SUPPLEMENTAL MATERIAL FOR FAIR FEATURE IMPORTANCE SCORES FOR INTERPRETING TREE-BASED METHODS AND SURROGATES

Anonymous authors
Paper under double-blind review

A PROOF OF PROPOSITION 1

Proof. By the Total Law of Expectation, we have that

$$\begin{aligned} Bias^{DP}(lev(t)) = \Bigg| \Bigg(& E(\hat{y}_i | i \in lev_\ell(t), z_i = 1) P(i \in lev_\ell(t) | z_i = 1) \\ & + E(\hat{y}_i | i \in lev_r(t), z_i = 1) P(i \in lev_r(t) | z_i = 1) \Bigg) \\ & - \Bigg(E(\hat{y}_i | i \in lev_\ell(t), z_i = 0) P(i \in lev_\ell(t) | z_i = 0) \\ & + E(\hat{y}_i | i \in lev_r(t), z_i = 0) P(i \in lev_r(t) | z_i = 0) \Bigg) \Bigg|. \end{aligned}$$

Notice that the expectation of $\hat{y}_i \in lev_\ell(t) = \pi_{lev_\ell(t)}$. Replacing the expectation terms with $\pi_{lev_\ell(t)}$ and $\pi_{lev_r(t)}$, we can see that

$$\begin{aligned} Bias^{DP}(lev(t)) &= \Big| \pi_{lev_\ell(t)} P(i \in lev_\ell(t) | z_i = 0) + \pi_{lev_r(t)} P(i \in lev_r(t) | z_i = 0) \\ &\quad - \pi_{lev_\ell(t)} P(i \in lev_\ell(t) | z_i = 1) - \pi_{lev_r(t)} P(i \in lev_r(t) | z_i = 1) \Big| \\ &= \Bigg| \frac{\sum_i \mathbb{1}_{\{z_i=1, i \in lev_\ell(t)\}} * \pi_{lev_\ell(t)} + \sum_i \mathbb{1}_{\{z_i=1, i \in lev_\ell(t)\}} * \pi_{lev_r(t)}}{\sum_i \mathbb{1}_{\{z_i=1, i \in lev(t)\}}} \\ &\quad - \frac{\sum_i \mathbb{1}_{\{z_i=0, i \in lev_\ell(t)\}} * \pi_{lev_\ell(t)} + \sum_i \mathbb{1}_{\{z_i=0, i \in lev_\ell(t)\}} * \pi_{lev_r(t)}}{\sum_i \mathbb{1}_{\{z_i=0, i \in lev(t)\}}} \Bigg|. \end{aligned}$$

Combining similar terms and simplifying, we have that

$$\begin{aligned} Bias^{DP}(lev(t)) = \Bigg| & \pi_{lev_\ell(t)} \left(\frac{\sum_i \mathbb{1}_{\{z_i=1, i \in lev_\ell(t)\}}}{\sum_i \mathbb{1}_{\{z_i=1, i \in lev(t)\}}} - \frac{\sum_i \mathbb{1}_{\{z_i=0, i \in lev_\ell(t)\}}}{\sum_i \mathbb{1}_{\{z_i=0, i \in lev(t)\}}} \right) \\ & + \pi_{lev_r(t)} \left(\frac{\sum_i \mathbb{1}_{\{z_i=1, i \in lev_r(t)\}}}{\sum_i \mathbb{1}_{\{z_i=1, i \in lev(t)\}}} - \frac{\sum_i \mathbb{1}_{\{z_i=0, i \in lev_r(t)\}}}{\sum_i \mathbb{1}_{\{z_i=0, i \in lev(t)\}}} \right) \Bigg|. \end{aligned}$$

□

The proof for $E\big[Bias^{EQOP}(lev(t))\big]$ is analogous to the proof for demographic parity.

For the multiclass classification case, let $\pi^m_{lev_\ell(t)}$ and $\pi^m_{lev_r(t)}$ denote a vector of length K, where π_k denotes the proportion of that class in the node.

Corollary 1. *Consider multiclass classification with probabilistic trees:*

$$
\begin{aligned}
Bias^{DP}(lev(t)) = \Bigg| & \pi^m_{lev_\ell(t)} \left(\frac{\sum_i \mathbb{1}_{\{z_i=1, i \in lev_\ell(t)\}}}{\sum_i \mathbb{1}_{\{z_i=1, i \in lev(t)\}}} - \frac{\sum_i \mathbb{1}_{\{z_i=0, i \in lev_\ell(t)\}}}{\sum_i \mathbb{1}_{\{z_i=0, i \in lev(t)\}}} \right) \\
& + \pi^m_{lev_r(t)} \left(\frac{\sum_i \mathbb{1}_{\{z_i=1, i \in lev_r(t)\}}}{\sum_i \mathbb{1}_{\{z_i=1, i \in lev(t)\}}} - \frac{\sum_i \mathbb{1}_{\{z_i=0, i \in lev_r(t)\}}}{\sum_i \mathbb{1}_{\{z_i=0, i \in lev(t)\}}} \right) \Bigg|
\end{aligned}
$$

B Additional Simulation Results

B.1 Classification Results

We evaluate our method on the same experiments as Figure 2 in the main paper but in Figure 5, $N = 500$ and in Figures 6 and 7, we use Equality of Opportunity as the fairness metric. In Figure 8, we consider a simulation with a large number of features with $p = 250$. In the large p simulation, there are 5 features in each group and we otherwise follow the same setting as our other classification simulations. In the correlated simulations, shown in Figure 9, we use an autoregressive design with $\mathbf{\Sigma}^{-1}_{j,j+1} = 0.5$ instead of $\mathbf{\Sigma} = \mathbf{I}$ in the uncorrelated p simulations. Similar to the results in the main paper, we see the correct magnitude and direction of the scores in all of the simulation scenarios.

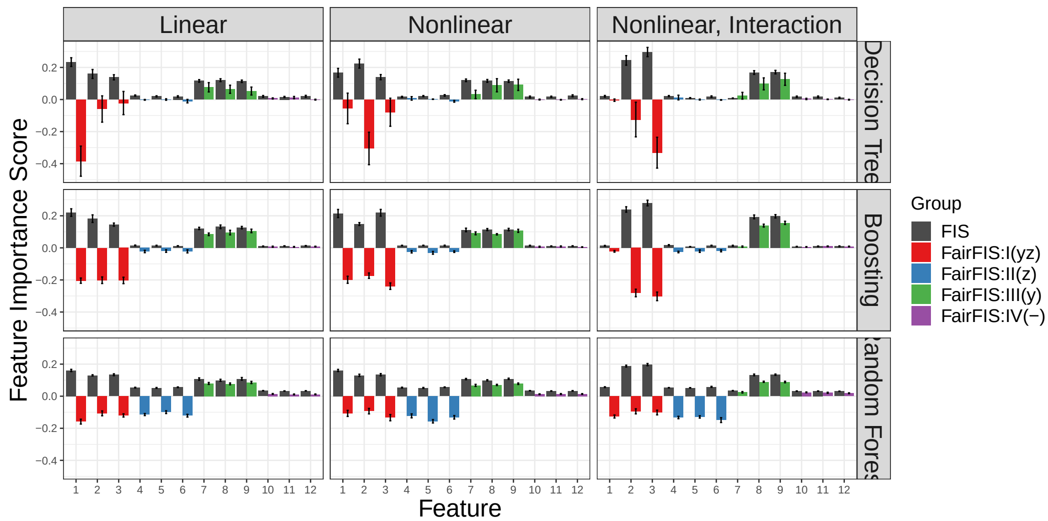

Figure 5: Classification *FIS* and *FairFIS* results for accuracy and Demographic Parity on three major simulation types that include a linear model (left), a non-linear additive model (middle), and a non-linear additive model with pairwise interactions (right), with $N = 500$ and $p = 12$. We examine a decision tree classifier, a boosting classifier, and a random forest classifier.

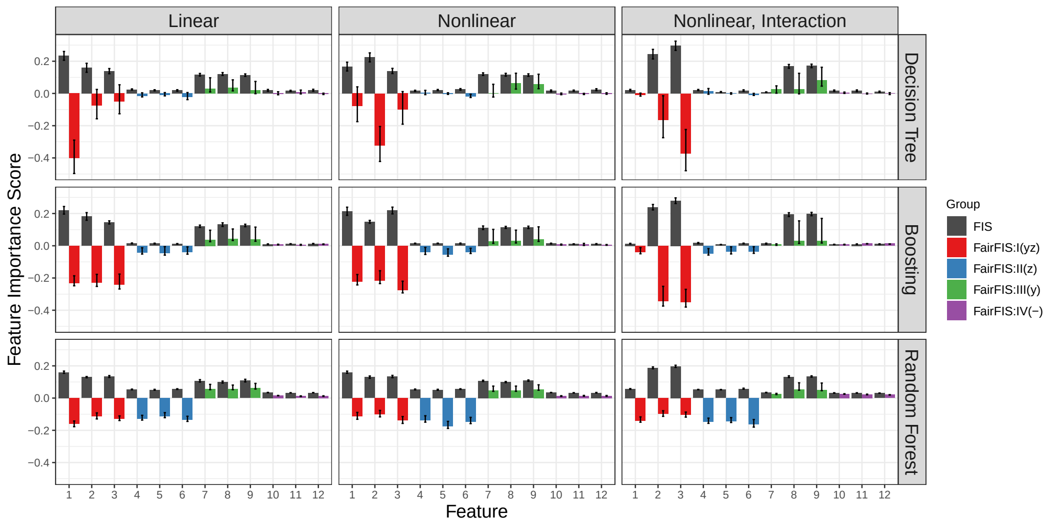

Figure 6: Classification *FIS* and *FairFIS* results for accuracy and Equality of Opportunity on three major simulation types that include a linear model (left), a non-linear additive model (middle), and a non-linear additive model with pairwise interactions (right), with $N = 500$ and $p = 12$. We examine a decision tree classifier, a boosting classifier, and a random forest classifier.

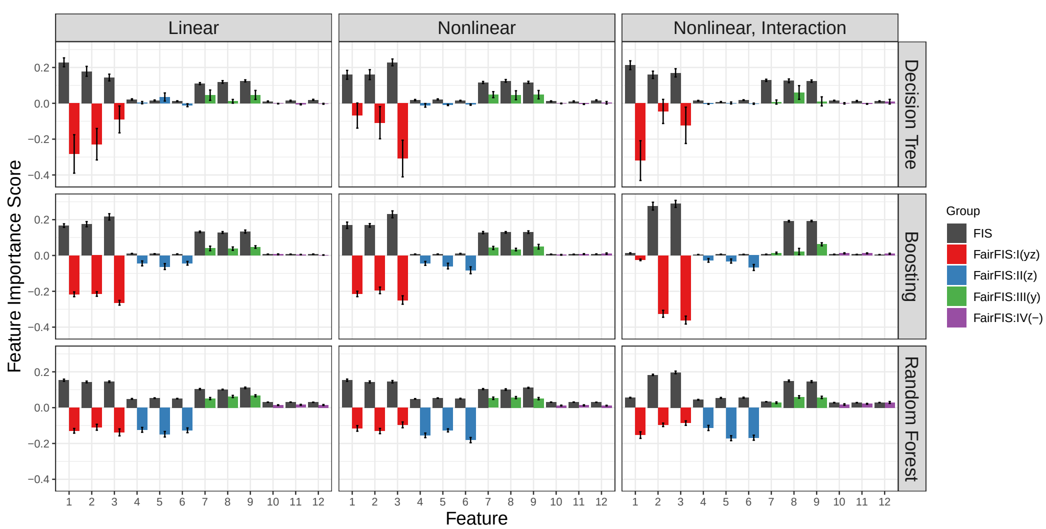

Figure 7: Classification *FIS* and *FairFIS* results for accuracy and Equality of Opportunity on three major simulation types that include a linear model (left), a non-linear additive model (middle), and a non-linear additive model with pairwise interactions (right), with $n = 1000$ and $p = 12$. We examine a decision tree classifier, a boosting classifier, and a random forest classifier.

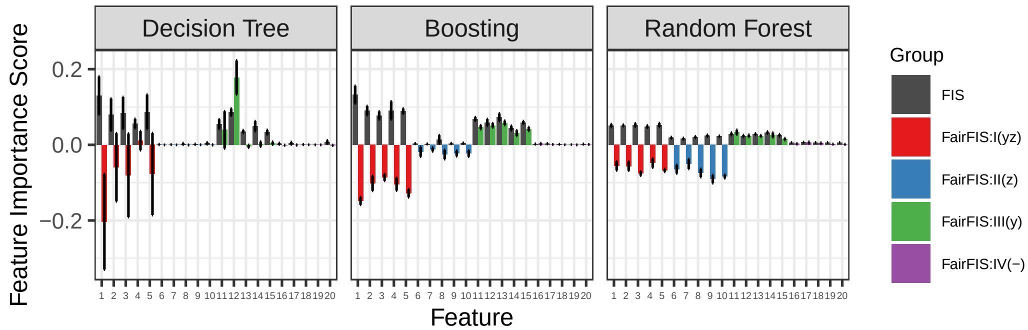

Figure 8: Large p classification *FIS* and *FairFIS* results for accuracy and Demographic Parity for a decision tree classifier, a boosting classifier, and a random forest classifier, with $N = 1000$ and $p = 250$. We show the *FIS* and *FairFIS* scores for the first 20 features.

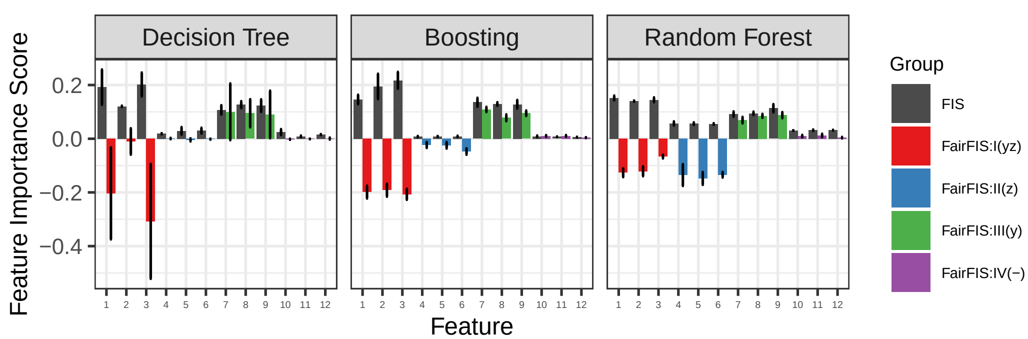

Figure 9: Correlated feature classification *FIS* and *FairFIS* results for accuracy and Demographic Parity for a decision tree classifier, a boosting classifier, and a random forest classifier, with $N = 1000$ and $p = 12$.

B.2 REGRESSION RESULTS

In Figures 10 and 11, we evaluate *FairFIS* results for Demographic Parity in the regression setting. Here, $\beta_j = 3$ for $j \in G_1$ or G_3 and $\beta_j = 0$ for $j \in G_2$ or G_4 and $\alpha_j = 0.4$ for $j \in G_1$ or G_2 and $\alpha_j = 0$ for $j \in G_3$ or G_4. All other aspects of the base simulation as described in the main paper remain the same. Similar to the results in the main paper and the additional classification results, the magnitudes and directions of the scores are as expected from the simulation design.

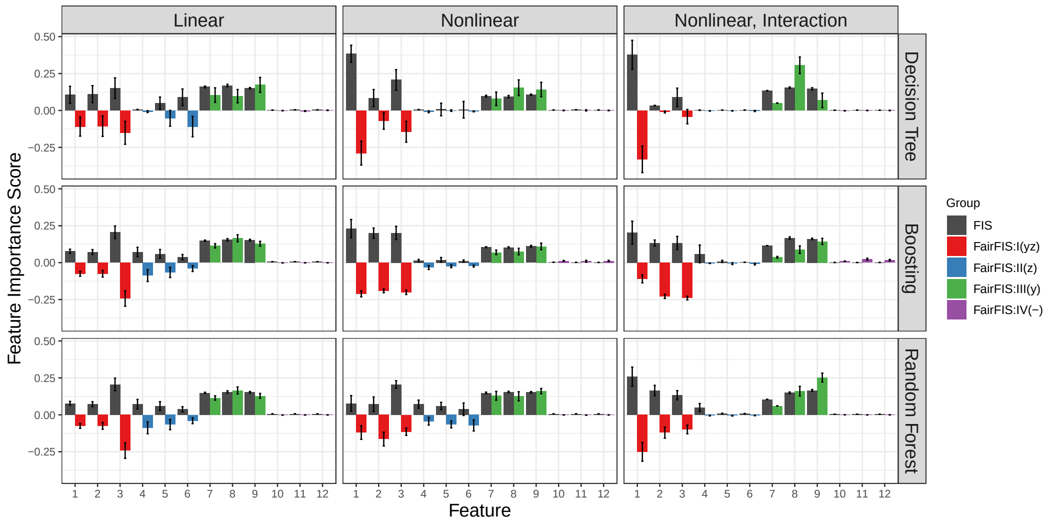

Figure 10: Regression *FIS* and *FairFIS* results for accuracy and Demographic Parity on three major simulation types that include a linear model (left), a non-linear additive model (middle), and a non-linear additive model with pairwise interactions (right), with $N = 500$ and $p = 12$. We examine a decision tree regressor, a boosting regressor, and a random forest regressor.

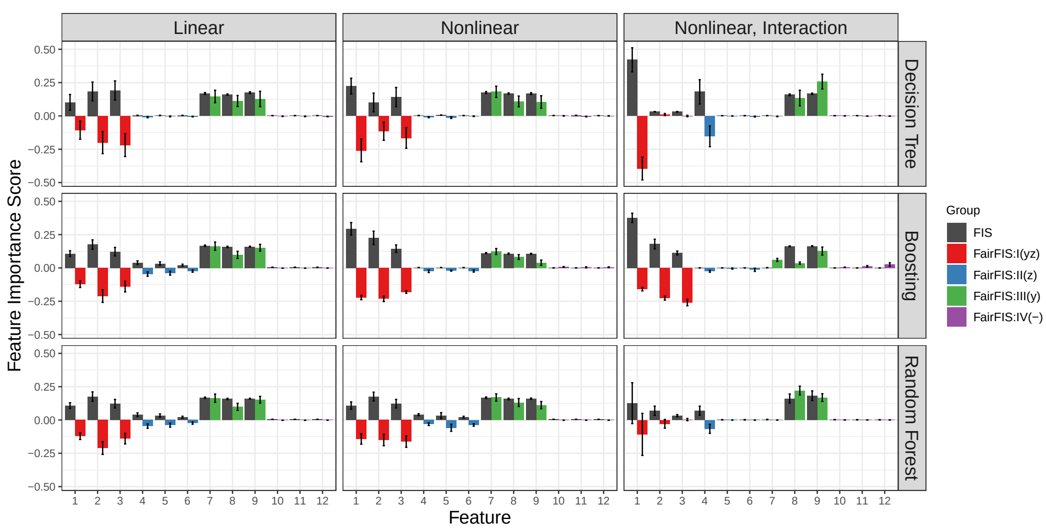

Figure 11: Regression *FIS* and *FairFIS* results for accuracy and Demographic Parity on three major simulation types that include a linear model (left), a non-linear additive model (middle), and a non-linear additive model with pairwise interactions (right), with $N = 1000$ and $p = 12$. We examine a decision tree regressor, a boosting regressor, and a random forest regressor. The magnitudes and directions of the *FairFIS* scores for each group align with what we would expect from the simulation setup, validating our method.

C Additional Results on Benchmark Datasets

We include the same experiment as Figure 3 from the main paper for the C & C dataset with Race as the protected attribute and the German dataset with Gender as the protected attribute in order to validate the use of global surrogates. We see that the magnitudes and the directions between the scores of the boosting classifier and the tree-based surrogate of the boosting classifier are similar.

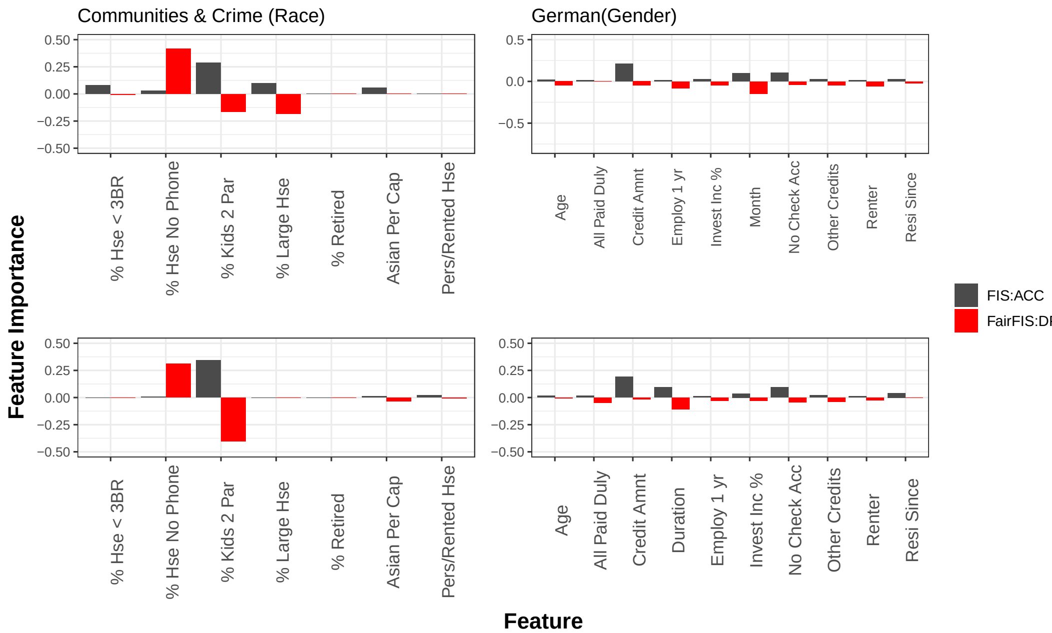

Figure 12: Global surrogate validation. The top row shows *FIS* and *FairFIS* results on a boosting classifier for the C & C dataset with Race as the protected attribute and the German dataset with Gender as the protected attribute.The bottom row shows *FIS* and *FairFIS* results for a tree-based surrogate of a boosting classifier. The scores between the top and bottom rows are similar in magnitude and direction, indicating that our scores are effective when used to interpret tree-based global surrogates.

In Figure 13, we explore the quality of *FairFIS* interpretations of tree-based surrogates of a deep learning model on the German dataset with Gender as the protected attribute and the Law School dataset with Race as the protected attribute. As shown in the main paper when discussing Figure 4, the *FairFIS* results provide reasonable feature interpretations in terms of fairness.

In order to validate using trees for interpretation versus model-specific interpretation, we compare *FIS* scores of a tree-based surrogate and Layerwise Relevance Propagation (LRP) for the Adult dataset with Gender as the protected attribute, the Law School dataset with Race as the protected attribute, the COMPAS dataset with Race as the protected attribute, and the German dataset with Gender as the protected attribute as shown in Figure 14. We implement LRP using the DeepExplain package with “elrp” set as the method name. We set the first layer of the MLP as the input layer and the last layer as the output. For all the datasets, we see that in general the magnitude of the importance scores for the tree surrogate and LRP surrogate are comparable. Specifically, both methods identify the same features as highly predictive, as reflected in the magnitude of the scores. As a result, we can validate that we can reasonably use trees for interpretation versus model-specific validation.

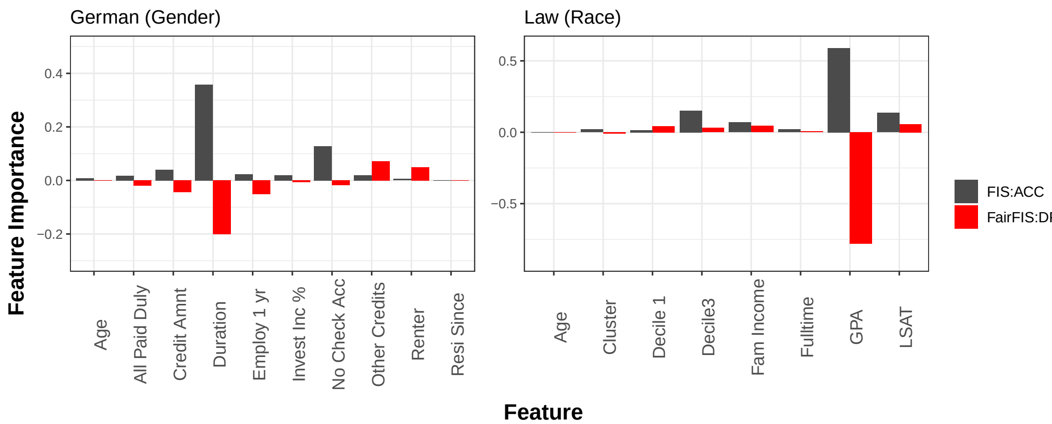

Figure 13: Importance scores for a tree-based surrogate of a deep learning model for the German dataset with Gender as the protected attribute (left) and Law School dataset with Race as the protected attribute (right).

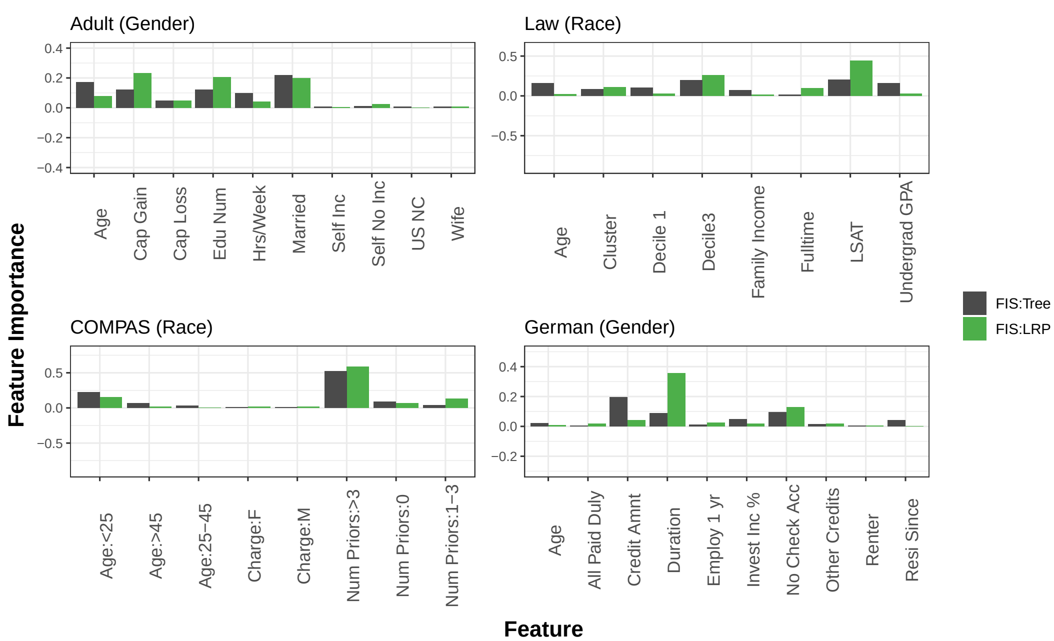

Figure 14: Validation for using trees as surrogates. For the Adult dataset with Gender as the protected attribute, the Law dataset with Race as the protected attribute, the COMPAS dataset with Race as the protected attribute, and the German dataset with Gender as the protected attribute, we show *FIS* scores for a tree-based surrogate of an MLP and an LRP surrogate. The magnitudes between the two methods are similar, validating we can use trees for interpreting deep learning models.