Fairness & Bias Assignment Report

Juejing Han
jhan446@gatech.edu

1 FAIRNESS & BIAS ASSIGNMENT REPORT

1.1 Explore the Dataset (Step 2)

Selected Dataset: Taiwan Credit Dataset

Number of Observations: 30000

Number of Variables: 24 (ID column not included)

Number of Variables Associated with Protected Class: 3 (**Per class instruction, Marital Status is categorized under the protected class of Familial Status**[1])

Table 1 — Variables Associated with Protected Class.

Variable	Protected Class	Legal Precedents/Laws
Sex (X2)	Sex	Equal Pay Act of 1963; Civil Rights Act of 1964, 1991
Age (X5)	Age	Age Discrimination in Employment Act of 1967 (over 40)
Marriage (X4)	Familial Status	Civil Rights Act of 1968

1.2 Defining Creditworthiness & Preparing Dataset (Step 3)

1.2.1 *Outcome Variable for Approving/Denying a Loan (Step 3.1)*

Outcome variable: Y (0 – Approved; 1 – Denied)[2]

1.2.2 *Creditworthiness Formula (Step 3.2)*

Default Payment, Credit Amount, Education, and Payment History have been selected to create the creditworthiness formula:

(If No_Default, then 40; else 0) + (If Credit_Amount > 500k, then 30; If 300k < Credit_Amount ≤ 500k, then 20; If 150k < Credit_Amount ≤ 300k, then 10; else 0) + (If Above_High_School, then 10; else 0) + (If No_Delayed_Payment, then 20; else 0)

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

2 https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients

1.2.3 *Protected Class Attribute & Privileged/Unprivileged Groups (Step 3.3 - 3.4)*

Select Protected Class Attribute: Sex

Privileged Group: Male

Unprivileged Group: Female

1.2.4 *Split Dataset (Step 3.5)*

Randomly split the original dataset into equally divided training and test sets, each containing 15000 observations. The analysis in the remainder of this report utilizes the training set. Table 2 demonstrates the number of members in the privileged and unprivileged groups.

Table 2 — Number of Members in Privileged and Unprivileged Groups.

	Privileged Group (Male)	**Unprivileged Group (Female)**
Training Set	5940	9060
Testing Set	5948	9052

1.3 Maximizing Profit (Step 4)

1.3.1 *Histogram for Creditworthiness (Step 4.1)*

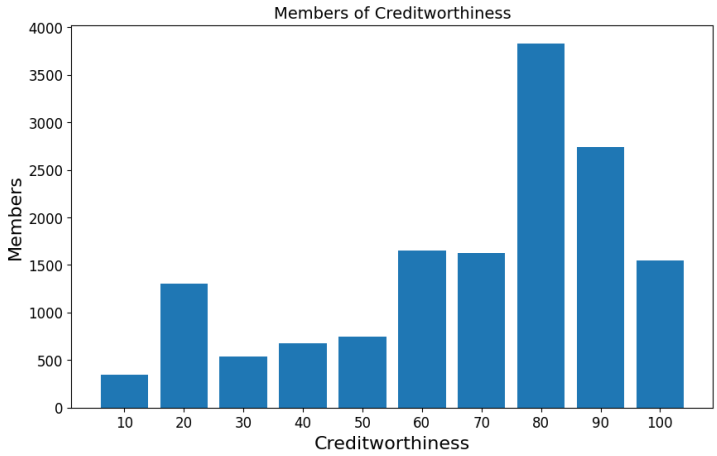

Figure 1 — Members of Creditworthiness

Fig. 1 illustrates the distribution of customers by creditworthiness. On the X-axis, each tick label represents a range of creditworthiness scores: '10' denotes scores from 0 to 10, '20' signifies scores from 10 to 20, and so on.

1.3.2 *Threshold for Loan Approval (Step 4.2 - 4.3)*

Outcome variable Y=0 indicates no default (loan approved), and Y=1 signifies default (loan denied). Set a threshold X for creditworthiness, such that loans are approved when creditworthiness ≥ X, and denies otherwise. X is determined by maximizing the profit, calculated based on the following rules:

1) A profit of 10 is earned when both Y and X approve a loan.

2) A profit of -5 is earned when Y approves but X denies a loan.

3) A profit of -3 is earned when Y denies but X approves a loan.

4) No profit is generated when both Y and X deny a loan.

Using a threshold of X=40 yields the highest profit of 114905.

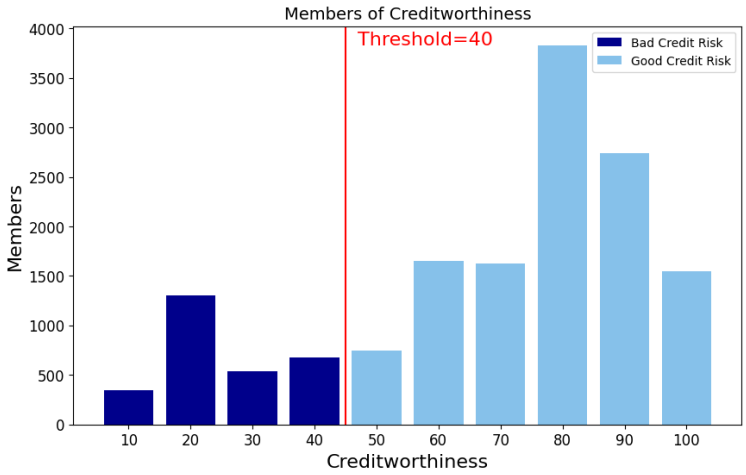

Figure 2 — Members of Creditworthiness

1.3.3 *Favorable vs Unfavorable Outcomes for Protected Class (Step 4.4)*

Table 3 — Favorable vs Unfavorable Outcomes for Privileged and Unprivileged Groups.

Outcome	Privileged Group (Male)	Unprivileged Group (Female)
Unfavorable (Declined)	1248	1618
Favorable (Approved)	4692	7442

Table 3 displays the favorable (loan approval) and unfavorable (loan denial) outcomes for the privileged and unprivileged groups while **applying the threshold of X=40**.

1.4 Fairness Metrics (Step 5)

1.4.1 *Select & Plot Fairness Metrics (Step 5.1 - 5.2)*

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

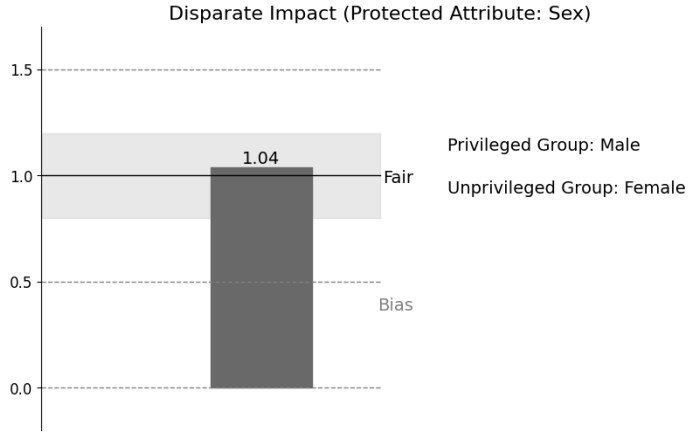

Figure 3 — Disparate Impact

Figure 4 — Statistical Parity Difference

DI measures the ratio of favorable outcome rates for unprivileged groups compared to that of privileged groups[3]. **To achieve fairness, its value should be 1.** DI > 1 indicates favoritism towards the unprivileged group, while DI < 1 signifies an advantage for the privileged group. In this report, **the acceptable range for DI value spans from 0.8 to 1.2, within which the bias is considered non-distinctive.**

SPD measures the difference in the rate of favorable outcomes between unprivileged and privileged groups[4]. **To achieve fairness, its value should be 0.** SPD > 0 indicates

3 https://dataplatform.cloud.ibm.com/docs/content/wsj/model/wos-disparate-impact.html

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

favoritism towards the unprivileged group, whereas SPD < 0 signifies an advantage for the privileged group. In this report, **the acceptable range for SPD value spans from -0.1 to 0.1, within which the bias is considered non-distinctive.**

Switch the role of privileged and unprivileged groups and calculate the difference for each metric (before-switch value minus after-switch value). According to the definitions of DI and SPD provided earlier, **the acceptable range for DI difference spans from -0.4 to 0.4, and the acceptable range for SPD difference ranges from -0.2 to 0.2. Within these intervals, bias is considered non-distinctive.** A positive difference indicates favoritism towards the unprivileged group, while a negative value signifies an advantage for the privileged group.

1.4.2 *Fairness Metrics & Bias Analysis (Step 5.3 - 5.4)*

Table 4 — Fairness Metrics (DI: Disparate Impact; SPD: Statistical Parity Difference).

Fairness Metric	Metric Value	Acceptable Range for Metric Value	Difference	Acceptable Range for Difference	Bias
DI	1.04	0.8 to 1.2	0.08	-0.4 to 0.4	No distinctive bias
SPD	0.03	-0.1 to 0.1	0.06	-0.2 to 0.2	No distinctive bias

Table 4 reveals that both DI difference and SPD difference indicate no distinctive bias because each metric falls within its acceptable range. Perfect fairness would be achieved with a DI difference of 0 or an SPD difference of 0, ensuring equal rates of positive outcomes for both privileged and unprivileged groups. According to the definition of metric difference in Section 1.4.1, a positive difference favors the unprivileged group, whereas a negative difference benefits the privileged group. In this analysis, the DI difference is 0.08 and the SPD difference is 0.06, indicating a slight advantage for the unprivileged group. **Therefore, the differences in the fairness metrics do not show distinctive bias but a slight bias against the privileged group.**

Meanwhile, as shown in Fig. 3 and 4, perfect fairness is represented by a DI of 1 or an SPD of 0, indicating equal positive outcome rates between privileged and unprivileged groups. In this analysis, DI is 1.04, and SPD is 0.03, suggesting that the positive outcome rates for the unprivileged group is slightly higher than the privileged group. Both DI and SPD values fall within their acceptable ranges (Table 4). **Therefore, both**

DI and SPD values do not show distinctive bias but a slight bias in favor of the unprivileged group.

1.5 Bias Mitigation (Step 6)

According to Section 1.4 (Step 5), there is no distinct bias. However, the Disparate Impact (DI) value is slightly above 1, and the DI difference is marginally greater than 0, signifying a slight advantage for the unprivileged group. In this section, DI is utilized as the fairness metric, and different thresholds are selected for privileged and unprivileged groups to mitigate bias by minimizing the differences in approval rates between the two groups.

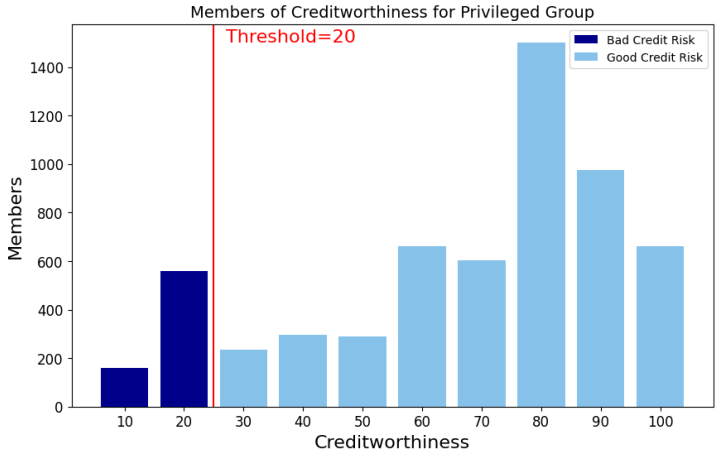

Figure 5— Members of Creditworthiness for Privileged Group

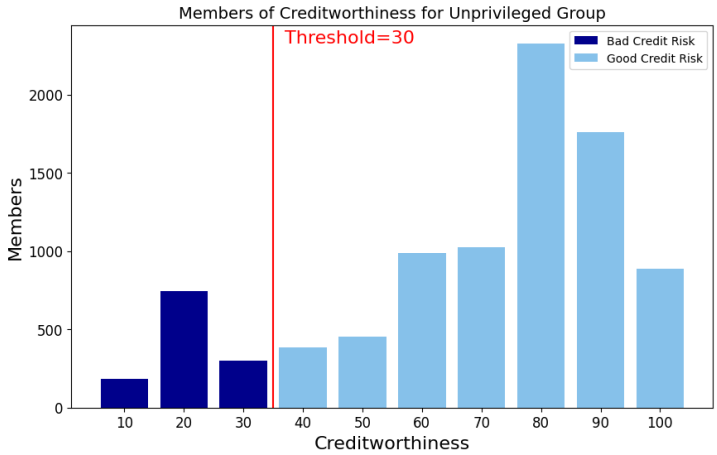

Figure 6— Members of Creditworthiness for Unprivileged Group

Threshold (X1) for privileged group: X1 = 20

Threshold (X2) for unprivileged group: X2 = 30

Disparate Impact (DI) value: 0.983

Disparate Impact (DI) difference: -0.034

Profit based on new thresholds: 112169

Table 5 — Favorable vs Unfavorable Outcomes for Privileged and Unprivileged Groups after Bias Mitigation.

Outcome	Privileged Group (Male)	Unprivileged Group (Female)
Unfavorable (Declined)	720	1234
Favorable (Approved)	5220	7826

1.6 Bias Mitigation Analysis (Step 7)

1) In Section 1.4 (Step 5), Disparate Impact (DI) and Statistical Parity Difference (SPD) are selected as the fairness metrics.

Initially, without altering the positions of the privileged and unprivileged groups, the DI and SPD values are 1.04 and 0.03, respectively. When the roles of these groups are reversed, the DI and SPD values adjust to 0.96 and -0.03, respectively, leading to a DI difference of 0.08 and an SPD difference of 0.06. While both metrics fall within the acceptable range, indicating no distinctive bias, these difference values suggest a marginally higher rate of positive outcomes for the unprivileged group compared to the privileged group, signifying a slight bias in favor of the unprivileged group.

The same conclusion can be drawn from the DI and SPD values. Since both DI and SPD values lie within their acceptable ranges, no significant bias is observed. However, a slight advantage for the unprivileged group is identified, indicating a slight bias against the privileged group.

2) The mitigation step in Section 1.5 (Step 6) **effectively reduces the slight bias that initially favored the unprivileged group, thereby addressing the disadvantage previously faced by the privileged group**. The DI difference changes from 0.08 (before mitigation, a 0.08 deviation from 0) to -0.034 (after mitigation, a 0.034 deviation from 0), effectively diminishing the bias against the privileged group.

Meanwhile, the DI value changes from 1.04 (before mitigation, a 0.04 deviation from 1) to 0.983 (after mitigation, a 0.017 deviation from 1), further demonstrating the mitigation's effectiveness.

Regarding the rate of favorable outcomes, both privileged and unprivileged groups receive an advantage because the reduction in threshold values leads to an increase in favorable outcomes for both groups.

Regarding the fairness metric, the privileged group gains an advantage as the metric moves from unfavorable to favorable. Conversely, this shift places **the unprivileged group at a disadvantage,** transitioning from a favorable to an unfavorable position. The rate of positive outcomes for the unprivileged group becomes lower than that of the privileged group, indicating a disadvantage for the unprivileged group.

3) Potential issues could emerge from the bias mitigation method used in this report.

a) The fairness metric in the bias mitigation method, Disparate Impact (DI), has its limitations. DI evaluates relative opportunities between two groups, but it does not address individual fairness. For instance, if a loan process approves all applicants from both privileged and unprivileged groups, leading to a 100% favorable outcome rate for both, DI would indicate perfect fairness, even though there may be a disproportionate rate of applications between the two groups. Relying solely on one metric to mitigate bias is insufficient. A more comprehensive approach, including various metrics and qualitative assessments, is essential for bias mitigation.

b) The mitigation method may introduce new biases. Setting different thresholds for groups raises fairness concerns, questioning why individuals from one group are subjected to higher thresholds than those from another. Such practices risk being perceived as discriminatory, potentially leading to overcorrection and the risk of reverse discrimination.

c) The mitigation method could have adverse consequences. Using different thresholds for different groups might compromise the fairness of the decision-making process and erode public trust. By lowering thresholds to meet a predefined notion of "fairness," there is a risk of approving loans for individuals who might struggle to fulfill their repayment obligations, thereby elevating financial risks. This could result in broader societal implications, including the potential for a financial crisis.