Bias Checkpoint Auditor

Comprehensive Audit Report

Run ID: c0fbb784
Generated: 2025-11-29 22:16:10

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

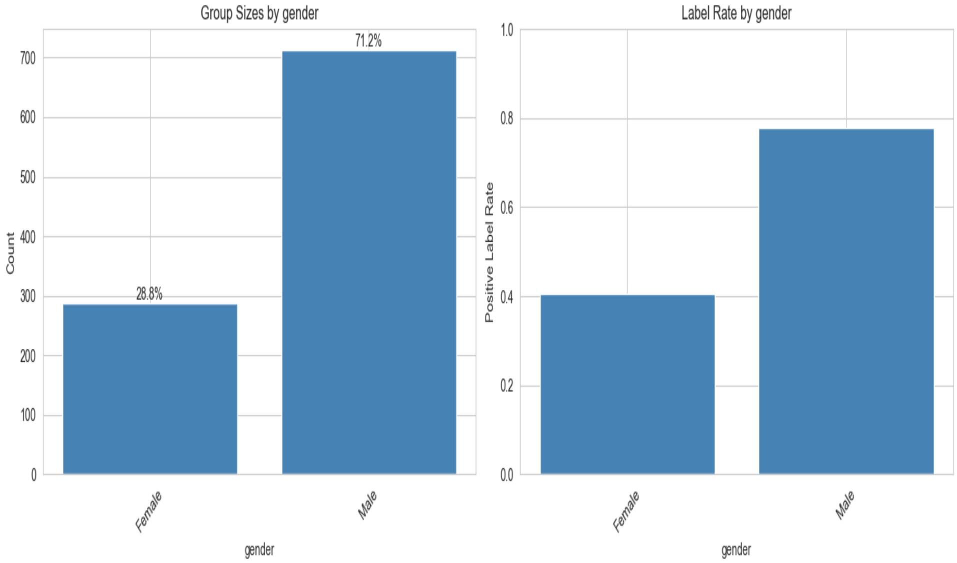
Stage	Status	Score	Issues
Data	Pass	0.00	None
Features	Pass	0.16	PROXY_FEATURES
Model	Fail	0.93	EQUALIZED_ODDS_VIOLATION, DEMOGRAPHIC_PARITY_V

Data Checkpoint Analysis

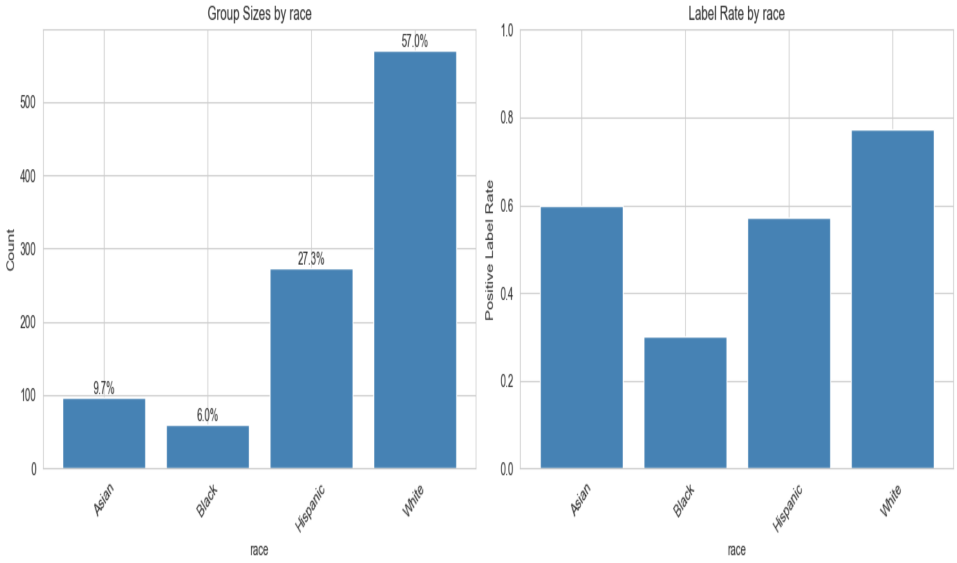
Bias Score: 0.00

Summary: No significant data bias detected.

Distribution: gender



Distribution: race



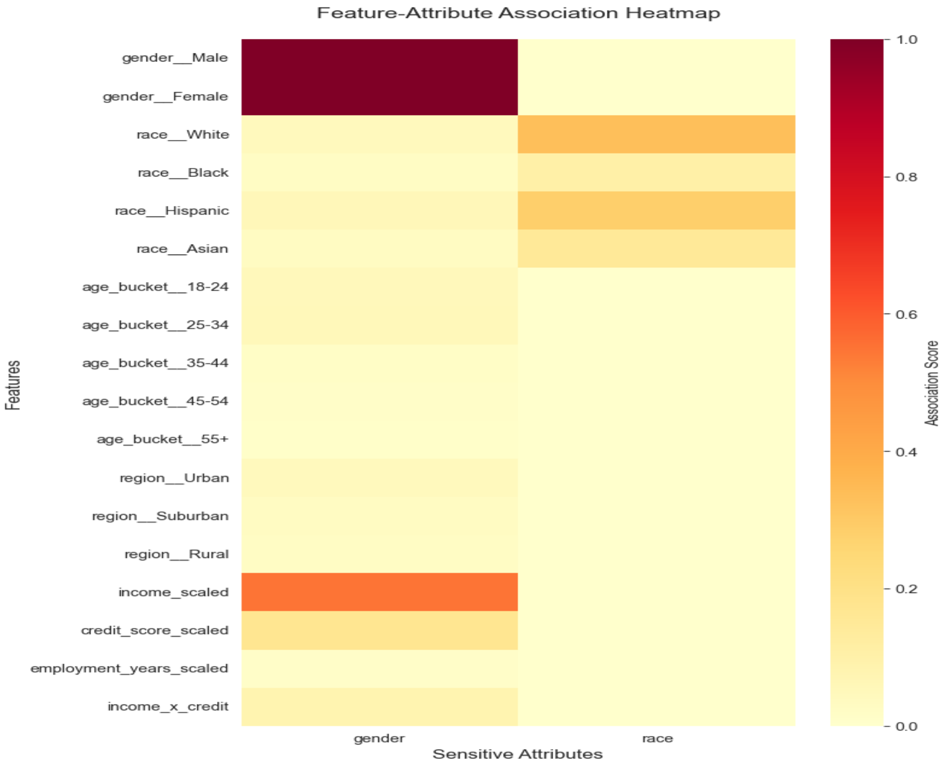
Feature Checkpoint Analysis

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



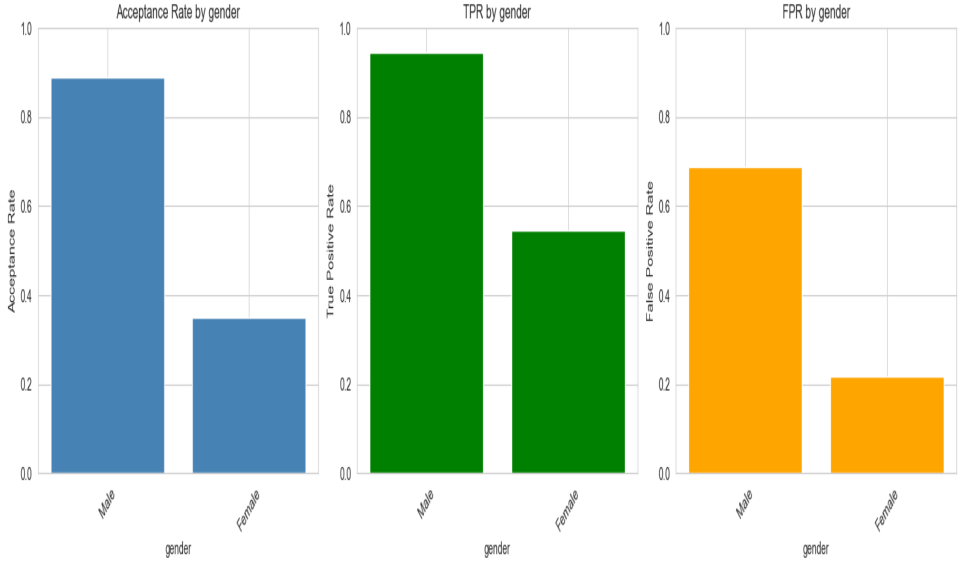
Model Checkpoint Analysis

Bias Score: 0.93

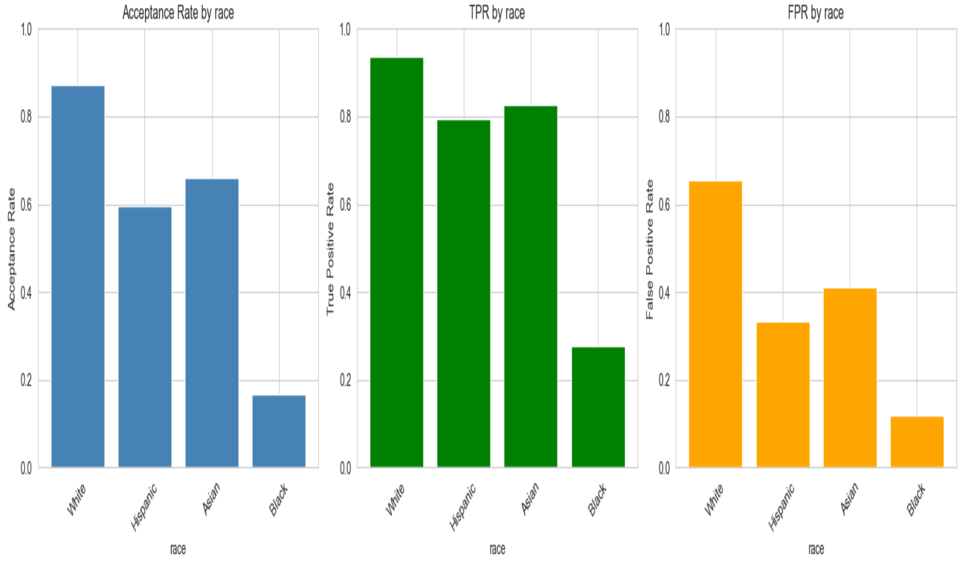
Summary: Fairness violations: demographic_parity, equal_opportunity, equalized_odds. Counterfactual sensitivity detected.

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