

# Bias Checkpoint Auditor

## Comprehensive Audit Report

**Run ID:** 877e4f55-7532-4adb-9e36-6d819297ee24  
**Generated:** 2025-11-29 22:33:06

# Executive Summary

Primary Bias Origin: MODEL

Bias primarily manifests in the model's predictions. Fairness violations: demographic\_parity, equal\_opportunity, equalized\_odds. Counterfactual sensitivity detected. The data and features may be relatively unbiased, but the model has learned to make discriminatory predictions.

## Checkpoint Summary

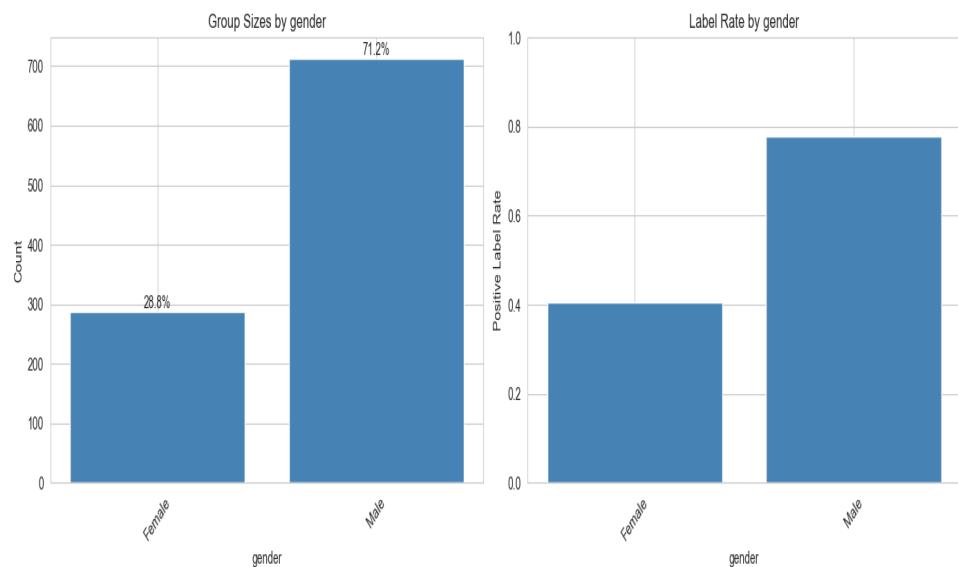
Stage	Status	Score	Issues
Data	Pass	0.00	None
Features	Pass	0.16	PROXY_FEATURES
Model	Fail	0.92	DEMOGRAPHIC_PARITY_VIOLATION, COUNTERFACTUAL_SENSITIVITY

# Data Checkpoint Analysis

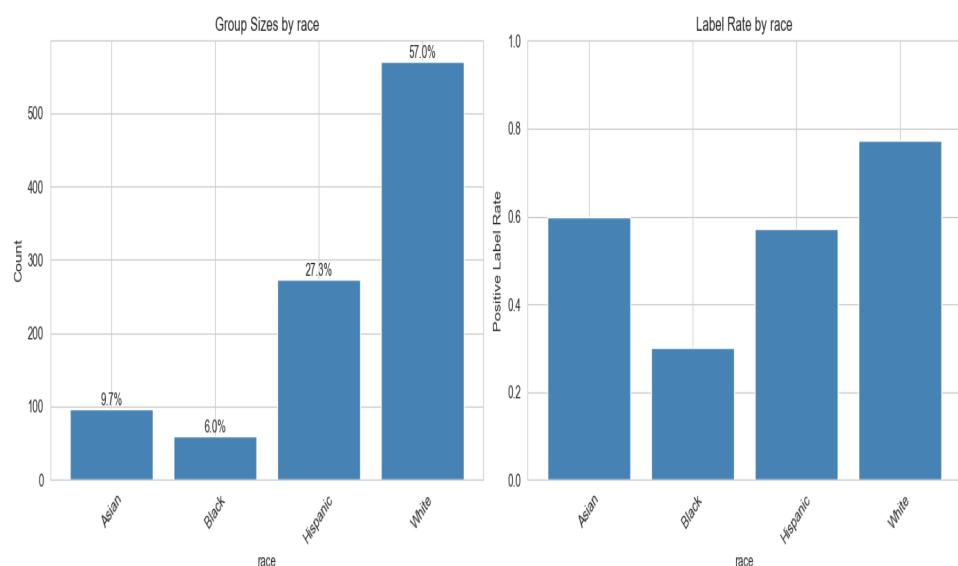
**Bias Score:** 0.00

**Summary:** No significant data bias detected.

## **Distribution: gender**



## **Distribution: race**



# Feature Checkpoint Analysis

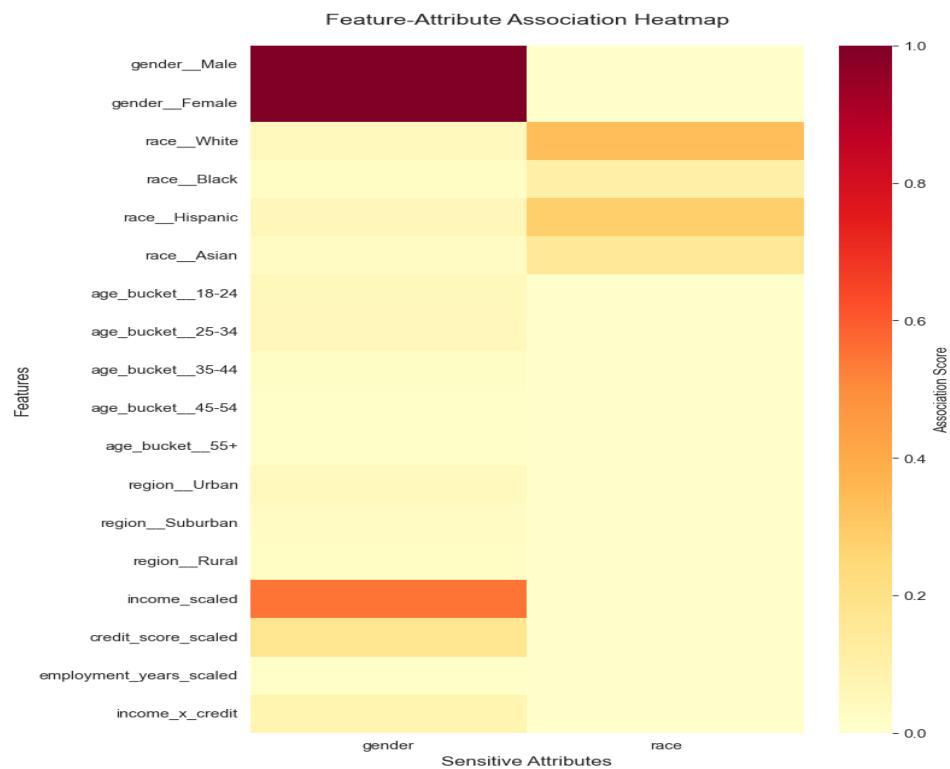
Bias Score: 0.16

Summary: 4 proxy features detected.

## Proxy Features Detected:

- gender: gender\_\_Male, gender\_\_Female, income\_scaled
- race: race\_\_White

## Feature-Attribute Association Heatmap



# Model Checkpoint Analysis

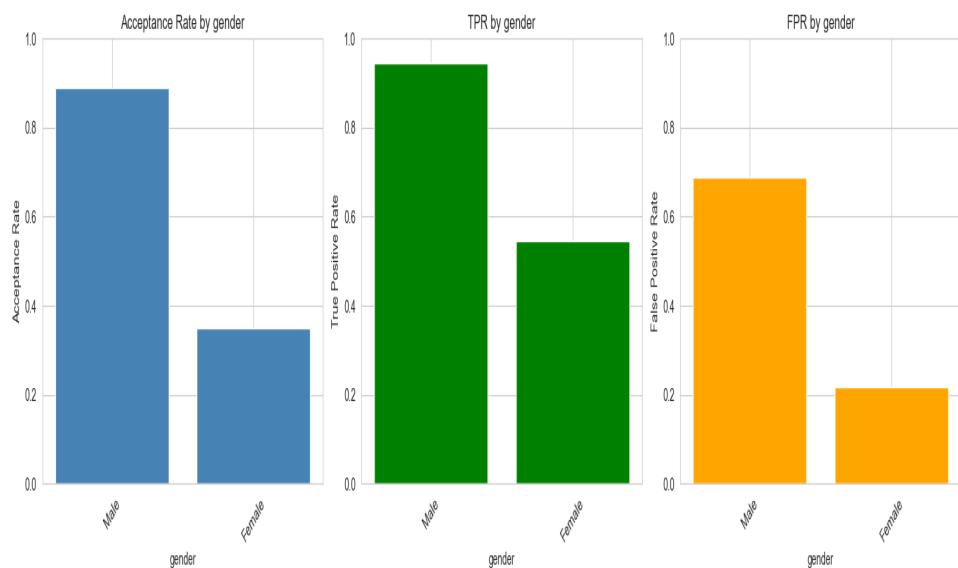
**Bias Score:** 0.92

**Summary:** Fairness violations: demographic\_parity, equal\_opportunity, equalized\_odds. Counterfactual sensitivity detected.

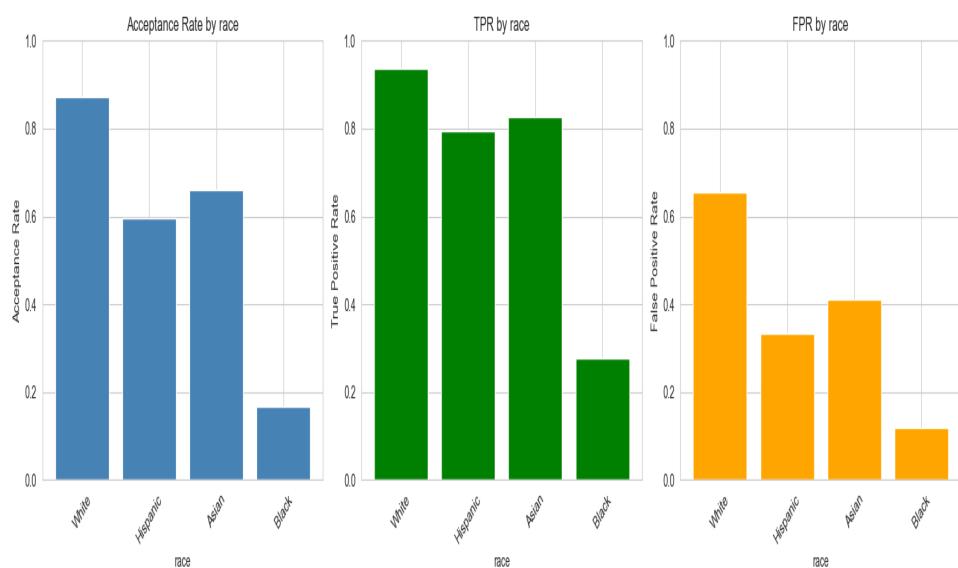
**Accuracy:** 0.772

**AUC:** 0.827

## Fairness Metrics: gender



## Fairness Metrics: race





## Recommended Fixes

### ***Features:***

1. Remove or regularize proxy features for 'gender': gender\_\_Male, gender\_\_Female, income\_scaled
2. Remove or regularize proxy features for 'race': race\_\_White

### ***Model:***

1. Consider threshold calibration per subgroup to reduce acceptance rate gaps.
2. Apply fairness-aware training methods or constraints to equalize TPR across groups.
3. Use post-processing techniques to equalize both TPR and FPR across groups.
4. Model predictions are sensitive to sensitive attributes. Consider removing these features or using fairness constraints.

## Appendix: Configuration

**Target Column:** label

**Sensitive Attributes:** gender, race

**Fairness Thresholds:**

- demographic\_parity\_diff\_threshold: 0.1
- equal\_opportunity\_diff\_threshold: 0.1
- equalized\_odds\_diff\_threshold: 0.1
- min\_group\_proportion\_threshold: 0.05
- min\_support\_for\_metrics: 30
- proxy\_corr\_threshold: 0.3
- counterfactual\_change\_threshold: 0.1
- label\_imbalance\_ratio\_threshold: 4.0