Brianna Capuano

12/1/2024

PODS Pascal

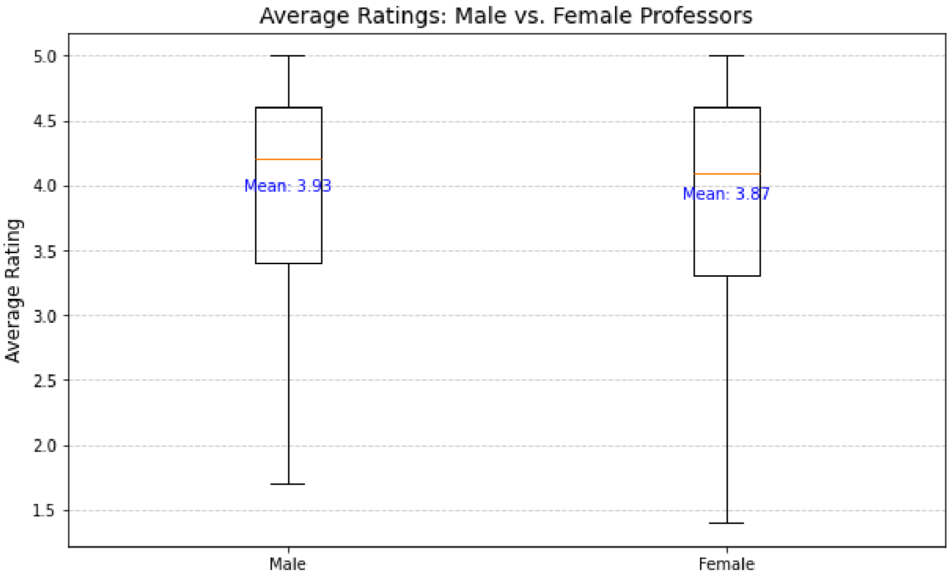
Capstone Project

Introduction: Preprocessing

In order to clean the data, I decided to set a threshold of the number of ratings to be above 5 (k>5). This way, professors with few ratings are excluded, but around 70% of the data remains. I chose 5 for k because I thought it was a good choice to ensure I didn't overfilter and that the averages would still be meaningful. For each question, I dropped missing values for the relevant columns in question. I also seeded the RNG as my N-number and also seeded numpy for reproducibility. I also set the random state to my N-number (12939850) when I did train and test split and for my logistic regression model.

Questions:

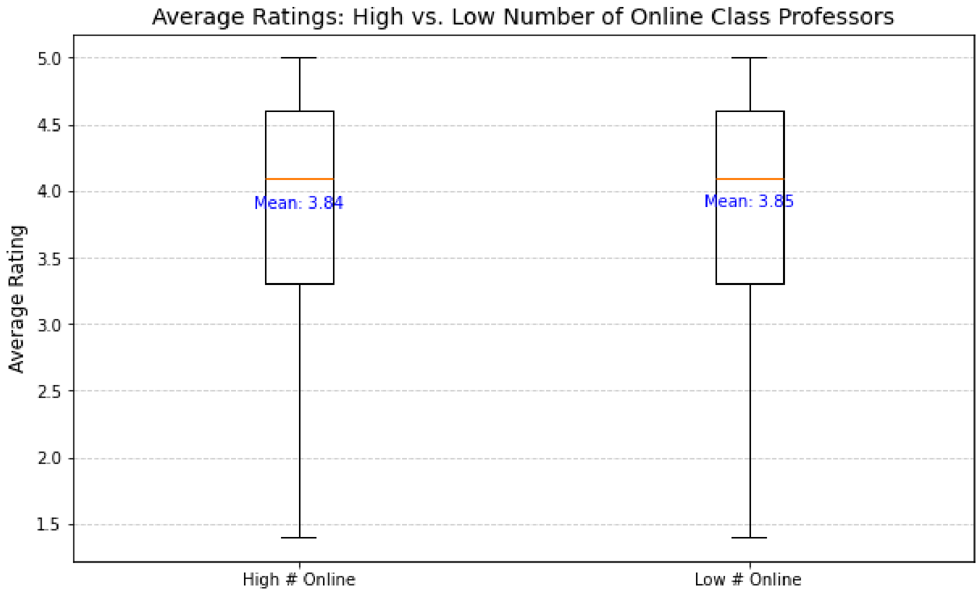
1. After doing a Mann-Whitney significance test, I found that there is evidence of a pro-male gender bias in the data set because I got a p-value of **~.0013** which is less than alpha at 0.005. This is the significance test visualization:



2. I got a p-value of **8.91 x 10^-8** after doing a spearman correlation and a p-value this low suggests a strong correlation between number of ratings and average rating.

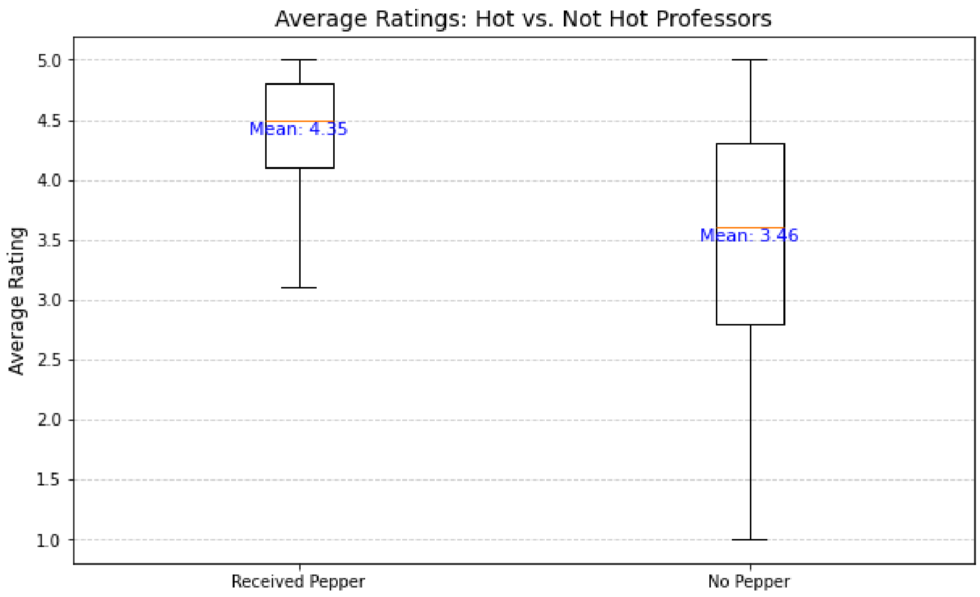
3. There is a strong negative relationship between average rating and average difficulty because after doing a spearman correlation test I got ~ **-0.612** for the correlation coefficient.
4. I performed a Mann-Whitney U Test after splitting the data into two groups, 1 with a high amount of online ratings and 1 with a low amount (split by median num of online ratings) and I compared the average ratings of both groups. The average ratings for both groups were relatively the same (around 3.8) and the p-value from the Mann Whitney test was ~ **0.286** which is greater than 0.005 alpha meaning there is not a significant difference in ratings of online vs not online. This means that whether a class is in-person or online, it doesn't seem to affect students' ratings of the professors.

   My significance test:



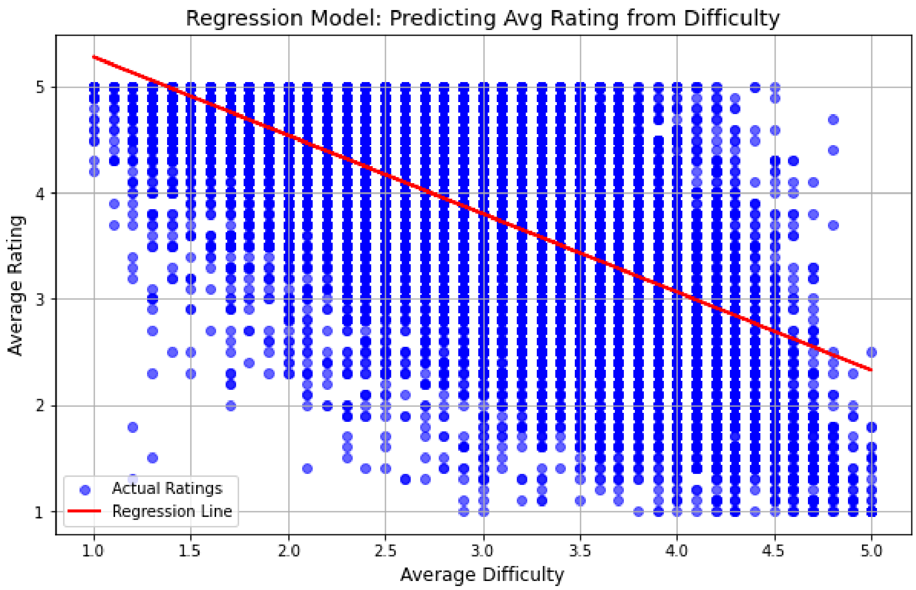
5. The relationship between the average rating and the proportion of people who would take the class the professor teaches again is a strong and positive correlation of **0.85**. This means the higher the rating of the class the more students want to take the class again.
6. To see whether professors who are "hot" receive higher ratings than those who are not, I did another Mann Whitney U test, by splitting the data into two groups, received a pepper versus didn't receive a pepper, and I got a p-value of ~ **0.0** which is <0.005 meaning there is enough evidence to reject the null, meaning there it is likely hot professors receive higher ratings.

   My significance test:



7. My predictor X was the average difficulty data and my target variable y was the average ratings. Then I ran a simple linear regression and predicted y using X. The regression equation outputs a coefficient of ~ **-0.739** which suggests that for each 1-unit increase in difficulty, the average rating decreases by approximately 0.739 units. R-squared is **~0.396** which means that about 39.6% of the variance in average ratings is explained by difficulty. My root mean squared error is **0.723**. This means that on average, the predictions deviate from the observed ratings by approximately 0.723 units. This model suggests a negative relationship between difficulty and average rating.

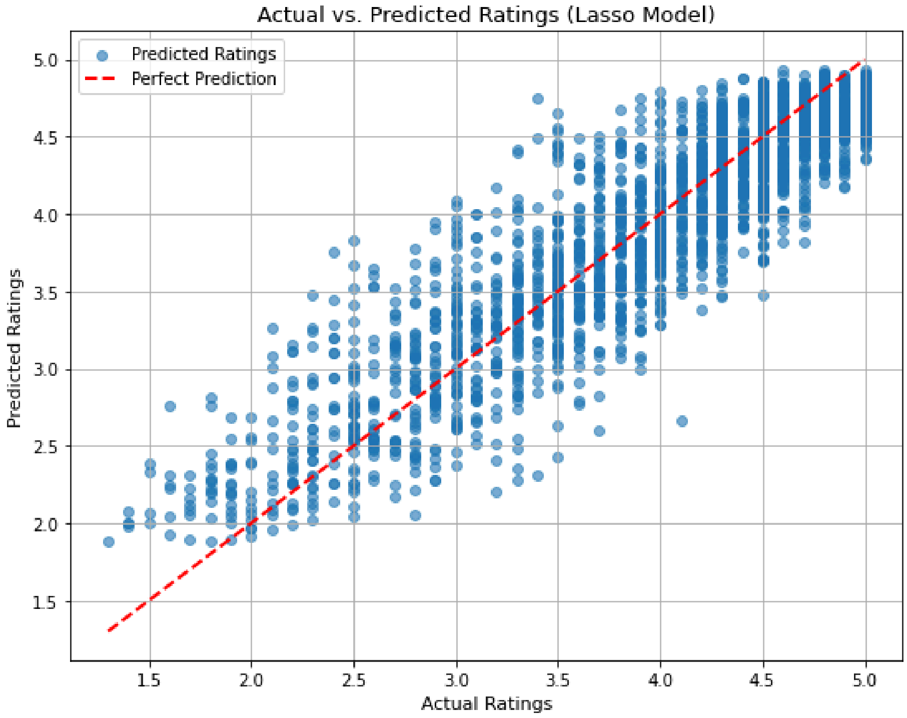
   My simple linear regression model:



You can see the general negative trend from the regression line, however, you can also clearly see a lot of the variability that is not captured by the model.

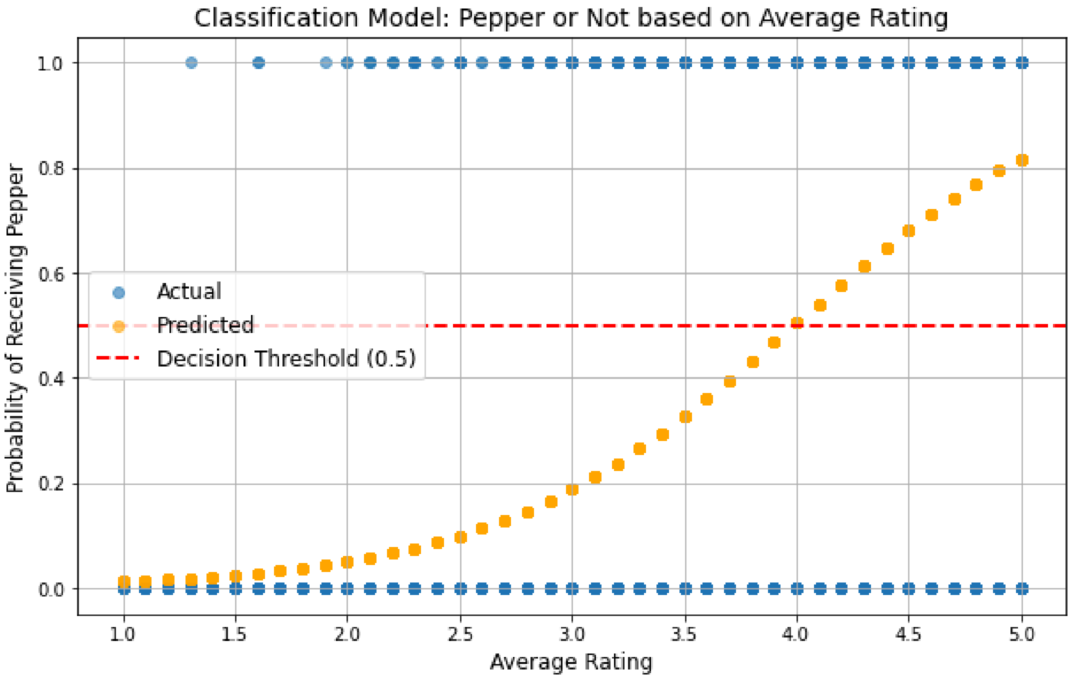
8. To build a regression model predicting average rating from all available factors, I first started with cleaning the data by dropping missing value rows. Then I defined X the predictor with all the factors and defined the target variable as the average ratings. To address collinearity, I used Lasso regression because it handles multicollinearity well (because the independent predictor variables aren't so independent) and regularizes the model to avoid any overfitting. It also removes redundant predictors by shrinking their coefficients to 0 which simplifies my model. R-squared was ~ **0.815** which means that approximately 81.5% of the variance in average ratings is explained by the predictors. The RMSE is ~ **0.353** which means that on average, the model's predictions deviate from the observed ratings by about 0.353 units. I also performed cross validation to find the best alpha which turned out to be 0.01. Lasso set the coefficients of male, female, and online class ratings to 0 indicating that they don't have a meaningful contribution to the average rating predictions. We can see that the significant predictors are average difficulty, received pepper, and proportion would take again. The average difficulty feature has -0.181 coefficient which means that average difficulty has the largest negative impact on average ratings. We also see that receiving a pepper feature has a coefficient of +0.154 which means being "hot" has a positive effect on ratings. The proportion-that-would-take-again feature had +0.026 coefficient which means this feature

has a smaller positive impact. When comparing this model to the "difficulty only" model, the Lasso model explains more than twice as much variance in avg_rating as the difficulty-only model and it has much lower prediction error when I compare R-squared and RMSE. Also when comparing individual betas, in the Lasso model, the effect of difficulty is much smaller than in the "difficulty only" model after including the other predictors. This shows that the importance of difficulty alone was overestimated in the simpler model. Here is a visualization of the Lasso model predictions:

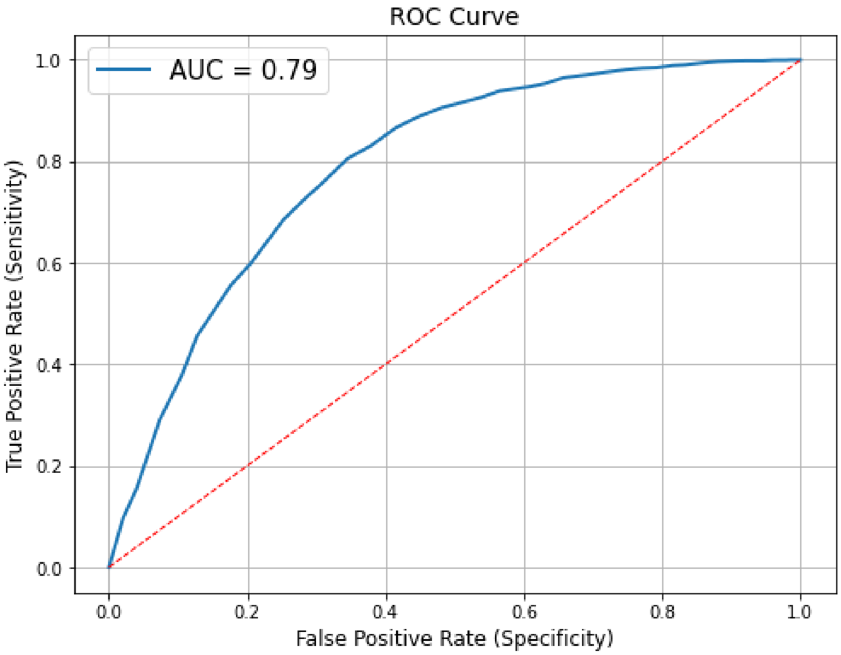


9. To build a classification model that predicts whether a professor receives a "pepper" from average rating only, I used a logistic regression model with class weight adjustment to address class imbalances because the class imbalance was not extreme. I also plotted the AU(RO)C curve to assess the quality of the classification model. I got a high AUC of **0.79** which shows that the model is effective at predicting the classification or receiving a pepper or not. The classification report showed me that the precision for "pepper" is slightly lower and this is probability due to the class imbalance (precision: no pepper = .81, pepper = .65). The pepper class is larger (class distribution: pepper = .56, no pepper = .44) but it was addressed by the class weight adjustment.

   My classification model:



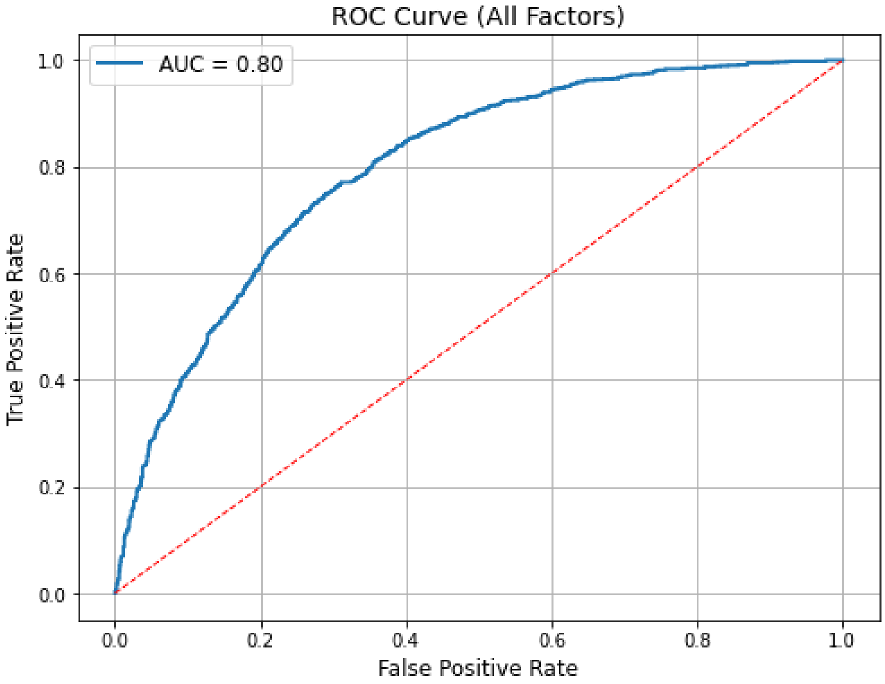
My AU(RO)C curve for my Logistic regression classification model:



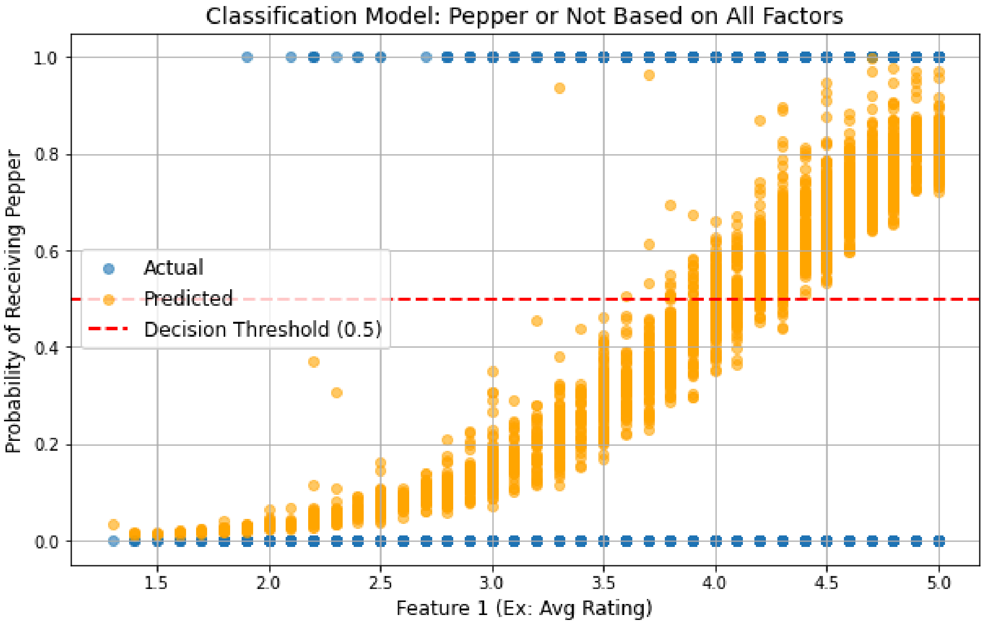
10. To build a classification model that predicts whether a professor receives a "pepper" from all available factors I split the data into train and test sets and did a logistic regression with class balancing. My X features were made up of all available factors and my y predictor target variable is received pepper. My AUC score is **0.80**. The predicted values

(orange dots) are distributed more widely than in the "average rating only" model, showing that this model uses additional features for better predictions.

My ROC Curve for all factors:



My classification model:


Classification Model: Pepper or Not Based on All Factors
Actual
Predicted
Decision Threshold (0.5)
Probability of Receiving Pepper
Feature 1 (Ex: Avg Rating)