Homework 2

Nisha Ramanna

April 4 2024

1 Problem 1

For this section, I watched the video "Algorithmic Transparency in the Public Sector". One responsible AI concern discussed in this lecture was the need for algorithmic transparency in the public sector. Specifically, this is needed to ensure transparency for decisions which affect individuals. There are multiple stakeholders discussed in this lecture. Mainly, the lecture focuses on how this issue affects government organizations and the general public. The lack of transparency of transparency can cause the general public to have

A lack of transparency in public sector algorithms means the general public is unsure how important decisions affecting their lives are being made. This could lead to a lack of trust in the government by the general public. Algorithms in the public sector which don't have a transparent decision-making process are considered black boxes.

The public sector has many incentives to be transparent. For one, the woman giving this lecture works for the UK government which is currently passing laws requiring some level of transparency for several AI models. Therefore, they may be legally required to ensure transparency. Beyond legality, the public sector

also has an ethical incentive to ensure transparency. By showing a commitment to interpret ability, the public sector could improve the public's trust in their automated decision-making processes.

2 Problem 2

2.1 Part A

Female and non-binary individuals will be harmed by this imputation method. That is because the mean value of experience in years for the overall population is lower than the mean value for specifically female and non-binary individuals. This will lead to some female and non-binary applicants appearing less experienced than they probably are.

2.2 Part B

Instead of imputing null values with the overall mean value for the experience feature, impute it with the mean value for the applicant's specific group (sex + gender). This way, certain applicants won't appear more/less experienced that they likely are.

2.3 Part C

The technical bias introduced by the data imputation method relates to pre-existing bias as it negatively effects female and non-binary individuals. This is a group that has historically been discriminated against, especially in hiring practices. Therefore, the technical bias against them, which leads to a lower chance of them being hired, reinforces this pre-existing bias.

The technical bias also leads to emergent bias as disadvantages the female and non-binary applicant group, causing them to be hired less, and leading

to further disparities between the number of males and female and non-binary individuals in the workforce.

3 Problem 3

3.1 Part A

For section 3, I imported the 20 newsgroup dataset. I then performed some preprocessing on it. Finally, a fit a TF-IDF vectorizer to the data and used the transformed data to train an SGDClassifier.

3.2 Part B

Below is the confusion matrix generated using the predictions of the SGD-Classifier:

TP: 276	FP: 43
FN: 4	TN: 394

I then chose 5 random documents (120, 331, 39, 636, 514) and generated SHAP explanations for them. They are displayed below:

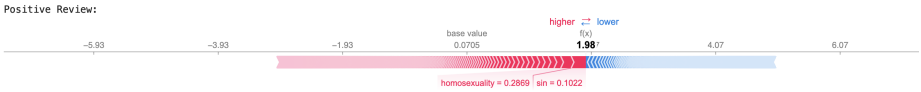

Figure 1: SHAP Explanation for Document 120

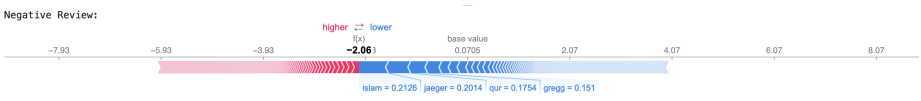

Figure 2: SHAP Explanation for Document 331

Figure 3: SHAP Explanation for Document 39

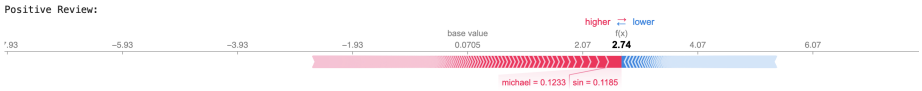

Figure 4: SHAP Explanation for Document 636

Figure 5: SHAP Explanation for Document 514

3.3 Part C

In this section, I analyzed my SGDClassifier. It had an accuracy of 93% and had misclassified 47 documents.

Then, I calculated the confidence error (conf_i) for each misclassified document. Below is the chart showing the distribution of conf_i for all misclassified documents:

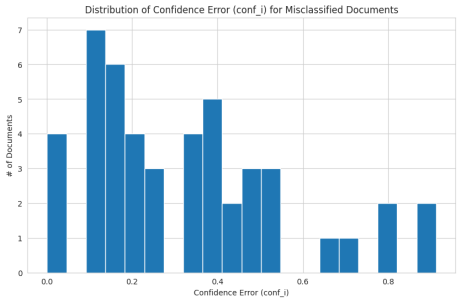

Figure 6: Distribution of Confidence Error (conf_i) for Misclassified Documents

I calculated two values for each word which contributed to the misclassification of a document. First, I computed the number of documents it helped misclassify (count_j). Then, I calculated the total weight of the word in all documents it helped misclassify (weight_j). Below is the distribution of count_j for all words that helped misclassify a document:

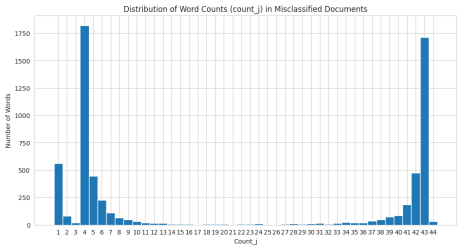

Figure 7: Distribution of Word Counts (count_j) in Misclassified Documents

Below is the distribution of weight_j for all words that helped misclassify a document:

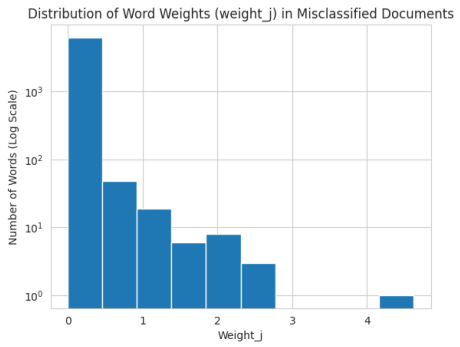

Figure 8: Distribution of Word Weights (weight_j) in Misclassified Documents

3.4 Part D

In part D, I was tasked with implementing a feature selection method to improve the accuracy of my SGDClassifier. Knowing SHAP values measure the

magnitude of a features effect on the classification of a document, I chose to sum the SHAP values for each feature accross all documents to get an idea of a features importance in prediction. I then chose to keep only the top 25% most important features. I retrained my SGD classifier using only these most influential features. My accuracy improved from 93.4% to 93.6% (wow!). Additionally, the number of misclassified documents decreased from 47 to 46. Specifically, document 20 was previously misclassified but was corrected after the feature selection process. Below is the SHAP explanation for this document before feature selection occurred:

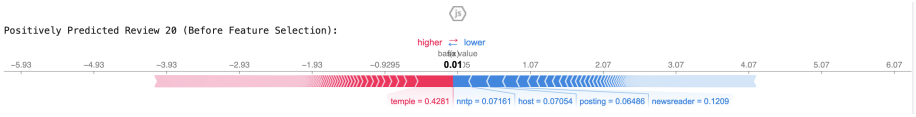

Figure 9: Document 20's SHAP Explanation Before Feature Selection

Here is the SHAP explanation for the document after feature selection:

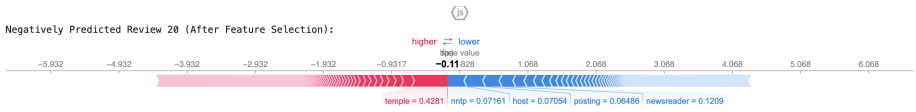

Figure 10: Document 20's SHAP Explanation After Feature Selection

Observing these two SHAP explanations, we can see that the five most important features in predicting document 20 remained the same before and after feature selection. Additionally, their contribution to the prediction stayed the same. However, the output changed from .01 to -.11. This is potentially because the interactions between these features has changed after feature selection, leading to a different end prediction.

4 Problem 4

4.1 Part A

The individual at position 100 was individual 56. According to scoring function score_function_WKHP, this individual was ranked 351. According to scoring function score_function_AGEP, this individual was ranked 223.

I then inspected the importance of all features to the ranking of individual 56 by the three scoring functions by using ShaRP waterfall plots. Here are some definitions for reference:

Feature 0 = Age

Feature 1 = Education Level

Feature 2 = Hours worked per week

Feature 3 = Race

Below is the waterfall plot displaying the individual feature contributions to the ranking of individual 56 by score_function_SCHL:

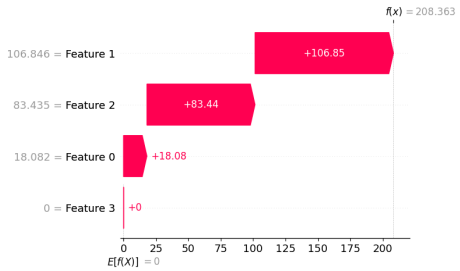

Figure 11: Individual Feature Contributions to Ranking of Individual 56 by score_function_SCHL

Here is the waterfall plot displaying the individual feature contributions to the ranking of individual 56 by score_function_WKHP:

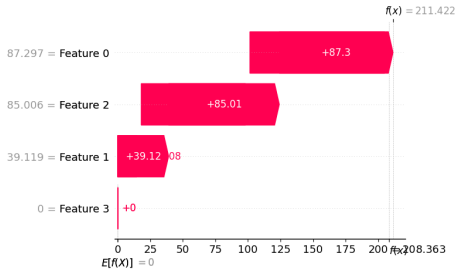

Figure 12: Individual Feature Contributions to Ranking of Individual 56 by score_function_WKHP

Below is the waterfall plot displaying the individual feature contributions to the ranking of individual 56 by score_function_AGEP:

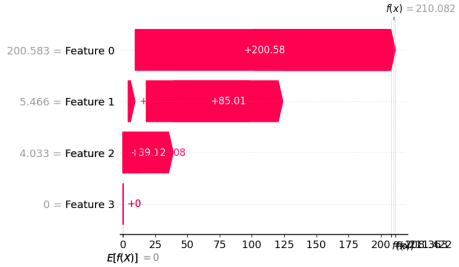

Figure 13: Individual Feature Contributions to Ranking of Individual 56 by score_function_AGEP

Clearly, all features do not contribute to the rankings' quality of information (QoI) equally. For score_function_WKHP and score_function_AGEP, age is the most important feature.

For score_function_SCHL, the feature "education level" is the most important feature. This makes sense as this function ranks individuals based on their education level. For similar reasons, "age" is the most important feature for score_function_AGEP. Confusingly, "age" is also the most important function for score_function_WKHP. "Hours worked per week", the feature this ranking

function is based on, is the second most important feature. However, the difference between these features contribution to the ranking of the individual in question is fairly minimal (2 points), so it doesn't appear that "age" is dominating over the "hours worked per week" when performing ranking for this scoring feature.

The least important feature is always "race", contributing nothing to the ranking of the individual regardless of the scoring function being used. This makes sense as "race" is our sensitive feature.

4.2 Part B

In this section, I first selected the top 20% of applicants and split them into white and non-white.I calculated the feature importance for both groups and visualized them in the box plots below:

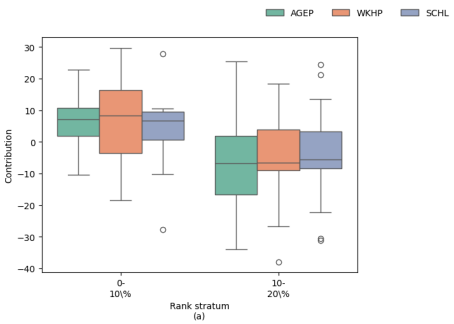

Figure 14: Boxplot Visualizing Feature Importance for White Individuals

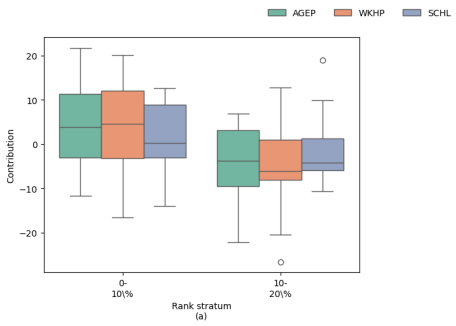

Figure 15: Boxplot Visualizing Feature Importance for Nonwhite Individuals

For white individuals, feature importance is different between the 0-10% and 10-20%. "Age" tends to have a more positive contribution in the top 10%, while it's contribution is usually negative for individuals in the 10-20%. The same is true for both "hours worked per week" and "education level". Based on the median line of the features, it appears that the three have about equal importance as one another in each stratum.

For nonwhite individuals, feature importance similarly differs between the 0-10% and 10-20%. Again, we see that all three features have a strong positive contribution for individuals in the 10-20th percentile, while the features generally contribute negatively for individuals in the 10-20th percentile. In the higher stratum, it appears that "education level" is less important than the other two features. However, in the lower stratum, their importances seem closer to equal.

(iv) Compare feature importances across racial groups. Which features are more / less important for each group? Hypothesize about why feature importance may be different across groups

Across racial groups, "age" and "education level" seem to have similar importances in the higher stratum. However, "education level" appears to be less important for nonwhite individuals than it is for white individuals. This is likely due to existing social circumstances which lead to lower return on investments

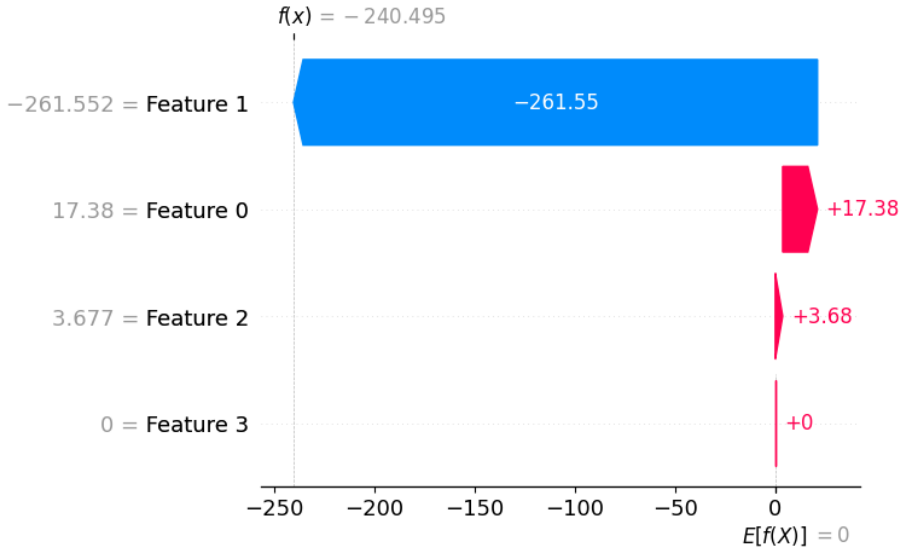

Figure 16: ShaRP Waterfall Plot for Individual Ranked 200

for education for nonwhite individuals. For example, they may have weaker networks than their white counterparts, making it difficult for them to leverage their educational experience for higher paying jobs. All features appear to have similar importances in the lower stratum across racial groups.

4.3 Part C

For the extra credit section, I examined the difference in ranking between the individual ranked 200 vs. 300 by score_function_SCHL. First, I created ShaRP waterfall plots to give insight into the features which contributed to the individual rankings:

Then, I created a graph to visualize why the individual ranked 200 was scored higher than the one ranked 300th. To achieve this, I calculated the difference between the 200th individual's feature importances and the 300th individual's feature importance. I then plotted these differences in the bar graph below:

The feature with the largest positive difference is "age". This indicates that

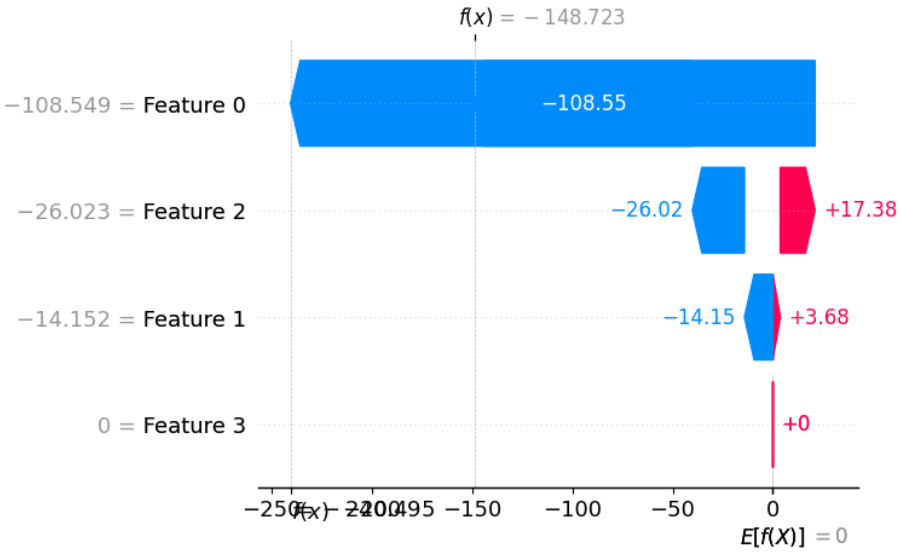

Figure 17: ShaRP Waterfall Plot for Individual Ranked 300

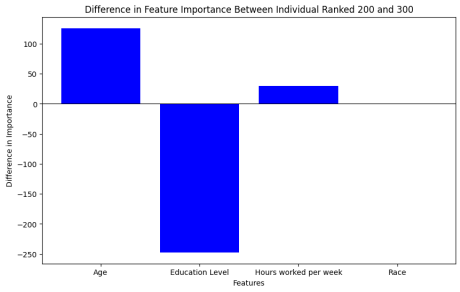

Figure 18: Difference in Feature Importance Between Individual Ranked 200 and 300

the choice to rank the 200th individual above the 300th individual relied heavily on the difference in their "age" feature. Interestingly, the largest negative difference is about double the largest positive difference. "Education level" is the feature associated with this difference. This indicates that despite "education level" contributing far stronger to the 300th ranked individual being classified in a positive manner, the 200th ranked individual still scored higher. This shows us that "education level" overall was not as important of a feature in the classification process as "age".