IDS Capstone Project

Group: CAP 12
Jayant Dabas, Rebecca Rinehart, Eric Zhao

Contributions:
Jayant Dabas contributed to questions 2, 4 and 7.
Rebecca Rinehart contributed to questions 5, 6, 8, and 10.
Eric Zhao contributed to questions 1, 3, 9, and the extra credit.

Preprocessing:

After merging the tables into one, we found that around 20,000 entries were null for several columns, most importantly for ‘avg_rating’ and ‘avg_difficulty’, which were central to our analysis. We chose to drop those rows due to the large sample size. Then, only ‘num_repeat’ contained null values, with 86.4% of the data null; Row-wise removal of any professors with this column null would yield a total sample size of 8,849. Additionally, plotting showed us a high correlation between ‘avg_rating’ and ‘num_repeat’, but since more than > 80% of the column is missing, we decided to drop this column entirely to prevent any bias where a model focuses too much on this single, highly sparse predictor.

To address data quality concerns with entries having the same boolean for both ‘is_male’ and ‘is_female’, we dropped those entries. To address multicollinearity concerns with the gender columns, we dropped the ‘is_female’ column. To reduce extreme averages based on a small number of ratings, we calculated the mean and median number of ratings, and found the mean to be 5.4 and the median 3.0. If we retained professors with 5 or more ratings, we would be left with only 28.22% of the original dataset. However, if we retained professors with 3 or more ratings, we would retain 45.08% of the original dataset, which we concluded to be an acceptable tradeoff as it still provides an acceptably large sample size of 40,528.

The tag frequency was normalized to represent a proportion– each tag was divided by the total number of tags that professor/entry had received. Each student can give up to 3 tags per professor, and professors with more ratings would naturally have more tags. Using proportions allowed the tags to be more comparable across professors. Lastly, for all modeling problems we seeded with Eric Zhao’s N-number, 19057054.

Q1: Is there evidence of a pro-male gender bias in the dataset?

H0: There is no evidence of pro-male gender bias in the dataset where male professors enjoy a boost in rating.

H1: There is evidence of pro-male gender bias in the dataset where male professors enjoy a boost in rating.

After the preprocessing, we split the dataset into two groups: average ratings of male professors and average ratings of female professors. A Mann-Whitney U test was then used to calculate the p-value and test statistic U. Since we are comparing the medians of two groups (male ratings vs female ratings), a one-tailed Mann-Whitney U test was preferred over a KS test or the KW test because it also allows us to gauge whether one of the rating distributions is greater than the other group, which is more in line with our research question of whether male professors have a boost in ratings, opposed to just observing if a difference exists. We have assumed that the ratings for professors are independent of each other.

After conducting a one-tailed Mann Whitney U test for ratings of male and female professors, a p-value of 2.062e-05 was reported where there were 15801 valid male professor ratings with median 4.2 and 13925 valid female professor ratings with a median of 4.1. According to figure 1, there is a visible separation in male professors receiving higher ratings when observing the proportions. With a large sample size like we have here, a noticeable difference in the groups' proportions can indicate significance.

Given that our p-value is 2.062e-05, which is significantly less than the significance level of 0.005, we drop the null hypothesis, and the results are not consistent with chance and there is evidence of pro-male gender bias in the dataset.

Q2: Is there a gender difference in the spread (variance/dispersion) of the ratings distribution?

H0: There is no difference in the spread (variance/dispersion) of the ratings of male and female professors.

H1: There is a significant difference in the spread (variance/dispersion) of the ratings of male and female professors.

To investigate gender differences in the spread (variance) of the ratings distribution, we considered multiple statistical methods. Initially, we explored Welch's t-test and permutation test. However, Welch's t-test was not used because it makes assumptions about the underlying data, such as normality, which may not hold in our case. Permutation test requires resampling, which makes it computationally expensive for large permutations. Instead, we opted for Levene's test, which is straightforward and specifically designed to assess whether the variances of two or more groups are equal. It doesn't assume the data follows a normal distribution, which makes it more flexible. Specifically, we separated professors into male and female groups, calculated the

observed variance, and then performed Levene's test on the "Average Rating" variable for the two groups to assess the equality of variances between them. Figure 2, shows that the average ratings do not follow a normal distribution further supporting our claim. The two groups consisted of 15801 male and 13925 female professors with an observed variance of 0.9132 and 1.0137, respectively. We obtained a p-value of 7.0077e-12 by performing Levene's test on this ratings data.

Given that our p-value is 7.0077e-12, which is significantly less than the significance level of 0.005, we drop the null hypothesis, and the results are not consistent with chance and there is evidence of difference in the spread of the ratings of male and female professors.

Q3: What is the likely size of both of these effects, as estimated from this dataset?

We estimated the effect size of the gender bias for professor ratings by comparing the average rating and variance between male and female professors using Cohen's d with 95% confidence interval. Cohen's d converts the difference of two groups into standard deviation units which allows for interpretability of the results and by using the pooled standard deviation, difference between groups and variability within groups are accounted for. This allows for a standardized approach to calculating the confidence interval. Cohen's d is typically designed for comparing means between two groups but we used it for comparing the variances because we have not covered alternative tests that would compare variances such as a f-test or Levene's test.

The effect size of the average rating between male and female professors is d=0.0599 [95% CI: 0.042708, 0.077098]. There is a small positive bias in favor of male professors based on comparing average ratings but it is important to note that the effect size is small so the difference between the two groups may not be important.
The effect size of the variance between male and female professors ratings is d=0.1054 [95% CI: 0.088210, 0.122616]. Initially, Cohen's d was found to be a negative value which means the variance of female professor ratings were greater. Therefore, male professors experience somewhat less variability in ratings compared to female professors based on comparing the variance of the ratings but it is important to note that the effect size is small so the difference between the two groups may not be important.

Q4: Is there a gender difference in the tags awarded by students? Comment on the 3 most gendered (lowest p-value) and least gendered (highest p-value) tags.
H0: There is no gender difference in the tags awarded by students between male and female professors.
H1: There is a gender difference in the tags awarded by students between male and female professors.

To assess whether there is a gender difference in the tags awarded by students, we tested each of the 20 tags using the Mann-Whitney U test, which is suitable for comparing the distribution of two independent groups (male and female professors) without assuming a normal distribution. Although the Chi-square test could also be used to compare frequencies, it treats tags as categorical data, which doesn't align well with the nature of the tags (numeric counts), making the Mann-Whitney U test a more appropriate choice.

For each of the 20 tags, we calculated the count of tags for male and female professors and used the two-tailed Mann-Whitney U test to assess if their distributions were significantly different. The two-tailed test was selected to check for differences in either direction (whether males or females received more tags), ensuring we account for any potential variations in either direction. A significance level of 0.005 was applied to identify gendered tags, with p-values below this threshold indicating significant differences.

The results in figure 4 showed that the tags "Hilarious," "Amazing lectures," and "Respected" had the lowest p-values of 1.21e-226, 1.55e-55, and 4.61e-42, respectively, indicating they are the most gendered, with significant differences between male and female professors.

Conversely, the tags "Tough grader," "Clear grading," and "Don't skip" had the highest p-values equal to 0.026, 0.083, and 0.092, respectively, suggesting these tags do not exhibit significant gender differences.

Q5: Is there a gender difference in terms of average difficulty?

H0: There is no difference in average difficulty between male and female professors.
H1: There is a difference in average difficulty between male and female professors.

We began by splitting the data into male professors and female professors based on the ‘is_male’ column. Because the difficulty score was presented in the dataset as an average, we assumed that the difficulty could be reduced to its sample mean. Although both groups visually appear to be normally distributed (see Figure 5), we chose to use Levene’s test to statistically test for whether the variances of the groups were similar or not. We selected Levene’s test as it does not require an assumption of normality, whereas other tests like f-test assumes normality, and Welch’s t-test assumes unequal variances (which was what we were hoping to answer). With a Levene’s test statistic of 0.0927 and a p-value of 0.7607, we concluded that the average difficulties for male and female professors have equal variances. Because we were comparing two sample means, did not know population parameters, no large inter-individual variability, and the groups had similar variances, we used an independent samples t-test. We found that with a t-test statistic of 0.1420 and a p-value of 0.8871 > alpha 0.005, the average difficulty scores of male and female professors are not significantly different. The results

are consistent with chance. Thus, we concluded that there is no gender difference in average difficulty.

Q6: What is the likely size of this effect at 95% confidence?

As in question 3, estimated the effect size of the gender bias for professor difficulty by comparing the average rating and variance of each group using Cohen's d with a 95% confidence interval. We calculated Cohen's d using each group's average and the pooled standard deviation of the two. From this, we estimate the effect size to be d = 0.001650 with 95% confidence that the true effect lies in the interval of [-0.02113, 0.02443]. With an effect size this small, the distribution of average difficulty between male and female professors nearly overlaps. The average difference in average difficulty is 0.165% of a standard deviation. See Figure 6 for a plot of the sampling distribution of effect size, with confidence intervals included.

Q7: Build a regression model predicting average rating from all numerical predictors. Which of these factors is most strongly predictive of average rating?

In this analysis, we built a Ridge regression model to predict the Average Rating of professors using all numerical predictors from the dataset. Ridge regression was chosen due to its ability to mitigate overfitting, especially in cases of multicollinearity among predictors. We used rmpCapstoneNum.csv file with 'avg_rating' as the target and remaining columns as predictors (features). In data preparation, we already addressed multicollinearity and bias by dropping 'is_female' and 'num_repeat' columns. However, in figure 7 (a), we can observe that 'num_repeat' is a strong predictor of 'avg_rating'. Thus, the model performance is expected to be quite low. Next, we split the data into training and test sets with a test size of 20%. To ensure Ridge regression works optimally, all predictors were standardized to have a mean of 0 and standard deviation of 1 using StandardScalar. We performed hyperparameter tuning to obtain the best value for alpha using GridSearchCV with 5 folds. Evaluating the best model with an alpha = 10 on a hold-out test set gave us r2 = 0.4226 and root mean square error (RMSE) score of 0.7496, meaning that the model's predictions typically deviate from the professor's true ratings by about 0.7496 of a rating point. The factor most strongly predictive of Average Rating is 'avg_difficulty' (because we dropped 'num_repeat') with a coefficient value of -0.51699854. The low score is expected due to dropping of a strong predictor and very sparse data with categorical variables as shown by the scatter plots. The dropping of column is justified due to 80% missing values, which would have introduced bias if kept (as explained in the preprocessing section).

Q8: Build a regression model predicting average ratings from all tags. Which of these tags is most strongly predictive of average rating? Comment on how this model compares to the previous one.

In order to address collinearity and overfitting concerns, we chose to use regularization with a Ridge Regression algorithm. We began by splitting the data into predictors (tag features) and the target (average rating). For the train-test split, we used a test size of 20%. We scaled the predictor features, as regularization assumes similarly scaled features, and scaling allows coefficients/weights to be directly compared. We performed hyperparameter tuning of alpha (which controls regularization strength) using GridSearchCV with 5 folds. Using the best model configuration with an alpha = 10, we found a cross validation coefficient of determination of r2 = 0.6465, where 64.65% of variance is explained by the model. We found a root mean square error (RMSE) of 0.5821, meaning that the model's predictions typically deviate from the professor's true ratings by about 0.5821 of a rating point. From the model coefficients, the top three strongest predictors of averaging rating were, in order, "Good feedback" (beta = 0.285), "Respected" (beta = 0.227), and "Amazing lectures" (beta = 0.226). In Figure 8, coefficient values for all 20 tags at alpha = 10 are plotted.

In comparison to the previous regression model which seeks to predict average rating based on numerical features (r2 = 0.4226, RMSE = 0.7496), this model based on tag features performs better (r2 = 0.6465, RMSE = 0.5821). This model is able to explain 22.39% more variance. This suggests that purely tag features are a better predictor of a professor' rating than purely numerical features.

Q9: Build a regression model predicting average difficulty from all tags. Which of these tags is most strongly predictive of average difficulty?

A ridge regression model was built and addressed collinearity concerns through L2 regularization in order to predict a professor's difficulty based on assigned tags. Data was first split on an 80:20 train test with the random state set to be Eric's student ID for reproducibility purposes. Features were then normalized with StandardScalar where the scalar was fit only on training data to avoid data leakage and will improve interpretability for the features. The model was then hyperparameter tuned using various alpha values ranging from 0.01 to 1000 with the alpha with the lowest test MSE tracked. Finally, the model used RidgeCV with the best alpha value and the r2 and RMSE performance metrics were also reported.

The model has a coefficient of determination of r2 = 0.4805 where approximately 48% of variance are explained by the model and a RMSE = 0.6191. Out of the alpha values [0.01, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.4, 0.5, 1.0, 10, 11, 15, 50, 100], alpha = 10 performed the best in balancing the bias and variance. The top 3 strongest predictors of difficulties based on their coefficient magnitudes were "Tough grader" (beta=0.342), "Clear grading (beta=0.106)", and "Hilarious" (beta=0.104).

Q10: Build a classification model that predicts whether a professor receives a "pepper" from all available factors.

We built a logistic regression model from the numerical and tag columns to predict pepper recipients; we did not include the qualitative columns around school or location. To determine target class imbalance, we plotted the number of professors in the sample with and without a pepper (see Figure 10 (a)), and found that only 37.8% of professors had received a pepper. This demonstrated class imbalance, which we would later address. We then split the data into predictors and the target ('is_pepper'). For the train-test split, we used a test size of 20%. We scaled all predictor features (see reasons for feature scaling above). We performed hyperparameter tuning with GridSearchCV with 5 folds to find the best C (inverse of regularization strength), class weight (as we wanted to address class imbalance), and solver type (algorithm used in the optimization problem. We defined the class weights to test as 'balanced', a ratio of 1:2 (no pepper:pepper), and a ratio of 1:3. We found the best parameters to be {'C': 0.01, 'class_weight': 'balanced', 'penalty': 'l2', 'solver': 'liblinear'}. The CV ROC-AUC score from the best model configuration was 0.7764, meaning that given a random professor with a pepper and a random professor without a pepper, the model will assign a higher probability of receiving a pepper to the true pepper recipient 77.64% of the time. From the model coefficients (Figure 10 (b)), we found that the strongest predictor by a large margin was average rating (beta = 0.966). The next two strongest predictors were the "Amazing lectures" tag (beta = 0.277) and "Inspirational" tag (beta = 0.236). See Figure 10 (c) for the ROC curve.

Extra Credit: Do professors from different regions (Northeast, Midwest, South, West, Non-US) and outside of the United States have a difference in difficulty?
H0: There is no difference in difficulty for professors of different regions.
H1: There is a difference in difficulty for professors of different regions.

After the preprocessing, we defined four regions that the census bureau has declared (West, Midwest, Northeast, South) with an additional region "Non-US" as there were professor ratings from outside of the states.Columns with the franchise keywords were first filtered out based on what franchise they are a part of. Afterwards, we chose to clean the data by removing empty and invalid ratings for each movie that is part of a franchise as we are interested in the actual ratings of each movie and do not want invalid ratings to skew our statistical analysis. A Kruskal Wallis (KW) test was then used to calculate the p-value and test statistic H. Since we are comparing the data ranks of three or more regions, a KW test was preferred over a KS test or the MW test because it also allows us to compare three or more groups. We have assumed that the ratings for professors are independent of each other.

After conducting the KW test that compared the average difficulty within each of the five regions, we found p-value = 4.796e-14. According to Figure 11 (a) and Figure 11 (b), we can visually see that non-US professors have a difficulty around a 3 and that northeast professors are given more higher difficulty ratings. A potential limitation is that

southern professors receive substantially more ratings than other regions at over 10000 and non-US professors only have a sample size of 1500. In addition, non-US categorizes professors from anywhere else in the world which may not be appropriate when comparing regions but the sample sizes if we were to split non-US would be even smaller.

Given that our p-value = 4.796e-14 is less than the significance level of 0.005, we conclude that we drop the null hypothesis as the results are not consistent with chance and there is a difference in difficulty for professors in different regions of US and non-US.

Appendix

Figure 1

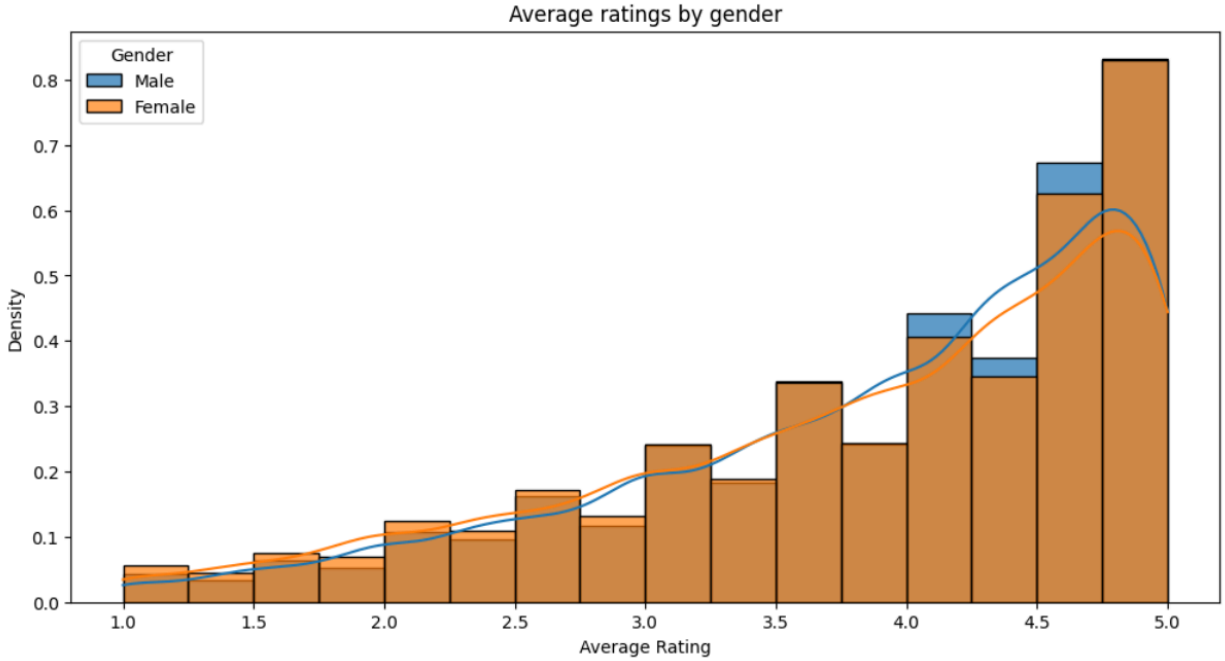

Figure 2

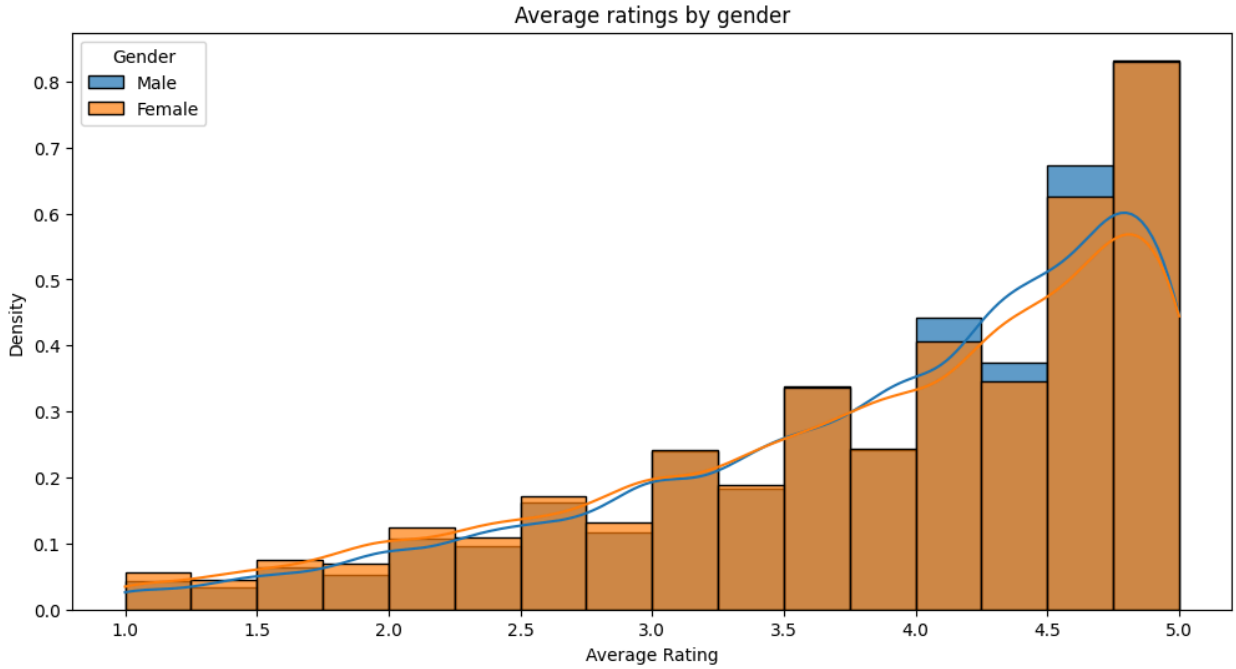

Figure 3 (a)

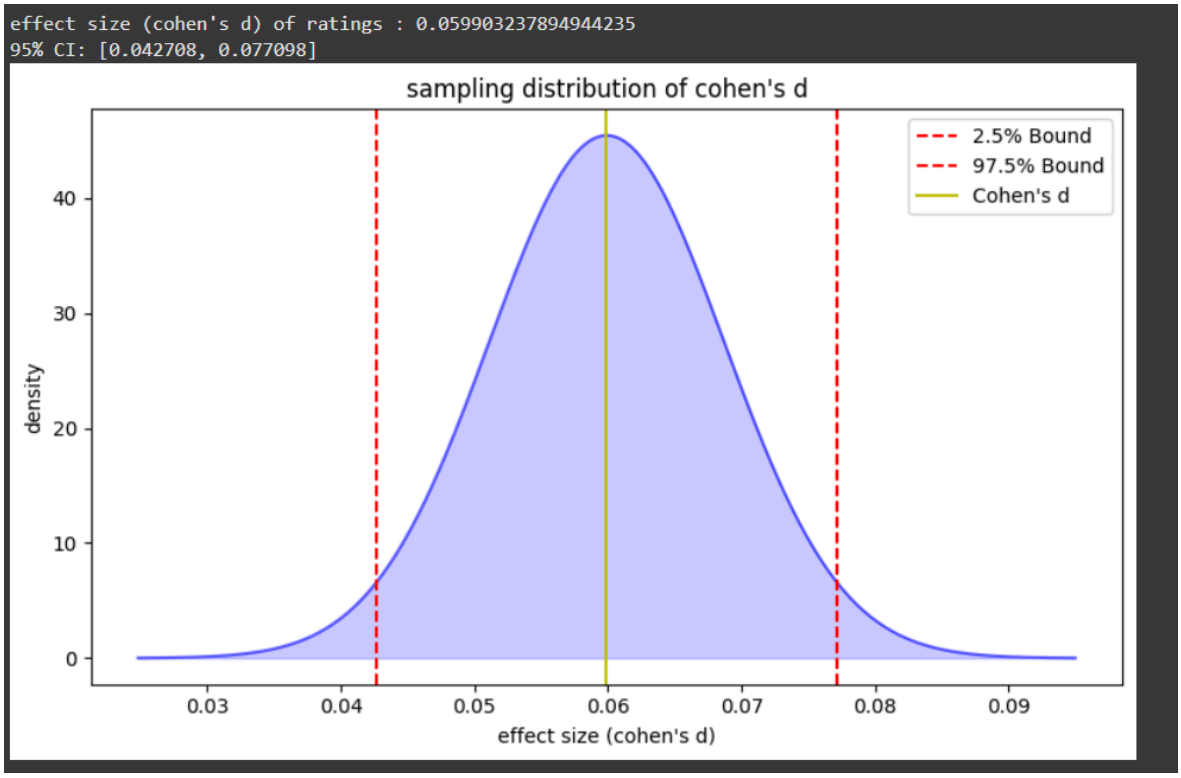

Figure 3 (b)

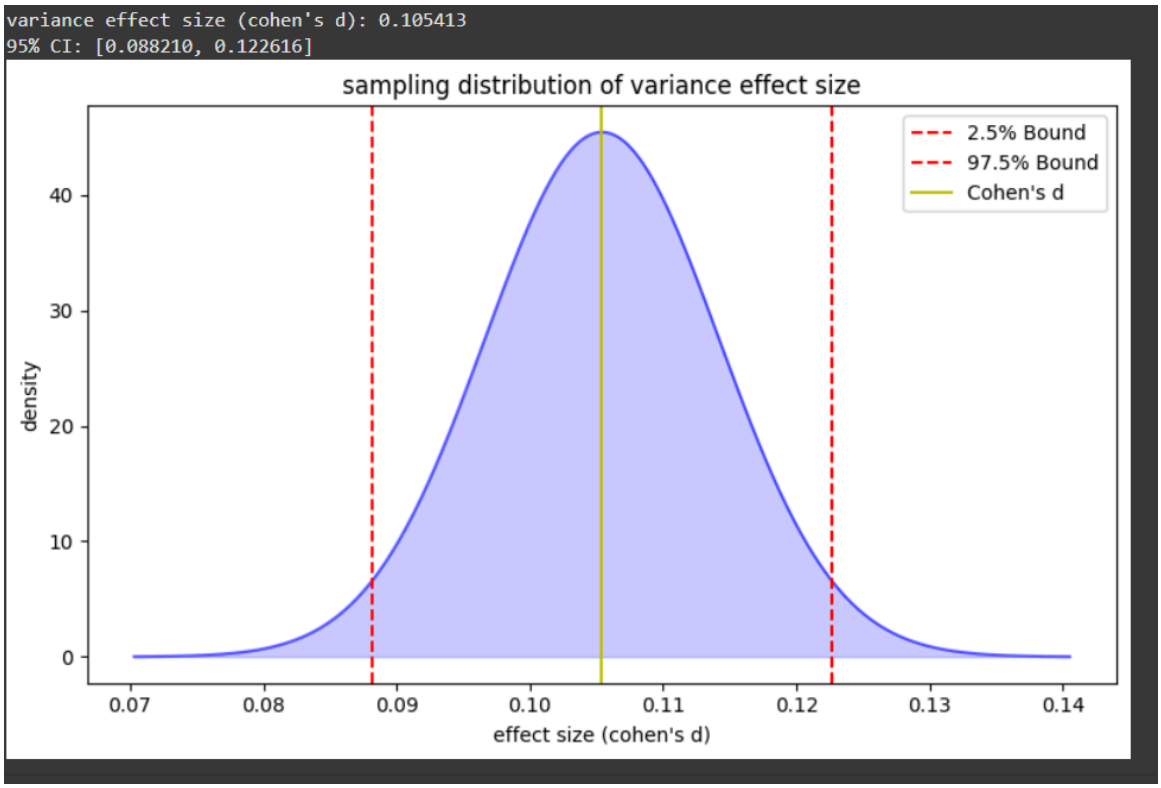

Figure 4

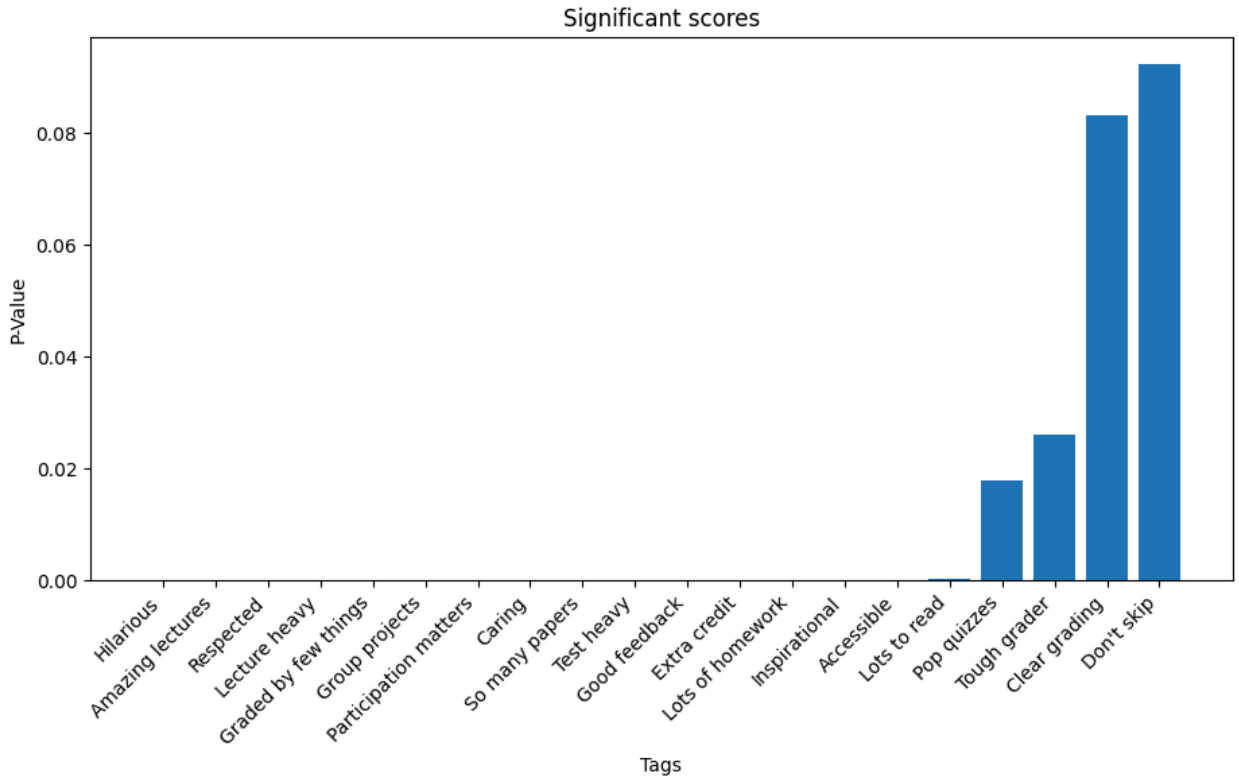

Figure 5

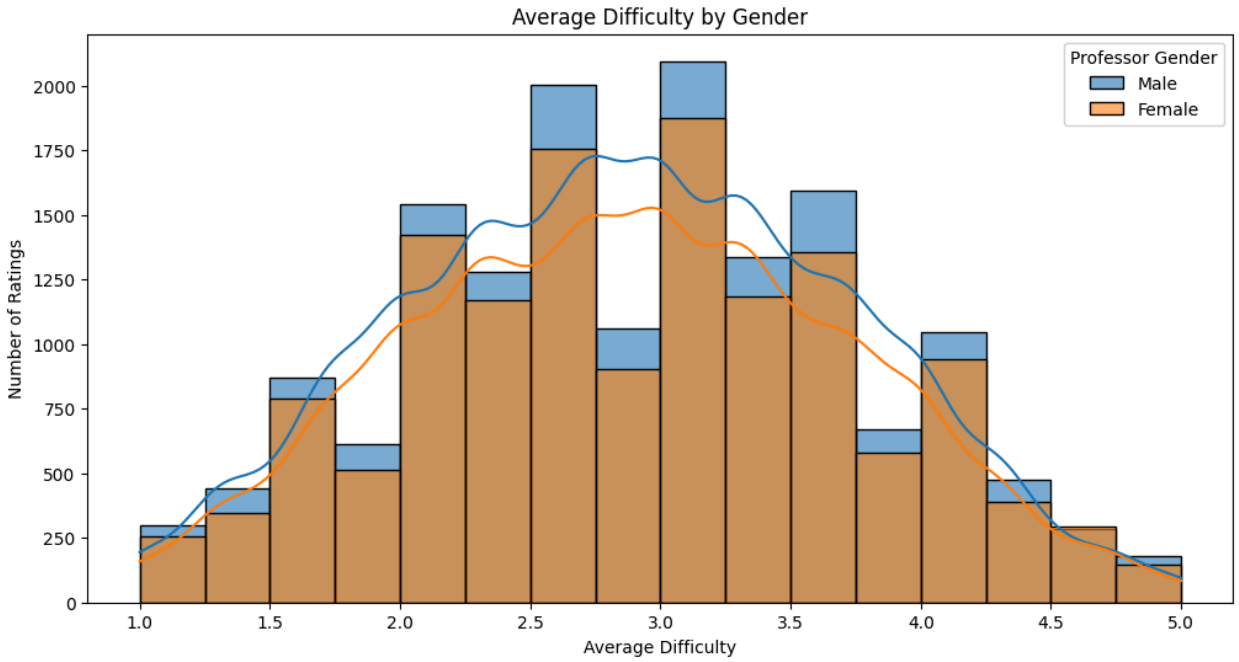

Figure 6

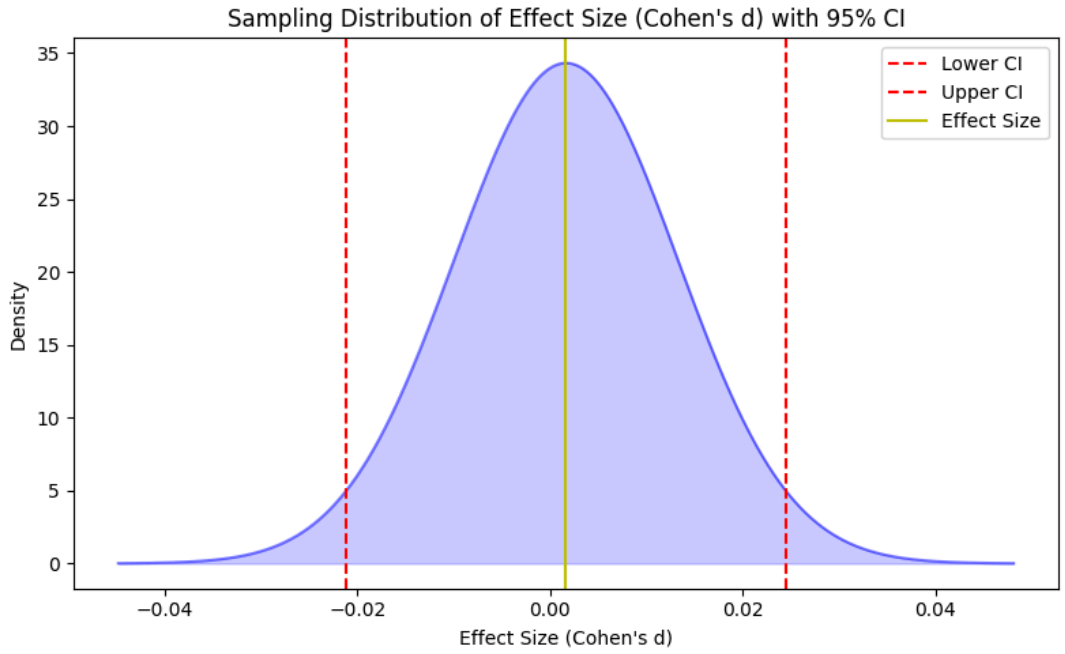

Figure 7 (a)

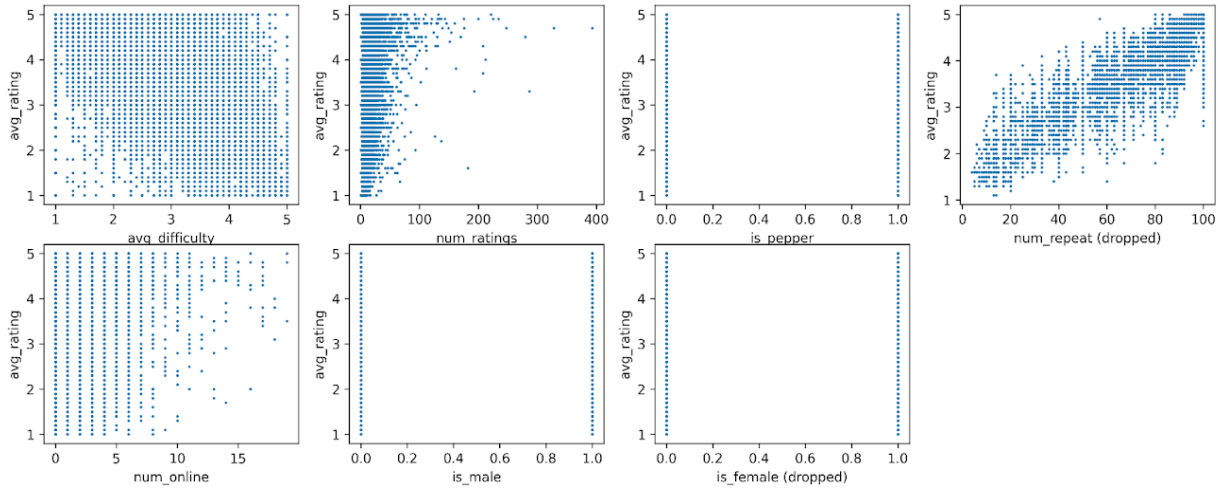

Figure 7 (b)

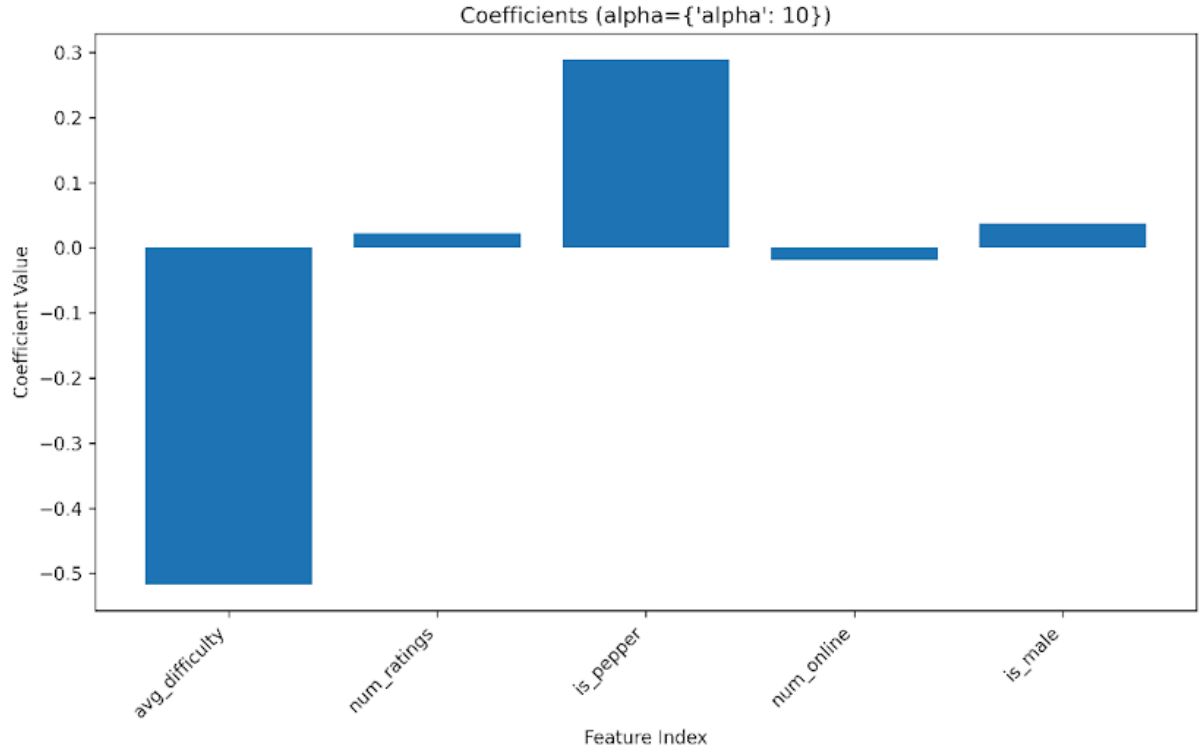

Figure 8

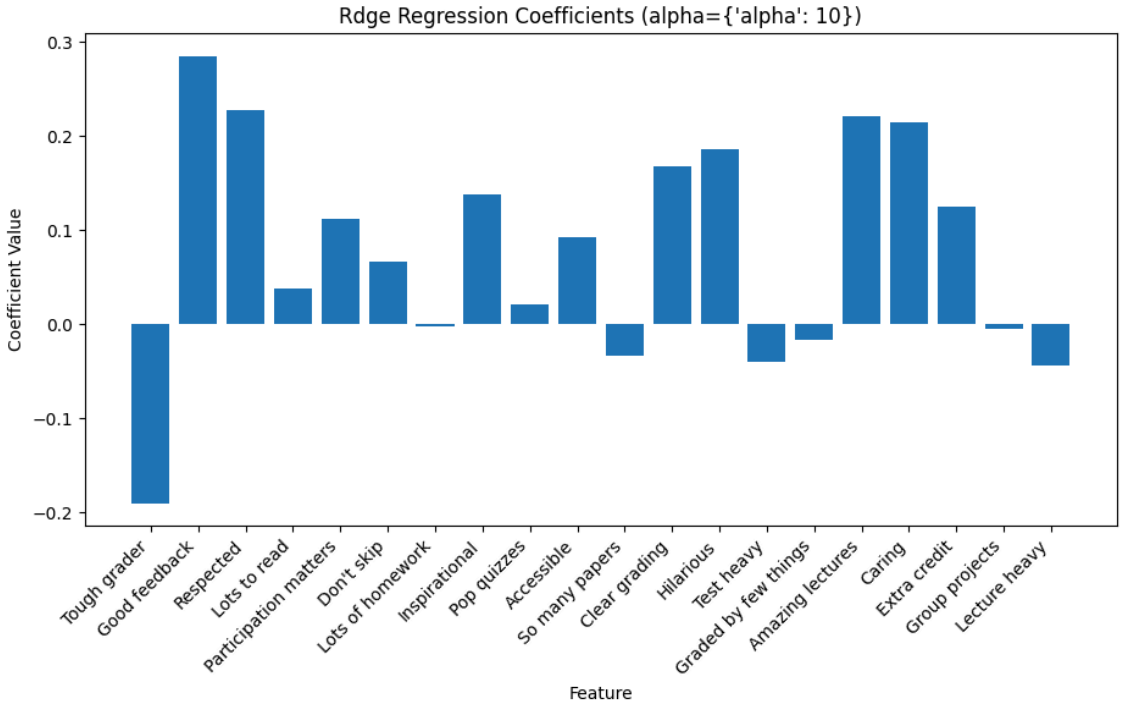

Figure 9

```
R²: 0.480534
RMSE: 0.619117
best alpha: 1.0

feature importance of tags:
            tag  coefficient
 Tough grader     0.342438
Clear grading     0.106459
    Hilarious     0.103522
       Caring     0.091430
   Test heavy     0.089690
Extra credit     0.084935
    Respected     0.066771
   Accessible     0.065874
   Don't skip     0.062176
Lots to read     0.059959
```

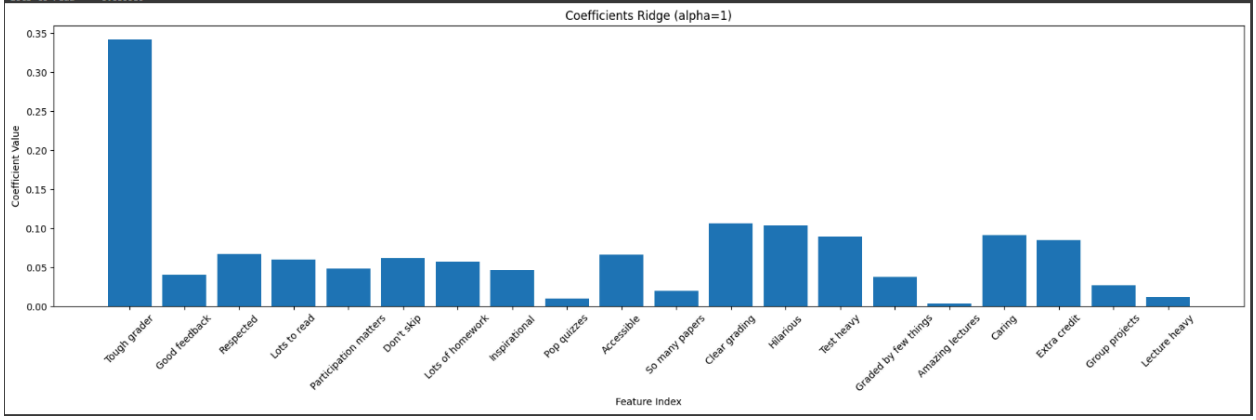

Figure 10 (a)

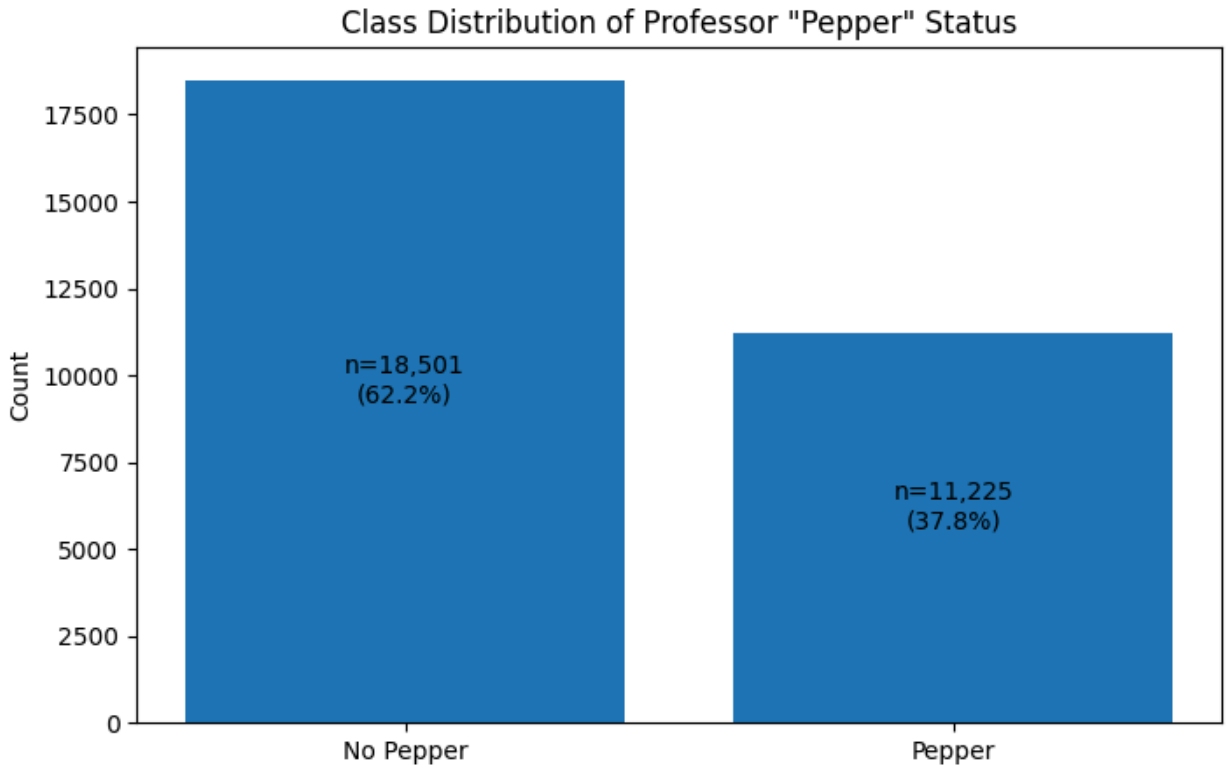

Figure 10 (b)

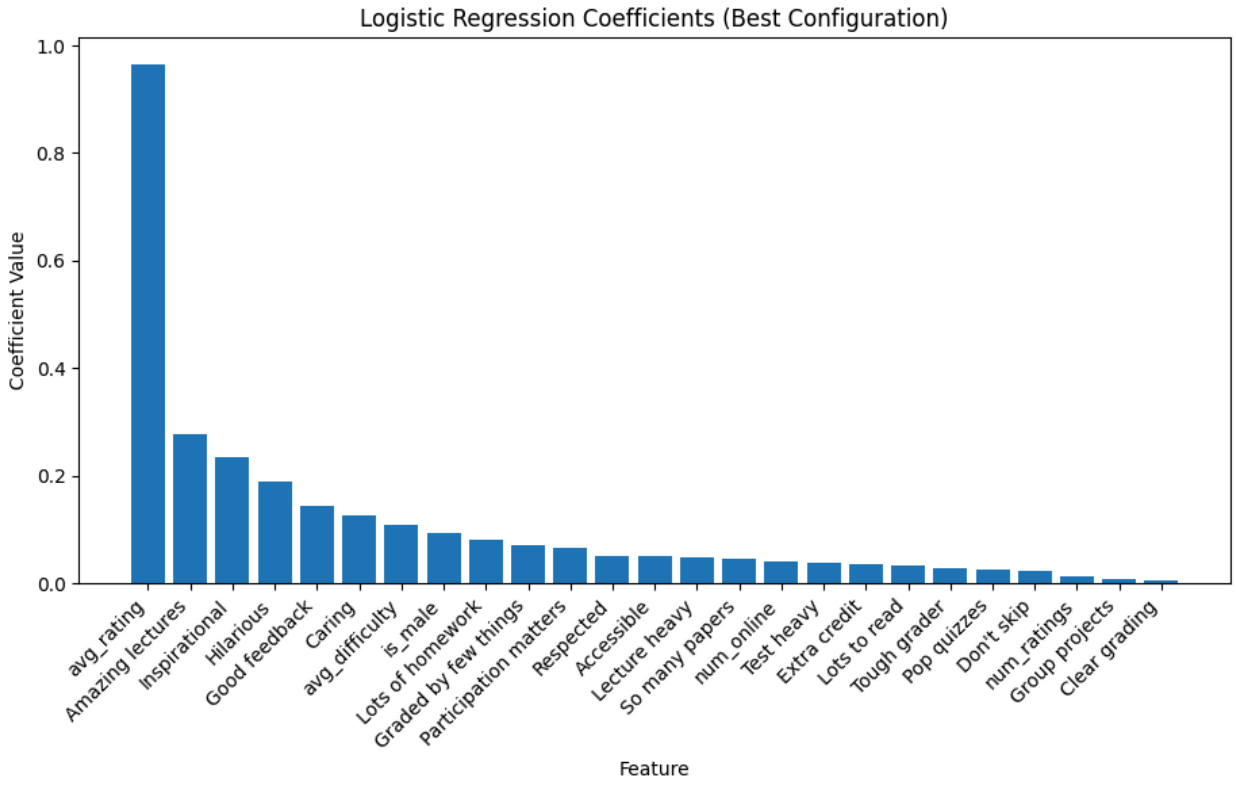

Figure 10 (c)

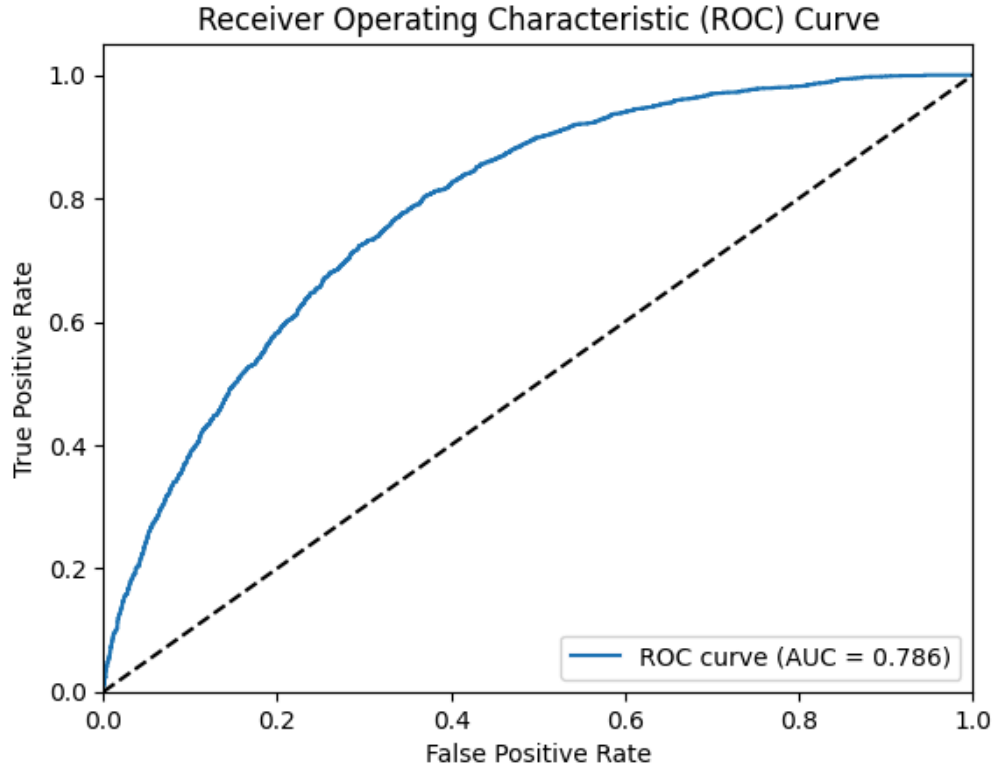

Figure 11 (a)

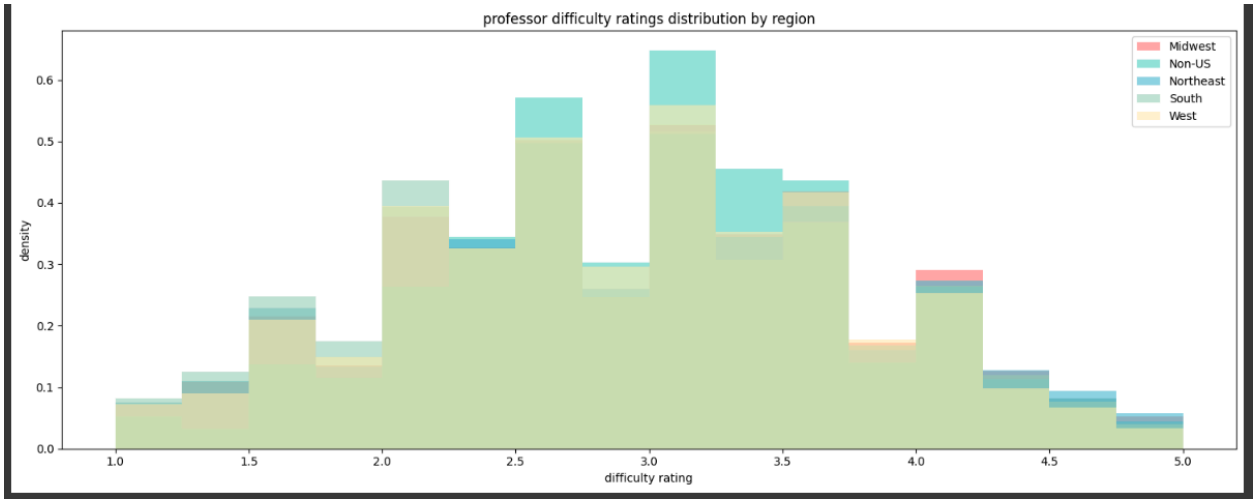

Figure 11(b)

```
median difficulty by region:
region
Midwest      3.0
Non-US       3.0
Northeast    2.9
South        2.8
West         2.9
Name: avg_difficulty, dtype: float64

Sample size by region:
region
South        10483
West          7520
Northeast     5244
Midwest       4959
Non-US        1520
Name: count, dtype: int64
```