Final Project Report

Juejing Han, Qian Fang
jhan446@gatech.edu, qianfang@gatech.edu

1 FINAL PROJECT REPORT

1.1 Dataset Selection (Step 1)

Selected Dataset: Credit Card Approvals[1]

Regulated Domain: Credit

Number of Observations: 690

Number of Variables: 16

In this report, "Approved" is designated as the first outcome variable and is referred to as **Dependent Variable 1**. Derived from variables within the dataset, "Y" has been chosen as the second outcome variable and is identified as **Dependent Variable 2**.

Dependent Variable 2 (Y) is calculated based on the formula:

Creditworthiness = (If Approved, then 40; else 0) + (If No_PriorDefault, then 20; else 0) + (If Income > 50k, then 30; If 5k < Income ≤ 50k, then 20; else 0) + (If CreditScore > 0, then 10; else 0). Creditworthiness < 60, Y = 0 (rejected); else Y = 1 (approved).

Number of Variables Associated with Protected Class: 5 (**Per class instruction, Marital Status is categorized under the protected class of Familial Status**[2])

Table 1 — Variables Associated with Protected Class.

Variable	Protected Class	Legal Precedents/Laws
Gender	Sex	Equal Pay Act of 1963; Civil Rights Act of 1964, 1991
Age	Age	Age Discrimination in Employment Act of 1967 (over 40)
Married	Familial Status	Civil Rights Act of 1968
Ethnicity	Race	Civil Rights Act of 1964, 1991
Citizen	Citizenship	Immigration Reform and Control Act

1 https://www.kaggle.com/datasets/samuelcortinhas/credit-card-approval-clean-data

2 https://edstem.org/us/courses/49501/discussion/4540513

1.2 Explore the Dataset (Step 2)

1.2.1 *Subgroups of Protected Class (Step 2.1 – 2.2)*

The protected class of Age contains continuous values ranging from 13.75 to 80.25 and is grouped into two subgroups: 40 & Under, and Above 40. The protected class of Race initially includes 5 subgroups: White, Black, Asian, Latino, and Other. In this report, Asian, Latino, and Other are consolidated into a single subgroup named "Other." The protected class of Citizenship initially includes 3 subgroups: ByBirth, ByOtherMeans, and Temporary. In this report, ByOtherMeans and Temporary are consolidated into a single subgroup named "OtherMeans."

Table 2 — Subgroups of Protected Class.

Protected Class	Subgroup	Numerical Value of Subgroup
Sex	Female	0
	Male	1
Age	40 & Under	0
	Above 40	1
Familial Status	Single/Divorced/etc.	0
	Married	1
Race	White	0
	Black	1
	Other (Asian, Latino, Other)	2
Citizenship	By Birth	0
	OtherMeans (By Other Means, Temporary)	1

1.2.2 *Select Protected Classes & Frequency (Step 2.3 – 2.5)*

Selected Protected Classes: Age and Familial Status

Table 3 — Frequency of Age with Dependent Variable 1.

Protected Class	Subgroup	Approved	Rejected
Age	40 & Under	221	327
	Above 40	86	56

Table 4 — Frequency of Familial Status with Dependent Variable 1.

Protected Class	Subgroup	Approved	Rejected
Familial Status	Single/Divorced/etc.	47	118
	Married	260	265

Table 5 — Frequency of Age with Dependent Variable 2.

Protected Class	Subgroup	Approved	Rejected
Age	40 & Under	173	375
	Above 40	59	83

Table 6 — Frequency of Familial Status with Dependent Variable 2.

Protected Class	Subgroup	Approved	Rejected
Familial Status	Single/Divorced/etc.	34	131
	Married	198	327

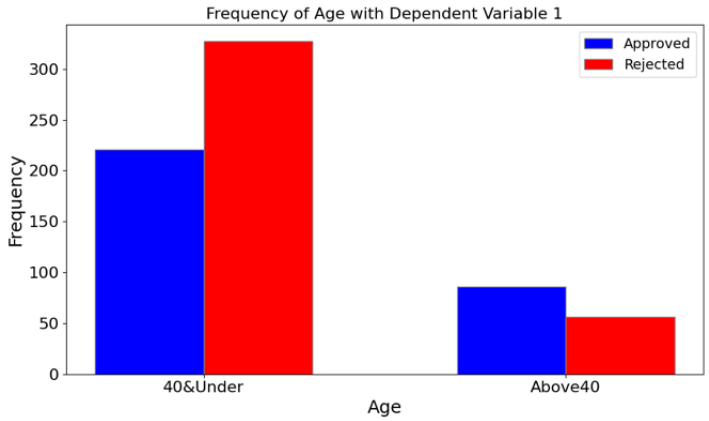

Figure 1 — Frequency of Age with Dependent Variable 1

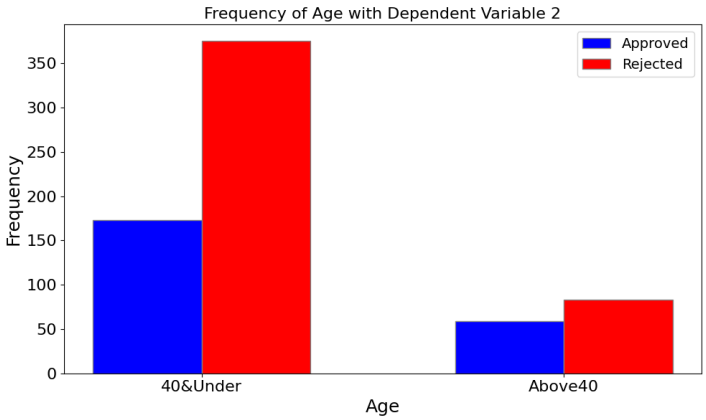

Figure 2 — Frequency of Age with Dependent Variable 2

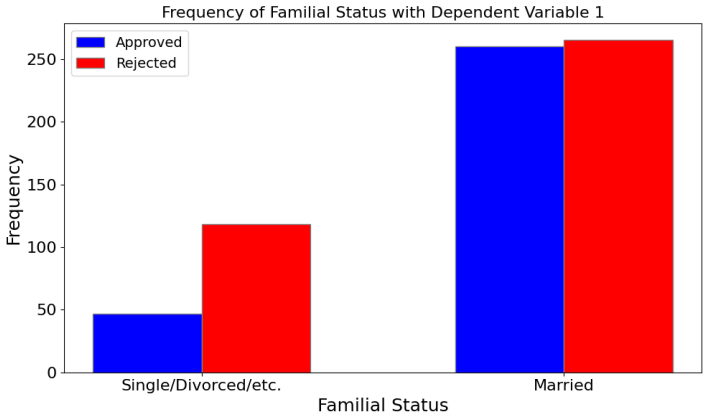

Figure 3 — Frequency of Familial Status with Dependent Variable 1

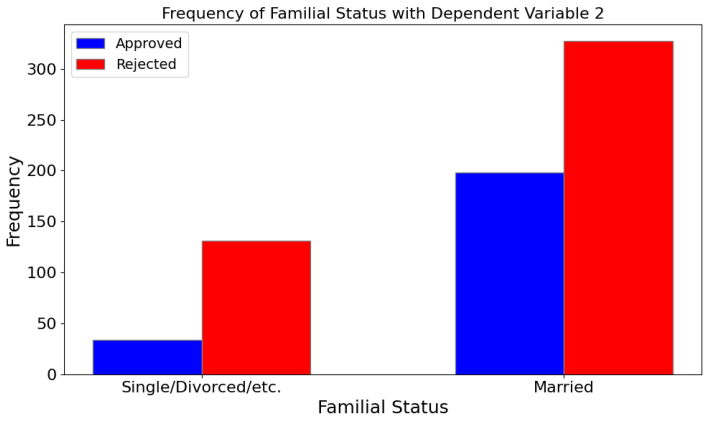

Figure 4 — Frequency of Familial Status with Dependent Variable 2

1.3 Fairness Metric & Bias Mitigation (Step 3)

1.3.1 *Privileged & Unprivileged Groups (Step 3.1)*

Table 7 demonstrates the privileged and unprivileged groups of the protected classes selected in Step 2.

Individuals over 40 are generally considered more financially stable than their younger counterparts, thus categorizing the "Above 40" subgroup as privileged and the "40 & Under" subgroup as unprivileged within the protected class of Age. Similarly, for the protected class of Familial Status, married individuals are presumed to have a stronger financial foundation. Consequently, the "Married" subgroup is deemed privileged, whereas the "Single/Divorced/etc." subgroup is classified as unprivileged.

Table 7 — Privileged & Unprivileged Groups of Protected Class.

Protected Class	Privileged Group	Unprivileged Group
Age	Above 40	40 & Under
Familial Status	Married	Single/Divorced/etc.

1.3.2 *Fairness Metrics (Step 3.2)*

Fairness metrics: Disparate Impact (DI) and Statistical Parity Difference (SPD) [3].

Disparate Impact (DI) quantifies the ratio of favorable outcome rates for the unprivileged group relative to the privileged group. To achieve fairness, the DI value should be 1. DI > 1 suggests a bias in favor of the unprivileged group, while DI < 1 indicates a bias in favor of the privileged group. In this analysis, a DI value within the range of 0.8 to 1.2 is considered acceptable, and the bias is deemed non-significant.

Statistical Parity Difference (SPD) quantifies the disparity in favorable outcome rates between the unprivileged and privileged groups. To achieve fairness, the SPD value should be 0. SPD > 0 indicates a bias in favor of the unprivileged group, whereas SPD < 0 signifies a bias in favor of the privileged group. In this analysis, an SPD value within the range of -0.1 to 0.1 is considered acceptable, and the bias is deemed non-significant.

Table 8 — Results of Fairness Metrics.

Protected Class	Dependent Variable 1		Dependent Variable 2	
	DI	SPD	DI	SPD
Age	0.67	-0.20	0.76	-0.10
Familial Status	0.58	-0.21	0.55	-0.17

3 https://www.mathworks.com/help/risk/explore-fairness-metrics-for-credit-scoring-model.html

Table 8 presents the values of fairness metrics for protected classes in relation to the dependent variables. The findings reveal that both DI and SPD indicate a distinctive bias against the unprivileged group across each protected class. The SPD value for the protected class of Age, when associated with Dependent Variable 2, stands at -0.1, positioning it at the threshold of indicating significant bias.

1.3.3 *Bias Mitigation (Step 3.3 – 3.4)*

Reweighting[4] is employed to the original dataset to address bias, with Dependent Variable 1 designated as the outcome variable.

Table 9 — Results of Fairness Metrics after Reweighing.

Protected Class	Dependent Variable 1		Dependent Variable 2	
	DI	SPD	DI	SPD
Age	0.73	-0.16	0.82	-0.07
Familial Status	0.65	-0.17	0.61	-0.15

Table 9 displays the outcomes of fairness metrics following bias mitigation efforts. Each metric progresses toward its respective ideal fairness benchmark (1 for DI and 0 for SPD), improving from their pre-mitigation standings. Notably, for the protected class of Age concerning Dependent Variable 2, both the DI and SPD values fall within their acceptable ranges.

1.4 Mitigating Bias (Step 4)

The original and transformed datasets are split into 80% training data and 20% testing data randomly. The training data is used to train a Random Forest classifier and the Dependent Variable 1 is selected as the output label (same as Step 3.3) for both datasets. Then the trained classifier is used to calculate the same fairness metrics DI and SPD as in Step 3.2 for the privileged and unprivileged groups of the two protected classes Age and Familiar Status.

The differences in the outcomes of privileged vs unprivileged groups for each fairness metric and dataset after training the classifier are presented in Tables 10 to 11.

4 https://aif360.readthedocs.io/en/latest/modules/algorithms.html - module-aif360.algorithms.preprocessing

Table 10 — Differences in outcomes of privileged vs unprivileged groups – Training Classifier on Original Dataset.

Protected Class	Dependent Variable	Fairness Metric	Value	Privileged/Unprivileged Group Difference
Age	1	DI	0.73	Favor privileged & bias against unprivileged
		SPD	-0.18	Favor privileged & bias against unprivileged
	2	DI	0.93	Slightly favor privileged & bias against unprivileged
		SPD	-0.03	Slightly favor privileged & bias against unprivileged
Familiar Status	1	DI	0.79	Favor privileged & bias against unprivileged
		SPD	-0.12	Favor privileged & bias against unprivileged
	2	DI	0.80	Slightly favor privileged & bias against unprivileged
		SPD	-0.08	Slightly favor privileged & bias against unprivileged

Table 11 — Differences in outcomes of privileged vs unprivileged groups – Training Classifier on Transformed Dataset.

Protected Class	Dependent Variable	Fairness Metric	Value	Privileged/Unprivileged Group Difference
Age	1	DI	0.81	Slightly favor privileged & bias against unprivileged
		SPD	-0.11	Favor privileged & bias against unprivileged
	2	DI	0.93	Slightly favor privileged & bias against unprivileged
		SPD	-0.03	Slightly favor privileged & bias against unprivileged
Familiar Status	1	DI	0.73	Favor privileged & bias against unprivileged
		SPD	-0.14	Favor privileged & bias against unprivileged
	2	DI	0.80	Slightly favor privileged & bias against unprivileged
		SPD	-0.08	Slightly favor privileged & bias against unprivileged

The fairness metrics results are listed in Tables 12 to 15. **Changes in fairness metrics below 0.05 are deemed minor; those between 0.05 and 0.2 (inclusive) are classified as considerable; and changes above 0.2 are considered significant**.

Table 12 — Age (Independent) to Dependent Variable 1 Fairness Metrics.

Dataset	DI	Change compared to previous	SPD	Change compared to previous
Original	0.67	NA	-0.20	NA
After Transforming	0.73	Considerable positive change	-0.16	Minor positive change
After Training Classifier on Original	0.73	Considerable positive change	-0.18	Minor positive change
After Training Classifier on Transformed	0.81	Considerable positive change	-0.11	Considerable positive change

Table 13 — Familiar Status (Independent) to Dependent Variable 1 Fairness Metrics.

Dataset	DI	Change compared to previous	SPD	Change compared to previous
Original	0.58	NA	-0.21	NA
After Transforming	0.65	Considerable positive change	-0.17	Minor positive change
After Training Classifier on Original	0.79	Significant positive change	-0.12	Considerable positive change
After Training Classifier on Transformed	0.73	Considerable positive change	-0.14	Minor positive change

Table 14 — Age (Independent) to Dependent Variable 2 Fairness Metrics.

Dataset	DI	Change compared to previous	SPD	Change compared to previous
Original	0.76	NA	-0.10	NA
After Transforming	0.82	Considerable positive change	-0.07	Minor positive change
After Training Classifier on Original	0.93	Considerable positive change	-0.03	Considerable positive change
After Training Classifier on Transformed	0.93	Considerable positive change	-0.03	Minor positive change

Table 15 — Familiar Status (Independent) to Dependent Variable 2 Fairness Metrics.

Dataset	DI	Change compared to previous	SPD	Change compared to previous
Original	0.55	NA	-0.17	NA
After Transforming	0.61	Considerable positive change	-0.15	Minor positive change
After Training Classifier on Original	0.80	Significant positive change	-0.08	Considerable positive change
After Training Classifier on Transformed	0.80	Considerable positive change	-0.08	Considerable positive change

1.5 Analysis (Step 5)

1.5.1 *Project Team Members*

Team member 1: Juejing Han – jhan446

Team member 2: Qian Fang – qfang36

1.5.2 *Fairness Metrics Graphs*

Figures 5 to 12 are from Step 3.2, illustrating the fairness metrics for the original dataset.

Figures 13 to 20 are from Step 3.4, demonstrating the fairness metrics for the transformed dataset.

Figures 21 to 28 are from Step 4.3, displaying the fairness metrics for the after-training classifier on the original dataset.

Figures 29 to 36 are from Step 4.6, showing the fairness metrics for the after-training classifier on the transformed dataset.

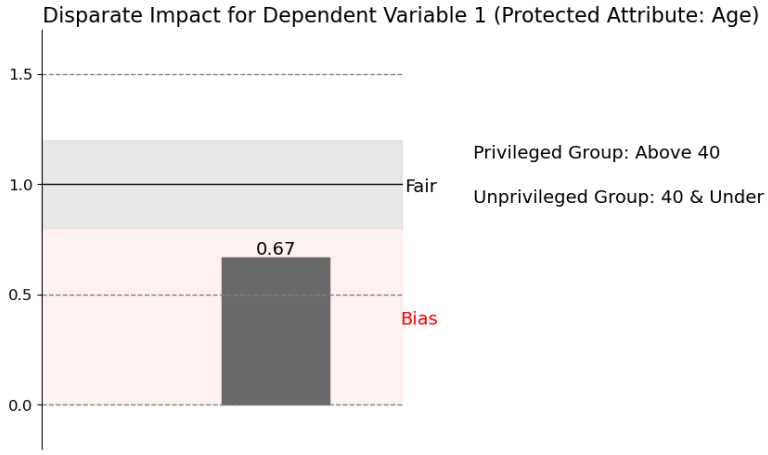

Figure 5 — Age to Dependent Variable 1 DI (Step 3.2)

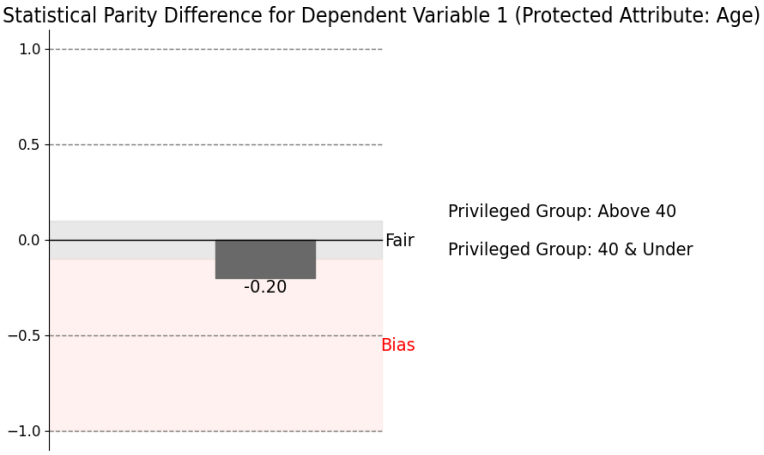

Figure 6 — Age to Dependent Variable 1 SPD (Step 3.2)

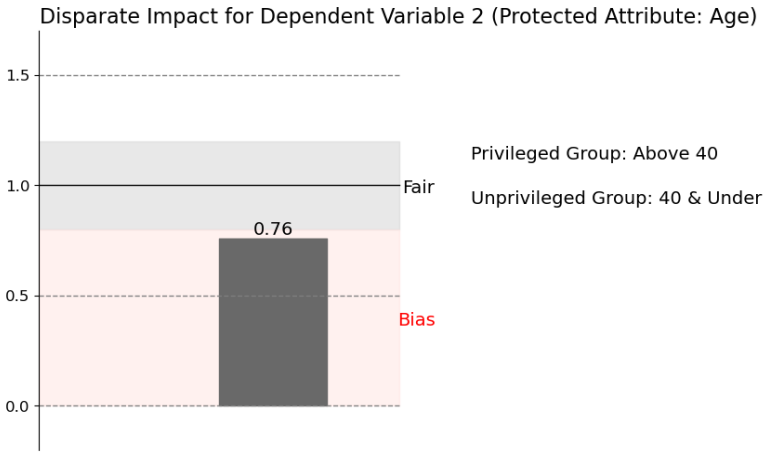

Figure 7 — Age to Dependent Variable 2 DI (Step 3.2)

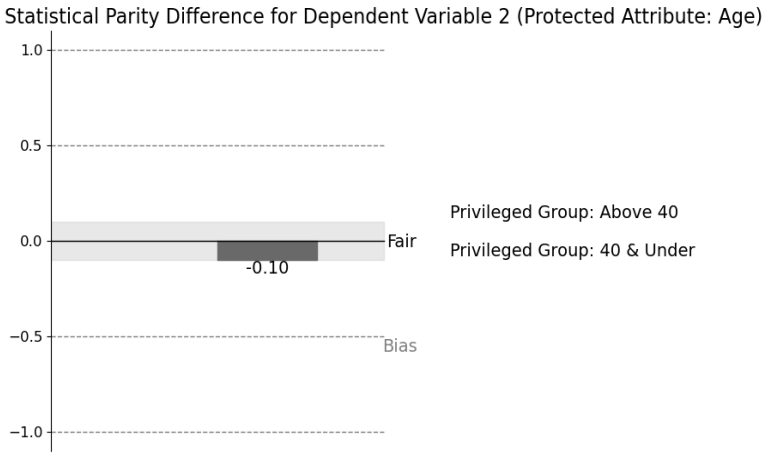

Figure 8 — Age to Dependent Variable 2 SPD (Step 3.2)

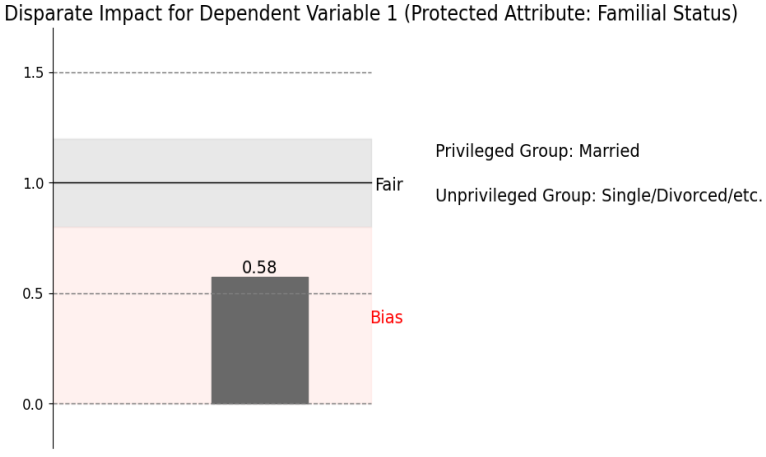

Figure 9 — Familial Status to Dependent Variable 1 DI (Step 3.2)

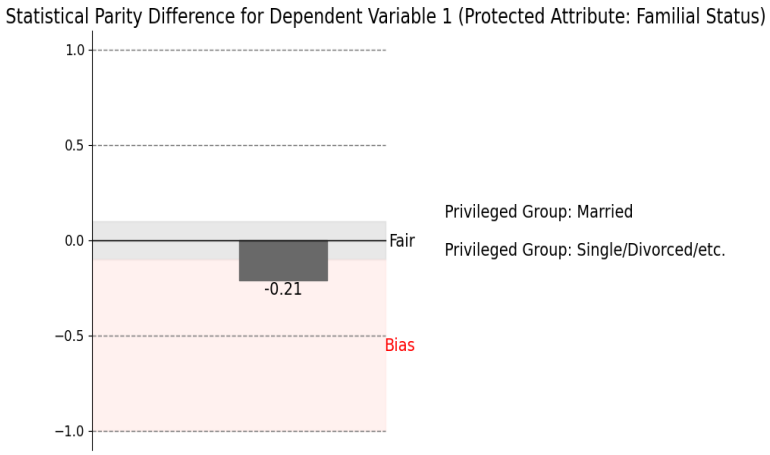

Figure 10 — Familial Status to Dependent Variable 1 SPD (Step 3.2)

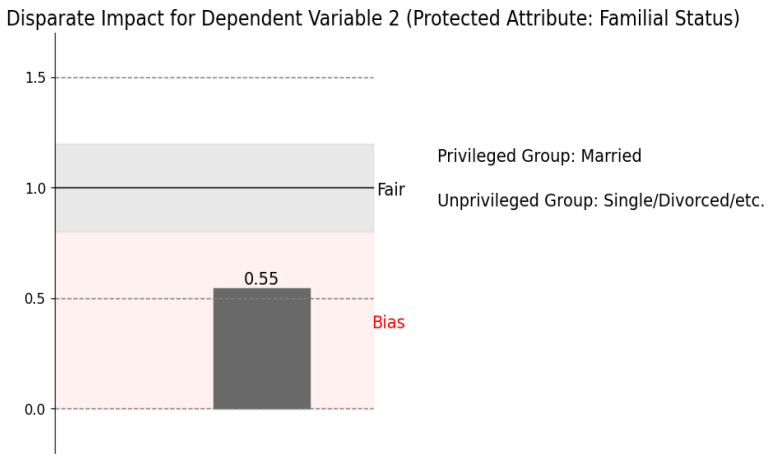

Figure 11 — Familial Status to Dependent Variable 2 DI (Step 3.2)

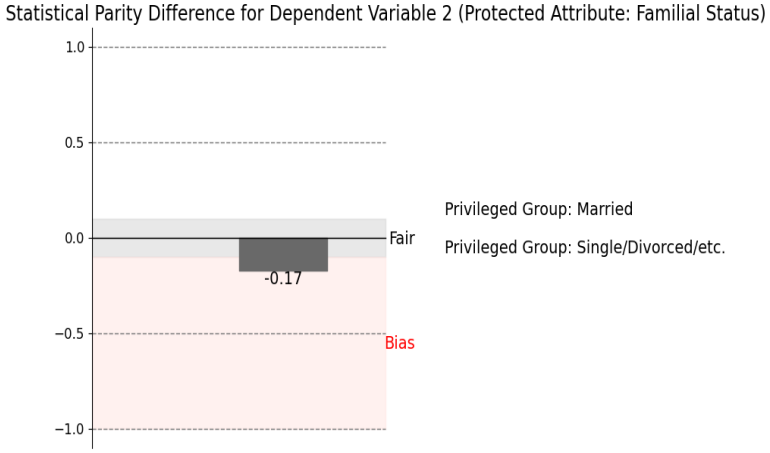

Figure 12 — Familial Status to Dependent Variable 2 SPD (Step 3.2)

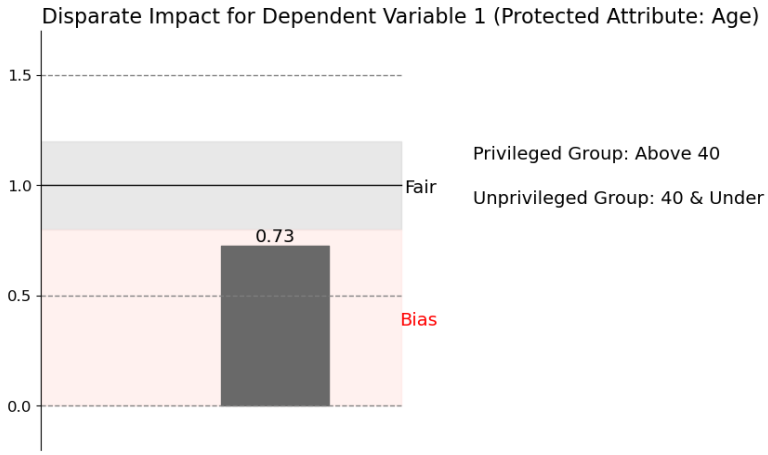

Figure 13 — Age to Dependent Variable 1 DI after Reweighing (Step 3.4)

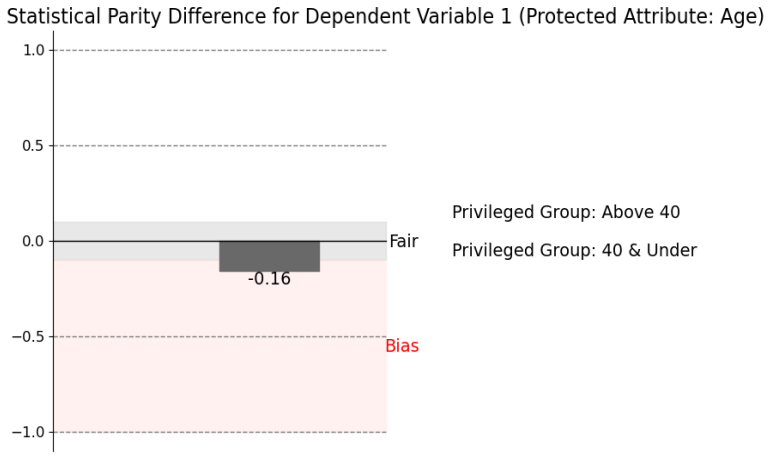

Figure 14 — Age to Dependent Variable 1 SPD after Reweighing (Step 3.4)

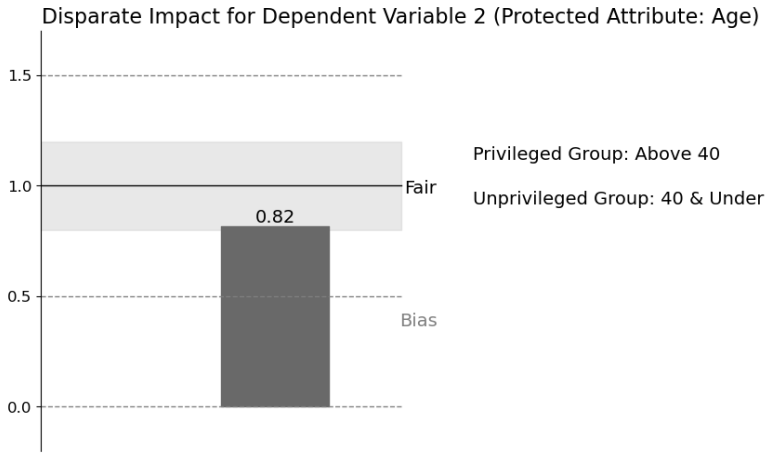

Figure 15 — Age to Dependent Variable 2 DI after Reweighing (Step 3.4)

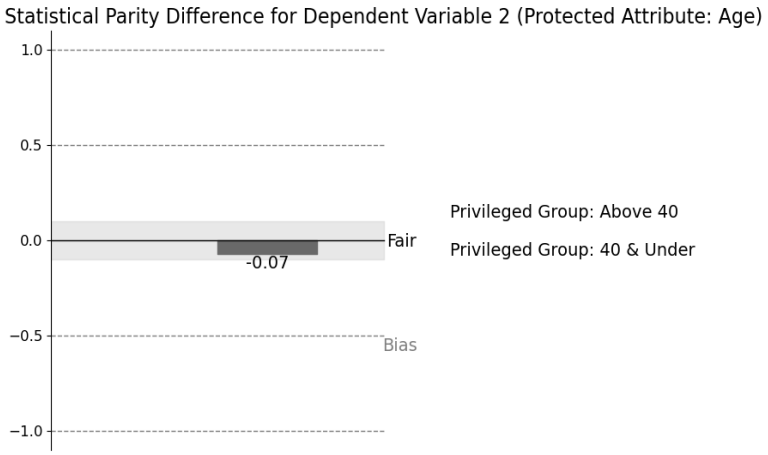

Figure 16 — Age to Dependent Variable 2 SPD after Reweighing (Step 3.4)

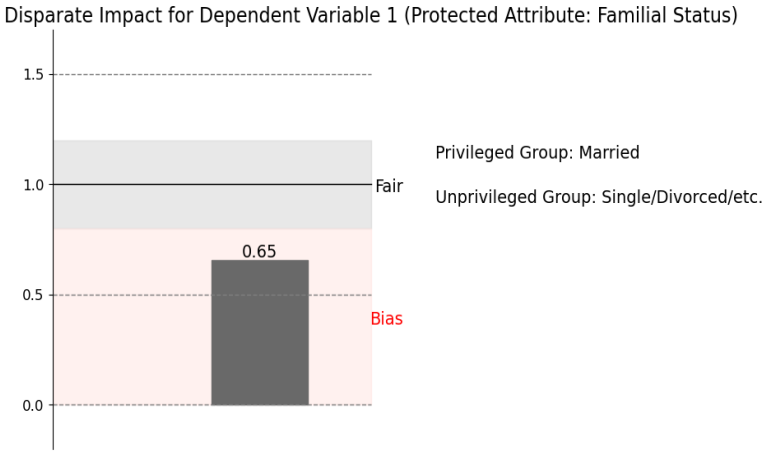

Figure 17 — Familial Status to Dependent Variable 1 DI after Reweighing (Step 3.4)

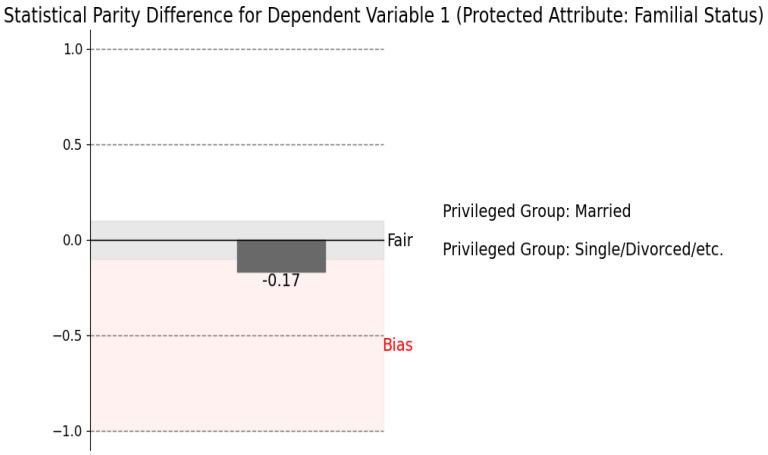

Figure 18 — Familial Status to Dependent Variable 1 SPD after Reweighing (Step 3.4)

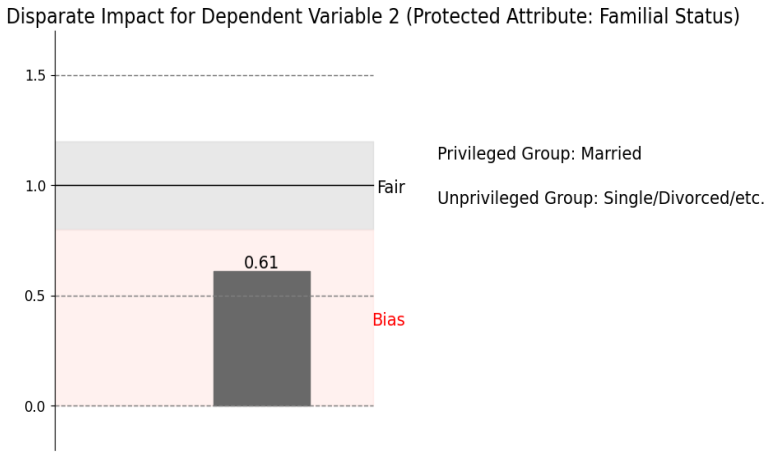

Figure 19 — Familial Status to Dependent Variable 2 DI after Reweighing (Step 3.4)

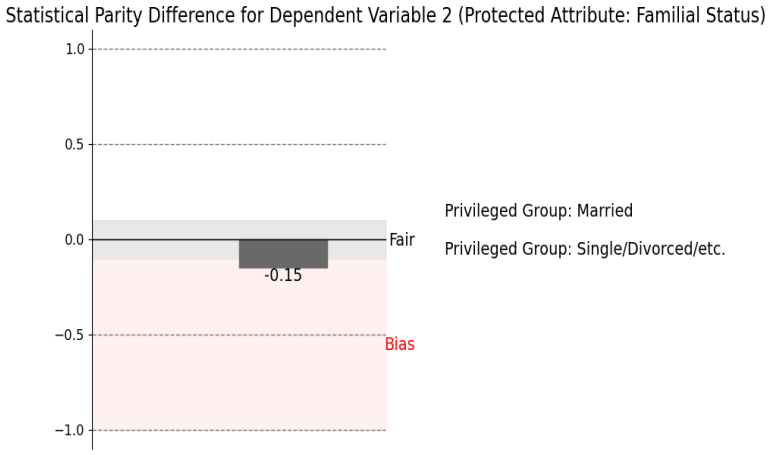

Figure 20 — Familial Status to Dependent Variable 2 SPD after Reweighing (Step 3.4)

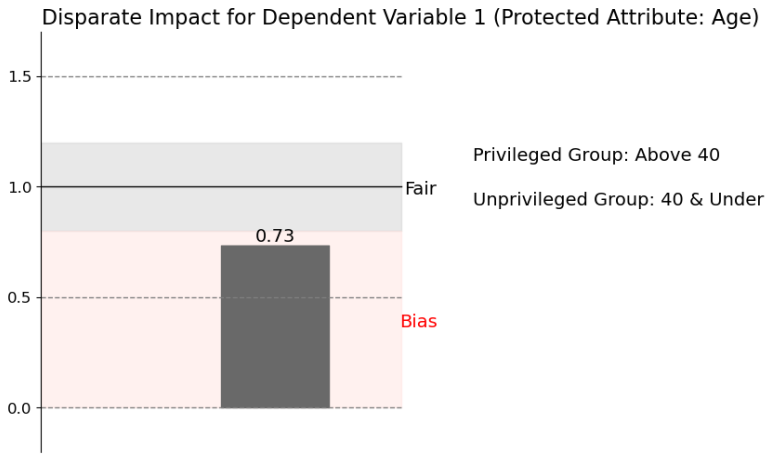

Figure 21 — Age to Dependent Variable 1 DI after Training Classifier on Original Dataset (Step 4.3)

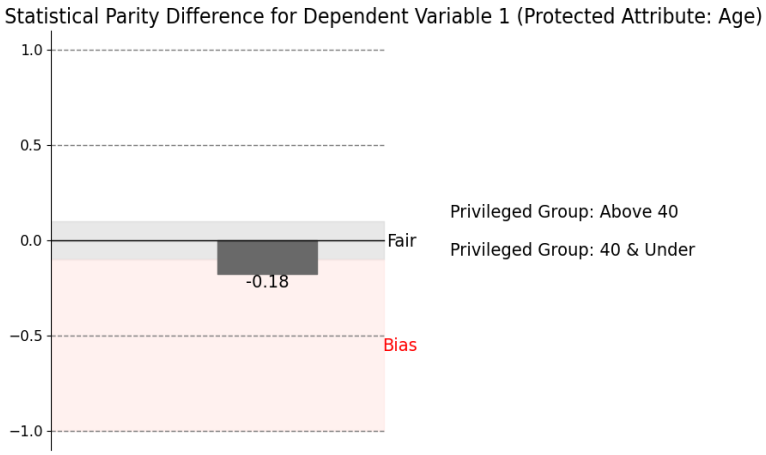

Figure 22 — Age to Dependent Variable 1 SPD after Training Classifier on Original Dataset (Step 4.3)

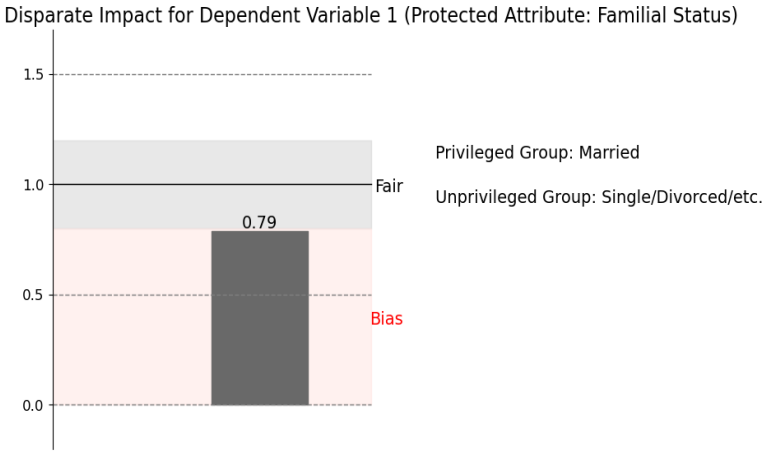

Figure 23 — Familial Status to Dependent Variable 1 DI after Training Classifier on Original Dataset (Step 4.3)

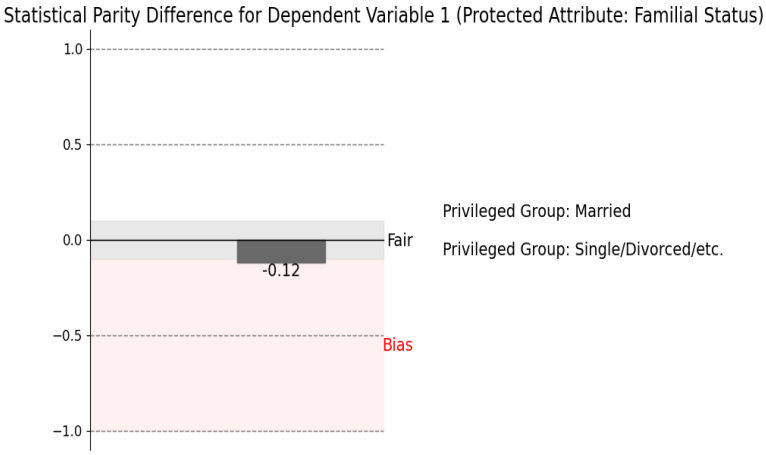

Figure 24 — Familial Status to Dependent Variable 1 SPD after Training Classifier on Original Dataset (Step 4.3)

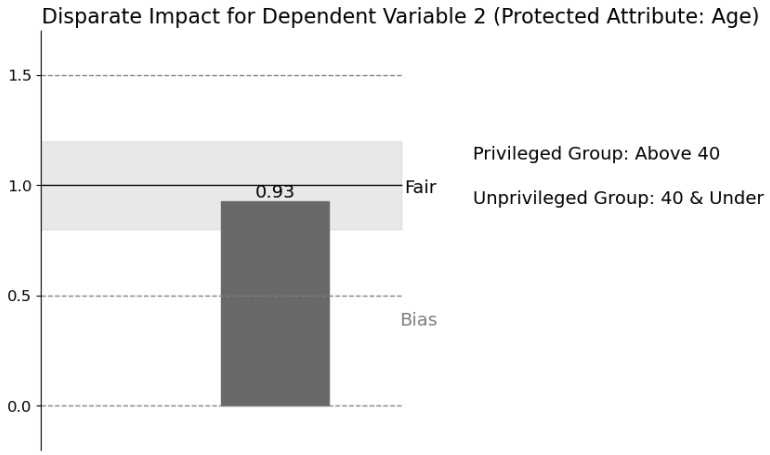

Figure 25 — Age to Dependent Variable 2 DI after Training Classifier on Original Dataset (Step 4.3)

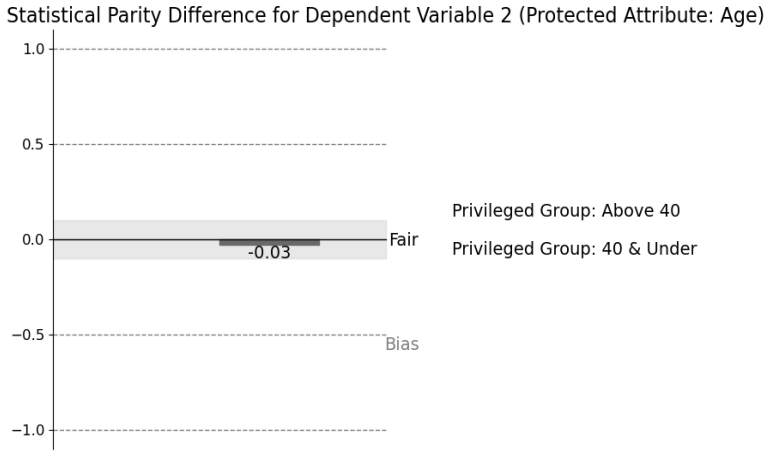

Figure 26 — Age to Dependent Variable 2 SPD after Training Classifier on Original Dataset (Step 4.3)

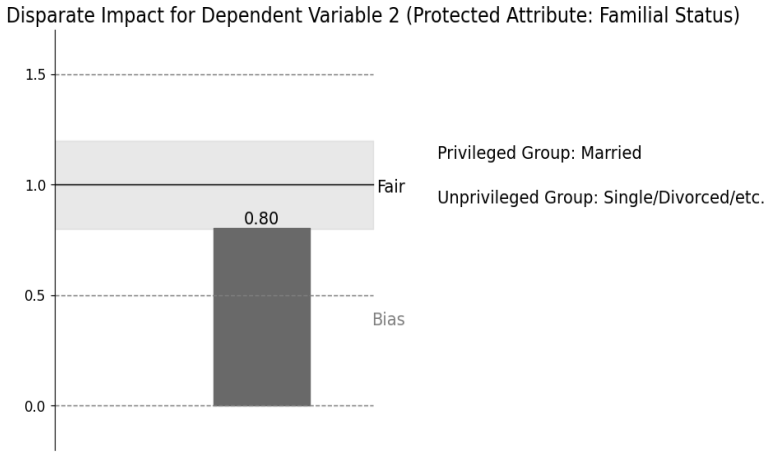

Figure 27 — Familial Status to Dependent Variable 2 DI after Training Classifier on Original Dataset (Step 4.3)

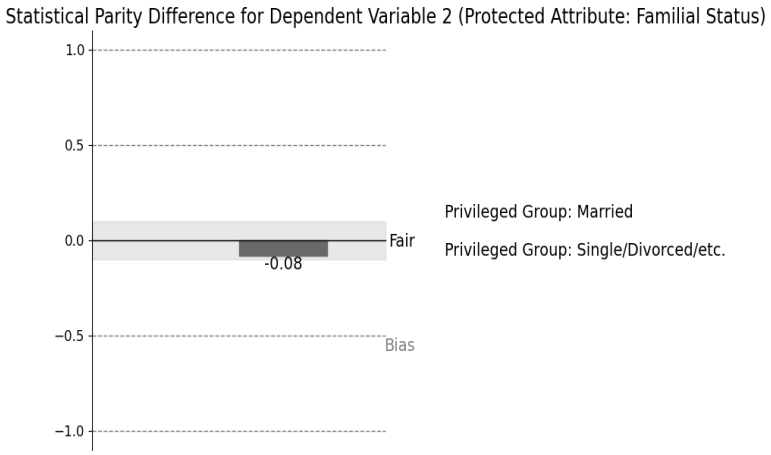

Figure 28 — Familial Status to Dependent Variable 2 SPD after Training Classifier on Original Dataset (Step 4.3)

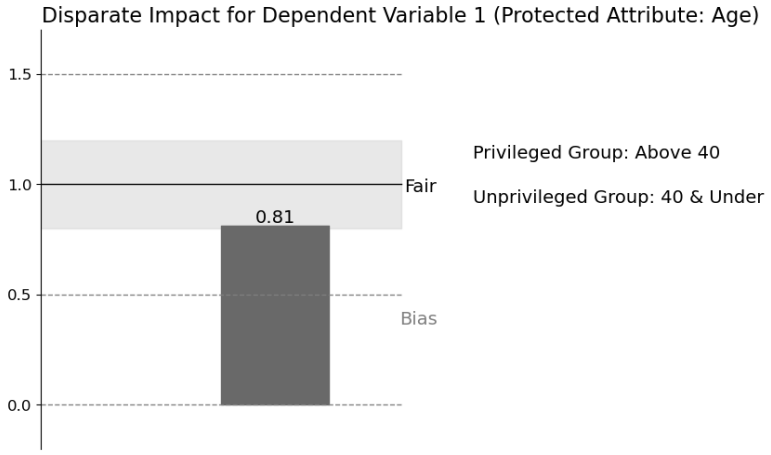

Figure 29 — Age to Dependent Variable 1 DI after Training Classifier on Transformed Dataset (Step 4.6)

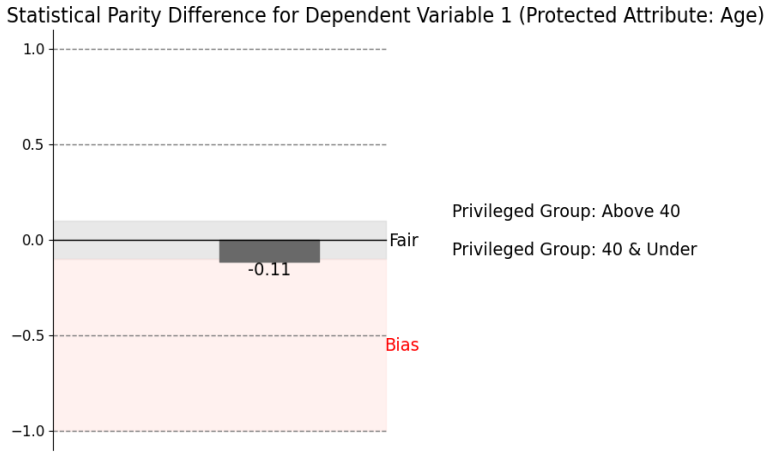

Figure 30 — Age to Dependent Variable 1 SPD after Training Classifier on Transformed Dataset (Step 4.6)

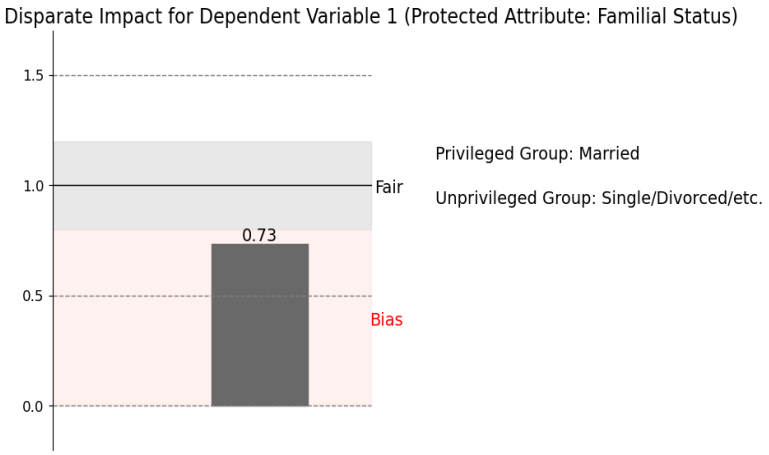

Figure 31 — Familial Status to Dependent Variable 1 DI after Training Classifier on Transformed Dataset (Step 4.6)

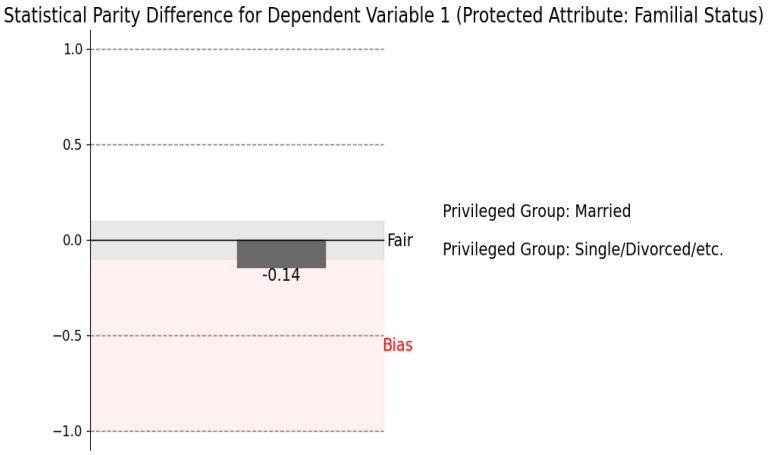

Figure 32 — Familial Status to Dependent Variable 1 SPD after Training Classifier on Transformed Dataset (Step 4.6)

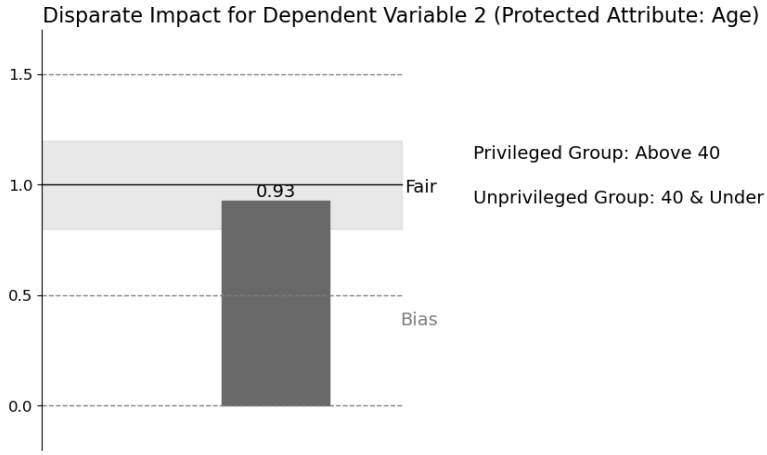

Figure 33 — Age to Dependent Variable 2 DI after Training Classifier on Transformed Dataset (Step 4.6)

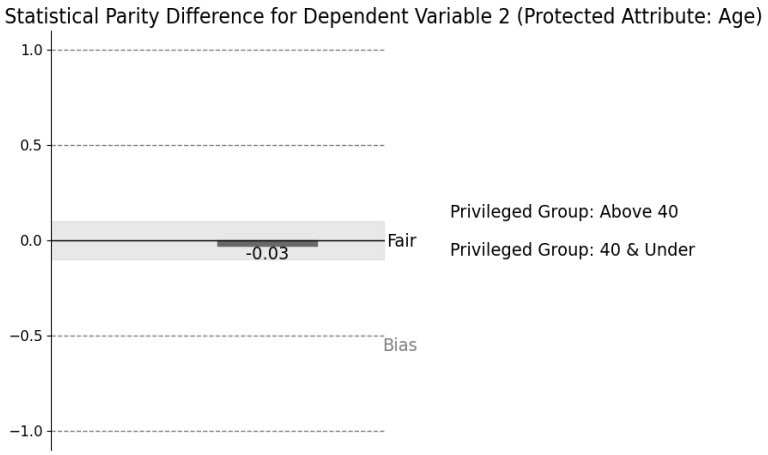

Figure 34 — Age to Dependent Variable 2 SPD after Training Classifier on Transformed Dataset (Step 4.6)

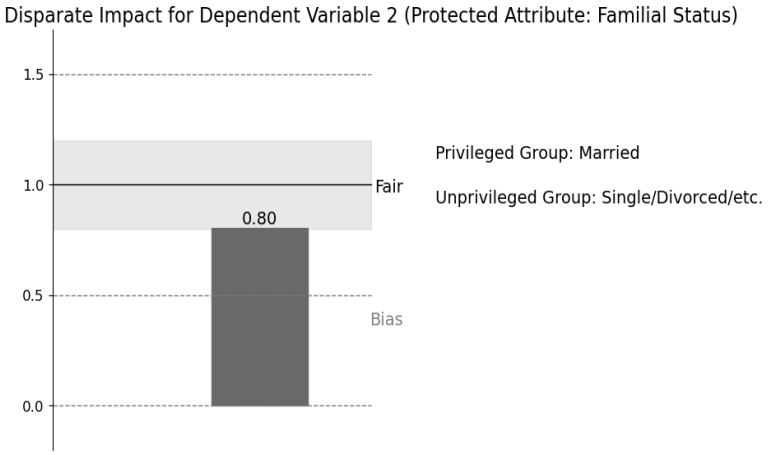

Figure 35 — Familial Status to Dependent Variable 2 DI after Training Classifier on Transformed Dataset (Step 4.6)

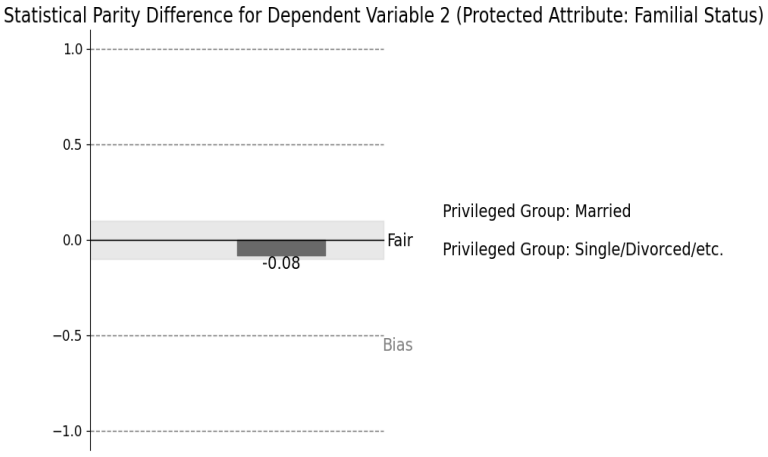

Figure 36 — Familial Status to Dependent Variable 2 SPD after Training Classifier on Transformed Dataset (Step 4.6)

In this study, both fairness metrics (DI and SPD) demonstrate effectiveness, with **SPD potentially offering superior insights compared to DI**. SPD measures a more absolute difference in the probability of favorable outcomes between different groups compared with DI. To illustrate, consider a scenario where the favorable outcome rate for the unprivileged group shifts from 30% to 40%, and the favorable outcome rate for the privileged group shifts from 60% to 80%. In this case, DI values remain constant (0.75), while SPD values change from -0.3 to -0.4. This is further exemplified in Table 12, where DI values stay at 0.73, and SPD values reveal a variance (0.16 versus 0.18).

Therefore, **SPD is considered better in this study**. It is a more direct measurement and easier to interpret the differences in approval rates between unprivileged and privileged groups, helping promote fairness and equal accessibility to credit opportunities.

1.5.3 *Analysis Answers*

Juejing Han (jhan446):

1) The approaches employed in this report (reweighing and Random Forest) have successfully mitigated bias. The fairness metrics, Disparate Impact (DI) and Statistical Parity Difference (SPD), showed positive changes toward their ideal fairness benchmarks (1 for DI and 0 for SPD) after bias mitigation. Compared to the original dataset, the statistical enhancements signify progress in reducing bias and prompting fairness.

2) The unprivileged groups within the selected protected classes of Age and Familial Status received positive advantages. Initially, the dataset indicated a bias against these groups, with DI values below 1 and SPD values below 0. After bias mitigation, both DI and SPD values shifted toward their ideal fairness benchmarks, illustrating a decrease in bias against the unprivileged groups.

3) Conversely, the privileged groups in the same protected classes were disadvantaged as they experienced a reduction in their previously advantageous position. Both DI and SPD values shifted toward their ideal fairness benchmarks, diminishing the initial bias in favor of the privileged groups.

4) The bias mitigation approaches utilized in this report have limitations, and potential issues might arise if these approaches are implemented to mitigate bias.

Reweighing aims to balance outcomes by treating the protected attribute and outcome label as independent[5], but this assumption may not always hold true. Its effectiveness depends on the data comprehensiveness and balance. If the dataset is severely skewed,

5 https://towardsdatascience.com/fairmodels-lets-fight-with-biased-machine-learning-models-f7d66a2287fc

reweighing might be ineffective. Furthermore, it has the potential to overcorrect, such as introducing reverse bias, affecting fairness towards other groups.

Mitigating bias with Random Forest requires meticulous data preprocessing, careful feature selection, and potential modification to the training process. Unbalanced data can exacerbate existing disparities, favoring the majority class unduly and compromising the model's ability to generalize well for the minority class. Moreover, the inherent complexity of its structure poses challenges to interpreting how decisions are made, which is crucial for identifying and addressing biases.

Qian Fang (qfang36):

The reweighting transformation method and the random forest classifier both worked to mitigate bias, because the fairness metrics all have considerable positive changes compared with the results of the original dataset.

Since the results of original dataset indicate that the privileged group received more favorable outcomes than the unprivileged group, after bias mitigation and/or training classifier, the unprivileged group received a positive advantage, while the privileged group was disadvantaged by these approaches.

If these methods were used to mitigate bias, there might be several issues. For reweighting, it assumes the estimates of group membership probabilities are accurate, which may not be always available; and the method is sensitive to data noise. For random forest training classifier, it is hard to interpret and might amplify bias if there is certain bias in the training data.