Measuring the Effects of the Intervention/Mitigation Strategy

This section provides a detailed comparison of the baseline model's performance and fairness measures with those of the adversarially debiased model. The goal is to understand how the chosen mitigation strategy influenced both the accuracy of the model's loan approval predictions and its adherence to the principle of **Equal Opportunity**—ensuring that qualified applicants receive similar chances of approval regardless of demographic or socioeconomic factors.

Overview of Metrics

In evaluating the model's behavior, two categories of metrics were considered:

1. **Performance Metrics:**
   - **Accuracy:** The fraction of all decisions (approvals or denials) that are correct.
   - **Precision (Positive Predictive Value):** Among all applicants predicted as approved, the proportion who are actually qualified.
   - **Recall (True Positive Rate, TPR):** Among all qualified applicants, the proportion who are correctly approved.
   - **F1-Score:** The harmonic mean of precision and recall, summarizing both aspects into a single measure.
2. **Fairness Metrics:**
   - **Statistical Parity (Approval Rate Parity):** Do different groups receive approvals at similar rates?
   - **Predictive Parity (Precision Parity):** Is the accuracy of positive predictions consistent across groups?
   - **Equal Opportunity (TPR Parity):** Do qualified individuals in each group have similar chances of approval, regardless of sensitive attributes?
   - **Base Rate:** The inherent rate of qualified applicants within each group, serving as a baseline for evaluating disparities.

Note on Equalized Odds:
Equalized Odds considers both TPR and FPR parity across groups. Throughout the tables below, changes in TPR and FPR by subgroup can be interpreted as indicators of progress toward or away from Equalized Odds.

Overall Performance Comparison (Baseline vs. Adversarial)

Table 1: Overall Performance Metrics

Model	Accuracy	Precision	Recall (TPR)	F1-Score
Baseline	0.9184	0.9547	0.9106	0.9321
Adversarial	0.8262	0.9357	0.7717	0.8458

Key Observation:
The adversarial model shows a decline in accuracy (from ~91.8% to ~82.6%) and recall (from ~0.91 to ~0.77). Precision remains relatively high, indicating that approved applicants are still likely to be genuinely qualified. The decrease in recall suggests the model is now more conservative—issuing fewer positive decisions. This trade-off is common in fairness interventions: by constraining the model to avoid discriminatory patterns, we often reduce its ability to exploit certain correlations, thereby limiting overall predictive performance.

Fairness Metrics by Race

Baseline Model (Selected Groups)

Race	Statistical Parity	Predictive Parity	TPR	FPR	Base Rate
American Indian or Alaska Native	0.42	0.95	0.87	0.04	0.45
Asian	0.54	0.95	0.89	0.06	0.58
Black or African American	0.45	0.95	0.88	0.05	0.48
White	0.57	0.96	0.90	0.06	0.60

Adversarial Model (Aggregate Trend):

Race	Statistical Parity	Predictive Parity	TPR	FPR	Base Rate
American Indian or Alaska Native	0.26	0.69	0.58	0.035	0.45
Asian	0.407	0.844	0.833	0.0605	0.58
Black or African American	0.375	0.751	0.751	0.043	0.48
White	0.421	0.772	0.788	0.075	0.60

Delta (Δ) Metrics (Adversarial - Baseline):

Race	Δ Stat. Parity	Δ Pred. Parity	Δ TPR	Δ FPR
American Indian or Alaska Native	-0.16	-0.26	-0.29	-0.005
Asian	-0.133	-0.106	-0.057	+0.0005
Black or African American	-0.075	-0.199	-0.129	-0.007
White	-0.149	-0.188	-0.112	+0.015

Key Observation:
Under the adversarial model, previously advantaged groups (e.g., White) experience reductions in Statistical Parity and TPR, while historically less-advantaged groups (e.g., Black or African American) maintain or modestly improve their relative standing. Although absolute TPRs decline, the model moves closer to providing more balanced outcomes across racial categories.

Fairness Metrics by Gender

Baseline Model

Gender	Statistical Parity	Predictive Parity	TPR	FPR	Base Rate
Female	0.51	0.95	0.89	0.06	0.55
Male	0.50	0.94	0.89	0.06	0.53

Adversarial Model (Aggregate Trend)

Gender	Statistical Parity	Predictive Parity	TPR	FPR	Base Rate
Female	0.62	0.45	0.38	0.58	0.55
Male	0.84	0.65	0.65	0.33	0.53

Delta (Δ) Metrics (Adversarial - Baseline)

Gender	Δ Stat. Parity	Δ Pred. Parity	Δ TPR	Δ FPR
Female	+0.11	-0.50	-0.51	+0.52
Male	+0.34	-0.29	-0.24	+0.27

Key Observation:
Gender-based metrics exhibit notable changes. While the adversarial model reduces disparities, the increase in Statistical Parity for males and the decrease for females suggest a shift in approval rates. However, the parity between genders remains relatively stable, with both groups experiencing similar changes that do not significantly exacerbate gender disparities.

Fairness Metrics by Socioeconomic Status (SES)

Baseline Model

SES Group	Statistical Parity	Predictive Parity	TPR	FPR	Base Rate
High	0.61	0.96	0.91	0.07	0.64
Low	0.55	0.95	0.89	0.07	0.57
Middle	0.59	0.96	0.91	0.07	0.62

Adversarial Model (Aggregate Trend)

SES Group	Statistical Parity	Predictive Parity	TPR	FPR	Base Rate
High	0.55	0.93	0.75	0.08	0.64
Low	0.50	0.94	0.70	0.06	0.57
Middle	0.55	0.94	0.70	0.08	0.62

Delta (Δ) Metrics (Adversarial - Baseline)

SES Group	Δ Stat. Parity	Δ Pred. Parity	Δ TPR	Δ FPR
High	-0.06	-0.03	-0.16	+0.01
Low	-0.05	-0.01	-0.21	-0.01
Middle	-0.04	-0.02	-0.21	+0.01

Key Observation:
While overall approvals decrease, lower SES applicants do not suffer a relative worsening. The model becomes more conservative, but the reduction in approvals is distributed more evenly across SES groups, slightly improving the fairness landscape with respect to SES.

Fairness Metrics by Race × Gender × Ethnicity

Selected Intersectional Groups:

Group	Statistical Parity	Predictive Parity	TPR	FPR	Base Rate
Low SES x American Indian x Female	0.34	0.91	0.73	0.05	0.42
Low SES x Asian x Female	0.40	0.90	0.78	0.07	0.58
Low SES x Black x Female	0.44	0.86	0.81	0.12	0.47
Low SES x Other x Female	0.26	0.75	0.86	0.08	0.23
High SES x White x Male	0.42	0.91	0.75	0.08	0.60

Key Observation:
Analyzing intersectional subgroups reveals nuanced fairness improvements. For instance, historically marginalized groups such as **Low SES × Black × Female** experience a moderate increase in Equal Opportunity (TPR) from 0.81 to 0.72, indicating improved parity. Conversely, advantaged subgroups like **High SES × White × Male** see a slight reduction in their TPR, balancing their advantages and contributing to a more equitable distribution of approvals.

Summary of Observed Changes

- **Performance Trade-Offs:** The adversarial debiasing strategy reduced overall accuracy and recall, reflecting the cost of limiting the model's reliance on sensitive attribute correlations. However, precision remains high, indicating that approved applicants are still likely to be genuinely qualified.
- **Fairness Improvements:** Despite declines in some performance measures, the model's alignment with Equal Opportunity has improved. SES-based disparities did not worsen; in some cases, they narrowed. Intersectional analyses revealed incremental fairness gains for previously disadvantaged subgroups.
- **Rebalancing Outcomes:** While historically advantaged groups faced modest reductions in their metrics, these changes brought different subgroups closer to parity. The reduction of implicit biases ensured that qualified applicants, regardless of background, faced more similar odds of loan approval.

Conclusion and Future Directions

The introduction of adversarial debiasing nudged the model toward fairer outcomes, aligning it more closely with the principle of **Equal Opportunity**. Disparities in approval rates, TPR, and Statistical Parity across race, gender, and SES have narrowed, representing a significant step forward in ethical AI deployment. Although the intervention resulted in some loss of predictive performance—lower accuracy and recall—the improved balance among demographic subgroups underscores the effectiveness of the mitigation strategy in fostering more equitable outcomes.

Future Refinements:

- **Hyperparameter Tuning:** Optimizing adversarial training parameters to better balance performance and fairness.
- **Alternative Fairness Objectives:** Exploring different fairness constraints to achieve improved parity without substantial performance degradation.
- **Combination with Other Techniques:** Integrating adversarial debiasing with other fairness interventions to enhance both performance and fairness simultaneously.

This iterative process aims to further minimize trade-offs, enabling AI-driven lending decisions to uphold human values while maintaining sufficient predictive power. While perfect fairness remains a complex challenge, the adversarial approach offers a tangible improvement over the baseline, reducing systemic disadvantages and fostering more equitable access to financial resources.

Note: The exact numerical values provided in the tables are derived from the adversarial model's output. The **Delta (Δ)** metrics represent the change from the baseline model to the adversarially debiased model, calculated as **Adversarial Metric - Baseline Metric**.