Question 1

I watched Natalia Domagala's lecture on algorithmic transparency in the public sector.

- The lecture expresses responsible AI concerns related to the use of algorithms in government decision-making, such as potential biases, lack of accountability, and the impact of automated decisions on individuals' rights and well-being.
- There can be various stakeholders impacted by the RAI concerns. For example, government agencies are responsible for implementing and deploying algorithms. Citizens and residents are directly impacted by algorithmic decisions such as social services, law enforcement, etc. Social organizations advocating for transparency and fairness and private companies providing algorithmic solutions to government entities are also impacted.
- Transparency ensures that citizens understand how decisions are made, fostering trust and accountability. Lack of transparency can lead to "black box" systems where the inner workings of algorithms remain hidden. This can hinder accountability and prevent users from challenging decisions. This lecture argues that the push for algorithmic transparency is essentially an effort to open up these "black boxes" of algorithmic decision-making in the public sector, ensuring that the processes and criteria behind automated decisions are clear and understandable to the public. This would build trust and allow for accountability. The lecture also details various measures and standards being implemented, especially in the UK, to promote this transparency.
- Data owners and companies are incentivized to ensure transparency due to legal requirements, public pressure, and ethical considerations. Vendors have a motivation to demonstrate their commitment to ethical service delivery. This helps in building public trust in their systems. These incentives shape the vendor's behavior by encouraging them to adopt transparent practices that not only enhance their reputation and trustworthiness but also contribute to the overall improvement and accountability of algorithmic systems in the public sector. Transparency also allows vendors to identify any potential problems with their tools early on, which can limit future negative impacts.
- Something else: The UK's algorithmic transparency standard aims to standardize information about algorithm use in the public sector.

Question 2

(a)
Given that the male group has a lower mean value than the female group and also has more null values, the impact of Alex's imputation method may give the male group an advantage and the female group a disadvantage. The imputation method will tend to inflate and overestimate the males while deflating and underestimating the females. This inflation may result in the model ranking male applicants higher based on artificially increased experience levels, potentially favoring them over female applicants who may have similar or even higher actual experience levels.

Plus, the fact that the male group has more null values means that a larger proportion of their experience values will be imputed using the overall mean value for the dataset. This means that a larger portion of the male group's experience data will be subject to inflation, further enlarging the inequality.

In summary, the combination of a lower mean value and a higher number of null values means that the male group may be advantaged and the female group may be disadvantaged by Alex's imputation method.

(b)
One alternative data imputation method that may help mitigate the potential unfairness is using a more targeted imputation approach based on relevant subgroups within the dataset. Instead of replacing missing values in the experience feature with the overall mean value for the dataset, we can impute missing values using the mean value of experience within each gender subgroup separately. This method can help preserve the relative experience levels within each gender group more accurately, potentially leading to fairer rankings for both male and female applicants.

(c)
The data imputation method described in (a), where missing values are replaced with the mean value of the respective gender subgroup, can introduce technical bias in MegaSoft's hiring process. This technical bias arises from the imputation process itself and can relate to pre-existing and emergent biases in the hiring example. Pre-existing biases are biases that exist within the dataset or the system before any analysis. In this case, if there are pre-existing biases related to gender in the hiring process or within the dataset itself, such as gender-based discrimination in the hiring process, the imputation method could inadvertently magnify these biases. On the other hand, emergent biases are biases that emerge as a result of the data analysis process itself. In this case, the imputation method could introduce emergent bias by systematically favoring one gender group over the other in the ranking process. For example, if the mean value for one gender is higher than the other due to systemic biases in education or career opportunities, the imputation method could artificially inflate the perceived qualifications of applicants from that gender group, leading to their preferential treatment in the hiring process.

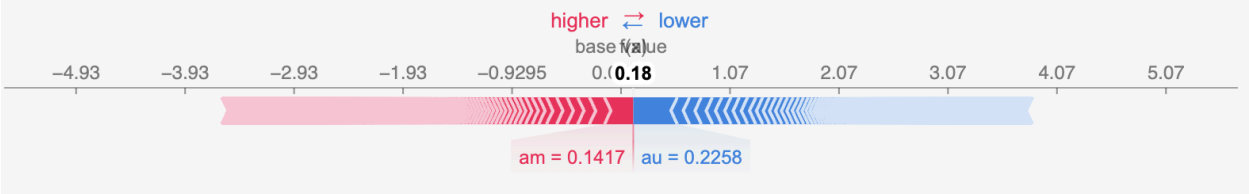
Question 3

(a)
I used the code to fetch the 20newsgroups data. I initialized a TF-IDF vectorizer on the data and transformed and split them into test and train data. I also extracted the feature names out and saved it as a variable for future uses such as graphing SHAP force plots. I then trained an SGD classifier. I set the random state to 1 to stay consistent with the first cell where the random seed is set to 1. I picked the loss function as logistic because SGD classifier works just like a logistic regression and we used logistic regression in our lab as an example. I then fit the classifier on the data and saved the prediction from test data as a variable for future uses.

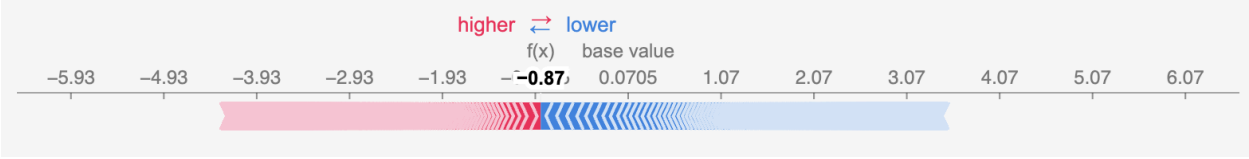
(b)
I generated a confusion matrix using sklearn.metrics.confusion_matrix and printed it out. As we can see, the count of correct Christian is 394. The count of correct Atheist is 276. The count of incorrect Christian (actually Atheist but predicted as Christian) is 43. The count of incorrect Atheist (actually Christian but predicted as Atheist) is 4.

	Actual Atheism	Actual Christian
Predicted Atheism	276	4
Predicted Christian	43	394

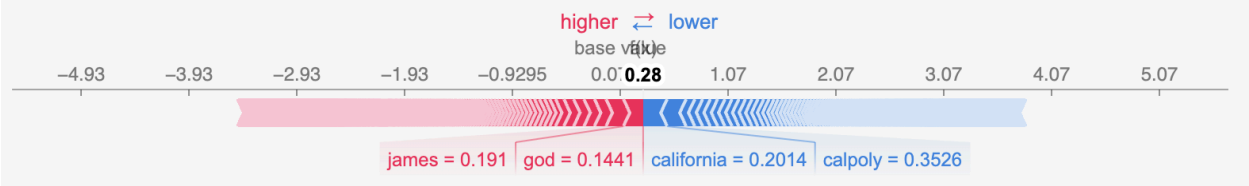
Further, I initialized a SHAP explainer to generate the SHAP values for the test data. I then found one document of each: correct Christian, correct Atheist, incorrect Christian, incorrect Atheist, and a random one. The correct Christian document has both actual class and predicted class as Christian. The index is 0. The SHAP force plot for this document is shown below. There are two significant features: “am” and “au”.



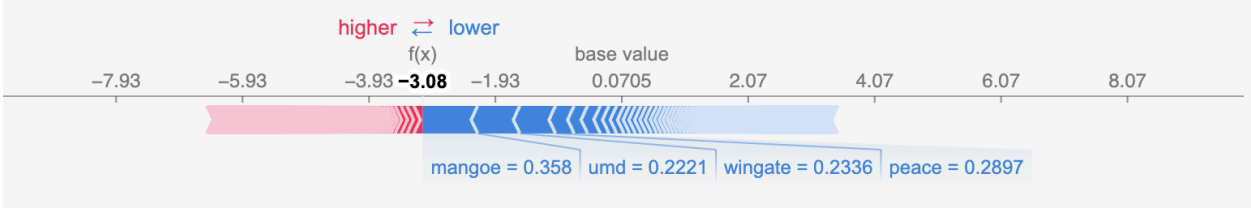
The correct Atheist document has both actual class and predicted class as Atheist. The index is 9. The SHAP force plot for this document is shown below. There is no significant feature in this document.



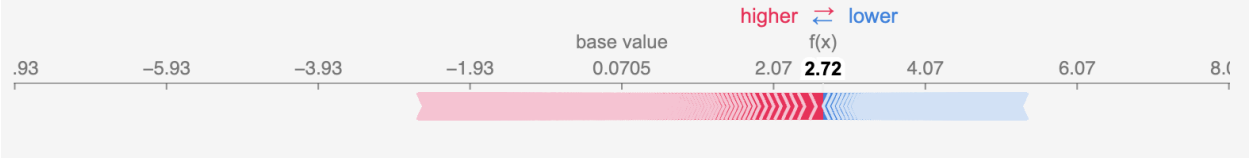
The incorrect Christian document has an actual class of Atheist but is misclassified as Christian. The index is 99. The SHAP force plot for this document is shown below. There are four significant features: “james”, “god”, “california”, and “calpoly”.



The incorrect Atheist document has an actual class of Christian but is misclassified as Atheist. The index is 54. The SHAP force plot for this document is shown below. There are four significant features: “mangoe”, “umd”, “wingate”, and “peace”.



The random one has both actual class and predicted class as Christian. The index is 37. The SHAP force plot for this document is shown below. There is no significant feature.

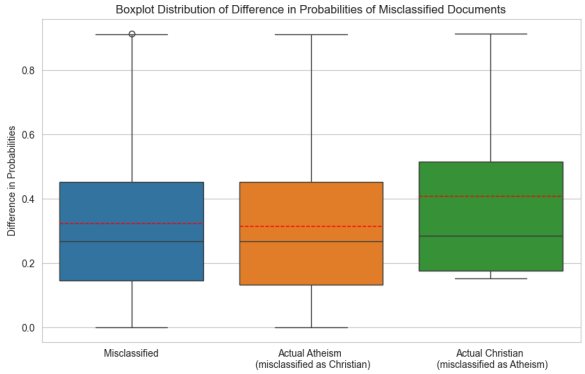


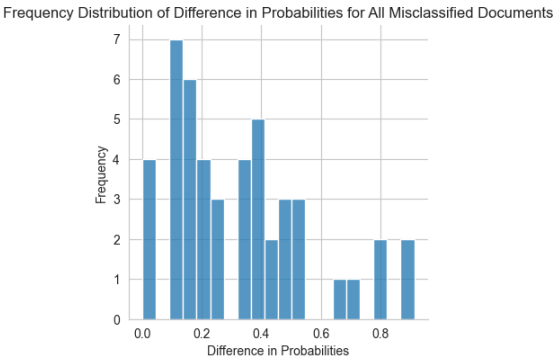
(c)
part (i)
I first reported the accuracy and number of misclassified documents.

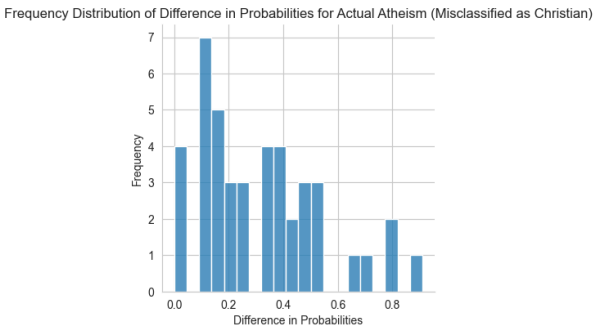
Accuracy: 0.9344490934449093

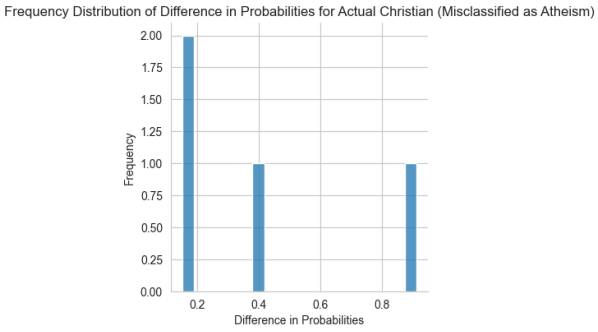
Number of misclassified documents: 47

part (ii)
I found the indices of all misclassified documents and saved it as a list. I then calculated the absolute value of the difference between the predicted probability between the two classes (Atheist and Christian) and saved them into a dictionary. This dictionary has the indices as the key and the absolute differences as the value. I also saved two new dictionaries for each single category of misclassification for drawing boxplots with three boxes for better comparison.





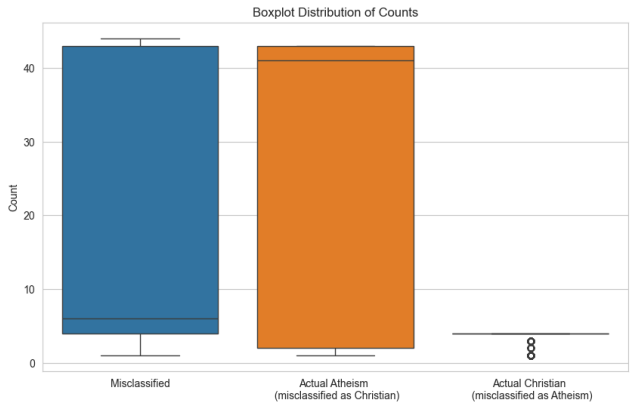


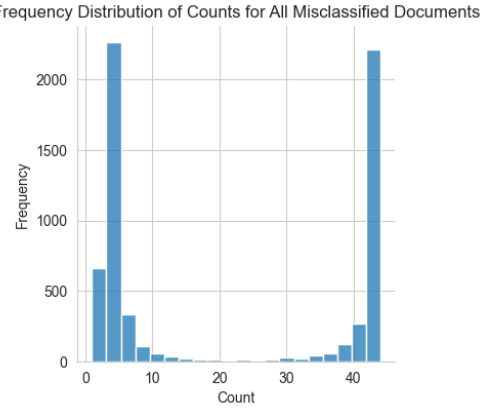


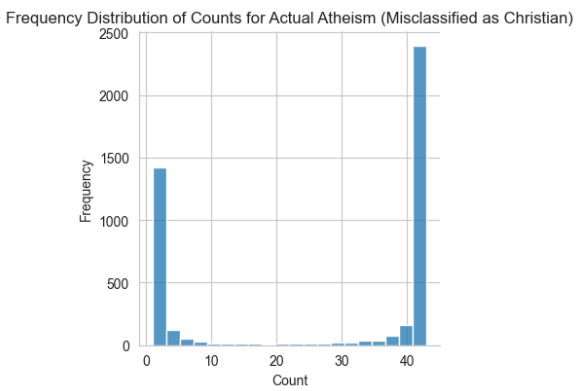
The boxplots show that among all misclassified documents, misclassified as Christian and misclassified as Atheist have a similar distribution. The misclassified as Atheist category has a slightly higher range, mean and median. Looking at the histograms, the misclassified as Christian category follows a similar distribution to overall. However, the misclassified as Atheist doesn't really show any patterns because it only has four values. There are only four documents in this category.

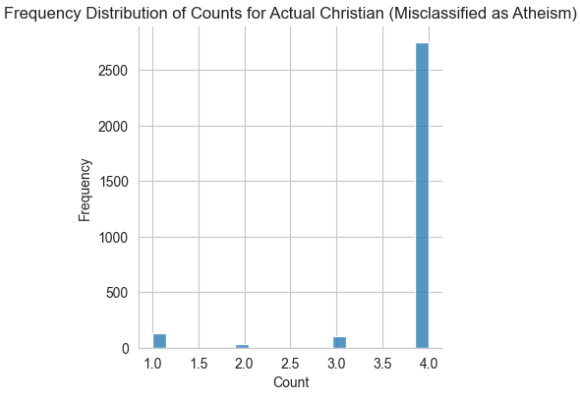
part (iii)
I found all words that contribute to the misclassification of documents. The specific logic is: for all misclassified documents, check if its actual class. If its actual class is Atheist, then any words that have a positive SHAP value is considered contributing to misclassification. If its actual class

is Christian, then any words with SHAP value negative. Like part (ii), I created three dictionaries to hold these info. The key is the word. The value is a list with the first element as the number of appearances of that word and the second element as the sum of the absolute SHAP values of that word. These two elements correspond to the count and weight we are observing here.

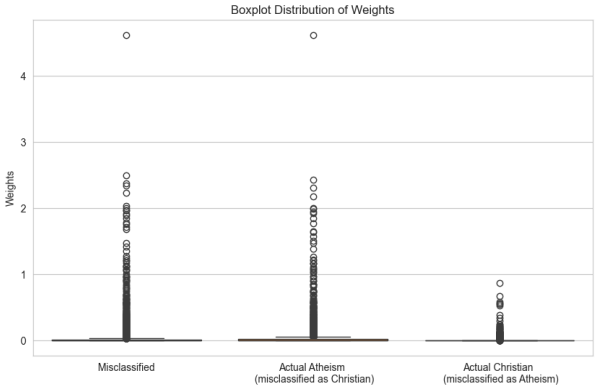


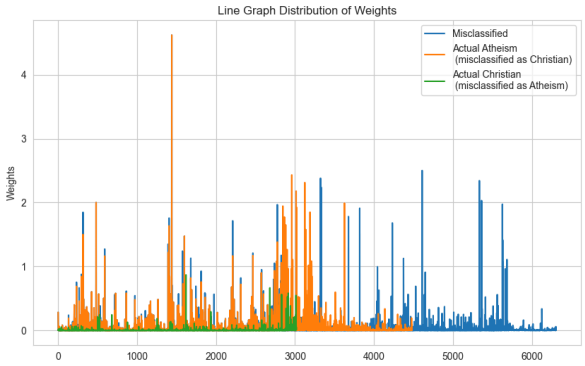


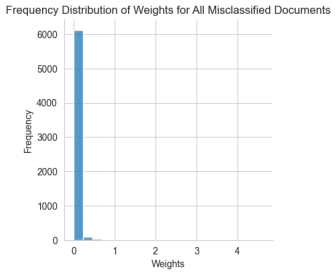


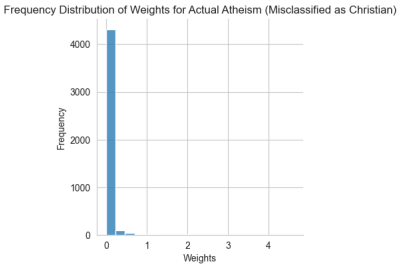


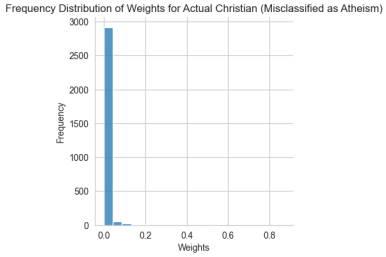
For the counts, we can see from the box plots that misclassified as Christian category has a similar spread to overall. Misclassified as Atheist category is very tightly spread and is close to 0. The difference in the median value for overall and misclassified as Christian shows this too. Because overall is a combination of both, the low median of misclassified as Atheist category together with the high median of misclassified as Chrisitian category make up for the overall median in the middle. In the histogram, we can validate this as well. The overal distribution and the misclassified as Christian category distribution are similar, following a U shape, which means that most words either appear a very small number of times or a very large number of times in contributing to misclassification. In other words, either a word barely contributes, or it contributes a lot. What's also worth noticing is the distribution of misclassified as Atheist. There are 2500+ words that appear 4 times in misleading to the prediction as Atheist. This can be seen from the overall distribution. The second left column that's abnormally high (not following the U shape pattern) is because of these words.











For weights, both the boxplots and the histograms show that all three follow the same pattern: most words have small sums of weights. The dots in the boxplots show the outliers. In this context, they are the words that have significant weights and thus have a significant impact on misclassifying the documents. The line graph shows that the misclassified as Atheist category has a smaller distribution than the misclassified as Christian category. It has both a smaller number of words and a smaller sum of weights. There are fewer high weights in this misclassified as Atheist category. The histograms also show this as we can see that, although all three have the same shape, the x-axises have different values. The misclassified as Atheist category have smaller weights.

(d)

My feature selection strategy is as follows. I first selected the top 10 words that have the largest sum of weights from part (c) and recorded their indices. I filtered the train and test data based on what indices to keep and what indices to remove. I also filtered the features. I fit the SGD classifier, saved the prediction, and calculated the accuracy again using the filtered data.

Accuracy: 0.9400278940027894

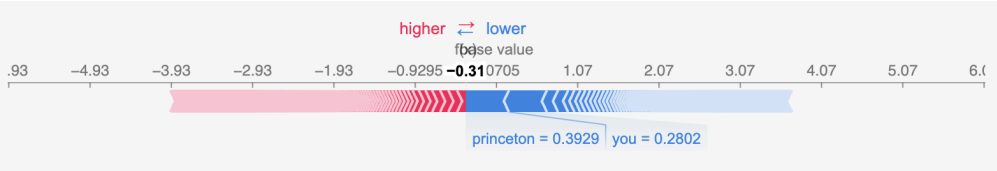
The accuracy improved from 0.934 to 0.94, a 0.006 improvement after my feature selection strategy. I also found two examples that are misclassified before and are correct after my feature selection strategy. The first one is:

Index: 62

Actual: christian

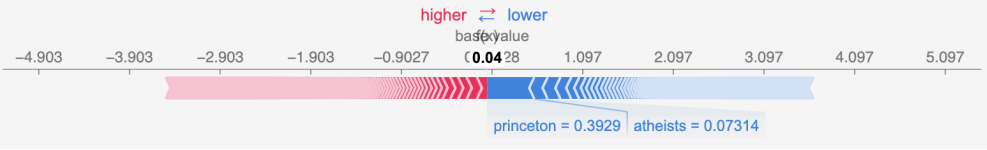
Before Feature Selection

Predicted: atheism



After Feature Selection

Predicted: christian



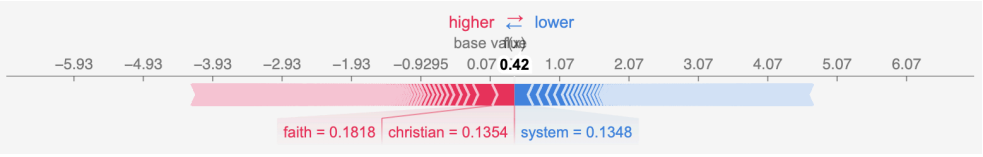
As we can see, before feature selection, it is predicted as Atheist but is actually Christian. The two significant features are “princeton” and “you”. After feature selection, it is predicted correctly as Christian and the two significant features become “princeton” and “atheists”. In both before and after, both features contribute to predicting to be Atheist. Therefore, my feature selection method is not perfect because there are still features left that contribute to misclassification. However, even though it’s not perfect, it already increase the accuracy by half percent and this example is already a correct classification. The second example is a stronger illustration of how my feature selection helps improve:

Index: 197

Actual: atheism

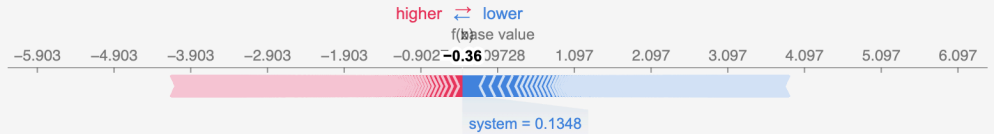
Before Feature Selection

Predicted: christian



After Feature Selection

Predicted: atheism



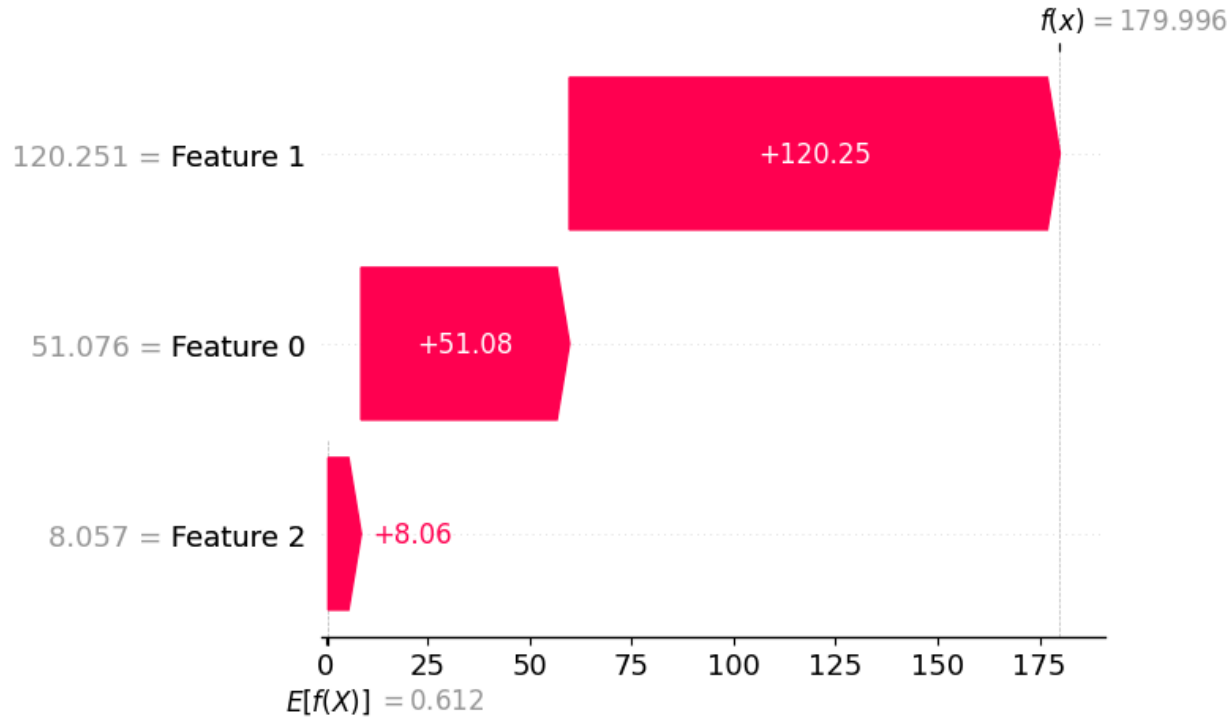
In this example, before feature selection, it is predicted as Christian but is actually Atheist. There are three significant features: “faith”, “christian”, and “system”. The first two contributes to predicting Christian (misclassification) and the last contributes to predicting Atheist. After feature selection, it is predicted correctly as Atheist and three significant features become one. The two words (“faith” and “christian”) that contribute to misclassification are now gone probably because I removed them in my feature selection. The classification is now correct and the significant feature is more normal.

Question 4

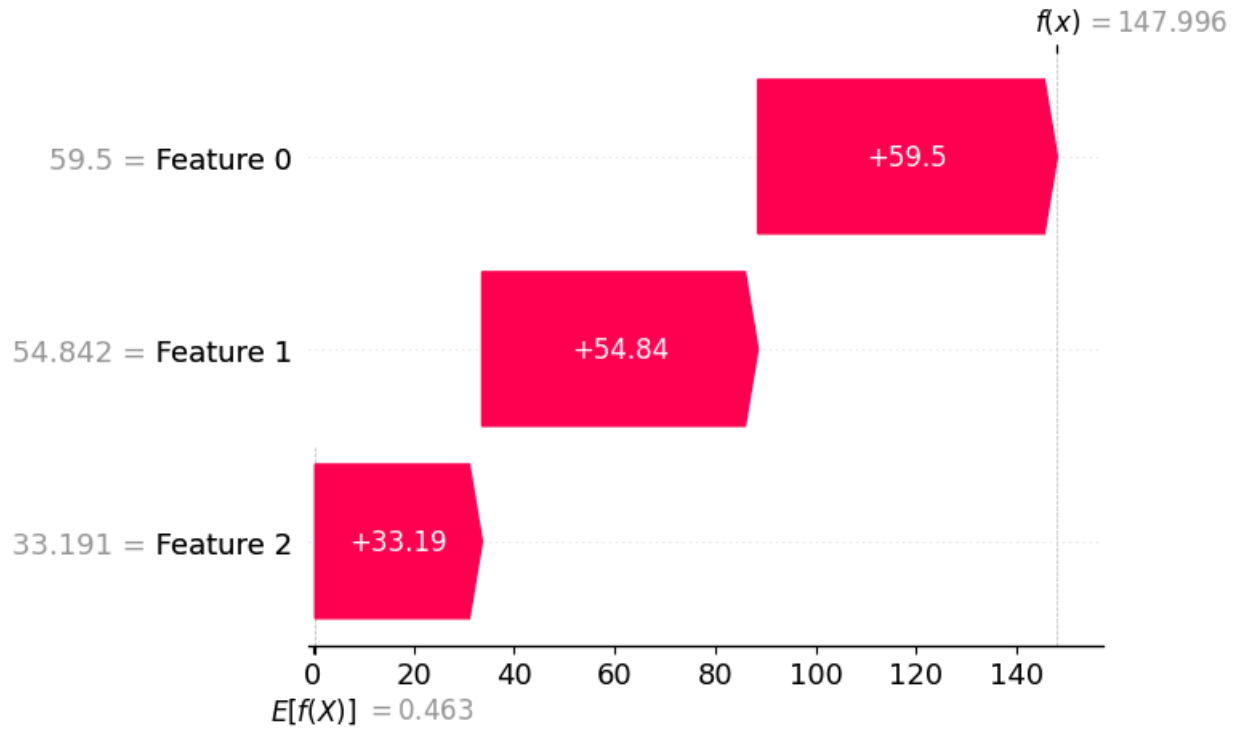
(a)
Part (i)
I first selected the 100th-ranked individual according to sharp_ranking_SCHL. I get that the index of that individual is 465. I then find that individual's rank according to the WKHP ranking function and the AGEP ranking function respectively. It turns out that, based on WKHP scoring, the individual is 128th. Based on AGEP scoring, the individual is 135th.

I repeated the above process for another individual ranked 66th according to SCHL scoring. I get that the index of that individual is 423. I then find that individual's rank according to the WKHP ranking function and the AGEP ranking function respectively. It turns out that, based on WKHP scoring, the individual is 201th. Based on AGEP scoring, the individual is 275th. We can see that as the SCHL rank goes up, both the WKHP and AGEP ranks go down. The difference becomes larger as the SCHL rank goes up.

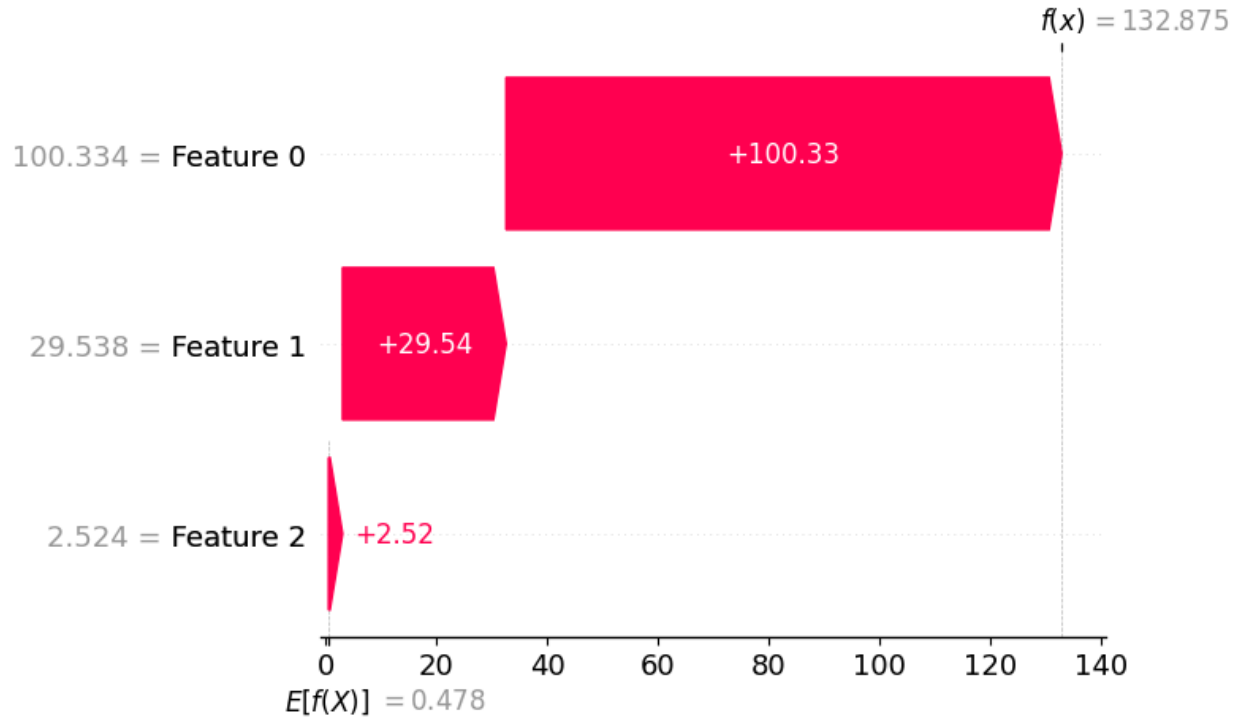
Part (ii)
I computed the importance of each feature for each of the 3 rankings. For the SCHL rankings shown below, none of the three features contribute to rank Qol equally. Feature 1 (school) is the most important.



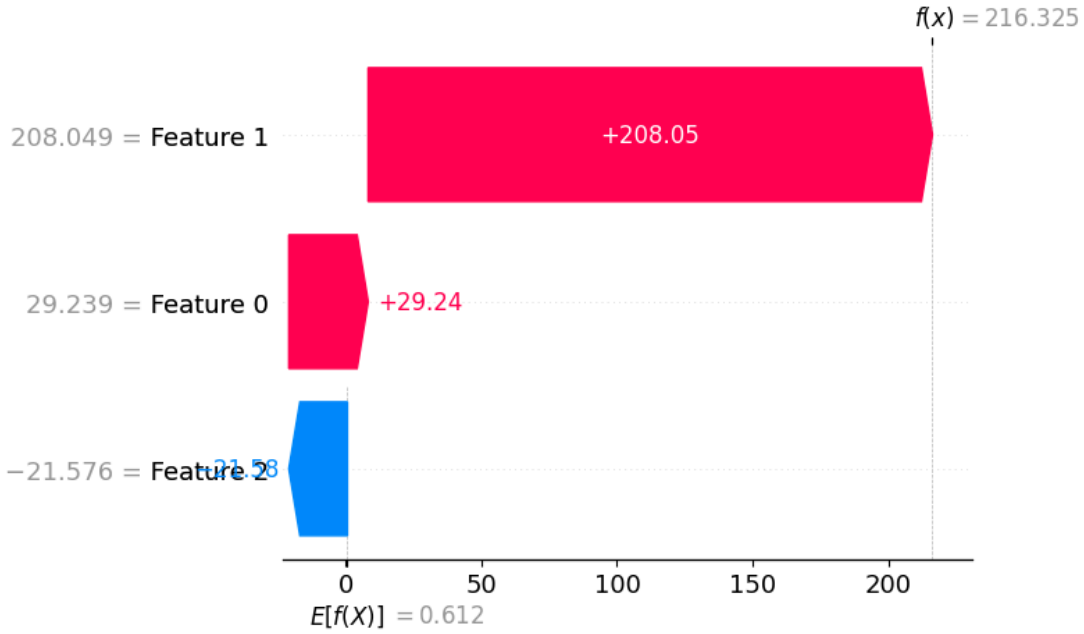
For the WKHP rankings shown below, feature 0 (age) and feature 1 (school) have similar contributions to rank Qol. Feature 0 (age) is the most important.



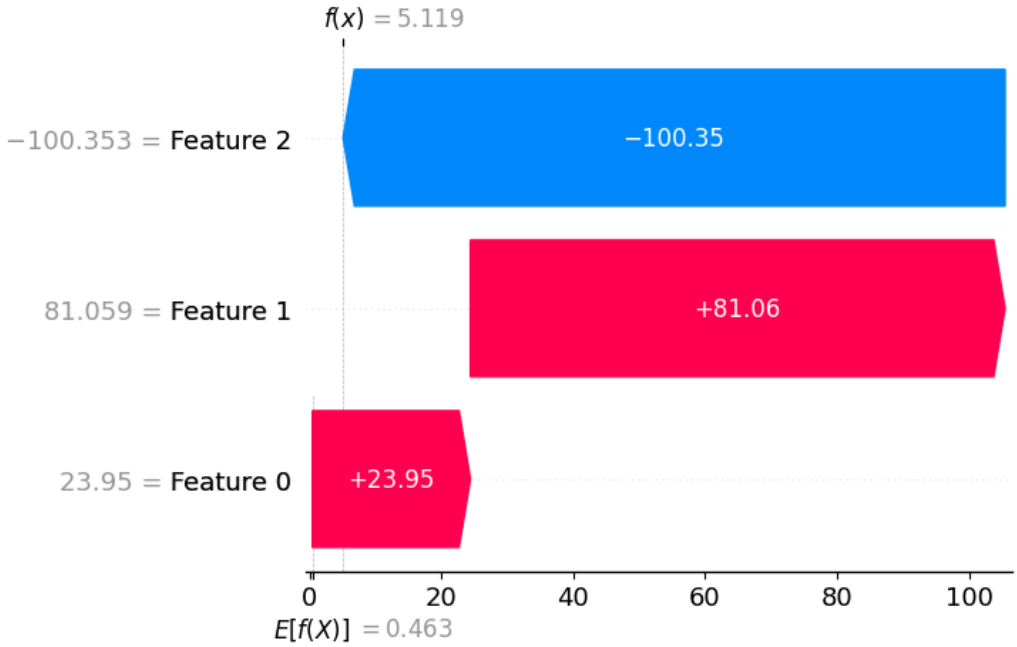
For the AGEP rankings shown below, none of the three features contribute to rank QoI equally. Feature 0 (age) is the most important.



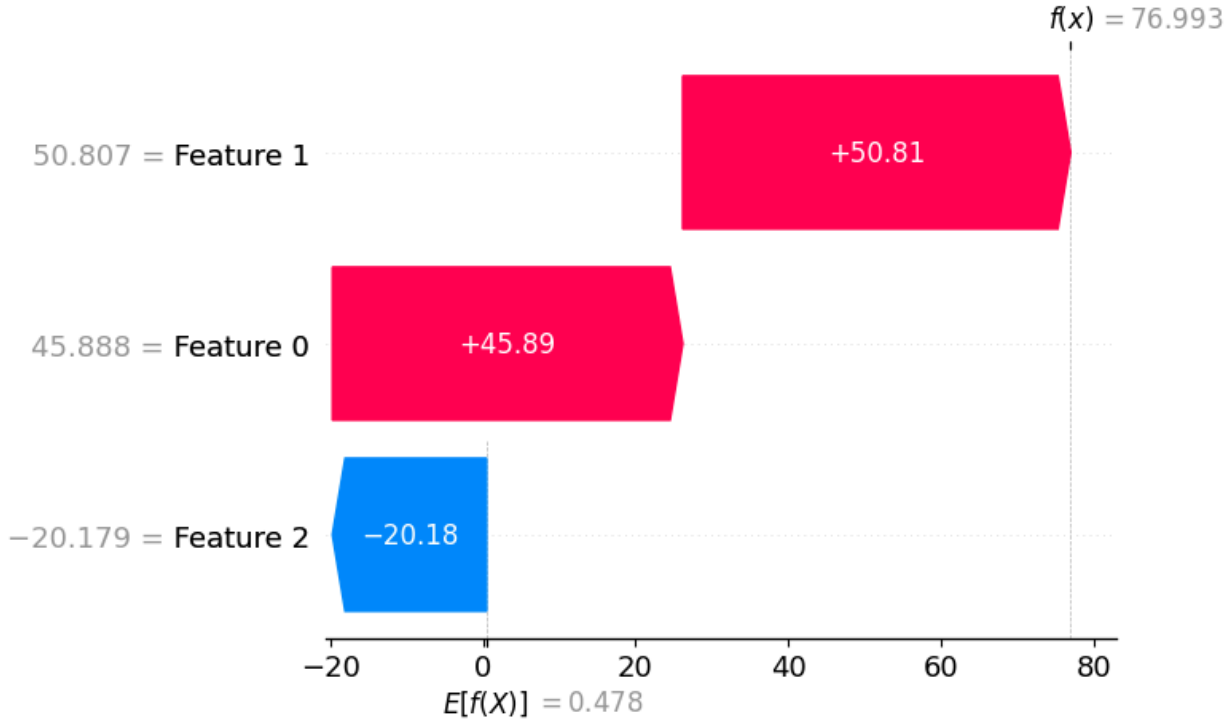
I repeated the above process for another individual ranked 66th according to SCHL scoring. I computed the importance of each feature for each of the 3 rankings. For the SCHL rankings shown below, none of the three features contribute to rank QoI equally. Feature 1 (school) is the most important. Feature 2 (work hour) is somewhat negative.



For the WKHP rankings shown below, none of the three features contribute to rank QoI equally. Feature 2 (work hour) is the most important (although negatively). The other two features (age and school) are positive but have smaller absolute values than feature 2 (work hour).



For the AGEP rankings shown below, none of the three features contribute to rank Qol equally. Feature 1 (school) is the most important. For the rest, feature 0 (age) is positive and feature 2 (work hour) is negative.

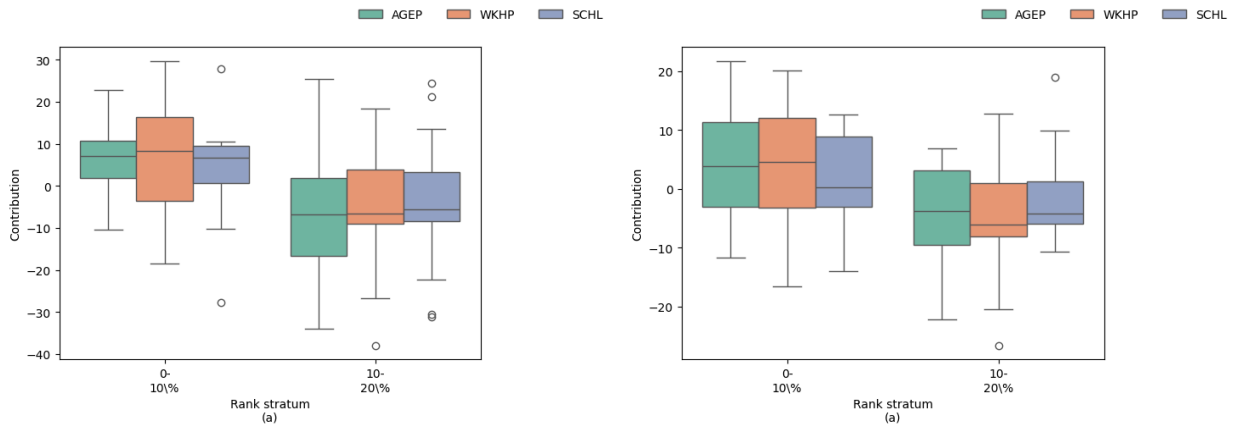


Such differences across the three ranking functions could be from the different scaling we did in making these functions. This difference also explains why SCHL rank goes up causes the other two ranks to go down. It's because different features play different amounts of roles in each ranking. A higher rank in SCHL would be mostly caused by a change in feature 1 whereas a higher rank in WKHP or AGEP mostly depends on feature 0. They can't be simultaneously increased.

(b)
Part (i)
I split the top 20% into two groups of whites and non-whites and stored them into two different dataframes. I defined two ShaRP objects, one for each group.

Part (ii)
I used the make_boxplot_top20 function on each of the dataframes to plot two boxplots for each ranking in each group.

Part (iii)

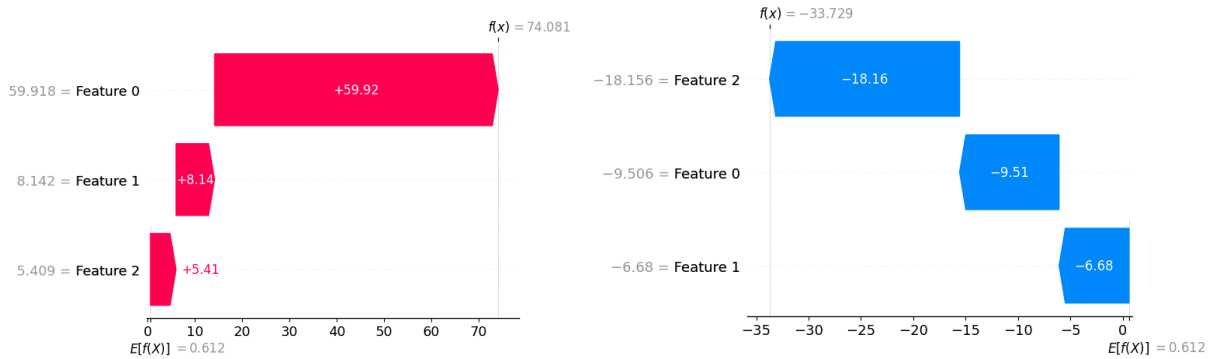


For the white group on the left, the feature importance is different across the two strata. In 0-10%, the median for all three rankings are positive whereas in 10-20%, the median for all three rankings are negative. In 0-10%, WKHP is the most important. In contrast, in 10-20%, AGEP is the most important. However, in both strata, the least important is SCHL.

For the non-white group on the right, the feature importance is different across the two strata. In 0-10%, the median for all three rankings are positive whereas in 10-20%, the median for all three rankings are negative. This time, in both strata, the most important is WKHP. The difference now lies in the least important feature. In 0-10%, SCHL is the least important. In contrast, in 10-20%, AGEP is the least important.

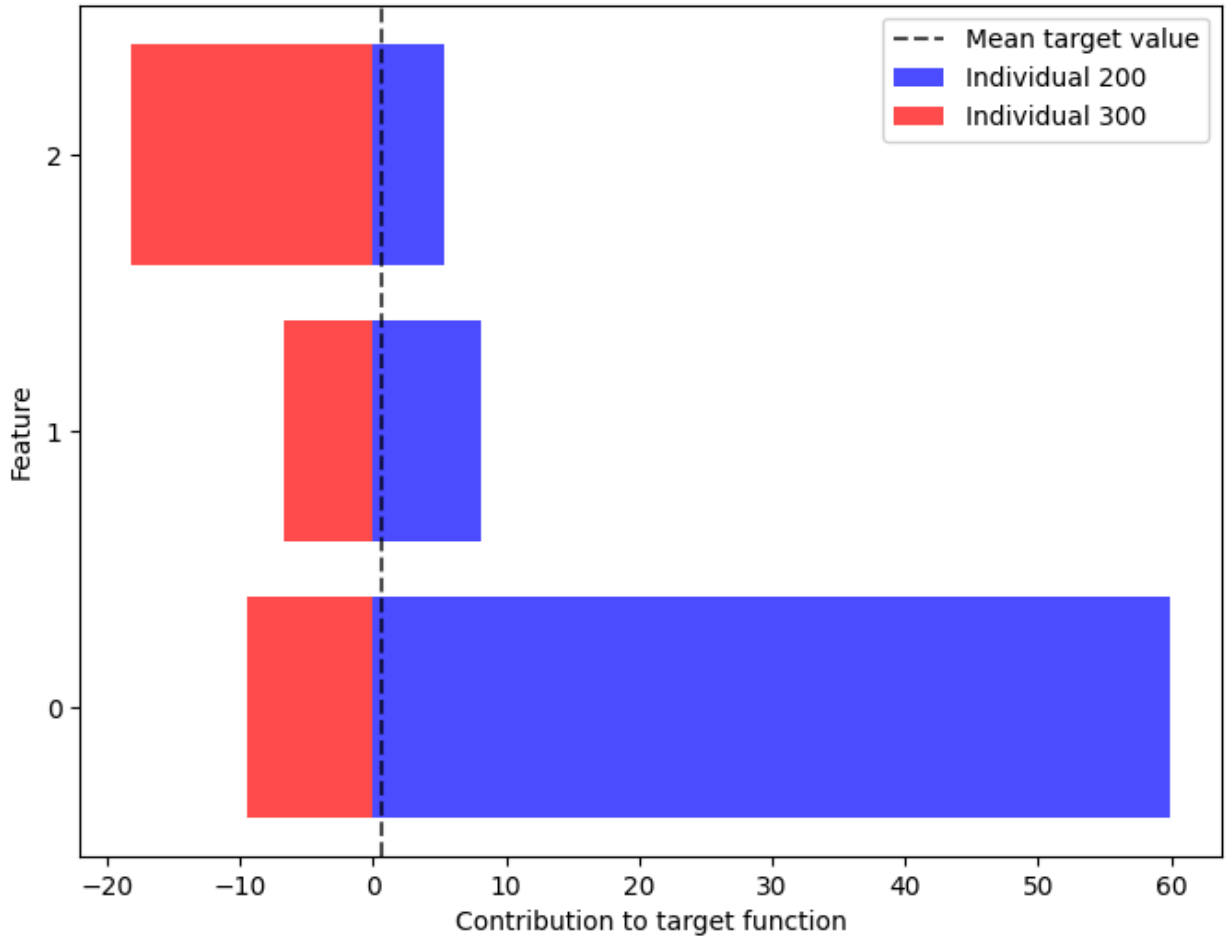
Part (iv)
For the white group, WKHP and AGEP are more important and SCHL the least. For the non-white group, SCHL and AGEP are less important and WKHP the most. A simple, straightforward hypothesis on why feature importance differs across groups can be that education may vary between racial groups.

(c)



I found the indices of the 200th and 300th ranking individual by the SCHL function and plotted a waterfall plot to show their respective feature importances. We can see that for the 200th individual, all three features are positive. The most important feature is feature 0 (age) and the

least is feature 2 (work hour). For the 300th individual, all three features are negative. The most important feature is feature 2 (work hour) and the least is feature 1 (school). The most important feature for 300th individual is the least for 200th.



I plotted a bar chart that shows each individual's each feature and how they differ. I also plotted the mean value line. Feature 0 (age) has the greatest difference. I can safely hypothesize that age plays the biggest role in causing someone to be ranked higher than others.