Group-specific Feature Importances

Falaah Arif Khan[1] , **Venetia Pliatsika**[2] ,
[1] NYU Center for Data Science, [2] NYU Tandon School of Engineering
{fa2161, venetia}@nyu.edu,

Abstract

Shapley values are commonly used to compute the feature importances (FI) of any black-box model. Despite their popularity, Shapley-based methods suffer from severe limitations which challenge their validity as an explanability tool. Specifically, when feature importance differs across subgroups in the data, global (population-level) FIs are no longer a reliable explanation for the model's predictions. In this work, we propose group-specific FIs as a meaningful construct in-lieu of population-level FIs, and provide a novel, efficient bottom-up algorithm that produces accurate and unbiased feature explanations for any group definition of the users' choice (Black women for example), with the additional capability of being able to provide accurate FIs for all ancestral groups (Black people, women, and the full population, in this example). We demonstrate our method's utility in recommending reliable recourse through experiments on real and synthetic datasets. Further our method has applications as a diagnostic tool for detecting unfairness by comparing group-specific FIs for different demographic groups in the dataset.

1 Introduction

As the size and complexity of machine learning (ML) models and data increase, so does the need to be able to explain the model's decisions. AI regulation already has established the *right to explanation* and there are local examples such as NYC Local Law 144 of 2021 requiring explanations and bias audits in the context of hiring already taking effect. For explainability, a leading approach is to explain the feature importance of a trained model. Arguably, the most popular among these explainability methods is Shapley values.

Shapley values (SV) were introduced as a profit allocation mechanism (13) in game theory. SVs in this context capture the marginal contribution of each player in all possible player subsets and assign their weighted sum as the player's payoff. SVs are used because of their mathematical properties and were hence adapted as an explainability method in ML. In this context, the game is the model, the team is one item of the dataset, and the players are the item's features. The contributions of the features are created per item (local explanations) and aggregated to identify the important features for the entire population (global explanations). Multiple methods have been developed for using Shapley values in ML applications, each measure a different characteristic function and have different mathematical guarantees (3; 10; 7; 5; 1; 11). The prevailing difference in the calculation of the characteristic function is whether they consider the marginal (3; 11) or the conditional distribution (5; 1) of the feature's contributions.

[a]

name	gender	salary	children	education	marital-status
Bob	m	5	2	4	M
Cal	m	8	1	7	U
Dia	f	4	0	4	M
Eli	f	4	2	2	U
Fay	f	4	1	7	M
Kat	f	4	0	3	U
Leo	m	4	0	3	M
Osi	m	7	1	7	U

[b]

v	i	S	$u \in \mathcal{D}$	U$_1$	U$_2$
Fay	salary	{edu, marital}	Bob	(f, 4, 1, **4**, **M**)	(f, **5**, 1, **4**, **M**)
Fay	salary	{edu, marital}	Cal	(f, 4, 1, **7**, **U**)	(f, **8**, 1, **7**, **U**)
Fay	salary	{edu, marital}	Dia	(f, 4, 1, **4**, **M**)	(f, **4**, 1, **4**, **M**)
Fay	salary	{edu, marital}	Eli	(f, 4, 1, **2**, **U**)	(f, **4**, 1, **2**, **U**)
Fay	salary	{edu, marital}	Kat	(f, 4, 1, **3**, **U**)	(f, **4**, 1, **3**, **U**)
Fay	salary	{edu, marital}	Leo	(f, 4, 1, **3**, **M**)	(f, **4**, 1, **3**, **M**)
Fay	salary	{edu, marital}	Osi	(f, 4, 1, **7**, **U**)	(f, **7**, 1, **7**, **U**)

Figure 1: [a] Dataset $\mathcal{D}$. [b] Frankenstein points for Fay when we are explaining the feature "salary", we use the coalition $\mathcal{S}$ of education-num and marital-status and we do exact computation i.e. use the entire dataset to create the Frankenstein points.

In this paper, we consider an adaptation of the marginal methods, specifically QII (3), and introduce group-wise feature importance explanations by conditioning on the groups. To estimate the marginal contributions, marginal methods create artificial datapoints, *Frankenstein points*, that consist of the combination of feature values of the item being explained and other items. To estimate the feature importance of one feature for one item, marginal methods take the difference between two Frankenstein points that differ only on this feature's value. Therefore, they measure how a specific feature value changes the model's behavior in a variety of settings.

Motivating Scenario: Recommending Recourse Our motivating example is a decision-maker, attempting to recommend recourse to an applicant, for a lending decision. Specifically, we are interested in explaining the lending decision of a black-box model, in terms of which features of the applicant led to an undesirable decision, so that they can go back and change those features to get the desired outcome (a positive lending decision). To do this, we seek a reliable and efficient explanation method, that computes meaningful feature importance for any query point.

In figure 1 we present an example table that shows how we can use SV and the Frankenstein points to provide feature importance for the specific decision. Let's assume we have dataset presented in [a] which includes the gender, salary, number of children, education in years, and marital status of eight people. Furthermore, let's say we want to explain why Fay was denied the loan. We create a Frankenstein datapoint for all different subsets of features S by merging Fay's features that are not in the subset S and the features that are in S of everyone else. We repeat this process for each feature. In [b], we show the Frankenstein datapoints that are created for the subset education and marital status when we explain the feature salary.

Research gap Shapley-based frameworks to compute feature importance have some well-documented limitations: A leading concern is the validity and meaningfulness of the feature importances that Shapley-based methods compute(6; 14). Specifically, interventional approaches used to construct Frankenstein points may result in data points that are not representative or feasible, and thereby the resulting feature explanations computed using them can be misleading. For instance, in figure 1[b], we see that multiple Frankenstein points, specifically the ones where Fay's features were merged with a male candidate's, have salary values different than 4. Looking at [a], we see that all women have a salary of 4 so arguably all such Frankenstein points constructed are infeasible. A further concern is that feature explanations can be misleading either by attributing importance to features not used by the model (conditional models) or by ignoring correlations (marginal models). Recent work has tried to address this issue by discussing guidelines by which to choose the method appropriate for the application (2) or arguing hybrid approaches that attempt to explain the features used by the model while keeping some of the correlations intact (9).

In this work, we take a group-fairness perspective to clarify the validity of Shapley-based feature explanations. Specifically, we posit that feature importances will differ for different subgroups in the dataset. Hence, providing a local explanation (for a single person or datapoint) is not meaningful, unless we make strong distributional assumptions, ie. that the query point and the sample points we aggregate over, come from the same distribution. Similarly, global explanation methods — which claim to summarize the feature importance for the full population, are not meaningful because the population is not one homogenous distribution, but instead a mixture of group-specific distributions.

We introduce the notion of a *group-specific* feature importance. We claim that it is only meaningful to explain which features are important within a particular subgroup and modify the QII framework from (3) to construct Frankenstein points only from samples within a chosen group. We formally propose a provably correct and efficient algorithm to compute group-specific feature importance, for any definition of a group appropriate to the data, including groups constructed based on protected attributes like gender, race, disability, etc. Our method is general and accommodates both local and global explanations. Specifically, if the entire population is treated as one homogeneous group, our group-specific feature importance is exactly global feature importance. On the other end of the spectrum, if each person is treated as constructing their own group (a maximally intersectional definition), then our group-specific feature importances are exactly local feature importances. We caution against either of these extremes for the following reasons: As mentioned before, population-level feature importance assumes that the data is homogenous, which is seldom true in practice. Hence, recommending recourse for a specific item based on feature importance at the population-level, instead of their specific sub-groups, could thereby result in an unreliable and ineffective recourse recommendation. On the other hand, defining each person as their own group makes our method unfeasible. Recall that we need to modify the QII framework to only use samples from within the group of interest. If the group of interest is too narrowly defined, then we will not have enough samples to compute a reliable (accurate) estimate.

Instead, our algorithm provides a minimum group size of m. For any group definition with at least m samples in each group, our algorithm is provably correct and accurate for any datapoint. Further, our algorithm is provably correct not just for the most fine-grained group definition, but for every ancestor group definition that precedes it. Specifically, if the user defines groups based on a collection of features (such as gender, race, disability), then if there are at least m samples in each group defined by the intersectional of these attributes, then our method is guaranteed to be ϵ-accurate for any ancestor group defined by those features as well (for our example, $\{$gender, race$\}$, $\{$race, disability$\}$, $\{$gender, disability$\}$, as well as binary $\{$gender$\}$, $\{$race$\}$ and $\{$disability$\}$.

Returning to our motivating scenario of recourse recommendation, our method allows a decision-maker to provide customized, and thereby more reliable, recommendations to applicants, based on their social group membership, thereby mindful of the unique challenges that come with it. Further, comparing feature importance at different 'levels' (child vs any ancestor) could be a useful diagnostic for detecting *unfairness*. For example, if feature importance is significantly different for women, Black women, and disabled Black women, then this might be an indication that disabled Black women need to put significantly more effort into improving their qualification (feature values) than other social groups, due to social inequality.

Summary of contributions Even though the ML community recognizes the necessity of analyzing ML models for unfairness between subgroups, limited effort has been put into the analysis of feature explanations for different subgroups. Additionally, the uncertainty of which Shapley value ought to be used per task further complicates the analysis of feature importance per subgroup because we might both want to know that the model directly uses an illegal attribute and also that it uses its proxies. Finally, there has been no analysis of the error bounds that takes into consideration the different possible subgroups. In this work, we formally propose a hybrid method similar to (9) that calculates Shapley values for distinct subgroups that can be identified with a method similar to (12), and guarantees the same error across all subgroups while keeping the number of computations required to calculate feature importance identical or lower than any other Shapley value method proposed.

2 Notation

Let $\mathcal{A}$ denote an ordered collection of features (or attributes), and let $\mathcal{D}$ denote a set of items (or points). We will consider that $\mathcal{A}$ is partitioned into two sets; the attributes that determine group membership (such as race, disability status, or gender) and relevant qualifications (all other attributes). We will denote as G the set of every subset of the group attributes.

An item $\mathbf{v}$ has $|\mathcal{A}| = d$ features $(v_1, \ldots, v_d) \in \mathbb{R}^d$. Our goal is to explain the importance of features in $\mathcal{A}$ to the outcome for $\mathbf{v}$, as determined by the black-box algorithm. We will do so using Shapley values.

$\mathbf{u} = (u_1, \ldots, u_d) \in \mathbb{R}^d$ is a point in $\mathcal{D}$ sampled by a mechanism $Samp$. We will denote by $\mathbf{U} \in \mathbb{R}^{d\times m}$ a matrix of m items $(\mathbf{u}^1, \mathbf{u}^2 \ldots \mathbf{u}^m)$. Let $Samp_g$ be the group-specific mechanism, which draws samples only from $\mathcal{D}_g$, instead of $\mathcal{D}$, where $\mathcal{D}_g$ are the samples from group $g \in G$. Note that this assumes that $\mathbf{v}$ is a member of group g. Unless otherwise noted, we use random sampling.

For a subset of features $\mathcal{B} \subseteq \mathcal{A}$, let $\mathbf{v}_{\mathcal{B}}$ denote a projection of $\mathbf{v}$ onto $\mathcal{B}$. When partitioning the features $\mathcal{A} = \mathcal{B} \cup \mathcal{C}$, with $\mathcal{B} \cap \mathcal{C} = \emptyset$, let $\mathbf{v}_{\mathcal{B}}\mathbf{U}_{\mathcal{C}}$ denote a vector of items in which each item takes on the values of the features in $\mathcal{B}$ from $\mathbf{v}$, and the values of the remaining $\mathcal{C}$ features from $\mathbf{U}$.

We denote by $QoI(\mathbf{U})$ the computation of the QoI (e.g. score) for each item in a sample $\mathbf{U} \subseteq \mathcal{D}$, and return a vector of real values. We compute the QoI on two sets of Frankenstein points ($\mathbf{U}_1$ and $\mathbf{U}_2$). These Frankenstein points are constructed by perturbing $\mathbf{v}$ based on $\mathbf{U}$, based on a set of features $\mathcal{S} \subseteq \mathcal{A}$, such that $\mathbf{U}_1 = \mathbf{v}_{\mathcal{A}\setminus\mathcal{S}}\mathcal{D}_{\mathcal{S}}$ and $\mathbf{U}_2 = \mathbf{v}_{\mathcal{A}\setminus\{\mathcal{S}\cup i\}}\mathcal{D}_{\mathcal{S}\cup i}$. Recall also that, since $\mathbf{U}_1$ and $\mathbf{U}_2$ are vectors of points constructed from the same sample $\mathbf{U}$, they are of the same size. To quantify feature importance to an item's QoI, we define $\tau(\mathbf{U}_1, \mathbf{U}_2)$ as our desired measured of difference (such as subtraction) between $QoI(\mathbf{U}_1)$ and $QoI(\mathbf{U}_2)$.

Finally, we denote as $\phi(\mathbf{v})$ the vector of Shapley values for each feature of $\mathbf{v}$, $\phi(i)$ the Shapley value of the i-th feature of $\mathbf{v}$ ($\mathbf{v}$ will be omitted when context renders it obvious) and $\phi_{\mathcal{S}}$ the Shapley value of the i-th feature of $\mathbf{v}$ over a subset $\mathcal{S} \subseteq \mathcal{A}$ of features.

3 A bottom-up algorithm for ϵ-accurate group-specific feature importances

The Quantitative Input Influence (QII) framework (3) uses Shapley values (SV) to $\epsilon - \delta$ approximate feature importance, as a way to explain the outputs of any black-box model. In this work, we modify the QII framework to use a group-specific sampler, satisfying our distributional assumption that the data is not a homogeneous iid sample, but rather a mixture of group-specific distributions.

We provide a complete summary of the QII framework in the Appendix and describe our proposed approach here.

Our method estimates $\epsilon - \delta$-accurate group-specific feature importance for all groups, based on a user-specified group definition. For example, let groups be defined by (race, gender), then $G = \{\text{Black women, White women, Black men, White men}\}$

$$\tilde{\phi}_{g,i}(v) = \sum_{S\subseteq A\setminus\{i\}} \frac{1}{m} \sum_{x\in \text{Samp}_g} \frac{1}{d\binom{d-1}{|S|}} \tau_S(x,v) \qquad (1)$$

where m is the number of samples, and Samp_g draws samples uniformly at random from group $g \in G$.

The algorithm can then efficiently answer queries for any ancestor group definition $g' \in G$ $\epsilon - \delta$-accurately, by linearly combining feature importance from descendant groups $g \in G$. For example, we could estimate the SVs for any $g' \in G'$ where $G' = \{\text{women, men}\}$ or $G' = \{\text{Black, White}\}$, as such:

$$\tilde{\phi}_{g'\in G,i}(v) = \sum_{g\in G} \frac{n_g}{n} \tilde{\phi}_{g,i}(v) \qquad (2)$$

where n_g is the number of samples in group g.

The algorithm is described in 1, an illustrative example is given in Figure 2, and a complete proof of its correctness and error guarantee is in section 4.

Algorithm 1 Approximate group-wise feature importance for per-item outcomes

Input: Dataset $\mathcal{D}_g$, item $\mathbf{v}$, sampling method $Samp_g$, number of samples m, quantity of interest $\tau()$
Output: Shapley values $\phi(\mathbf{v})$ of $\mathbf{v}$'s features

```
: $\phi_g(i) = 0 \quad \forall i$
for $i \in \mathcal{A}$ do
    for $\forall \mathcal{S} \subseteq \mathcal{A} \setminus \{i\}$ do
        $\mathbf{U} = Samp_g(\mathcal{D}_g \setminus \mathbf{v}, m)$
        $\mathbf{U}_1 = \mathbf{v}_{\mathcal{A}\setminus\mathcal{S}}\mathbf{U}_{\mathcal{S}}$
        $\mathbf{U}_2 = \mathbf{v}_{\mathcal{A}\setminus\{\mathcal{S}\cup i\}}\mathbf{U}_{\mathcal{S}\cup i}$
        $\phi_{\mathcal{S}} = \frac{\tau(\mathbf{U}_1,\mathbf{U}_2)}{m}$
        $\phi_g(i) = \phi_g(i) + \frac{1}{d}\frac{1}{\binom{d-1}{|S|}}\phi_{\mathcal{S}}$
    end for
end for
return $\phi_g = [\phi_g(1), \ldots, \phi_g(d)]$
```

[a]

name	gender	salary	children	education	marital-status
Bob	m	5	2	4	M
Cal	m	8	1	7	U
Dia	f	4	0	4	M
Eli	f	4	2	2	U
Fay	f	4	1	7	M
Kat	f	4	0	3	U
Leo	m	4	0	3	M
Osi	m	7	1	7	U

[b]

v	i	S	m	$u \in \mathcal{D}$	U$_1$	U$_2$
Fay	salary	{edu, marital}	2	Dia	(f,4,1,**4**,**M**)	(f,**4**,1,**4**,**M**)
Fay	salary	{edu, marital}	2	Eli	(f,4,1,**2**,**U**)	(f,**4**,1,**2**,**U**)

Figure 2: (a) Dataset $\mathcal{D}$. (b) Frankenstein points for Fay when we are explaining the feature ”salary”, we use the coalition $\mathcal{S}$ of education and marital-status, and we do the approximate computation to create the Frankenstein points and only sample $m = 2$ data points from the same demographic group as the item (here, women).

4 Theoretical Analysis

We begin this section by defining the Shapley values for QII, using the notion of QoI. Without loss of generality, let groups be defined on the basis of gender, g={M (male), F (female)}

Definition 1. *We define the Shapley value ϕ of an item v as*

$$\phi_i(v) = \frac{1}{d}\sum_{S\subseteq A\setminus\{i\}} \frac{1}{\binom{d-1}{|S|}}.\frac{1}{n}\sum_{x\in D}\tau_S(x,v) \tag{3}$$

where $\tau_S(x,v) = f(v_S x_{\overline{S}}) - f(v_{S\cup\{i\}} x_{\overline{S\cup\{i\}}})$ is the quantity of interest(QoI).

Lemma 1. *The Shapley value of the union of any two subgroups is a linear combination of the individual Shapley values of the subgroups.*

Let $D = D_M \cup D_F$ and $n_M = |D_M|$, $n_F = |D_F|$. Then,

$$\phi_i(v) = \frac{n_M}{n}\phi_{M,i}(v) + \frac{n_F}{n}\phi_{F,i}(v) \tag{4}$$

Proof.

$$\phi_i(v) = \sum_{S\subseteq A\setminus\{i\}} \frac{1}{n}\sum_{x\in D}\frac{1}{d\binom{d-1}{|S|}}.\tau_S(x,v) \tag{5}$$

$$= \sum_{S\subseteq A\setminus\{i\}} \frac{1}{n}[\sum_{x\in D_M}\frac{1}{d\binom{d-1}{|S|}}.\tau_S(x,v) \tag{6}$$

$$+ \sum_{x\in D_F}\frac{1}{d\binom{d-1}{|S|}}.\tau_S(x,v)] \tag{7}$$

$$= \sum_{S\subseteq A\setminus\{i\}} [\frac{n_M}{n}(\frac{1}{n_M}\sum_{x\in D_M}\frac{1}{d\binom{d-1}{|S|}}.\tau_S(x,v)) \tag{8}$$

$$+ \frac{n_F}{n}(\frac{1}{n_F}(\sum_{x\in D_F}\frac{1}{d\binom{d-1}{|S|}}.\tau_S(x,v))] \tag{9}$$

$$= \frac{n_M}{n}(\sum_{S\subseteq A\setminus\{i\}}\frac{1}{n_M}\sum_{x\in D_M}\frac{1}{d\binom{d-1}{|S|}}.\tau_S(x,v)) \tag{10}$$

$$+ \frac{n_F}{n}(\sum_{S\subseteq A\setminus\{i\}}\frac{1}{n_F}\sum_{x\in D_F}\frac{1}{d\binom{d-1}{|S|}}.\tau_S(x,v))] \tag{11}$$

$$= \frac{n_M}{n}\phi_{M,i}(v) + \frac{n_F}{n}\phi_{F,i}(v) \tag{12}$$

□

We have shown that ancestor SV can be computed by linearly combining the SVs of their children. We will show that the approximation error is more than ϵ with probability δ in every such constructed SV if it $\epsilon - \delta$ for all its children. The QII paper provides a proof that this holds for each SV directly computed.

Theorem 1. *For $m \geq \mathcal{O}(\frac{1}{\epsilon^2})$ samples, the bottom-up algorithm 1 is $\epsilon - \delta$-accurate for all ancestor groups of g.*

$$\tilde{\phi}_i(v) = \frac{n_M}{n}\tilde{\phi}_{M,i}(v) + \frac{n_F}{n}\tilde{\phi}_{F,i}(v) \tag{13}$$

where

$$\tilde{\phi}_{g,i}(v) = \sum_{S\subseteq A\setminus\{i\}}\frac{1}{m}\sum_{x\in Samp(D_g)}\frac{1}{d\binom{d-1}{|S|}}\tau_S(x,v) \tag{14}$$

Proof. To prove the claim, it suffices to show that each $\tilde{\phi}_{g,i}(v)$ is $\epsilon - \delta$-accurate. To do this, we prove the following lemmas. □

Lemma 2 (Unbiasedness)**.** *The SVs our method calculates by combining the SVs children equals the SV that we would calculate directly.*

$$\mathbb{E}[\tilde{\phi}_i(v)] = \phi_i(v) \tag{15}$$

Proof. We begin by proving the lemma for the SV of each feature i for an item v at the population level.

$$\mathbb{E}[\tilde{\phi}_i(v)] = \mathbb{E}[\frac{n_M}{n}\tilde{\phi}_{M,i}(v) + \frac{n_F}{n}\tilde{\phi}_{F,i}(v)] \tag{16}$$

$$= \frac{n_M}{n}\mathbb{E}[\tilde{\phi}_{M,i}(v)] + \frac{n_F}{n}\mathbb{E}[\tilde{\phi}_{F,i}(v)] \tag{17}$$

$$= \frac{n_M}{n}\mathbb{E}[\phi_{M,i}(v)] + \frac{n_F}{n}\mathbb{E}[\phi_{F,i}(v)] \tag{18}$$

$$= \phi_i(v) \tag{19}$$

To show that this is true we need to calculate the expected values of the SV per group. We have that:

$$\tilde{\phi}_{M,i}(v) = \sum_{S\subseteq A\setminus\{i\}}\frac{1}{m}\sum_{x\in D_M}\frac{1}{d\binom{d-1}{|S|}} \tag{20}$$

$$\cdot \mathbb{1}[x \in \text{Samp}(D_M)] \cdot \tau_S(x,v) \tag{21}$$

$$= \sum_{S\subseteq A\setminus\{i\}}\frac{1}{m}\sum_{x\in D_M}\frac{1}{d\binom{d-1}{|S|}} \tag{22}$$

$$\cdot \mathbb{1}[x \in \text{Samp}(D_M)] \cdot \tau_S(x,v) \tag{23}$$

We now calculate the expected value of those quantities.

$$\begin{align}
\mathbb{E}[\tilde{\phi}_{M,i}(v)] &= \sum_{S\subseteq A\setminus\{i\}} \frac{1}{m} \sum_{x\in D_M} \frac{1}{d\binom{d-1}{|S|}} \tag{24}\\
&\cdot \mathbb{E}[\mathbb{1}[x \in \text{Samp}(D_M)].\tau_S(x,v)] \tag{25}\\
&= \sum_{S\subseteq A\setminus\{i\}} \frac{1}{m} \sum_{x\in D_M} \frac{1}{d\binom{d-1}{|S|}} \tau_S(x,v) \tag{26}\\
&\cdot \mathbb{E}[\mathbb{1}[x \in \text{Samp}(D_M)]] \tag{27}\\
&= \sum_{S\subseteq A\setminus\{i\}} \frac{1}{m} \sum_{x\in D_M} \frac{1}{d\binom{d-1}{|S|}} \tau_S(x,v) \cdot \frac{m}{n_M} \tag{28}\\
&= \phi_{M,i}(v) \tag{29}
\end{align}$$

□

Lemma 3 (Variance)**.** *The estimated SV deviates from the exact SV by atmost ϵ, with high probability*

$$|\phi_i(v) - \tilde{\phi}_{g,i}(v)| < \epsilon \qquad \textit{wp 0.9} \tag{30}$$

Proof. First, we will use the SV definition in the following form:

$$\begin{align}
\phi_{g,i}(v) &= \sum_{S\subseteq\mathcal{A}\setminus i} \sum_{x\in\mathcal{D}_g} \frac{1}{nd\binom{d-1}{|S|}} \tau_S(v,x) \tag{31}\\
&= \frac{1}{n} \sum_{x\in\mathcal{D}_g} \sum_{S\subseteq\mathcal{A}\setminus i} \frac{1}{d\binom{d-1}{|S|}} \tau_S(v,x) \tag{32}
\end{align}$$

To help with the proof, we define a vector $\vec{b}$ as follows:

$$\begin{align}
\vec{b} := [&\sum_{S\subseteq\mathcal{A}\setminus i} \frac{1}{d\binom{d-1}{|S|}} \tau_S(v,x_1) \tag{33}\\
&, \sum_{S\subseteq\mathcal{A}\setminus i} \frac{1}{d\binom{d-1}{|S|}} \tau_S(v,x_2), \ldots \tag{34}\\
&, \sum_{S\subseteq\mathcal{A}\setminus i} \frac{1}{d\binom{d-1}{|S|}} \tau_S(v,x_n)] \tag{35}
\end{align}$$

So now we have,

$$\phi_{g,i}(v) = \frac{1}{n} \sum_{i=1}^{n} b_i \tag{36}$$

We also define a vector $\vec{c}$ that contains the value b_i or 0 in position i depending on whether we sampled item i or not. So now, we have:

$$\tilde{\phi}_{g,i}(v) = \frac{1}{m} \sum_{i=1}^{m} c_i \tag{37}$$

And we can calculate the expected value of $\tilde{\phi}_i(v)$ as follows:

$$\begin{align}
\mathbb{E}[\tilde{\phi}_{g,i}(v)] &= \mathbb{E}[\frac{1}{m} \sum_{i=1}^{m} c_i] \tag{38}\\
&= \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}[c_i] \tag{39}\\
&= \phi_{g,i}(v) \tag{40}
\end{align}$$

Because:

$$\mathbb{E}[c_i] = \sum_{j=1}^{n} \Pr[c_i = b_j] b_j = \phi_{g,i}(v) \tag{41}$$

And we will also compute the variance.

$$\text{Var}[\tilde{\phi}_{g,i}(v)] = \frac{1}{m^2} \sum_{i=1}^{m} \text{Var}[c_i] \tag{42}$$

Where,

$$\begin{align}
\text{Var}[c_1] &= \mathbb{E}[c_1^2] - (\mathbb{E}[c_1])^2 \tag{43}\\
&= \sum_{j=1}^{n} \Pr[b_j^2 = c_1^2] b_j^2 - \phi_i(v)^2 \tag{44}\\
&= \frac{1}{n} \sum_{j=1}^{n} b_j^2 - \phi_i(v)^2 \tag{45}\\
&\le \frac{1}{n} \sum_{j=1}^{n} b_j^2 \tag{46}\\
&= \frac{1}{n} \|\vec{b}\|_2^2 \tag{47}\\
& \tag{48}
\end{align}$$

Finally, we can show that the error between the exact SV and the one we construct from the subgroups is small with high probability.

$$\begin{align}
|\phi_{g,i}(v) - \tilde{\phi}_{g,i}(v)| &\le C\sqrt{\text{Var}[\tilde{\phi}_{g,i}(v)]} \qquad \text{wp } 0.9 \tag{49}\\
&\le \frac{C}{\sqrt{mn}} \|\vec{b}\|_2 < \frac{C}{\sqrt{m}} \tag{50}\\
&< \varepsilon \qquad \text{if } m > \frac{C}{\varepsilon^2} \tag{51}
\end{align}$$

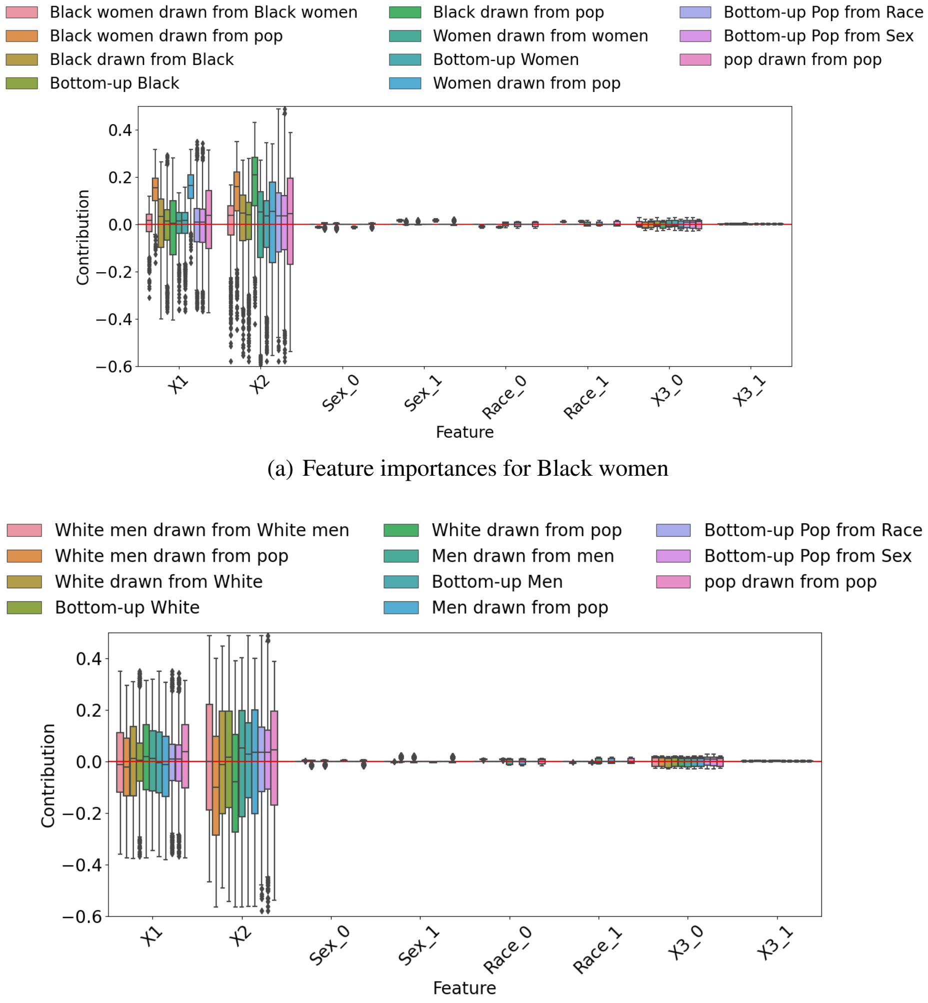
□

The claim of the theorem follows from these lemmas.

5 Experiments

We now compare the group-specific FIs computed by our bottom-up algorithm with the group-level and population-level FIs computed by QII, with experiments on real and synthetic datasets.

We implemented the QII framework, and extended it to sample from a specific group instead of over the full population. The quantity of interest we chose is the predicted probability of the positive class. We trained a $LogisticRegression$



(a) Feature importances for Black women

(b) Feature importances for White men

Figure 3: Synthetic: Comparing group-wise FI computed by different methods for [a] Black Women [b] White Men

as our black-box model, and used the $predict\_proba()$ function of the classifier to compute the quantity of interest. We set $m = 150$ which corresponds to $\epsilon = 0.1$ and $\delta = 0.1$.

We defined groups by an intersection of gender and race, and calculated the feature importance for all subgroups including ancestors, specifically: (Black men, Black women, White men, White women, Black people, White people, Men, and Women). In order to compare our method with the group-specific feature importances of QII, for each group we estimated two FIs: one by sampling only from that group (our bottom-up method) and one by sampling from the full population (QII). Lastly, we also calculated the SVs for the entire population, in order to validate our claim that global FIs do not capture group-specific FIs. This resulted in 17 experiments. Each experiment ran for approximately 30 minutes on 2021 Macbook M1 Pro with 16Gb memory.

We ran our experiments on two datasets: a synthetic dataset, and Folktables; a real-world fairness benchmark dataset. We present the results in figure 3 and 4. In each subplot, ie. for each intersectional group, we compare the following FI methods:

1. **intersectional groups drawn from intersectional group**: this is our bottom-up algorithm, applied to the intersectional group (gender, race)
2. **intersectional groups drawn from pop**: this is the QII estimate for the intersectional group (gender, race)
3. **single-attribute (race) group drawn from single-attribute (race) group**: this is our bottom-up algorithm, applied to the single-attribute group (race)
4. **bottom-up for single-attribute(race) group**: this is the estimate for the single-attribute (race) group computed as a linear combination of the intersectional group estimates
5. **single-attribute (race) group drawn from population**: this is the QII estimate for the single-attribute (race) group

6. **single-attribute (gender) group drawn from single-attribute (gender) group**: this is our bottom-up algorithm, applied to the single-attribute group (gender)
7. **bottom-up for single-attribute (gender) group**: this is the estimate for the single-attribute (race) group computed as a linear combination of the intersectional group estimates of the bottom-up algorithm
8. **single-attribute (gender) drawn from population**: this is the QII estimate for the single-attribute (gender) group
9. **bottom-up for population, aggregated over race**: this is the estimate for the population-level FIs computed as a linear combination of the single-attribute group (race) estimates of the bottom-up algorithm
10. **bottom-up for population, aggregated over gender/sex**: this is the estimate for the population-level FIs computed as a linear combination of the single-attribute group (gender) estimates of the bottom-up algorithm
11. **population-level drawn from the population**: this is the QII estimate for the global/population-level FIs

5.1 Synthetic

To construct the data, we sampled $n = 60,000$ from the data-generating process described below. We modified the procedure from (8) to include multiple features that depend on protected group membership, specifically, gender/sex(S) and race(R). 80% of the samples are used to train a $LogisticRegression$ classifier (with the default hyperparameter settings), and the remaining 20% are for testing. The trained model has an accuracy of 0.7613.

$$\begin{gathered}
S \sim \text{Bernoulli}(0.2) \\
R \sim \text{Bernoulli}(0.4) \\
X_1 | S = 1 \text{ (Women)} \sim 20 * \text{Beta}(2, 7) \\
X_1 | S = 0 \text{ (Men)} \sim 20 * \text{Beta}(2, 2) \\
X_2 | R = 1 \text{ (Black)} \sim 20 * \text{Beta}(2, 5) \\
X_2 | R = 0 \text{ (White)} \sim 20 * \text{Beta}(2, 2) \\
X_3 | G = 1 \text{ and } R = 1 \sim \text{Bernoulli}(0.7) \\
X_3 | G = 0 \text{ or } R = 0 \sim \text{Bernoulli}(0.4) \\
z_1 = 1/(1+exp-(0.2*X_1+0.3*X_2+0.2*X_3-\text{Normal}(5, 0.5)) \\
Y_1 | X_1, X_2, X_3 \sim \text{Bernoulli}(z_1)
\end{gathered}$$

We use our method to explain the predictions of the trained model on the test set in Figure 3. Due to space constraints, we report only the intersectionally privileged and disadvantaged groups — namely White men and Black women, in Figure 3, and defer the remaining plots to the appendix.

Results We can see that the group-specific feature importances for Black Women and White Men are starkly different. Specifically, feature $X1$ and $X2$ have large positive contribution for the former, and a negative contribution for the latter. Further the population-level estimate is closer to the trend for Black women. Hence, if recourse was suggested to all applicants based only on the population-level results (from QII) then recourse would be less effective for White Men than Black Women.

5.2 Real: Folktables

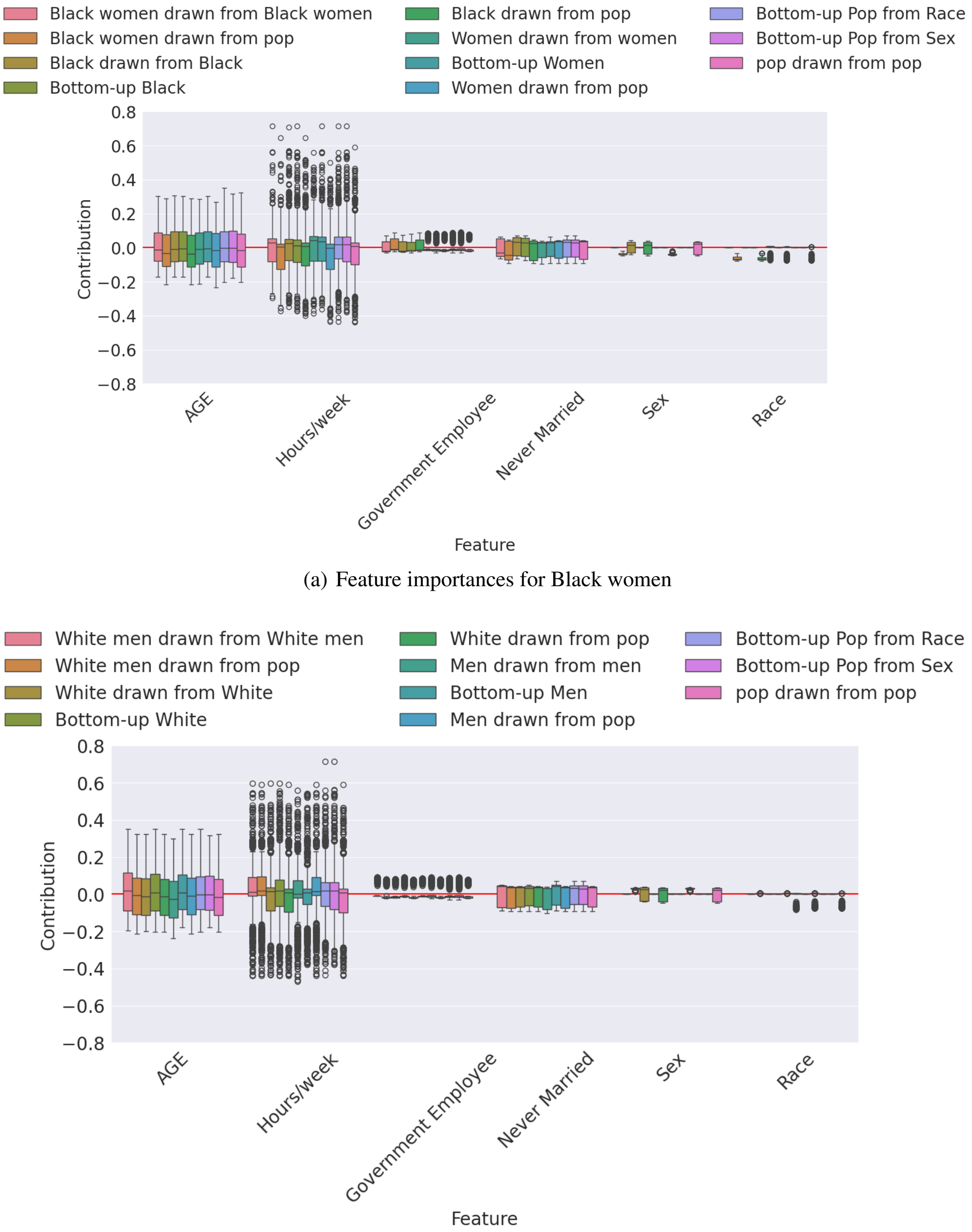
Folktables (4) is a dataset consisting of US Census data. It consists of several datasets that formulate different tasks. We used the ACSIncome dataset, whose aim is to predict whether the income of USA-based adults is over \$50,000 USD. It consists of several features, namely Age, Worker Class, Education attainment, Marital Status, Occupation, Place of birth, Relationship to the person submitting the census, Usual hours worked per week, Sex, and Race. We downloaded 200,000 datapoints from the California 2018 census.

For the purposes of this experiment, and due to the exponential nature of the SV, we dropped many features and combined the values of others into larger groups to reduce the cardinality of the feature. Specifically, we dropped, Occupation, Place of birth, Relationship to the person submitting the census, and Education attainment. Then we combined the Marital Status into three categories (*Married/Widowed*, *Divorced/Separated*, and *Never Married*) from the original five. Finally, we combined the class of worker into four categories (*Employee*: Employee of a private for-profit/Employee of a private not-for-profit, *Government Employee*: Local government employee/State government employee/Federal government employee, *Self-employed*: Self-employed in own not incorporated business, professional practice, or farm/Self-employed in own incorporated business, professional practice or farm, and finally *Unemployed*: Working without pay in family business or farm/Unemployed and last worked 5 years ago or earlier or never worked instead) of the original nine. From the Race feature, we kept only White and Black people, to reduce cardinality. Finally, the Sex feature is binary, and we retain it as is. For our analysis, we chose to use the features Sex and Race as those that define the subgroups. While we run our analysis on all these features, here we present only the features that differ the most between groups due to the lack of space. To make the computations easier, we also dropped any datapoint that contained NULLs, resulting in 68,661 datapoints out of which 64,131 are White and 4,530 are Black. The Sex feature is more balanced.

We used 80% of the samples to train a $LogisticRegression$ classifier, and the remaining 20% are for testing. The accuracy we achieved is 0.7631 and is as expected for this task.

Results We show one plot for each maximally intersectional group (Black men, Black women, White men, White women) and present the feature importance for each ancestor group. Therefore, we can compare the feature importance between maximally intersectional groups (by comparing different plots) and for each ancestor (by looking at each plot). Due to space constraints, we report only the intersectionally privileged and disadvantaged groups — namely White men and Black women, in Figure 4, and defer the remaining plots to the appendix.

Comparing across plots, we can see that the method attributes high importance to the hours worked for both Black and White men when the feature importance is calculated at the Men group level. However, for White men, the feature importance is high when the feature importance is calculated in the White men subgroup. This indicates that the feature



(a) Feature importances for Black women

(b) Feature importances for White men

Figure 4: Folktables: Comparing group-wise feature importances computed by different methods for [a] Black Women [b] White Men

has high importance for White men and if Black Men were explained at the Men level, they would incorrectly be recommended to increase their work hours. However, we see that bottom-up Men do not overestimate the importance of this feature due to the balanced nature of the aggregation we are proposing.

There are many other results of this nature, for instance, Black Women have a positive effect on income when they are Government employees and that can be seen only when we draw directly from them or the Black population. Black Men, however, do not have the same feature importance when the features are drawn from their population directly. Indicating that this effect is caused in the Black subgroup by Black Women.

6 Conclusion

In this work we introduced the notion of *group-specific* feature importance, asserting that global feature importance is only a meaningful construct under strong distributional assumptions that do not hold in practice. Instead, we provide an efficient and provably accurate algorithm, that computes ϵ-accurate FIs for any group-definition of the user's choice. Our method is invariant to the group-definition, and provides ϵ-accurate estimates for all ancestral groups in addition to the defined group. Through experiments, we demonstrate how our approach computes more meaningful estimates of FIs, that can guide reliable recourse recommendations and can be used to detect unfairness.

References

[1] AAS, K., JULLUM, M., AND LØLAND, A. Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *Artificial Intelligence 298* (2021), 103502.

[2] CHEN, H., JANIZEK, J. D., LUNDBERG, S., AND LEE, S.-I. True to the model or true to the data?, 2020.

[3] DATTA, A., SEN, S., AND ZICK, Y. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE Symposium on Security and Privacy (SP)* (2016), pp. 598–617.

[4] DING, F., HARDT, M., MILLER, J., AND SCHMIDT, L. Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems 34* (2021), 6478–6490.

[5] FRYE, C., ROWAT, C., AND FEIGE, I. Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. *Advances in Neural Information Processing Systems 33* (2020), 1229–1239.

[6] KUMAR, I. E., VENKATASUBRAMANIAN, S., SCHEIDEGGER, C., AND FRIEDLER, S. Problems with shapley-value-based explanations as feature importance measures. In *International Conference on Machine Learning* (2020), PMLR, pp. 5491–5500.

[7] LIPOVETSKY, S., AND CONKLIN, M. Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry 17*, 4 (2001), 319–330.

[8] LIPTON, Z., MCAULEY, J., AND CHOULDECHOVA, A. Does mitigating ml's impact disparity require treatment disparity? *Advances in neural information processing systems 31* (2018).

[9] LIU, T., AND UNGAR, L. Towards cotenable and causal shapley feature explanations. In *AAAI 2021 Workshop: Trustworthy AI for Healthcare* (2021).

[10] LUNDBERG, S. M., AND LEE, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems 30* (2017).

[11] MERRICK, L., AND TALY, A. The explanation game: Explaining machine learning models using shapley values. In *Machine Learning and Knowledge Extraction: 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020, Dublin, Ireland, August 25–28, 2020, Proceedings 4* (2020), Springer, pp. 17–38.

[12] PASTOR, E., DE ALFARO, L., AND BARALIS, E. Identifying biased subgroups in ranking and classification. *arXiv preprint arXiv:2108.07450* (2021).

[13] SHAPLEY, L. S., ET AL. A value for n-person games.

[14] SUNDARARAJAN, M., AND NAJMI, A. The many shapley values for model explanation. In *International conference on machine learning* (2020), PMLR, pp. 9269–9278.

A Appendix

A.1 QII Framework

The QII (3) framework begins by defining a "Quantitity of Interest" (QoI or τ), which is the output of the black-box model that we are trying to explain. QoIs include individual predictions for a single sample, predictions aggregated for a specific demographic group, or the disparity in outcomes between groups. The framework introduces several *metrics*, namely unary, set, and marginal QII, to evaluate the importance of a feature i to a specific QoI output of the black-box model.

Consider that we are trying to explain the outcome for an item **v**. To measure the QoIs over a range of possible feature value configurations, QII constructs new items (Frankenstein points) from the feature vector of **v** and one or more randomly sampled datapoints. Finally, to combine these measurements into a feature importance score, it employs SV — a popular aggregation technique from game theory. To compute exact feature importance, we must aggregate QoIs over all data points in the dataset (for more details see Algorithm 2 and Figure 1). For a large dataset with many features, this grows exponentially in the number of features and is thereby prohibitively computationally expensive. Instead, the same paper proposes an ϵ-δ approximation, that requires at least $m \geq \log(2/\delta)/2\epsilon^2$ samples.

A.2 QII-Exact

The exact QII calculates the feature importance by constructing Frankenstein datapoints using the entire dataset.

Algorithm 2 (Exact) feature importance for per-item outcomes

```
Input: Dataset $\mathcal{D}$, item $\mathbf{v}$, $\tau()$
Output: Shapley values $\phi(\mathbf{v})$ of v's features
: $\phi(i) = 0 \quad \forall i$
for $i \in \mathcal{A}$ do
    for $\forall \mathcal{S} \subseteq \mathcal{A} \setminus \{i\}$ do
        $\mathbf{U}_1 = \mathbf{v}_{\mathcal{A} \setminus \mathcal{S}} \mathcal{D}_{\mathcal{S}}$
        $\mathbf{U}_2 = \mathbf{v}_{\mathcal{A} \setminus \{\mathcal{S} \cup i\}} \mathcal{D}_{\mathcal{S} \cup i}$
        $\phi_{\mathcal{S}} = \frac{\tau(\mathbf{U}_1, \mathbf{U}_2)}{|\mathcal{D}|}$
        $\phi(i) = \phi(i) + \frac{1}{d} \frac{1}{\binom{d-1}{|\mathcal{S}|}} \phi_{\mathcal{S}}$
    end for
end for
return $\phi = [\phi(1), \ldots, \phi(d)]$
```

A.3 QII-Approximate

The QII approximate computation is calculating the SVs using an i.i.d. sample from the population.

Algorithm 3 Approximate feature importance for per-item outcomes

```
Input: Dataset $\mathcal{D}$, item $\mathbf{v}$, sampling method $Samp$, number of samples $m$, $\tau()$
Output: Shapley values $\phi(\mathbf{v})$ of v's features
: $\phi(i) = 0 \quad \forall i$
for $i \in \mathcal{A}$ do
    for $\forall \mathcal{S} \subseteq \mathcal{A} \setminus \{i\}$ do
        $\mathbf{U} = Samp(\mathcal{D} \backslash \mathbf{v}, m)$
        $\mathbf{U}_1 = \mathbf{v}_{\mathcal{A} \setminus \mathcal{S}} \mathbf{U}_{\mathcal{S}}$
        $\mathbf{U}_2 = \mathbf{v}_{\mathcal{A} \setminus \{\mathcal{S} \cup i\}} \mathbf{U}_{\mathcal{S} \cup i}$
        $\phi_{\mathcal{S}} = \frac{\tau(\mathbf{U}_1, \mathbf{U}_2)}{m}$
        $\phi(i) = \phi(i) + \frac{1}{d} \frac{1}{\binom{d-1}{|\mathcal{S}|}} \phi_{\mathcal{S}}$
    end for
end for
return $\phi = [\phi(1), \ldots, \phi(d)]$
```

Using the running example, the Frankensein datapoints that will be constructed looks like the figure 5 [b].

[a]

name	gender	salary	children	education	marital-status
Bob	m	5	2	4	M
Cal	m	8	1	7	U
Dia	f	4	0	4	M
Eli	f	4	2	2	U
Fay	f	4	1	7	M
Kat	f	4	0	3	U
Leo	m	4	0	3	M
Osi	m	7	1	7	U

[b]

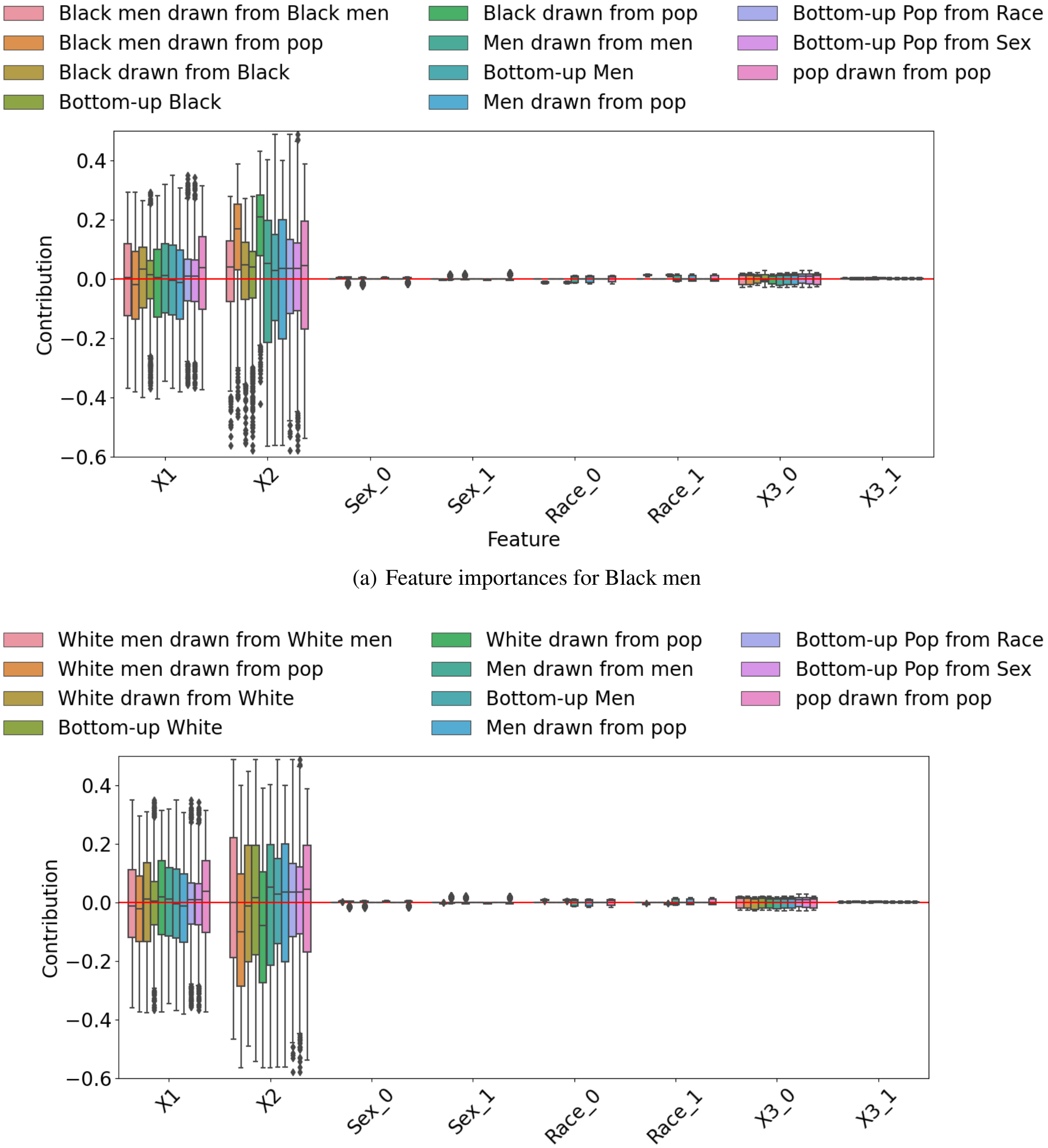
v	i	S	m	$u \in \mathcal{D}$	U_1	U_2
Fay	salary	{edu, marital}	4	Cal	(f,4,1,**7**, **U**)	(f,**8**,1,**7**,**U**)
Fay	salary	{edu, marital}	4	Kat	(f,4,1,**3**, **U**)	(f,**4**,1,**3**,**U**)
Fay	salary	{edu, marital}	4	Leo	(f,4,1,**3**, **M**)	(f,**4**,1,**3**,**M**)
Fay	salary	{edu, marital}	4	Osi	(f,4,1,**7**, **U**)	(f,**7**,1,**7**,**U**)

Figure 5: [a] Dataset $\mathcal{D}$. [b] Frankenstein points for Fay when we are explaining the feature "salary", we use the coalition $\mathcal{S}$ of education and marital-status and we do approximate computation with $m = 4$ sampling from the entire dataset (instead of the group).

A.4 Experimental results

Synthetic dataset : We report the group-wise feature importance computed by our approach and the QII framework, for [a] Black Men, [b] White women in in Figure 6.

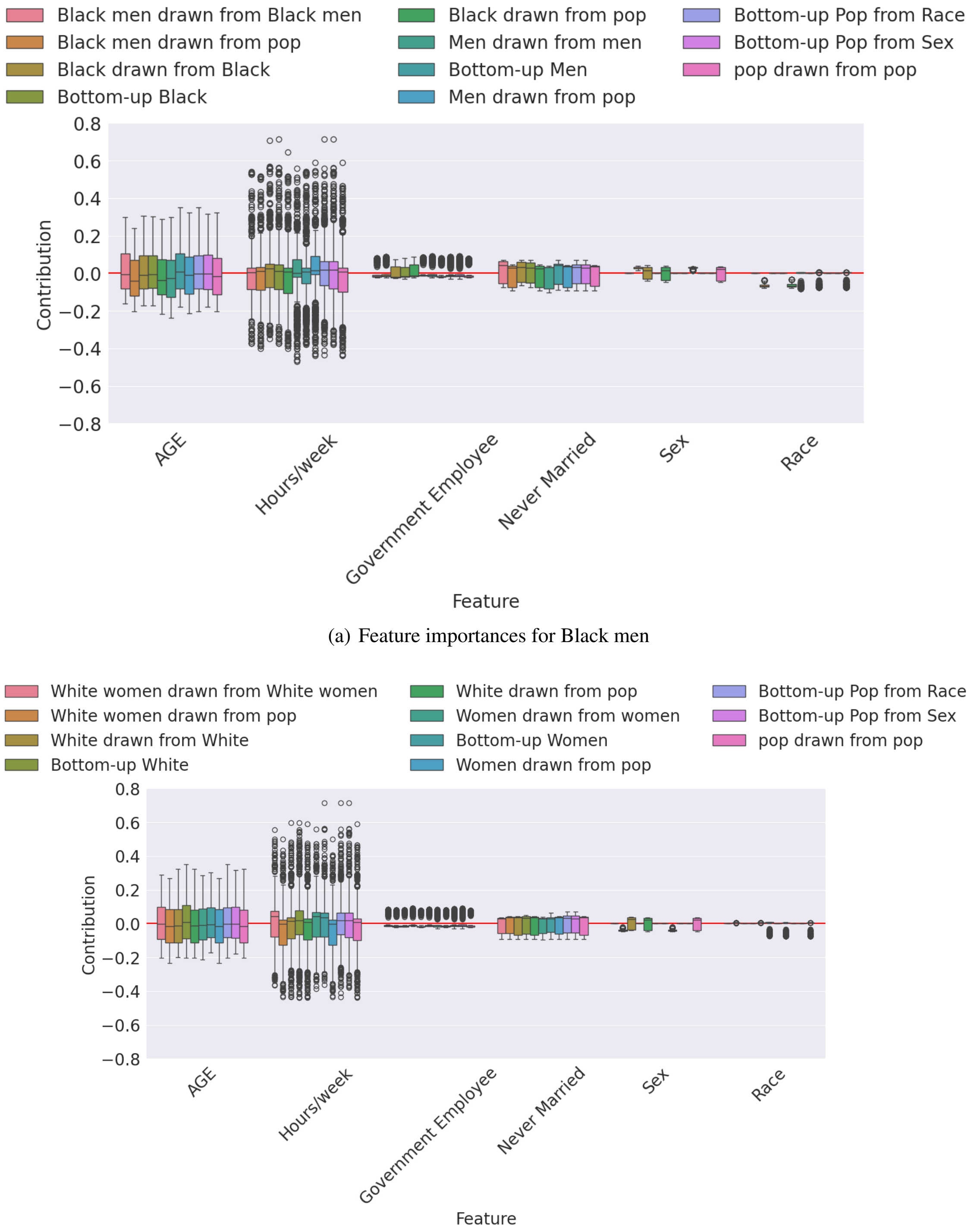
Folktables : We report the group-wise feature importance computed by our approach and the QII framework, for [a] Black Men, [b] White women in Figure 7.



(a) Feature importances for Black men

(b) Feature importances for White women

Figure 6: Synthetic: Comparing FIs computed by different methods for [a] Black Men, [b] White women



(a) Feature importances for Black men

(b) Feature importances for White women

Figure 7: Folktables: Comparing FIs computed by different methods for [a] Black Men, [b] White Women