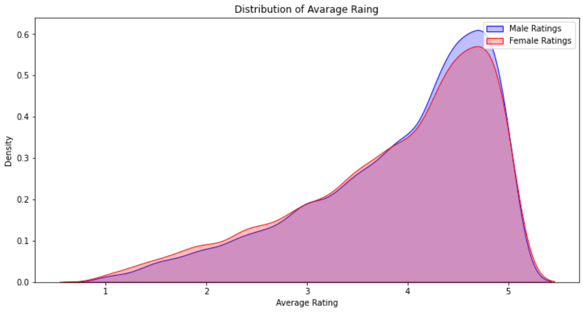
Principle of Data Science: Capstone Project

Data Preprocessing:

I start by adding names to corresponding columns in order to easier retrieve data and checked number of nan value for each column. I noticed that key columns Average Rating, Average Difficulty, Number of Ratings, Number of online ratings, and Received a Pepper all had 19,889 rows missing values and those nan values took place in the same rows. Since these rows lacked meaningful information which we need to analyze in each question, I dropped them. For the column Proportion Take Again, nearly 86% of the values were missing, but other columns of same rows contained valuable data. I filled the missing values of "Proportion take again" with the median to prevent the loss of valuable data in the other column of the same row and ensure no error arose in the analysis (In some questions, the value of "Proportion take again" will be filtered to use the original value).

To mitigate the influence of extreme average ratings, I set a threshold of 5 for the number of ratings by the reasons as follow. I do so because accepting all data will result in extreme number of ratings like 1, which potentially caused by students' bias. Then, I tested the weighting approach, where the proportion of total ratings for each professor scaled ratings. However, this method introduced confounding effects because the subsequent evaluation of the rating" may be confounded by the additional information from the number of ratings. Therefore, setting a threshold of at least five ratings per professor ensured a more reliable dataset by reducing bias from extreme values while retaining sufficient data for meaningful analysis.

Question 1:



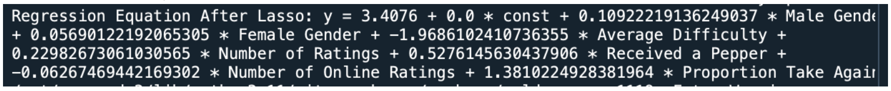
1

To evaluate whether male professors enjoy a boost in ratings due to gender bias, I extracted the male and female professor's Average rating (where the columns male =1 or female = 1, since I noticed there are professor either not indicated being male or female). We evaluate the distribution of "Average Rating" for male and female groups, and as shown on the graph 1, the result of Shapiro-Wilk test (P-value = 0) which confirms the hypothesis that the distribution of two groups is non-normal. Also, combine with the facts that "Average Rating" is ordinal data, two sets are not equal size, and distribution for both group is non-normal, I used the Non-parametric Mann-Whitney U-test, a nonparametric test suited for comparing two groups with potentially unequal sample sizes and non-normal distributions.

The null hypothesis is "there is no difference in average ratings between male and female professors." After running the Mann-Whitney U-test, the test statistic was 50901762.5, and the p-value was 0.00083876, which is below the alpha threshold of 0.005. This result indicates we should drop the null hypothesis a statistically significant difference in average ratings between the two groups, with male professors having higher average ratings than female professors.

However, while the statistical significance of this difference is clear, its interpretation requires careful consideration. Since we cannot guarantee that other factors (e.g., teaching difficulty, number of ratings, etc.) are equally distributed between male and female professors, A significant difference does not necessarily imply gender bias or discrimination. To better understand the contribution of gender relative to other factors, I conducted a multiple regression analysis with "Average Rating" as dependent variable, two Gender as two of the independent variables, and alongside other potential predictors.

Before running this multiple regression, I normalized the scale of each factor to ensure the interpretability and comparability with each other, and I also applied Ridge regression to prevent collinearity.

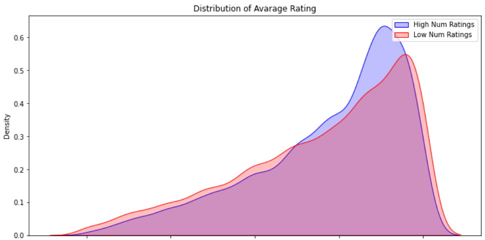


```
Regression Equation After Lasso: y = 3.4076 + 0.0 * const + 0.10922219136249037 * Male Gende
+ 0.05690122192065305 * Female Gender + -1.9686102410736355 * Average Difficulty +
0.22982673061030565 * Number of Ratings + 0.5276145630437906 * Received a Pepper +
-0.06267469442169302 * Number of Online Ratings + 1.3810224928381964 * Proportion Take Agai
```

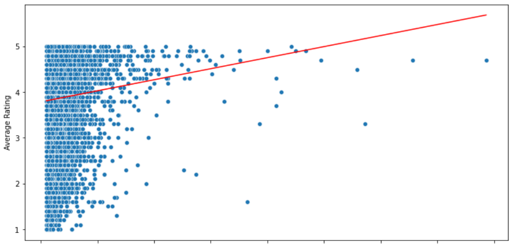
And according to the Regression Equation after normalization and Ridge regression in the graph above, we can notice that the coefficient (0.1092) of Male Gender and (0.0569) of Female Gender on "Average Rating" is minimal compared to other factors, being a male only improved 0.1092 unit in Average rating and being a female only improves (0.0569), their contributions to Average Rating are close). Contrast to the other factors, for example, the negative effect (-1.9686) brings by Average Difficulty, meaning an increase in a unit of Average Difficulty will reduce 1.9686 unit in Average rating, contributes much more to the Average rating.

Thus, it is reasonable to say that though being a male do associate with having a higher "Average Rating" than being a female, but the effect is small compared to the influence by other variables.

Question 2:



3



4

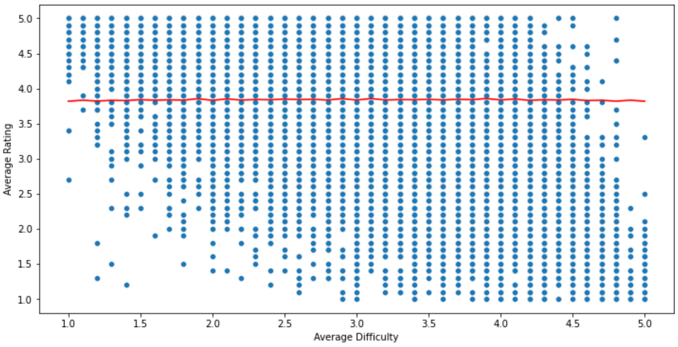
To examine the effect of experience on teaching quality, I used the number of ratings as a proxy for experience and split the professors into two groups based on the median number of ratings. Professors with the number of ratings above the median were classified as the high-experience group, while those below or equal to the median formed the low-experience group. We first state the null hypothesis test that "there is no statistically significant difference in average ratings between the low-experience and high-experience groups." And since the distributions of both groups were non-normal, as confirmed by graph 3 and Shapiro-Wilk tests (p-value = 0), I applied the non-parametric test Mann-Whitney U-test.

The test returned a U-statistic 82,000,665 and p-value 0.00052132 which is below the alpha threshold of 0.005. This result indicates a statistically significant difference in average ratings between the two groups, with experienced professors receiving higher average ratings.

Then, to quantify the strength of this effect, I performed a linear regression between X: “Number of Ratings” and Y: “Average Rating” (I did not do test-train split here as I was not making a predictive model, my goal is making a descriptive Analysis model). The regression yielded a positive slope of 0.0048, indicating that as the number of ratings increases, the average rating slightly improves. However, the effect is not strong as a 100-point increase in the number of ratings results in only a 0.48-point increase in the average rating. This is also reflected in the low R^2 value 0.003, suggesting experience is not a strong predictor of teaching quality.

In conclusion, while there is evidence that experience impacts teaching quality, the actual effect of Number of ratings alone is not strong.

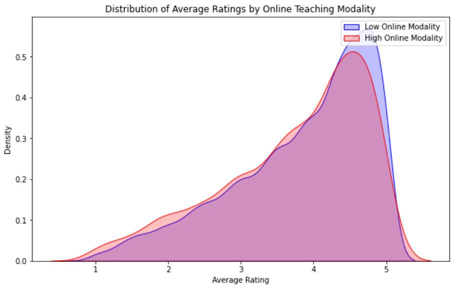
Question 3:



5

To examine the relationship between Average Rating and Average Difficulty, I first conducted EDA to have a sense of the general trend of the data set, and then I began with a linear regression model to quantify the effect of Average difficulty on Average rating. The model shows a correlation coefficient -0.619 meaning they have a moderate strong negative relationship, and the slope negative -0.526992, meaning a change in a unit of Average difficulty will reduce 0.526992 in Average rating. But the regression yielded an R^2 value of 0.3832, indicating that only 38.32% of the variance in Average Rating is explained by Average Difficulty. Also, according to the above scatter plot with fitted regression line, our fitted linear regression line does not capture the pattern well. combining with low R^2 and the plot, I assume there is a non-linear relationship between Average Rating and Average Difficulty and thus apply the Spearman correlation which measures monotonic relationships. The Spearman correlation coefficient was -0.6027, indicating a moderately strong negative monotonic relationship between the two variables. The p-value derived from this model was 0, confirming that this relationship is statistically significant.

Question 4:

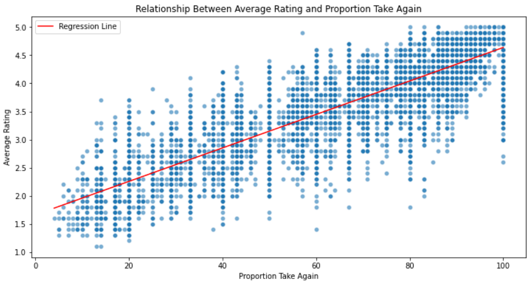


6

To examine whether professors who teach more online classes receive higher ratings, I first split the dataset into two groups based on the median number of online ratings. Noticeably, the median value was 0, meaning a large proportion of professors had no online ratings and I found out that only 10.59% of professors we tested received at least 2 online rating. Based on this observation, it is reasonable to define professors with at least 2 Number of Online Ratings were classified as teaching "a lot" of online classes (High online modality group) as they are small portion of total number of professors we tested, and to create a meaningful contrast, those with Number of Online Ratings 0 were classified to teaching "few" online classes (Low online modality group) as making the group less likely to include professors with moderate number of online rating (1).

Next, we assume the null Hypothesis that "there is no difference in the average ratings between professors who teach more online classes and those who teach fewer or no online classes." And I choose the Mann-Whitney U-test because the both groups do not have equal sample size and failed the Shapiro-Wilk test for normality with p_value 0.0 indicating the non-normal distribution; the distribution graph 6 also confirms this fact. Since we are comparing whether one group tends to have systematically higher or lower ratings rather than examining the overall shape of the distributions, the Mann-Whitney U-test, a non-parametric test, is suitable for this scenario. The test returned a U-statistic of 29199517.5 and a p-value of 8.911e-09, which is below the alpha threshold of 0.0025 (this is a two-tail test – either lower or higher, so divide 0.005 threshold by 2). And then I found out the median Average Rating for professors teaching less online classes was higher (Median = 4.1) compared to those teaching more online classes (Median = 4.0). This confirms the alternative hypothesis that "there is a difference in ratings between professors who teach more online classes and those who teach fewer or none; specifically, professors who teach less online classes receive higher Average Rating than the professor group who teach more than one online class, suggesting teaching less online courses is associated with higher student ratings.

Question 5:

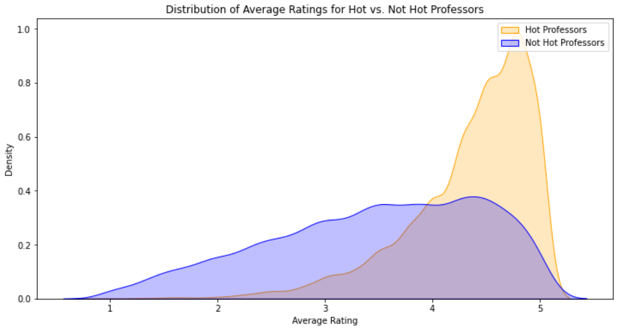


7

To find the relationship between Average Rating and the Proportion Take Again, we first revisited the column Proportion Take Again. During preprocessing, missing values in this column were filled with the median; however, for this analysis, I used the original, non-imputed data to ensure accuracy. Then I dropped the rows with missing values in either “Proportion Take Again” or “Average Rating” since we only investigating the relationship between those.

Starting analysis, I first conducted EDA to have a sense of the general pattern of the data set, and I it is very clear that data is following and positive linear relationship. To test my assumption, I applied a linear regression model with “Proportion Take Again” as the independent variable (X-axis) and Average Rating as the dependent variable (Y-axis). The results indicated an R^2 value of 0.77502, suggesting that the linear model explains 77.502% (majority) of the variance in “Average Rating,” and the correlation coefficient 0.8804 meaning there exist a strong positive relationship between “Proportion Take Again” and “Average Rating.” The regression coefficient (slope) was 0.0298, meaning that a one percentage increase in the proportion of students willing to take the class again is associated with an increase of 0.0298 in the “Average Rating.”

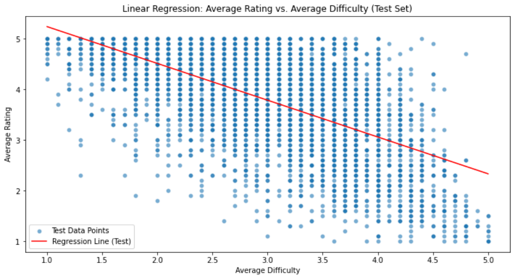
Question 6:



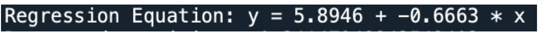
8

To test whether “hot” professor receive higher “Average Rating” compares to those who are not, I split the data base into two groups: hot professors: those who received a "pepper" (Received a Pepper=1); not hot professors: those who did not receive a "pepper" (Received a Pepper=0). Then I state the null hypothesis that the distributions of ratings for 'hot' and 'not-hot' professors are the same in terms of central tendency, meaning there is no systematic difference in ratings between these two groups." To test this, I assess the normality of two group by Shapiro-Wilk test (p_value =0) and visualized their distributions using density plots 8; the result indicated two group are not normally distributed, so I use the Mann-Whitney U-test to determine if the “hot” professor group has a higher average rating than the “not hot” professor group. The test returns the U-statistic of 121910581.5 and a p-value of 0.0, which is significantly below the alpha threshold of 0.005. Additionally, the median average rating for the hot professor group was 4.5, compared to 3.6 for the not-hot group. Therefore, the result provides strong statistical evidence to drop the null hypothesis and support the alternative hypothesis that “there is a difference in Average rating between the “hot” professor group and the not hot professor group where the hot professor tends to have a higher Average rating than the not hot professor.

Question 7:



9

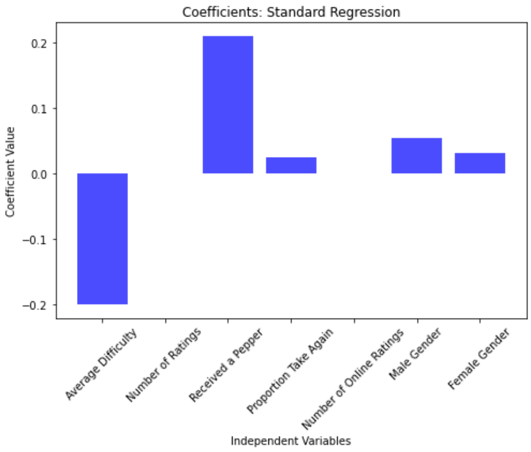


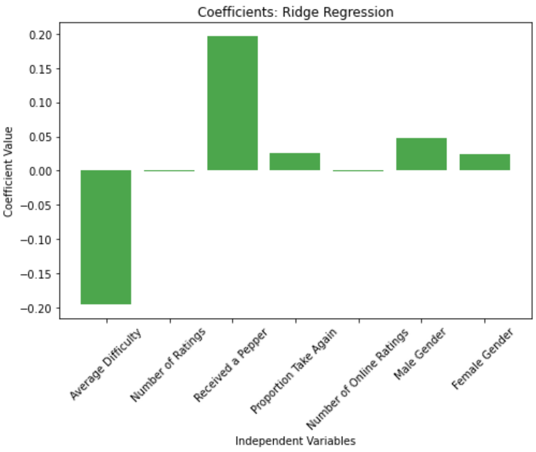
```
Regression Equation: y = 5.8946 + -0.6663 * x
```

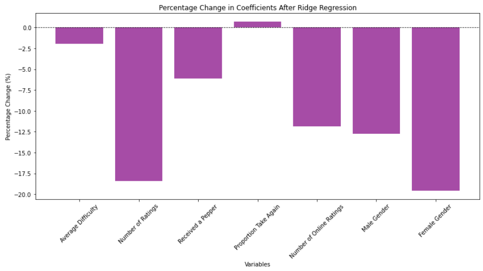
To build a regression model predicting Average Rating from Average Difficulty, I readjusted data set by only include the professor having value on all factors (drop Nan base on question 8); now, two questions are analyzed by the same data set, and this ensures the comparability. Then I set Average Difficulty as the independent variable (X) and Average Rating as the dependent variable (Y).

Since the goal is to create a predictive model, we need to ensure the model generalizes well and is not overfitting; thus, I split the dataset into a training set (80%) and a test set (20%). Then I trained the Linear regression predictive model on the training set and evaluated its performance on both the training and test sets (Analysis will be according to the test group). The model returns a coefficient -0.6663, meaning each unit increase in Average difficulty will result 0.6663 decrease in Average rating. The model also yields RMSE 0.6860 for the training set and 0.6875 for the test set, which are very close, indicating the model generalizes well to unseen data and is not overfitting. We take the RMSE of the test group to explain its meaning, and this suggest on average the predicted Average Rating deviates approximately 0.6875 score from the actual Average Rating on a scale of 1 to 5. The R^2 values of the training set (0.3415) and the test set (0.3454) were low, meaning Average Difficulty only explains small portion, around 34.5%, of variance in Average Rating. The plot above also suggested this single factor regression line is not capturing the pattern well, and with those reason combined, I was incentivized to consider "Average Difficulty" alone is not a strong predictor of Average Rating. This finding suggests the need to include additional independent variables to improve the model's explanatory performance.

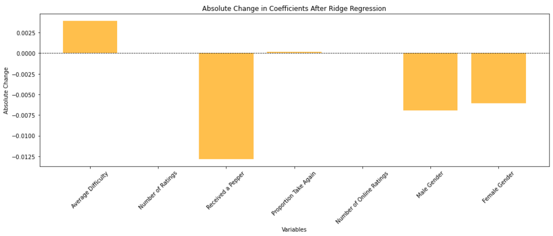
Question 8:



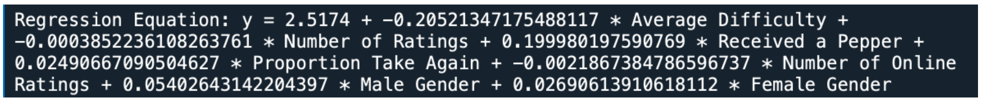




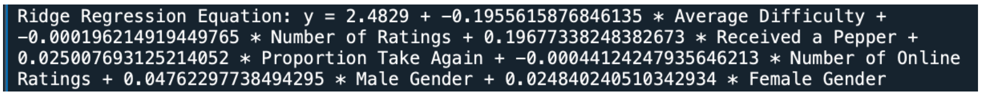
11



12



```
Regression Equation: y = 2.5174 + −0.20521347175488117 ∗ Average Difficulty +
−0.0003852236108263761 ∗ Number of Ratings + 0.199980197590769 ∗ Received a Pepper +
0.02490667090504627 ∗ Proportion Take Again + −0.0021867384786596737 ∗ Number of Online
Ratings + 0.05402643142204397 ∗ Male Gender + 0.02690613910618112 ∗ Female Gender
```



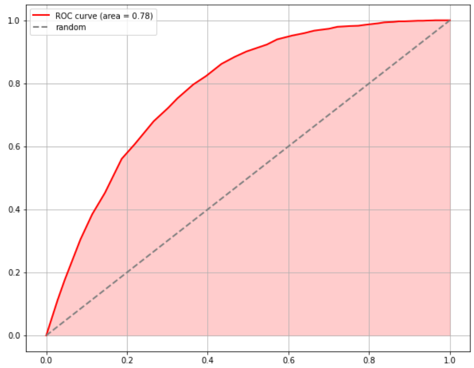
```
Ridge Regression Equation: y = 2.4829 + −0.1955615876846135 ∗ Average Difficulty +
−0.000196214919449765 ∗ Number of Ratings + 0.19677338248382673 ∗ Received a Pepper +
0.025007693125214052 ∗ Proportion Take Again + −0.00044124247935646213 ∗ Number of Online
Ratings + 0.04762297738494295 ∗ Male Gender + 0.024840240510342934 ∗ Female Gender
```

To build a regression model predicting average ratings, I included "Average Difficulty," "Number of Ratings," "Received a Pepper," "Proportion Take Again," "Number of Online Ratings," "Male Gender," and "Female Gender" as independent variables (X) and "Average Rating" as the dependent variable (Y) using the same data set as Q7. To prevent overfitting, I split the data into training (80%) and testing (20%) sets.

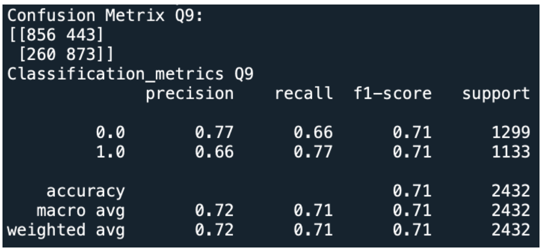
After training a linear regression model on the training set, I evaluated its performance and observed potential collinearity issues among predictors. Since I believe all factors are relevant, I employed Ridge regression (L2 regularization) which penalizes large coefficients and thereby reducing the impact of collinearity. I plotted coefficients before and after the Ridge regression in graph 10 above. Also, since the coefficients in the standard regression model are small in magnitude, the percentage and absolute changes in coefficients by Ridge regression (visualized in the graphs above) shows by the graph 11&12 offer a clearer perspective. Significant reductions in the coefficients for "Female Gender," "Male Gender," "Number of Ratings," and "Number of Online Ratings" suggest these predictors were highly collinear with other variables, diminishing them reduces collinearity and retaining the predictive power of model.

We calculated the R^2 and RMSE for both the training and test datasets. The R^2 values were 0.8101 for the training group and 0.8096 for the test group, while the RMSE values were 0.3707 and 0.3612, respectively. The close alignment of the R^2 and RMSE between the training and test groups suggests that the model generalizes well to unseen data and is not overfitting. Compared to the single-factor model (Average Difficulty as the sole predictor) in Question 7, where the test group's R^2 was 0.3454, the R^2 of the multiple regression model increased significantly to 0.8096. This indicates that the multiple regression model explains 80.96% of the variance in Average Rating, demonstrating incorporating additional factors significantly improves explanatory power. The RMSE of the test group decreased from 0.6875 in the single-factor model to 0.3612 in the multiple regression model. This reduction indicates that the predictions of the multiple regression model are much closer to the actual Average Rating values and also suggests improved predictive performance of the model. On a scale of 1 to 5, the predicted Average Rating deviates by only approximately 0.3612 from the true values.

Question 9:



[13]



```
Confusion Metrix Q9:
[[856 443]
 [260 873]]
Classification_metrics Q9
              precision    recall  f1-score   support

         0.0       0.77      0.66      0.71      1299
         1.0       0.66      0.77      0.71      1133

    accuracy                           0.71      2432
   macro avg       0.72      0.71      0.71      2432
weighted avg       0.72      0.71      0.71      2432
```

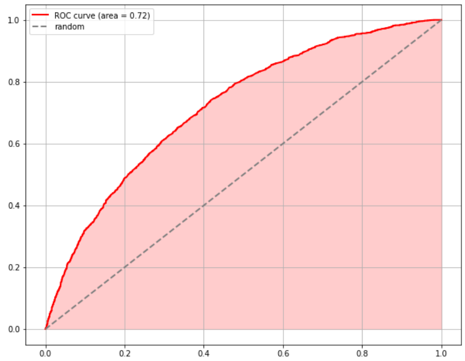
[14]

To predict whether a professor “receives a pepper" (1) or not (0) based solely on their average rating, I built a classification model using logistic regression, which is appropriate for binary classification tasks. Like Question 7 and 8 I readjusted data set by only include the professor having value on all factors (drop nan base on question 10); two questions are analyzed by the same data set, ensuring the comparability. To prevent overfitting and ensure the model generalizes well to unseen data, I split the dataset into an 80% training set and a 20% test set. The dependent variable ("Received a Pepper") is imbalanced, with 53.40% of professors not receiving a pepper (0) and 46.60% receiving one (1). To address this imbalance, I employed Synthetic Minority Oversampling Technique (SMOTE), which resampled the training set to equalize the proportions of the two classes and balance the classes to 50% each.

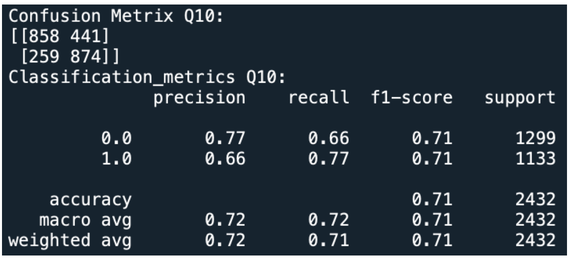
The logistic regression model was trained on the balanced dataset and evaluated on the test set. The model's performance was assessed using several key metrics, including the Area Under the Receiver Operating Characteristic Curve (ROC AUC), accuracy, precision, recall, and F1-score. The model achieved a test ROC AUC of 0.7807, indicating the model has a 78.07% chance of correctly distinguishing between a randomly chosen professor who received a pepper and one who did not. The training set ROC AUC (0.7880) is close to the test group ROC AUC (0.7807), suggesting the model generalizes well without overfitting. For professors predicted to receive a pepper (Class 1), the precision was 0.66, meaning 66% of those predicted as pepper-receivers were correctly classified, while 34% were false positives. The recall for Class 1 was 0.77, showing the model successfully identified 77% of actual pepper-receivers, missing only 23%. For professors predicted not to receive a pepper (Class 0), the precision was 0.77, meaning 77% of those predicted not to receive a pepper were correctly classified, with 23% misclassified. The recall for Class 0 was 0.66, showing the model correctly identified 66% of professors who did not receive a pepper, leaving 34% as false negatives. Overall, the model achieved an accuracy of 0.71, correctly predicting 71% of all cases.

Conclusively, the model achieves a high ROC AUC and balanced precision and recall values. It shows good predictive ability for identifying pepper-receivers based on average ratings.

Question 10:



15



```
Confusion Metrix Q10:
[[858 441]
 [259 874]]
Classification_metrics Q10:
              precision    recall  f1-score   support

         0.0       0.77      0.66      0.71      1299
         1.0       0.66      0.77      0.71      1133

    accuracy                           0.71      2432
   macro avg       0.72      0.72      0.71      2432
weighted avg       0.72      0.71      0.71      2432
```

16

To predict whether a professor “receives a pepper" (1) or not (0) based on all available factors, I built a classification model using logistic regression, as it is well-suited for binary classification problems. The independent variables included "Average Rating," "Male Gender," "Female Gender," "Average Difficulty," "Number of Ratings," "Number of Online Ratings," and "Proportion Take Again." The dependent variable was "Received a Pepper."

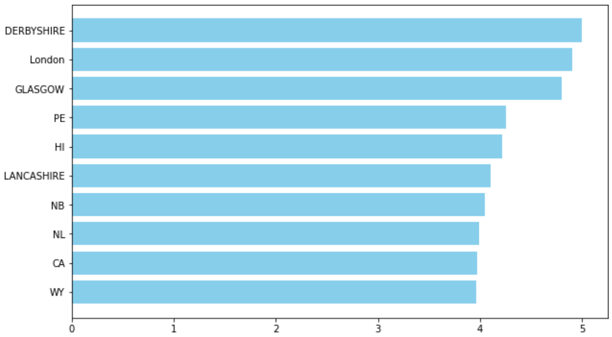
Similar to Question 9, I first split the dataset into an 80% training set and a 20% test set to prevent overfitting and ensure the model generalizes well to unseen data. The dataset displayed class imbalance, with 53.40% of professors not receiving a pepper and 46.60% receiving one. To address this imbalance, I applied Synthetic Minority Oversampling Technique (SMOTE) to the training set, balancing the two classes to 50% each.

The logistic regression model was trained on the balanced dataset. The model evaluated on the test set using metrics such as the Area Under the Receiver Operating Characteristic Curve (ROC AUC), precision, recall, and F1-score. The model achieved a test ROC AUC of 0.7827, indicating a 78.27% probability of correctly distinguishing between a professor who received a pepper (1) and one who did not (0). The test ROC AUC is slightly higher than the “Average Rating” only model (0.7807), suggesting the inclusion of additional factors slightly improves predictive power while reaffirming that “Average Rating” remains the most significant factor in this classification model.

For professors predicted to receive a pepper (Class 1), the precision was 0.66, meaning 66% of those predicted as pepper-receivers were correctly classified, while 34% were false positives. The recall for Class 1 was 0.77, showing the model successfully identified 77% of actual pepper-receivers, missing 23%. For professors predicted not to receive a pepper (Class 0), the precision was 0.77, meaning 77% of those predicted not to receive a pepper were correctly classified, with 23% misclassified. The recall for Class 0 was 0.66, showing the model correctly identified 66% of professors who did not receive a pepper, leaving 34% as false negatives. Overall, the model achieved an accuracy of 0.71, correctly predicting 71% of all cases.

In conclusion, the logistic regression model incorporating all factors achieved slightly better predictive power (ROC AUC: 0.7827) compared to the single-factor mode. The model demonstrates balanced precision and recall and generally having a good predictive power of whether a professor receives a pepper.

Extra Credit:



For extra credit, I investigate whether there are significant differences in "Average Rating" between states. First, I merged the qualitative and numerical datasets and grouped the data by state to compute the mean "Average Rating" for each state. The resulting bar chart suggests potential differences in average ratings among states, as some states exhibit clear gaps in their mean ratings.

However, further exploration revealed that some states had very small sample sizes, with as few as one or two records. I realized it is not reasonable to reduce the data to sample means. To address this, I employed the non-parametric Kruskal-Wallis test, which is appropriate for comparing more than three groups when the data may not meet the assumptions of normality or equal variances.

The null hypothesis states that there is no difference in "Average Rating" among the states. The test yielded a statistic of 227.24 and a corresponding p-value of 2.66e-19, which is significantly smaller than the alpha value of 0.005. Therefore, we drop the null hypothesis and conclude that there is strong statistical evidence of the existence of differences in "Average Rating" between states.