Introduction to Data Science: Capstone Project[*]

Tien-Leng Wu[†], Neil Chen[‡]

Abstract

It is our honour to have the opportunity to assess professor effectiveness as part of our capstone project. This report begins by introducing the methods we used to handle missing data and discusses specific "considerations" regarding certain columns in the dataset. Subsequently, we address each question individually using our proposed methods, explaining the rationale behind and implementation of these methods for each question. This capstone project was completed by Tien-Leng Wu and Neil Chen. Both contributed to data preprocessing and report composition; Tien-Leng Wu focused on statistical inference questions, while Neil Chen was responsible for machine learning questions.

0. Data Preprocessing: Missing Data and Considerations

Before answering the questions, we merge the rmpCapstoneQual.csv, rmpCapstoneNum.csv, and rmpCapstoneTags.csv into one dataset named rmp.csv and then we move on to data preprocessing.

0.0 Missing Data

We observed that there are no missing values in the tag columns or in the male and female gender columns. However, there are 19,889 missing values in the qualitative data and in most of the numerical data columns, except for "The proportion of students that said they would take the class again", which contains 77,733 missing values. For all analyses, we decided to exclude rows with missing data. We believe it is not reasonable to impute these values using methods such as mean imputation, random imputation, or machine learning algorithms like KNN. Excluding missing data, in our view, best preserves the original distribution of the dataset. Lastly, due to the large number of missing values in the column "The proportion of students that said they would take the class again," we have excluded this column from all analyses.

0.1 Considerations: Gender

We observed that the dataset contains separate columns for male and female genders. Typically, we assume that gender is binary; in other words, if the value in the male column is 1, the value in the female column should be 0, and vice versa. However, we found numerous instances where both the male and female columns are either 1 or 0. We refer to this issue as the "Inconsistent Gender Problem" and propose the following two methods to address it in our analyses.

0.1.1 Consistent Gender

One approach to addressing the inconsistent gender problem is to discard rows with inconsistent gender values. This means retaining only rows where either the male column is 1 and the female column is 0, or the male column is 0 and the female column is 1 for our analyses. By adopting this method, we operate under the assumption that it is not possible to determine a professor's gender when the gender data is inconsistent.

0.1.2 Random Imputation

Another approach to addressing this issue is to ensure all rows have "consistent genders." To achieve this, we randomly impute gender values for rows with inconsistencies. Our algorithm first clears the values in both the male and female columns for rows with inconsistent gender data. Then, it imputes the male column with either 0 or 1 with equal probability (0.5), and assigns the female column a value of 1 - (male gender). This method preserves the gender distribution of rows with "consistent genders" while minimizing data loss, as we do not discard these rows but instead impute their values.

[*] Group: CAP62

[†] N-number: N19716300

[‡] N-number: N14204806

0.2 Considerations: Average Ratings & Average Difficulty

The average rating is more meaningful if it is based on more ratings. However, many average ratings in the dataset are derived from only a few ratings, which may reduce their reliability. To address this issue, we propose the following three methods. Note that it is analogous when handling average difficulty.

0.2.1 Threshold Setting

This method is straightforward: we set a threshold for the number of ratings. If the number of ratings for a row is less than or equal to the threshold, we discard the row; otherwise, we retain it. We set the threshold at 5, as we believe an average rating is sufficiently representative if it is based on at least 6 ratings.

0.2.2 Bayesian Adjustment

This method adjusts the average ratings of all rows with the following formula:

$$r_{adj,i} = \frac{n_i \cdot r_i + C \cdot \frac{1}{N} \sum_{i=1}^{N} r_i}{n_i + C} \qquad \forall i = 1, ..., N.$$

where n_i is the number of ratings for row i, r_i is the average rating for row i and N is total the number of rows. C is a hyperparameter representing the strength of prior mean, which is given by $\frac{1}{N} \sum_{i=1}^{N} r_i$. Note that we set C to be 5, as we believe this is a reasonable and balanced choice.

0.2.3 Threshold Setting + Bayesian Adjustment

This method combines the two approaches described above. If the number of ratings for a row is less than or equal to the threshold, we apply the Bayesian adjustment instead of discarding the data. Otherwise, we retain the data.

0.3 Considerations: Tags

Note that a student can award up to 3 to a professor in a given rating. This means that we cannot just use the raw number of tags for anything meaningful, as a professor with more ratings will have received more tags, everything else being equal. Therefore, it is essential to normalize the tags in some way. We propose the following three methods to address this issue.

0.3.1 Normalization with Number of Ratings

This method simply divides the raw number of tags by the number of ratings, which can be expressed as follows:

$$t_{norm,i,j} = \frac{t_{i,j}}{n_i} \qquad \forall i = 1, ..., N, j = 1, ..., K.$$

where $t_{i,j}$ is the raw number of tags for row i and column j, and n_i is the number of ratings for row i. N is the total number of rows and K is the number of tag columns.

0.3.2 Bayesian Adjustment

Similar to how we adjust the average ratings, we apply the following formula to adjust the number of tags:

$$t_{adj,i,j} = \frac{t_{i,j} + C \cdot \frac{\frac{1}{N} \sum_{i=1}^{N} t_{i,j}}{\frac{1}{N} \sum_{i=1}^{N} n_i}}{n_i + C} \qquad \forall i = 1, ..., N, j = 1, ..., K.$$

where n_i is the number of ratings for row i, $t_{i,j}$ is the raw number of tags for row i and column j, N is the total number of rows and K is the number of tag columns. C is a hyperparameter and we set it to be 1, which we believe is a reasonable choice.

0.3.3 Percentage Normalization

This method transforms the raw number of tags into percentages by dividing the raw number of tags by the total number of tags for each row. This can be expressed as follows:

$$t_{per,i,j} = \frac{t_{i,j}}{\sum_{j=1}^{K} t_{i,j}} \qquad \forall i = 1, ..., N, j = 1, ..., K.$$

where $t_{i,j}$ is the raw number of tags for row i and column j, N is the total number of rows and K is the number of tag columns. If a professor has no tags, applying this normalization method would result in $t_{per,i,j} = \frac{0}{0}$, which is undefined. In such cases, we preserve the raw numbers, which should all be zero.

Before ending this section, we hereby make the following declaration:

We set the seed for the random number generator using Tien-Leng Wu's N-number, 19716300, at the start of our code and each time we apply randomization.

1. Question 1

In this question, we aim to determine whether there is evidence of a pro-male gender bias.

Null hypothesis: There is no gender bias in average ratings.
Alternative hypothesis: There is gender bias in average ratings.

We propose three different tests for our analysis. Since we use two methods for handling the inconsistent gender problem and three methods for addressing average ratings, we will have $2 \times 3 = 6$ results for each test. Before running the tests, we plot the distribution of average ratings for both male and female professors across the 6 datasets. All distributions appear to be left-skewed. Due to space limitations, we present only one example figure on the last page (Figure (a)).

1.1 *t*-test

The following table shows the results of the t-test. The numbers in the table represent p-values.[*] We believe caution is needed when interpreting the results from the t-test as the data is left-skewed. Therefore, we propose additional methods, as outlined below. Note that we assume unequal variance in the average ratings between male and female.

	Threshold Setting	Bayesian Adjustment	Both
Consistent Gender	9.289396691873744e-05*	1.5408312112707627e-12*	8.696747852603423e-11*
Random Imputation	0.0005158851145964263*	1.6153117115957482e-10*	8.696747852603423e-11*

1.2 Mann-Whitney *U* test

The following table shows the results of the Mann-Whitney U test. We believe the results from the Mann-Whitney U test are more robust than those from the t-testt, as it does not require the assumption of normality.

	Threshold Setting	Bayesian Adjustment	Both
Consistent Gender	0.0007859717675877128*	4.628039829156059e-12*	7.604974340022987e-12*
Random Imputation	0.0022674953090889255*	2.8582140073117083e-11*	7.604974340022987e-12*

1.3 Permutation test

The following table shows the results of the permutation test. We set the number of iterations to 11,260, as we believe this is sufficient to generate robust results without running the tests for an excessive amount of time.

	Threshold Setting	Bayesian Adjustment	Both
Consistent Gender	8.880994671403197e-05*	0.0*	0.0*
Random Imputation	0.0004440497335701599*	0.0*	0.0*

We reject the null hypothesis; all results show statistical significance.

2. Question 2

In this question, we aim to determine whether there is a gender difference in the spread (variance/dispersion) of the ratings distribution.

[*] If a p-value has a star in the upper-right corner, it indicates statistical significance. This notation applies to all tables in this report.

Null hypothesis: There is no gender bias in spread in average ratings.
Alternative hypothesis: There is gender bias in spread in average ratings.

We will use two methods for our analysis, so we will obtain 4 results for each test.

2.1 Fligner-Killeen test

The following table shows the results of the Fligner-Killeen test.

	Threshold Setting	Bayesian Adjustment	Both
Consistent Gender	2.392205517461604e-07*	0.001225831846932436*	0.005607079489954568
Random Imputation	2.495563773410372e-08*	1.5840643928736017e-05*	0.005607079489954568

2.2 Brown–Forsythe test

The following table shows the results of the Brown–Forsythe test. We believe that both the Fligner-Killeen and Brown-Forsythe tests provide robust results, as they do not require the data to be normally distributed.

	Threshold Setting	Bayesian Adjustment	Both
Consistent Gender	8.798664061895703e-05*	0.0007525319733344336*	0.20410038683129852
Random Imputation	0.000465957880934212*	0.0032834711618745863*	0.20410038683129852

Most of the results show statistical significance.

3. Question 3

We construct 95% confidence intervals based on the results in Question 1 and 2 for this question. All confidence intervals are generated using bootstrapping, a simple yet powerful method. Note that we set the number of iterations for the bootstrap to be 11,260, as we believe this is sufficient to generate robust results without excessively prolonging the duration.

3.1 Confidence Interval for Gender Bias

Before moving on to the confidence intervals for gender bias, we want to clarify that although the permutation test and bootstrapping are quite similar, they have a key difference: the way they sample. The permutation test samples **without** replacement, while bootstrapping samples **with** replacement. In Python, we use numpy.random.shuffle for the permutation test and numpy.random.choice for bootstrapping.

3.1.1 *t*-test

The following table shows the 95% confidence intervals for the *t*-test.

	Threshold Setting	Bayesian Adjustment	Both
Consistent Gender	(1.9315800070599691, 5.854920847826368)	(5.119439312592283, 9.006680824054813)	(4.523681246818701, 8.406826822210116)
Random Imputation	(1.5521566724101168, 5.437109410175222)	(4.419444475405626, 8.348431742274423)	(4.523681246818701, 8.406826822210116)

3.1.2 Mann-Whitney *U* test

The following table shows the 95% confidence intervals for the Mann-Whitney *U* test.

	Threshold Setting	Bayesian Adjustment	Both
Consistent Gender	(27777833.1125, 28813282.1375)	(347046200.85, 353715431.4625)	(346884387.6625, 353577172.5625)
Random Imputation	(53091103.9625, 54741991.1375)	(622928976.0625, 636562608.05)	(346884387.6625, 353577172.5625)

3.1.3 Bootstrapping

The following table shows the 95% confidence intervals using bootstrapping. Unlike the permutation test, which samples without replacement and returns p-values, bootstrapping samples with replacement and constructs confidence intervals.

	Threshold Setting	Bayesian Adjustment	Both
Consistent Gender	(0.029088290666055273, 0.08823884771512618)	(0.019743314588215644, 0.03482805504082166)	(0.022658263549309186, 0.04209168412446838)
Random Imputation	(0.020236501791138152, 0.07101749318511262)	(0.015038811011852815, 0.028380094926704468)	(0.022658263549309186, 0.04209168412446838)

3.2 Confidence Interval for Gender Bias in Spread

We have shown the confidence intervals for gender bias. Now, we will move on to the confidence intervals for gender bias in spread.

3.2.1 Fligner-Killeen test

The following table shows the 95% confidence intervals for the Fligner-Killeen test.

	Threshold Setting	Bayesian Adjustment	Both
Consistent Gender	(5.8903193813421275, 50.22233710812007)	(1.600157349315613, 29.248001781988847)	(0.08511615092210546, 20.713733635096546)
Random Imputation	(3.2594280864825707, 54.53584899828578)	(1.6563481096100527, 39.72195414624868)	(0.08511615092210546, 20.713733635096546)

3.2.2 Brown–Forsythe test

The following table shows the 95% confidence intervals for the Brown–Forsythe test.

	Threshold Setting	Bayesian Adjustment	Both
Consistent Gender	(3.6625088289514847, 34.33658305834512)	(1.9346031924886093, 28.560011096017085)	(0.00492023394710357, 10.358334194571478)
Random Imputation	(2.269842846874529, 28.592511950546978)	(1.0108802701456114, 23.81856091740735)	(0.00492023394710357, 10.358334194571478)

4. Question 4

We aim to determine if there is a gender difference in the tags awarded by students.

Null hypothesis: There is no gender difference in tags.
Alternative hypothesis: There is gender difference in tags.

As in previous analyses, we propose three different tests, resulting in 6 outcomes for each test. Before running the tests, we plot the distributions of the 20 tags for both male and female professors across the 6 datasets. Most distributions appear to be right-skewed. Due to space limitations, we present only one example figure on the last page (Figure (b)). Additionally, we will not present p-values for all tags also due to space constraints; instead, we will report which tags exhibit statistical significance and highlight the three most gendered and least gendered tags.

4.1 *t*-test

Here, we identify the tags that exhibit statistical significance across the 6 results for the *t*-test. However, caution is needed when interpreting these results, as the data is right-skewed. More robust methods will be presented in subsequent sections. Note that we assume unequal variance for all tags between male and female professors.

- Consistent Gender ⊗ Number of Ratings: Only “Pop quizzes!” is **not** statistically significant.
  - Top 3 Most Gendered Tags: Hilarious, Amazing lectures, Caring
  - Top 3 Least Gendered Tags: Lots to read, Don’t skip class or you will not pass, Pop quizzes!
- Consistent Gender ⊗ Bayesian Adjustment: Only “Pop quizzes!” is **not** statistically significant.

  - Top 3 Most Gendered Tags: Hilarious, Amazing lectures, Caring
  - Top 3 Least Gendered Tags: Lots to read, Inspirational, Pop quizzes!

- Consistent Gender $\otimes$ Percentage: Only "Lots to read", "Pop quizzes!", and "Don't skip class or you will not pass" are **not** statistically significant.
  - Top 3 Most Gendered Tags: Hilarious, Amazing lectures, Respected
  - Top 3 Least Gendered Tags: Lots to read, Pop quizzes!, Don't skip class or you will not pass

- Random Imputation $\otimes$ Number of Ratings: Only "Pop quizzes!" is **not** statistically significant.
  - Top 3 Most Gendered Tags: Hilarious, Amazing lectures, Caring
  - Top 3 Least Gendered Tags: Test heavy, Inspirational, Pop quizzes!

- Random Imputation $\otimes$ Bayesian Adjustment: Only "Pop quizzes!" is **not** statistically significant.
  - Top 3 Most Gendered Tags: Hilarious, Amazing lectures, Caring
  - Top 3 Least Gendered Tags: Lots to read, Inspirational, Pop quizzes!

- Random Imputation $\otimes$ Percentage: Only "Tough grader", "Pop quizzes!", "Lots to read", "Clear grading", and "Don't skip class or you will not pass" are **not** statistically significant.
  - Top 3 Most Gendered Tags: Hilarious, Amazing lectures, Respected
  - Top 3 Least Gendered Tags: Lots to read, Clear grading, Don't skip class or you will not pass

4.2 Mann-Whitney *U* test

Here, we identify the tags that exhibit statistical significance across the 6 results for the Mann-Whitney U test. We believe the results from the Mann-Whitney U test are more robust than those from the t-test, as it does not require the assumption of normality.

- Consistent Gender $\otimes$ Number of Ratings: Only "Clear grading" is **not** statistically significant.
  - Top 3 Most Gendered Tags: Hilarious, Amazing lectures, Lecture heavy
  - Top 3 Least Gendered Tags: Pop quizzes!, Don't skip class or you will not pass, Clear grading

- Consistent Gender $\otimes$ Bayesian Adjustment: Only "Inspirational", "Accessible", and "Test heavy" are **not** statistically significant.
  - Top 3 Most Gendered Tags: Hilarious, Participation matters, Caring
  - Top 3 Least Gendered Tags: Inspirational, Accessible, Test heavy

- Consistent Gender $\otimes$ Percentage: Only "Don't skip class or you will not pass" and "Clear grading" are **not** statistically significant.
  - Top 3 Most Gendered Tags: Hilarious, Amazing lectures, Lecture heavy
  - Top 3 Least Gendered Tags: Tough grader, Don't skip class or you will not pass, Clear grading

- Random Imputation $\otimes$ Number of Ratings: Only "Don't skip class or you will not pass" and "Clear grading" are **not** statistically significant.
  - Top 3 Most Gendered Tags: Hilarious, Amazing lectures, Lecture heavy
  - Top 3 Least Gendered Tags: Tough grader, Don't skip class or you will not pass, Clear grading

- Random Imputation $\otimes$ Bayesian Adjustment: Only "Inspirational", "Accessible", and "Test heavy" are **not** statistically significant.
  - Top 3 Most Gendered Tags: Hilarious, Participation matters, Amazing lectures
  - Top 3 Least Gendered Tags: Inspirational, Accessible, Test heavy

- Random Imputation $\otimes$ Percentage: Only "Lots to read", "Tough grader", "Don't skip class or you will not pass", and "Clear grading" are **not** statistically significant.
  - Top 3 Most Gendered Tags: Hilarious, Amazing lectures, Lecture heavy
  - Top 3 Least Gendered Tags: Tough grader, Don't skip class or you will not pass, Clear grading

4.3 Permutation test

Here, we identify the tags that exhibit statistical significance across the 6 results for the permutation test. We set the number of iterations to 11,260, as we believe this is sufficient to generate robust results without running the tests for an excessive amount of time.

- Consistent Gender ⊗ Number of Ratings: Only “Pop quizzes!” is **not** statistically significant.
  - Top 3 Most Gendered Tags: Tough grader, Good feedback, Respected
  - Top 3 Least Gendered Tags: Lots to read, Don’t skip class or you will not pass, Pop quizzes!
- Consistent Gender ⊗ Bayesian Adjustment: Only “Pop quizzes!” is **not** statistically significant.
  - Top 3 Most Gendered Tags: Tough grader, Good feedback, Respected
  - Top 3 Least Gendered Tags: Group projects, Lecture heavy, Pop quizzes!
- Consistent Gender ⊗ Percentage: Only “Pop quizzes!” and “Don’t skip class or you will not pass” are **not** statistically significant.
  - Top 3 Most Gendered Tags: Tough grader, Good feedback, Respected
  - Top 3 Least Gendered Tags: Lots to read, Pop quizzes!, Don’t skip class or you will not pass
- Random Imputation ⊗ Number of Ratings: Only “Pop quizzes” is **not** statistically significant.
  - Top 3 Most Gendered Tags: Tough grader, Good feedback, Respected
  - Top 3 Least Gendered Tags: Test heavy, Inspirational, Pop quizzes!
- Random Imputation ⊗ Bayesian Adjustment: Only “Pop quizzes!” is **not** statistically significant.
  - Top 3 Most Gendered Tags: Tough grader, Good feedback, Respected
  - Top 3 Least Gendered Tags: Lecture heavy, Inspirational, Pop quizzes!
- Random Imputation ⊗ Percentage: Only “Tough grader”, “Pop quizzes!”, “Lots to read”, “Clear grading”, and “Don’t skip class or you will not pass” are **not** statistically significant.
  - Top 3 Most Gendered Tags: Good feedback, Respected, Participation matters
  - Top 3 Least Gendered Tags: Lots to read, Clear grading, Don’t skip class or you will not pass

5. Question 5

This question is quite similar to Question 1, but here we apply the tests with respect to average difficulty.

Null hypothesis: There is no gender bias in average difficulty.
Alternative hypothesis: There is gender bias in average difficulty.

As before, we will have 6 results for each test. Before running the tests, we plot the distribution of average difficulty for both male and female professors across the 6 datasets. Surprisingly, all distributions appear to be normally distributed. Due to space limitations, we present only one example figure on the last page (Figure (c)).

5.1 *t*-test

The following table shows the results of the *t*-test. Given that the data is normally distributed, the *t*-test results are considered credible. Note that we assume unequal variance in the average ratings between male and female.

	Threshold Setting	Bayesian Adjustment	Both
Consistent Gender	0.9257469292043111	0.9986284959760907	0.986268532118078
Random Imputation	0.5126814457485944	0.5406470566602406	0.986268532118078

5.2 Mann-Whitney *U* test

The following table shows the results of the Mann-Whitney *U* test. We believe the results from the Mann-Whitney *U* test have slightly lower power than those of the *t*-test, a parametric test, because the data is normally distributed.

	Threshold Setting	Bayesian Adjustment	Both
Consistent Gender	0.9680622117890496	0.8755744641786454	0.771115373953872
Random Imputation	0.43953891968306924	0.6841447800447791	0.771115373953872

5.3 Permutation test

The following table shows the results of the permutation test. As usual, we set the number of iterations to 11,260, as we believe this is sufficient to generate robust results without running the tests for an excessive amount of time.

	Threshold Setting	Bayesian Adjustment	Both
Consistent Gender	0.9253108348134991	0.9986678507992895	0.9865896980461811
Random Imputation	0.5170515097690941	0.5466252220248667	0.9865896980461811

We do not reject the null hypothesis; none of the results show statistical significance.

6. Question 6

We construct 95% confidence intervals based on the results from the previous question. The logic is the same as in Question 3: we use bootstrapping to generate all of the confidence intervals.

6.1 *t*-test

The following table shows the 95% confidence intervals for the *t*-test.

	Threshold Setting	Bayesian Adjustment	Both
Consistent Gender	(-2.0561969645968947, 1.8723988031980752)	(-1.9651862421147126, 1.9125073598032778)	(-1.9509667300057192, 1.9472522813803745)
Random Imputation	(-1.2887372415682734, 2.5905499573024016)	(-1.345821726096362, 2.5817924214449395)	(-1.9509667300057192, 1.9472522813803745)

6.2 Mann-Whitney *U* test

The following table shows the 95% confidence intervals for the Mann-Whitney *U* test.

	Threshold Setting	Bayesian Adjustment	Both
Consistent Gender	(26919920.25, 27949635.325)	(334893472.475, 341620695.875)	(334653267.9, 341407546.4125)
Random Imputation	(52122269.7875, 53774428.5375)	(606116146.9875, 619747458.8625)	(334653267.9, 341407546.4125)

6.3 Bootstrapping

The following table shows the 95% confidence intervals using bootstrapping. As usual, unlike the permutation test, which samples without replacement and returns p-values, bootstrapping samples with replacement and constructs confidence intervals.

	Threshold Setting	Bayesian Adjustment	Both
Consistent Gender	(-0.02687585723978555, 0.02449788319019501)	(-0.006619716876197168, 0.00645693838712662)	(-0.008544810236832234, 0.008518660803525212)
Random Imputation	(-0.014304286171902224, 0.02866594303816732)	(-0.003938812563448357, 0.007555843296914652)	(-0.008544810236832234, 0.008518660803525212)

7. Question 7

In this question, we run regression models to predict average ratings using all numerical predictors. To address collinearity concerns, we implement three regularization techniques, which offer robust solutions. Additionally, to prevent overfitting, we use cross-validation. Unlike previous analyses, our setup for the 6 datasets is slightly different. Instead of discarding inconsistent gender data and using only consistent gender data, we retain the entire dataset unchanged. Moreover, when applying Bayesian adjustment, both average ratings and average difficulty are adjusted accordingly. We set the train-test ratio to 8:2 for every training session. Lastly, we standardize all predictors using StandardScaler from sklearn.preprocessing before each training.

7.1 LASSO

The following table shows the R^2 and RMSE of LASSO regression model.[†] We apply 5-fold cross-validation and utilize the default settings of sklearn.linear_model.LassoCV to search for the optimal alpha value, as we believe this configuration is effective for identifying the best parameters.

	Threshold Setting	Bayesian Adjustment	Both
Unchanged Gender	0.5175, 0.6435	0.4252, 0.3411	0.4394, 0.4430
Random Imputation	0.5143, 0.6456	0.4234, 0.3416	0.4378, 0.4437

"Average Difficulty" is the strongest predictor for every scenario.

7.2 Ridge

The following table shows the R^2 and RMSE of Ridge regression model. We apply 5-fold cross-validation and search for the optimal alpha value using numpy.logspace(-6, 6, 13) with sklearn.linear_model.RidgeCV, as we believe this configuration is effective for identifying the best parameters.

	Threshold Setting	Bayesian Adjustment	Both
Unchanged Gender	0.5175, 0.6435	0.4252, 0.3411	0.4394, 0.4430
Random Imputation	0.5143, 0.6456	0.4234, 0.3416	0.4378, 0.4437

"Average Difficulty" is the strongest predictor for every scenario.

7.3 Elastic Net

The following table shows the R^2 and RMSE of Elastic Net regression model. We apply 5-fold cross-validation utilizing the default settings of sklearn.linear_model.ElasticNetCV to search for the optimal alpha value and [.1, .5, .7, .9, .95, .99, 1] for searching for the best l1_ratio, as we believe this configuration is effective for identifying the best parameters.

	Threshold Setting	Bayesian Adjustment	Both
Unchanged Gender	0.5175, 0.6435	0.4252, 0.3411	0.4394, 0.4430
Random Imputation	0.5143, 0.6456	0.4234, 0.3416	0.4378, 0.4437

"Average Difficulty" is the strongest predictor for every scenario.

8. Question 8

In this question, we build regression models to predict average ratings using all tags as predictors. Similar to the previous question, we apply the three regularization techniques and use cross-validation. As before, when applying Bayesian adjustment, both average ratings and average difficulty are adjusted accordingly. We set the train-test ratio to 8:2 for every training session. Lastly, we standardize all predictors using StandardScaler from sklearn.preprocessing before each training.

8.1 LASSO

The following table shows the R^2 and RMSE of LASSO regression model. All model settings are the same as in the previous question.

[†] Each cell presents the following order: R^2, RMSE.

	Threshold Setting	Bayesian Adjustment	Both
Number of Ratings	0.7573, 0.4563	0.5564, 0.2996	0.5363, 0.4030
Bayesian Adjustment	0.7565, 0.4571	0.6280, 0.2744	0.6064, 0.3713
Percentage	0.7419, 0.4706	0.5492, 0.3021	0.5272, 0.4069

"Good feedback" is the strongest predictor in all scenarios where the threshold setting is applied, while "Average Difficulty" is the strongest predictor in all other scenarios. We observe that the R^2 is higher and the RMSE is lower for every scenario compared to the previous ones.

8.2 Ridge

The following table shows the R^2 and RMSE of Ridge regression model. All model settings are the same as in the previous question.

	Threshold Setting	Bayesian Adjustment	Both
Number of Ratings	0.7573, 0.4564	0.5564, 0.2996	0.5362, 0.4030
Bayesian Adjustment	0.7565, 0.4572	0.6279, 0.2744	0.6064, 0.3713
Percentage	0.7423, 0.4703	0.5491, 0.3021	0.5272, 0.4069

"Good feedback" is the strongest predictor in all scenarios where the threshold setting is applied, while "Average Difficulty" is the strongest predictor in all other scenarios. We observe that the R^2 is higher and the RMSE is lower for every scenario compared to the previous ones.

8.3 Elastic Net

The following table shows the R^2 and RMSE of Elastic Net regression model. All model settings are the same as in the previous question.

	Threshold Setting	Bayesian Adjustment	Both
Number of Ratings	0.7573, 0.4564	0.5564, 0.2996	0.5363, 0.4030
Bayesian Adjustment	0.7565, 0.4572	0.6280, 0.2744	0.6064, 0.3713
Percentage	0.7419, 0.4706	0.5492, 0.3021	0.5272, 0.4069

"Good feedback" is the strongest predictor in all scenarios where the threshold setting is applied, while "Average Difficulty" is the strongest predictor in all other scenarios. We observe that the R^2 is higher and the RMSE is lower for every scenario compared to the previous ones.

9. Question 9

This question is very similar to the previous one, but here we build the models to predict average difficulty instead of average ratings.

9.1 LASSO

The following table shows the R^2 and RMSE of LASSO regression model. All model settings are the same as in the previous question.

	Threshold Setting	Bayesian Adjustment	Both
Number of Ratings	0.6008, 0.5045	0.4477, 0.2878	0.4511, 0.3746
Bayesian Adjustment	0.6023, 0.5035	0.4853, 0.2779	0.4887, 0.3616
Percentage	0.5847, 0.5145	0.4274, 0.2931	0.4316, 0.3813

"Tough grader" is the strongest predictor in all scenarios where the threshold setting is applied, while "Average Rating" is the strongest predictor in all other scenarios.

9.2 Ridge

The following table shows the R^2 and RMSE of Ridge regression model. All model settings are the same as in the previous question.

	Threshold Setting	Bayesian Adjustment	Both
Number of Ratings	0.6008, 0.5044	0.4477, 0.2878	0.4512, 0.3746
Bayesian Adjustment	0.6023, 0.5035	0.4853, 0.2779	0.4887, 0.3616
Percentage	0.5848, 0.5144	0.4274, 0.2931	0.4316, 0.3813

"Tough grader" is the strongest predictor in all scenarios where the threshold setting is applied, while "Average Rating" is the strongest predictor in all other scenarios.

9.3 Elastic Net

The following table shows the R^2 and RMSE of Elastic Net regression model. All model settings are the same as in the previous question.

	Threshold Setting	Bayesian Adjustment	Both
Number of Ratings	0.6008, 0.5045	0.4477, 0.2878	0.4511, 0.3746
Bayesian Adjustment	0.6023, 0.5035	0.4853, 0.2779	0.4887, 0.3616
Percentage	0.5846, 0.5146	0.4274, 0.2931	0.4316, 0.3813

"Tough grader" is the strongest predictor in all scenarios where the threshold setting is applied, while "Average Rating" is the strongest predictor in all other scenarios.

10. Question 10

For this question, we build three classification models to predict whether a professor receives a "pepper" based on all available factors, including both tags and numerical data. We retain the values unchanged in both male and female columns for this question. Moreover, when applying Bayesian adjustment, both average ratings and average difficulty are adjusted accordingly. We set the train-test ratio to 8:2 for every training session. To address class imbalance concerns, we apply SMOTE (Synthetic Minority Over-sampling Technique), which we believe is effective in mitigating class imbalance. We standardize all predictors using StandardScaler from sklearn.preprocessing before each training. We employ GridSearchCV from sklearn.model_selection with 5-fold cross-validation to find the best combination of parameters, which will be specified below, for each model. All the ROC curves will be presented on the last page (Figure (d) - (l)).

10.1 Logistic Regression

The following shows the results of Logistic Regression obtained from the best combinations of parameters.[‡] To find the optimal results with parameter tuning, we set the parameters as follows before training: C: [0.01, 0.1, 1, 10, 100], penalty: ["l1", "l2"], solver: ["liblinear"]. Lastly, we set the maximum number of iterations to be 11,260 for the solver to converge.

- Number of Ratings ⊗ Threshold Setting: 0.5369, 0.8505, 0.0895, 0.1620
- Number of Ratings ⊗ Bayesian Adjustment: 0.5000, 0.5000, 0.0000, 0.0001
- Number of Ratings ⊗ Both: 0.5001, 0.7600, 0.0004, 0.0008
- Bayesian Adjustment ⊗ Threshold Setting: 0.5519, 0.8496, 0.1260, 0.2195
- Bayesian Adjustment ⊗ Bayesian Adjustment: 0.5016, 0.9278, 0.0036, 0.0071
- Bayesian Adjustment ⊗ Both: 0.5129, 0.8862, 0.0295, 0.0571
- Percentage ⊗ Threshold Setting: 0.6191, 0.7825, 0.3298, 0.4640
- Percentage ⊗ Bayesian Adjustment: 0.4999, 0.4918, 0.0072, 0.0141
- Percentage ⊗ Both: 0.5065, 0.6261, 0.0324, 0.0616

10.2 Random Forest

The following shows the results of Random Forest obtained from the best combinations of parameters. To find the optimal results with parameter tuning, we set the parameters as follows before training: n_estimators: [50, 100, 200], max_depth: None, 10, 20, 30], min_samples_split: [2, 5, 10].

- Number of Ratings ⊗ Threshold Setting: 0.5198, 0.8143, 0.0512, 0.0964
- Number of Ratings ⊗ Bayesian Adjustment: 0.5197, 0.8002, 0.0526, 0.0987

[‡] Each bulletpoint presents the following order: Accuracy, Precision, Recall, F-1 Score.

- Number of Ratings $\otimes$ Both: 0.5259, 0.8187, 0.0665, 0.1229
- Bayesian Adjustment $\otimes$ Threshold Setting: 0.5749, 0.6698, 0.2953, 0.4099
- Bayesian Adjustment $\otimes$ Bayesian Adjustment: 0.5141, 0.7862, 0.0388, 0.0740
- Bayesian Adjustment $\otimes$ Both: 0.5123, 0.7574, 0.0361, 0.0689
- Percentage $\otimes$ Threshold Setting: 0.5157, 0.7749, 0.0444, 0.0839
- Percentage $\otimes$ Bayesian Adjustment: 0.5136, 0.7928, 0.0369, 0.0705
- Percentage $\otimes$ Both: 0.5191, 0.8046, 0.0503, 0.0947

10.3 XGBoost

The following shows the results of XGBoost obtained from the best combinations of parameters. To find the optimal results with parameter tuning, we set the parameters as follows before training: learning_rate: [0.01, 0.1, 0.2], n_estimators: [50, 100, 200], max_depth: [3, 5, 7], scale_pos_weight: [1, $\frac{|y_{bal}|}{\sum_{i=1}^{N} y_{bal,i}}$], where the numerator denotes the cardinality of the balanced labels and the denominator denotes the sum of the values of the balanced labels.

- Number of Ratings $\otimes$ Threshold Setting: 0.5057, 0.7510, 0.0170, 0.0332
- Number of Ratings $\otimes$ Bayesian Adjustment: 0.6028, 0.7497, 0.3087, 0.4374
- Number of Ratings $\otimes$ Both: 0.6579, 0.7427, 0.4833, 0.5856
- Bayesian Adjustment $\otimes$ Threshold Setting: 0.5804, 0.7300, 0.2551, 0.3781
- Bayesian Adjustment $\otimes$ Bayesian Adjustment: 0.5391, 0.8254, 0.0991, 0.1770
- Bayesian Adjustment $\otimes$ Both: 0.5308, 0.8000, 0.0823, 0.1492
- Percentage $\otimes$ Threshold Setting: 0.5067, 0.7045, 0.0231, 0.0447
- Percentage $\otimes$ Bayesian Adjustment: 0.5614, 0.7020, 0.2135, 0.3274
- Percentage $\otimes$ Both: 0.5931, 0.7260, 0.2991, 0.4237

Extra

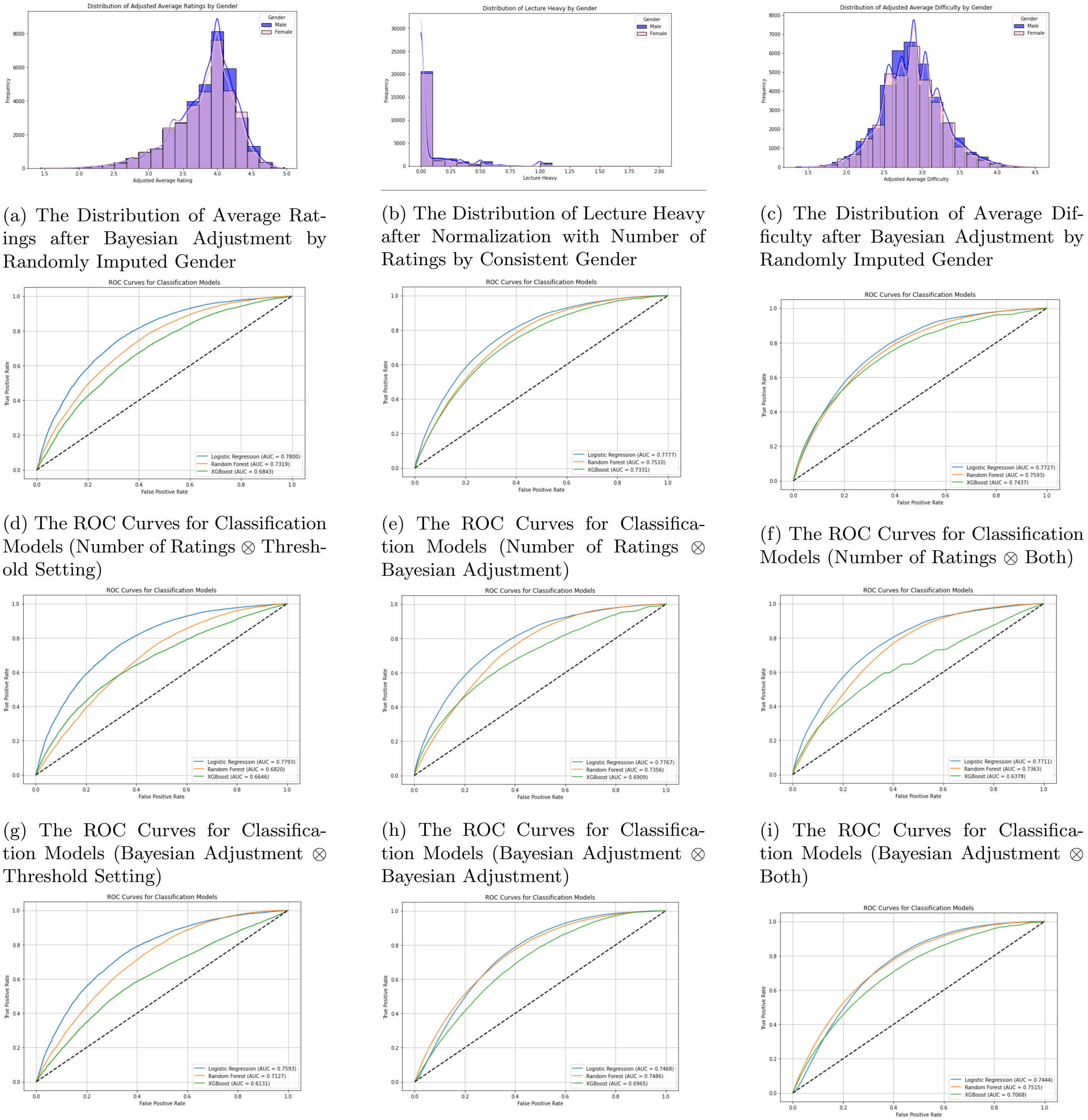
For extra credit, we examine whether there are differences in average ratings across majors.

Null hypothesis: There is no difference in average ratings between majors.
Alternative hypothesis: There are differences in average ratings between majors.

To test this, we handle the average ratings using the threshold setting method with the threshold set to 50, resulting in 77 different majors. We then apply ANOVA to obtain the results. The p-value is 0.0012720565558368155, indicating a statistically significant.

Discussion and Conclusion

In this report, we have explored a variety of models in both statistical inference and machine learning to different datasets generated using various data preprocessing techniques. We are confident that we have derived robust results and valuable insights. Due to space limitations, there are still some technical details regarding both the settings and results could not be included in this report, but they can be found in the accompanying code files. In future work, we aim to further explore the qualitative data and uncover additional insights. We hope this report provides a comprehensive overview of the analysis conducted, and we look forward to refining our approach and exploring new dimensions of the data moving forward.



(a) The Distribution of Average Ratings after Bayesian Adjustment by Randomly Imputed Gender

(b) The Distribution of Lecture Heavy after Normalization with Number of Ratings by Consistent Gender

(c) The Distribution of Average Difficulty after Bayesian Adjustment by Randomly Imputed Gender

(d) The ROC Curves for Classification Models (Number of Ratings ⊗ Threshold Setting)

(e) The ROC Curves for Classification Models (Number of Ratings ⊗ Bayesian Adjustment)

(f) The ROC Curves for Classification Models (Number of Ratings ⊗ Both)

(g) The ROC Curves for Classification Models (Bayesian Adjustment ⊗ Threshold Setting)

(h) The ROC Curves for Classification Models (Bayesian Adjustment ⊗ Bayesian Adjustment)

(i) The ROC Curves for Classification Models (Bayesian Adjustment ⊗ Both)

(j) The ROC Curves for Classification Models (Percentage ⊗ Threshold Setting)

(k) The ROC Curves for Classification Models (Percentage ⊗ Bayesian Adjustment)

(l) The ROC Curves for Classification Models (Percentage ⊗ Both)