a lot of classes online receive higher or lower ratings than those who don't?**

Similar to how I cleaned the data initially, I separated professor ratings into two categories: 1) professors who received five or more online ratings & 2) professors who received no online ratings. This separates those who mainly teach online from those who likely don't. I used a **Welch's t-test** for a comparison of the average ratings between these groups, and here are the values I got:

- **Mean rating of online professors: 3.75**
- **Mean rating of offline professors: 3.85**
- **t-statistic: -2.73**
- **p-value: 0.0064**

Although professors who don't teach online tend to get slightly higher ratings, the difference **wasn't statistically significant at the significance level of 0.05. Hence, we fail to reject the null hypothesis**. This indicates no substantial evidence that professors teaching more online classes sufficiently affect their average ratings. Below is a **histogram comparing these ratings.** While most professors generally receive higher ratings, **offline professors are more frequently rated 4.0 or higher than online professors.**

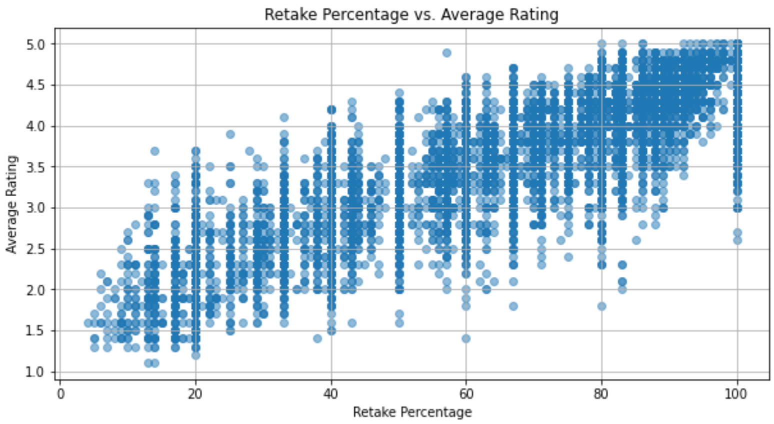


5. **What is the relationship between the average rating and the proportion of people who would take the class the professor teaches again?**

I used `retake` and `average_rating` to analyze this relationship. First, I limited the analysis to only those with valid `retake` values, as many professors had NaN values in their columns. Then, I conducted a Pearson correlation test to see the linear relationship between `retake` and `average_rating`. These were the numbers that I got:

- **Pearson correlation coefficient: 0.8804**
- **p-value: < 0.001**

This shows a **robust, statistically significant positive relationship between both variables.** Professors more likely to have their classes retaken tend to **receive much *higher average ratings***. The **strength of the relationship between `retake` and `average_rating` suggests that retake likelihood strongly predicts student satisfaction and perceived teaching quality**. Below is a scatter plot I designed to show the correlation between both variables.

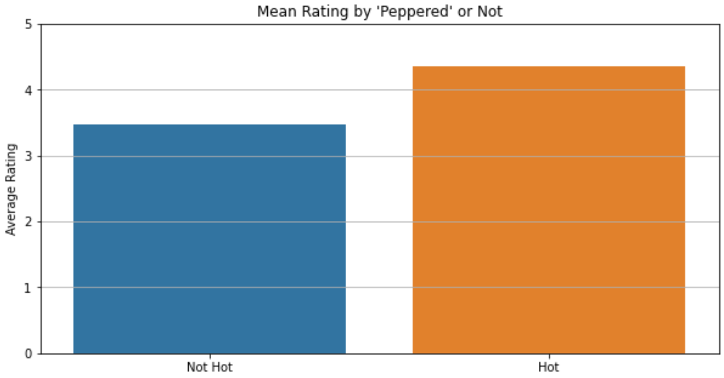


6. **Do professors who are "hot" receive higher ratings than those who are not?**

To see whether professors marked by students as 'hot' (as shown by the `pepper` column, one being hot) receive higher ratings, I **compared the average ratings of professors with and without the one in their rows.** I did a Welch's t-test here, as `pepper` is a binary variable and we are comparing two independent groups here: 1) professors with the `pepper` and 2) professors without. These two groups **likely have unequal variances and sample sizes**, making Welch's t-test more appropriate than a regular t-test.

- **Mean rating of `hot` professors: 4.36**
- **Mean rating of `un-peppered` professors: 3.47**
- **t-statistic: 90.32**
- **p-value: 0.0**

These values show a **large, statistically significant difference between the two groups in average ratings.** On average, professors **marked as `hot` received higher ratings** than those not. Given the **highly minuscule p-value and the magnitude of this difference**, there is **strong evidence that being rated as `hot` *is associated with higher professor ratings.*** Below is a barplot that I made, showing the average rating of professors with/without the `hot` rating.

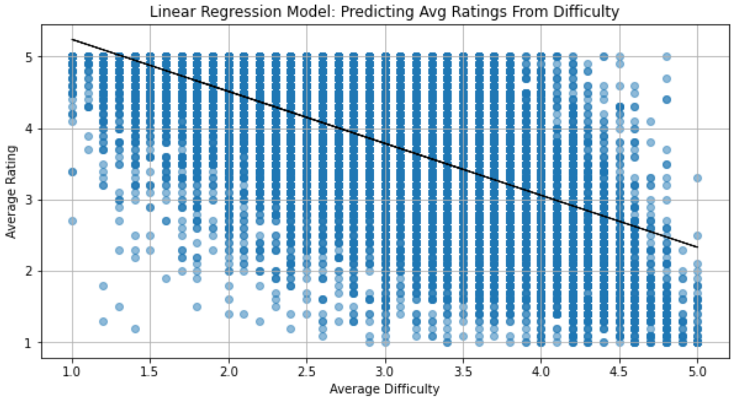


7. **Build a regression model predicting average rating from difficulty (only).**

I trained this model using scikit-learn and evaluated it using **R^2 and RMSE**. The regression equation that I got was: average_rating = 5.97 + -0.73 * average_difficulty. As far as model performance goes:

- **R^2: 0.383**, as the model approximately explains 38.3% of the variance in the professor's ratings.
- **RMSE: 0.744,** on average, predictions deviate from actual ratings by about 0.74 points.

This model indicates a moderate, negative correlation between both variables. **As difficulty increases, professor ratings tend to decrease in response.** I constructed a scatter plot with a regression line to visualize the general relationship between variables and the trend.



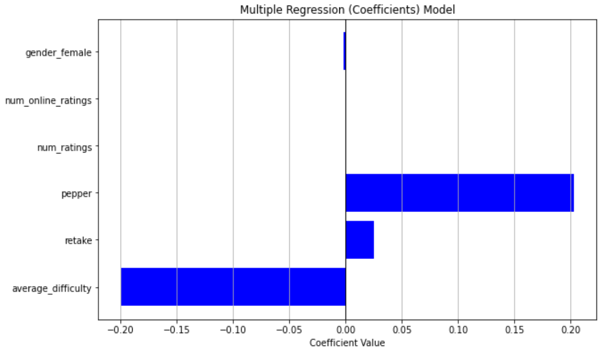
8. **Build a regression model predicting average rating from all available factors.**

Building upon my model in Q7, I built a multiple regression model using all available predictors: **`average_difficulty`, `retake`, `pepper`, `num_ratings`, `num_online_ratings`, and `gender_female`.** To avoid collinearity, I converted the gender variable to a dummy column and dropped `gender_male` (since it was used in a previous analysis). In terms of model performance:

- **R^2: 0.809**, which means that the model **explains 80.9% of the variance in average_rating**, which is a substantial improvement in comparison to the model in Q7 (R^2: 0.383)
- **RMSE: 0.369**, indicating that the average error in predicting ratings is significantly lower than in **Q7 (RMSE: 0.744)**, showing that t**his model has more substantial predictive power.**

In terms of interpreting coefficients (**`num_ratings` & `num_online_ratings both had minimal and negative values, indicating that both have very little effect/influence)** :

- **`average_difficulty` (-.200):** As difficulty increases, ratings tend to decrease for professors (even while controlling for other variables).
- **`retake` (0.025):** Professors for whose students said yes to retaking their class were rated slightly higher.
- **`pepper` (0.203):** Professors marked as hot receive significantly higher average ratings.
- **`gender_female` (-0.0017):** Female professors receive slightly lower ratings than male professors. I created a **horizontal bar plot of the regression coefficients, showing the direction and importance of each predictor variable.**



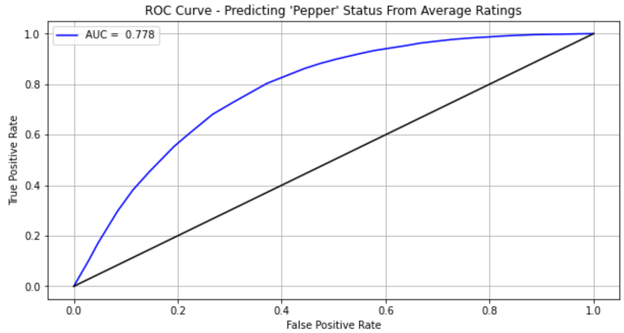
9. **Build a classification model that only predicts whether a professor receives a "pepper" from the average rating.**

First, I built a binary classification model that predicts whether a professor receives a `pepper` solely based on their average rating. The model was trained on **`LogisticRegression`** with the **`average_rating` variable as the sole predictor.** Since most professors have not received a `pepper`, I incorporated the parameter of **`class_weight=balanced`** to reduce bias towards the ruling majority and improve fairness across non-pepper and pepper cases. In terms of model performance:

- **Accuracy: 70%**
- **AUC: 0.778**
- **Confusion Matrix: [[9241 5469]**
  [2097 8561]]

The results indicate that the **average rating strongly predicts whether a professor has `pepper` status. For my visualization, I plotted a ROC curve to evaluate the model's ability to distinguish** between professors with and without the `pepper` across different thresholds. Since the AUC value I obtained was **0.778,** my model has a 77.8% chance of assigning a higher predicted pepper probability to the `peppered` professor.

Below is the visualization that I built, an **ROC curve that reflects the model's overall classification performance.**

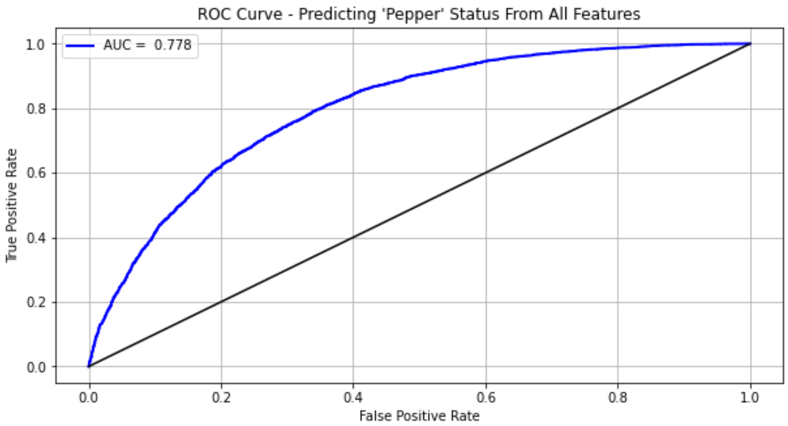


10. **Build a classification model that predicts whether a professor receives a "pepper" from all available factors.**

To build on the model built in Q9, I built a logistic regression classification model to predict whether a professor receives a "pepper" using all available predictor variables, including: **["average_rating", "average_difficulty", "retake", "num_ratings", "num_online_ratings", "gender_female"].** Since the data was imbalanced, I trained the model using `class_weight=balanced` to remove bias and treat both classes more equally. In terms of model performance:

- **Accuracy: 72%**
- **AUC-ROC: 0.798**
- **Confusion Matrix: [[4270 2224]**
  ` **[1179 4487]]**

Regarding performance compared to Q9, the Q10 model has improved in performance, proving that additional factors like `average_difficulty` and `retake` provide substantial predictive power. I visualized an ROC curve to evaluate the model's classification performance at different thresholds—a value of 0.798 suggests that the model is highly effective at distinguishing between professors with and without having received a `pepper`.


ROC Curve - Predicting 'Pepper' Status From All Features
AUC = 0.778
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0