MACS 30135 Graded Exercise 1

Samantha Jingyi Yom (CNET ID: samanthayom)

The code for this analysis is available in this GitHub repository.

Exercise 1

Here, I predict an individual's participation at national elections from their interest in politics, engagement in political activities and perceptions of democracy, authoritarianism, governance and political ideologies. For this classification task, I employ the decision tree and CatBoost supervised learning algorithms, using data from the World Values Survey Wave 7[1] (Haerpfer et al., 2022). The decision tree and random forest models (hereafter Model 1 and Model 2) obtained an accuracy score of 0.639 and 0.665, respectively.

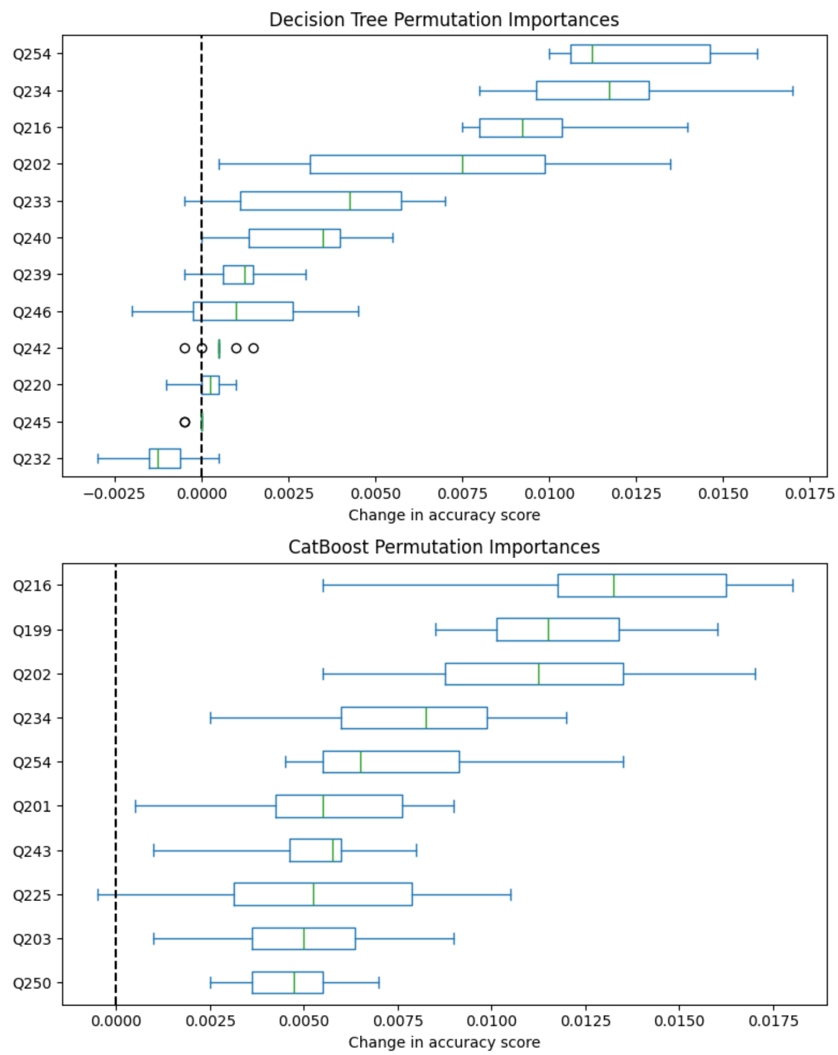

Figure 1. Partial Dependence Plot for Model 1 and 2

[1] The mapping between each question ID and the corresponding question can be found in the WS7 Codebook Variable Report.

The permutation feature importance plots identified 12 and 10 features that were important for predicting participation in national elections for Model 1 and 2, respectively. That is, these features increase the model's prediction error when it is permuted. While Q254, Q234 and Q216 were the three most important features in Model 1's predictions, Model 2 found Q216, Q199 and Q202 as the top three predictors. Only 4 features were common between both plots: Q202, Q254, Q216 and Q234. This suggests that both models depend on a slightly different set of features to make their predictions. This is expected, given how different models are likely to capture different aspects of the data in learning patterns.

Next, we can compare the permutation feature importance with the SHAP values summary plots to see if the features identified as most important are consistent across both methods.

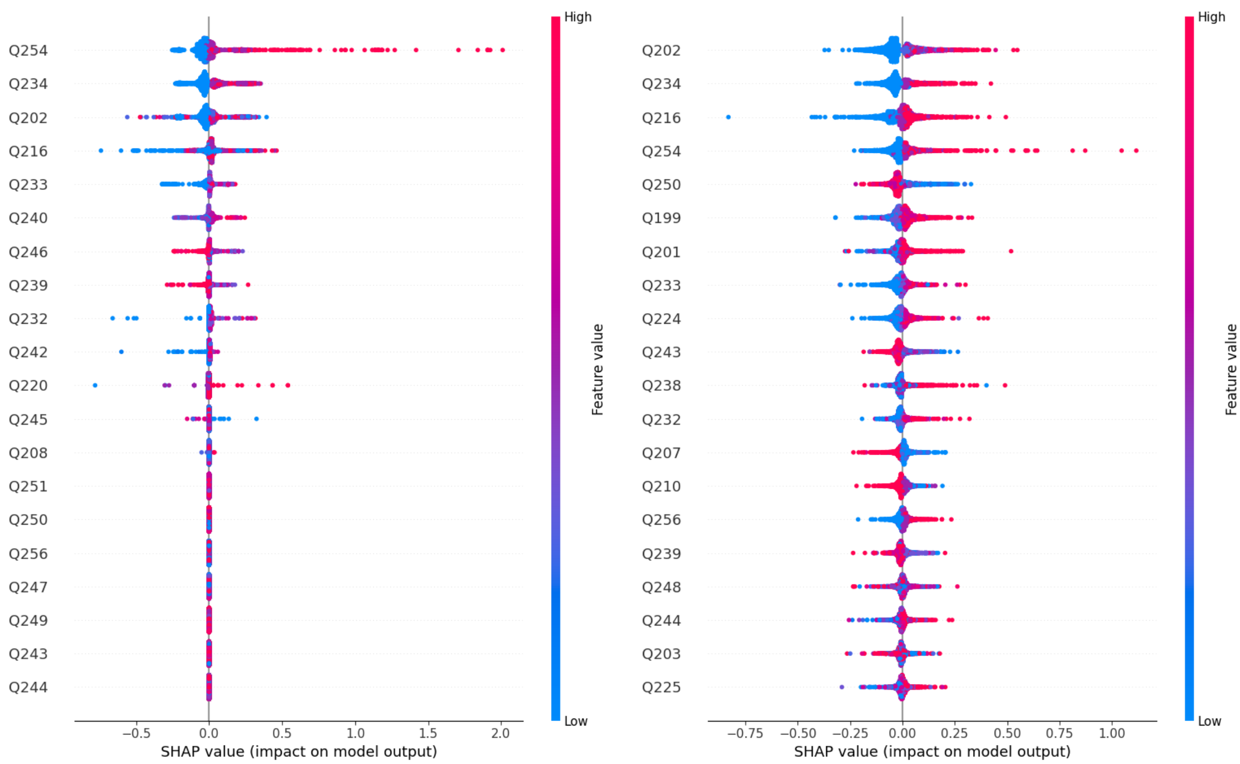

Figure 2. SHAP values summary plots for Model 1 and 2

Although there are some variations in the rank order, the features identified for both models in the SHAP values summary plots are similar to that in the permutation feature importance plots. It is worth noting that, while more features are shown in the plot for Model 1 in Figure 2, this is likely because the summary plot displays a default of 20 features. The influence of the last few features has a close to zero impact on the model's predictions. This demonstrates that both techniques have detected similar features as having the largest influence of electoral participation. Both plots also align in terms of the degree of impact each feature has on their respective model's predictions. For example, the influence of each feature is smaller and more consistent across features in Model 2, which is clearly illustrated in the beeswarm plot.

The permutation feature importance and SHAP values summary plot have indicated a relationship between several features and an individual's participation in national elections. However, we want to take a closer look at the exact form of this relationship. We do this using Partial Dependence Plots (PDPs) and Accumulated Local Effects (ALE) plots.

Figures 3 and 4 show the PDP and ALE plot for Q254, which has been identified earlier as the most important feature for Model 1. Q254 is a five-level discrete variable representing an individual's national pride, with a value of 1 indicating the highest level of pride. The PDP indicates national pride has a significantly larger influence on electoral participation when individuals report having lower levels of it. This is corroborated in the ALE plot, where high levels of national pride have little effect on Model 1's predictions. Therefore, the relationship between national pride and electoral participation is non-linear.

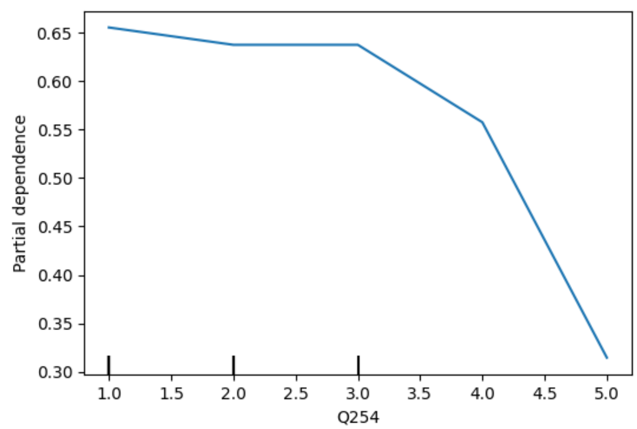

Figure 3. PDP of Q254 for Model 1

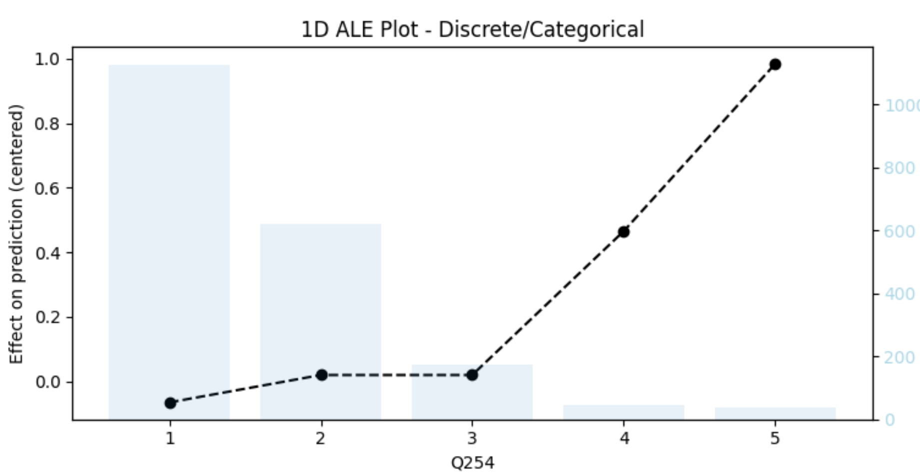

Figure 4. ALE plot of Q254 for Model 1

For Model 2, I create PDPs and ALE plots for Q216 (i.e., engagement in social activism by encouraging others to vote) and Q202 (i.e., frequency of obtaining information from TV news), the features identified to be the most important from the permutation feature importance and SHAP values summary plots, respectively.

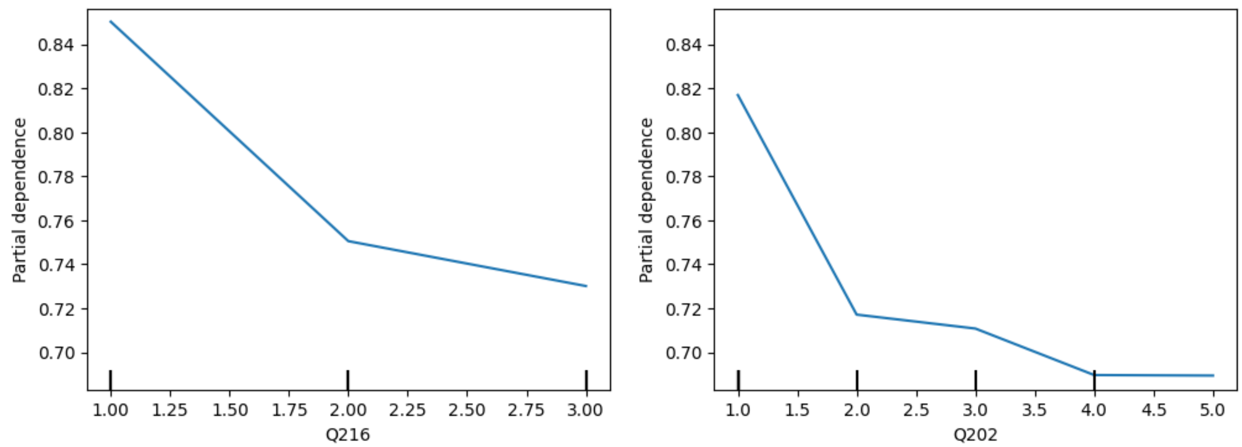

Figure 5. PDPs of Q216 and Q202 for Model 2

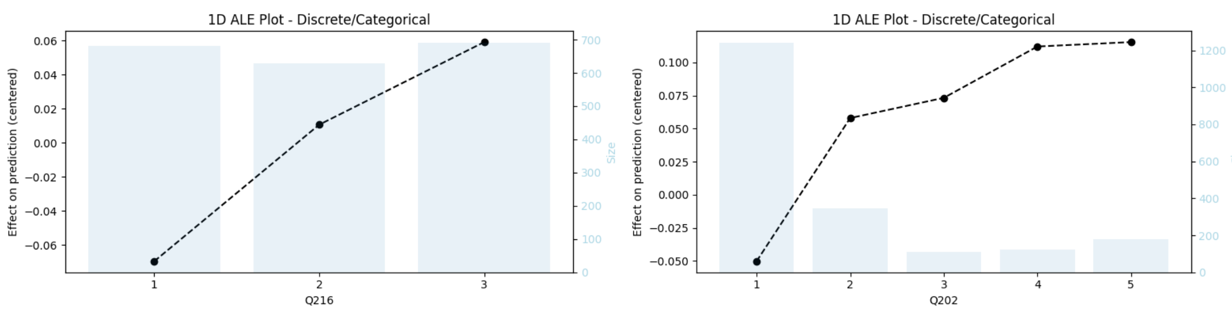

Figure 6. 1-D ALE plots of Q216 (left) and Q202 for Model 2

The PDPs in Figure 5 show a non-linear relationship between Q216 and Q202, respectively, and the model's prediction. More specifically, differences in lower values of each of these features have a stronger effect on electoral participation, particularly for Q202. Once again, this is corroborated with the 1-D ALE plots, as shown in Figure 6.

Considering how the direction of these relationships is the same for both features, we also obtain a 2-D ALE plot (see Figure 7 below) to investigate a possible correlation between them and their interaction effect on predictions. From this plot, we observe that for higher levels of Q216, changes in the level of Q202 have a larger effect on Model 2's predictions. Changes in the level of Q202 at low levels of Q216 have little impact on electoral participation.

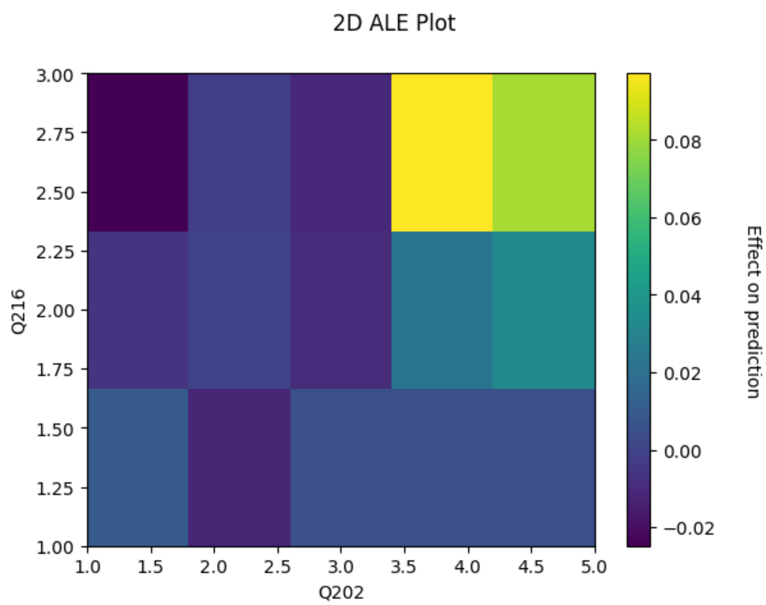

Figure 7. 2-D ALE plots of Q216 (left) and Q202 for Model 2

Exercise 2(a)

The importance of each feature in predicting whether, in the US, an individual's income exceeds $50k seems to vary depending on how importance is measured, as observed from the plots in Figures 8, 9 and 10. Features are ranked based on their F-score. However, in comparing feature importance, the absolute F-scores are not inherently meaningful. They are only meaningful when compared between relative to other features, which is what we focus on in this analysis.

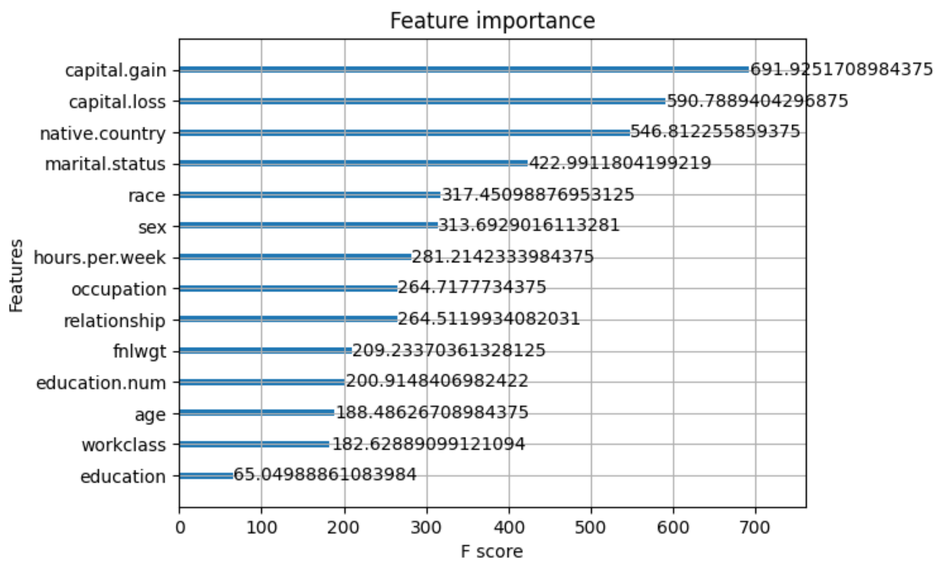

Figure 8. Feature importance measured by cover

In Figure 8, feature importance is calculated by its "cover", that is, how often a feature is used in splits, weighted by the number of data points passing through. There appears to be a gradual decline in importance across the ranked features, though less important features have relatively similar levels of importance. This suggests that, while they do contribute to the model's predictions, the individual impact of each of these latter features is more evenly distributed. The model's predictions seem to rely heavily on a few features, namely an individual's capital gain, capital loss, and native country. Given how they are closely related with income levels, one would expect that features like capital gain and loss would influence a large proportion of data in splits.

When importance is instead measured by the average reduction in loss when a feature is used for splitting, the ranked features and their relative importance are starkly different. Figure 9 reveals that an individual's "relationship" (i.e., status within their family) as the most critical predictor of their income level, with a significantly higher F-score than all other features. This means that the "relationship" feature contributes significantly toward improving the model's prediction accuracy. The relationship between an individual's status within the family and their income level seems less obvious, though it is easy to imagine scenarios in which this has an impact on earning potential due to factors such as caregiving responsibilities.

On the other hand, the remaining features have much lower F-scores, with those at the lower end contributing little to the model's predictions. This could point to how other meaningful features (e.g., capital gain, capital loss) are underrepresented when we feature importance is measured in relation to model accuracy.

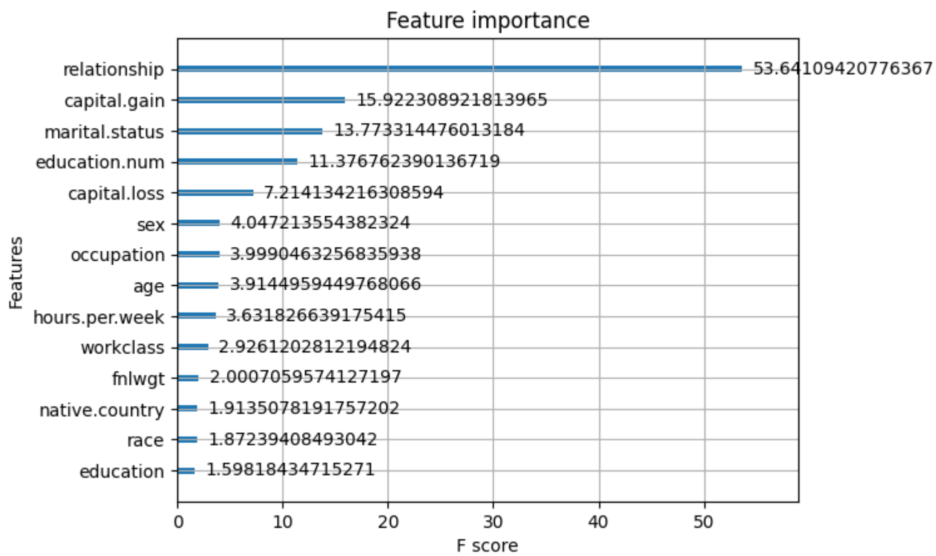

Figure 9. Feature importance measured by gain

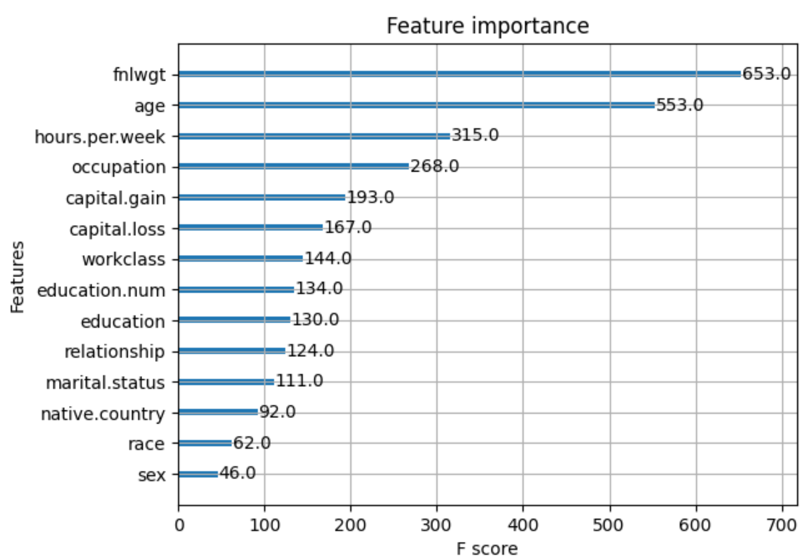

Figure 10. Feature importance measured by weight

Figure 10 shows how important each feature is based on how often it appears in splits across trees. Once again, the results are different from both of the earlier plots. Here, two features have a significantly higher F-score: final weight and age, while there is a gradual decline across the remaining features. Both final weight and age were not highly ranked features in either of the earlier importance plots. It is worth noting that just because a feature appears frequently in splits, by extension, is ranked as more important here, does not mean that it has an equally significant contribution to the model's prediction accuracy. Furthermore, we observe that final weight, age and hours per week (i.e., the top three most important features) are all continuous variables. Unlike categorical (e.g., occupation, education) and binary (e.g., sex) variables, continuous variables can be used in many more splits because they are not limited to a set of unique values.

In conclusion, these plots show us that feature importance can vary depending on what aspect of the model's decision-making we care about. For example, features that frequently appear in splits may not contribute meaningfully to the model's performance, but are rank higher due to the nature of the variable and the consequences this has on the partitioning of trees.

Exercise 2(b)

SHAP can be used to provide both global and local explanations for how much each feature contributes to the models' predictions. The impact of each feature is compared based on its SHAP value. Once again, it is worth noting that the absolute values are not meaningful on their own, only when compared to that of other features. First, we take a closer look at a few global explanations.

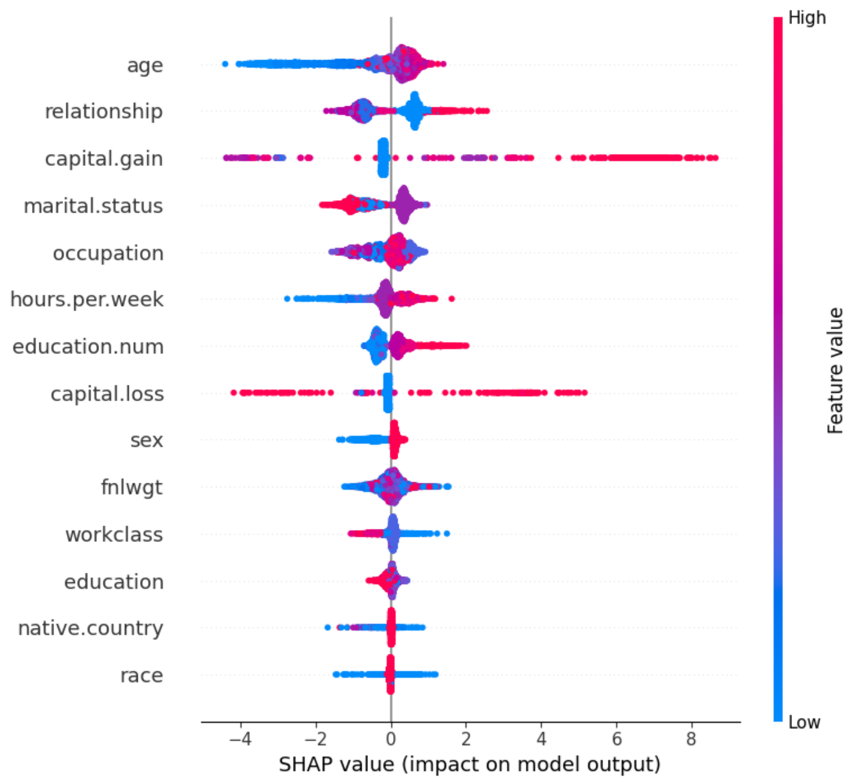

Figure 11. SHAP summary plot

The SHAP summary plot in Figure 11 illustrates not only feature importances, but also feature effects. The order of the ranked features is different from those we obtained in Exercise 2. Although we know features are ranked by descending order, it is hard to tell from this summary plot how exactly their importance differs. That said, there are several insights we can glean from the relationship between a feature's SHAP value and its value.

First, we observe that lower values for age have a more significant contribution towards the model's predictions, particularly that income levels are below $50k. Second, when it comes to capital gain and loss, only non-zero values influence the model's predictions, with their impact exhibiting a wide range of variation. However, given that all non-zero values are deemed "high" relative to zero, we cannot identify the exact form of the relationship between both capital gain and loss, and income levels from this plot. Lastly, it seems clear that specific values of categorical variable are associated with different direction and degree of impact on model output. For example, one sex appears to have much less impact predictions that the other. Certain values for relationship have a much more varied impact on predictions than others.

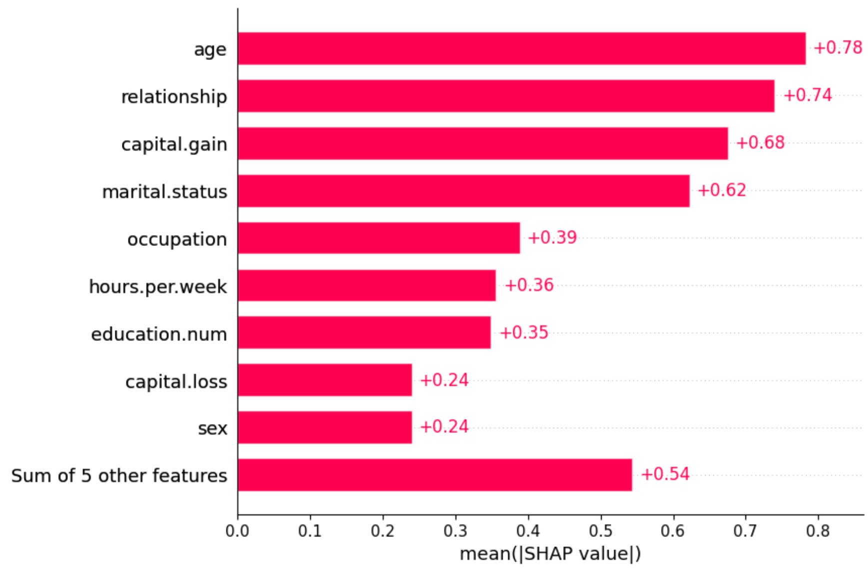

Figure 12. SHAP feature importance

Unlike the summary plot, the feature importance plot in Figure 12 provides a clearer view of how each feature compares with regard to their importance. There seem to be four features, that is, age, relationship capital gain and marital status, that contribute most significantly to improving model performance.

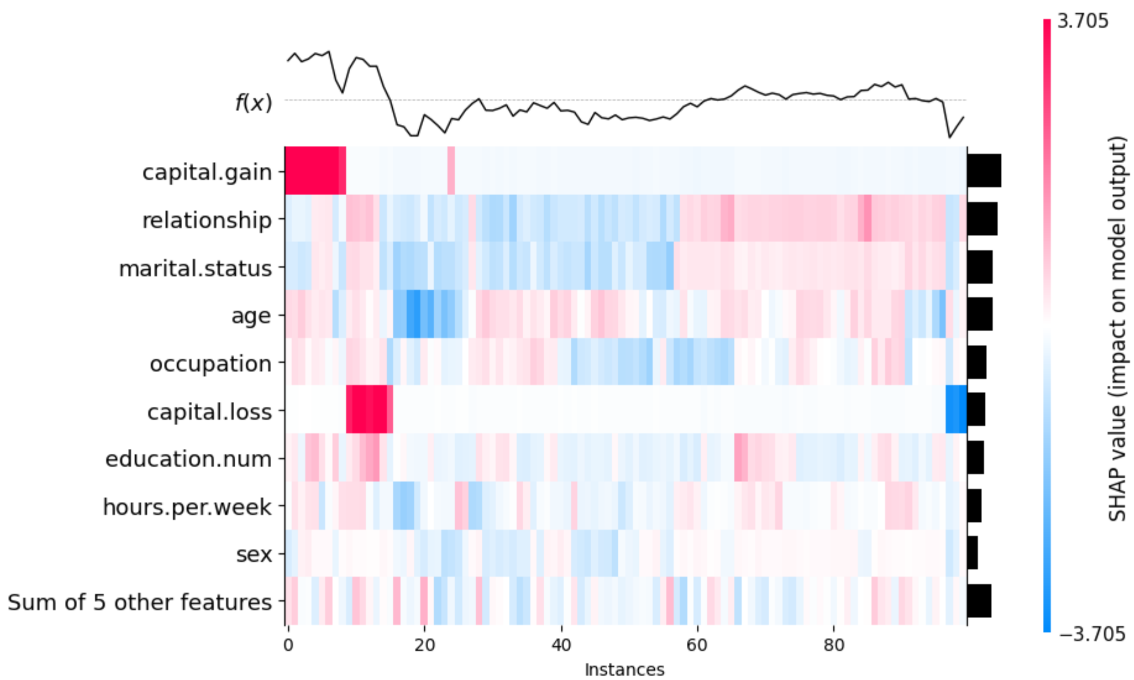

Figure 13. SHAP heat map

The SHAP heat map (Figure 13) illustrates the impact of each feature at the instance level. It should be noted that the inconsistency in feature importance rankings between this and earlier plots can be attributed to the fact that the heat map was generated using a random sample of 100 instances. That said, its level of granularity offers additional insight that was missing from the other plots. For one, we see high levels of income (i.e., above $50k) are predicted mainly by capital gain or loss, as seen from the dark red bars. On the other hand, low levels of income are predicted mainly by capital loss and age, though to a smaller extent.

Now, we focus on a specific feature: age. Since it was identified as the most important feature, we explore its feature dependence using the scatterplot in Figure 14. Our earlier observations about age in the summary plot are corroborated here. In addition, we see how that the probability of having an income greater than $50k increases significantly with age, but only for the range between 20 and 40. This probability actually decreases beyond the age of 60, albeit less significantly. The decline is expected since people are likely to leave the workforce beyond that age.

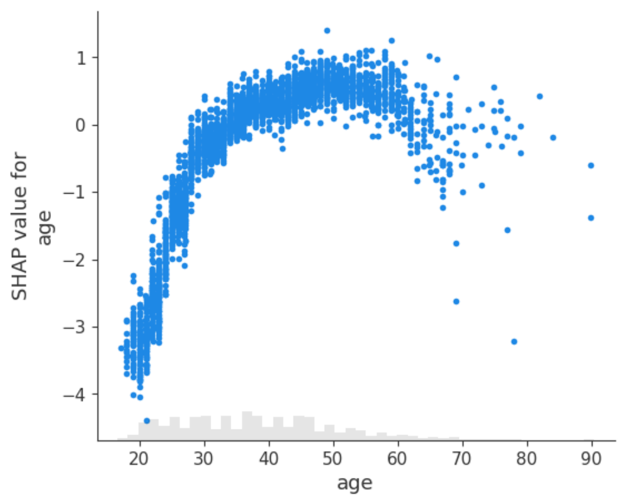

Figure 13. SHAP feature dependence plot for age

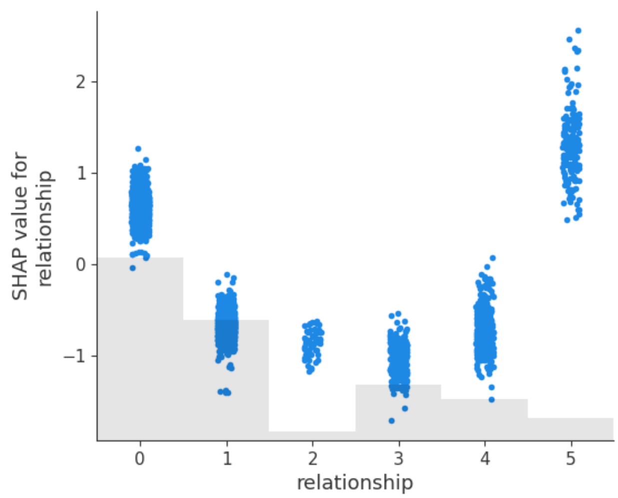

Figure 14. SHAP feature dependence plot for relationship

Earlier plots were unable to provide much insight into the relationship between specific values of categorical variables and income levels. This is made clear in Figure 14, where we present a feature dependence plot for the second most important feature: relationship. The plot reveals that being a wife (i.e., relationship value of 5), rather than a husband (i.e., relationship value of 0), is more strongly associated with a higher probability of having an income above $50k. In contrast, all other family statuses are more likely to correlate with earning below $50k.

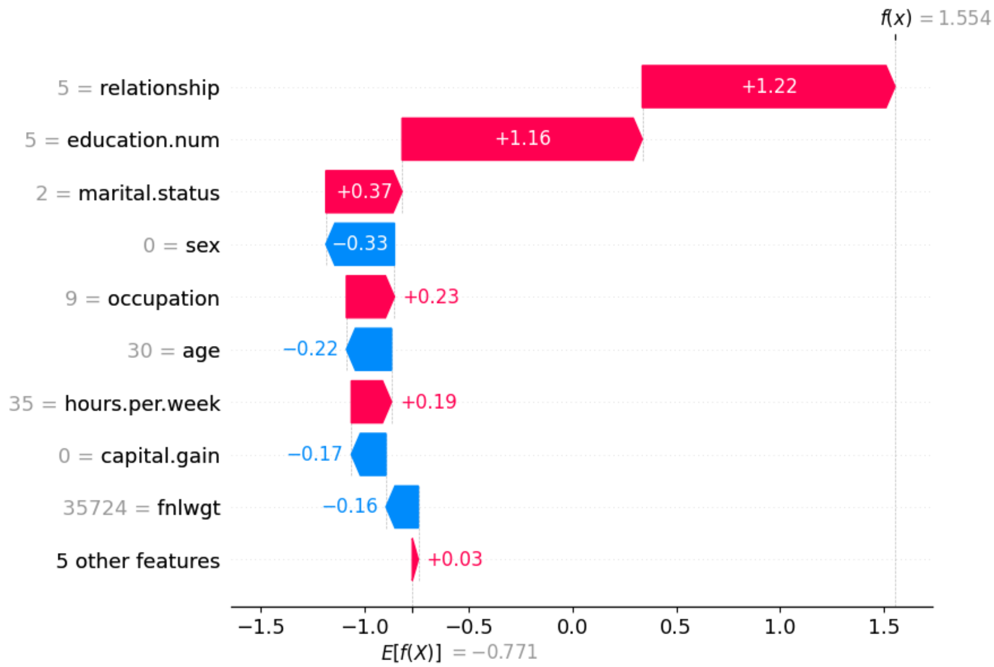

Figure 15. SHAP waterfall plot

Finally, we create a waterfall plot see Figure 15) to obtain a local explanation for an individual prediction. As an example, we consider that of a 30-year-old female master's student. Being a wife significantly increases this individual's predicted probability of earning more than $50k, though more interestingly, this impact is comparable to that of their education level—not even their earning a master's.

References

Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano J., M. Lagos, P. Norris, E. Ponarin & B. Puranen (eds.). 2022. World Values Survey: Round Seven - Country-Pooled Datafile Version 5.0. Madrid, Spain & Vienna, Austria: JD Systems Institute & WVSA Secretariat. doi:10.14281/18241.24