Universidad del Valle de Guatemala
Responsible AI
Sección 10

Identificación y Mitigación de Sesgos en Modelos de Machine Learning

Andres Quezada 21085
Javier Ramirez 21600
Javier Chavez 21016
Mario Cristales 21631

Guatemala, 9 de julio de 2025

Exploratory Data Analysis (EDA)

The dataset contains 18,316 records with demographic, criminal history, and risk score information. Key columns include race, sex, age, priors_count, decile_score, and the targets is_recid and is_violent_recid. Initial cleaning ensured only binary targets were used, and categorical fields were normalized.

Key observations include:

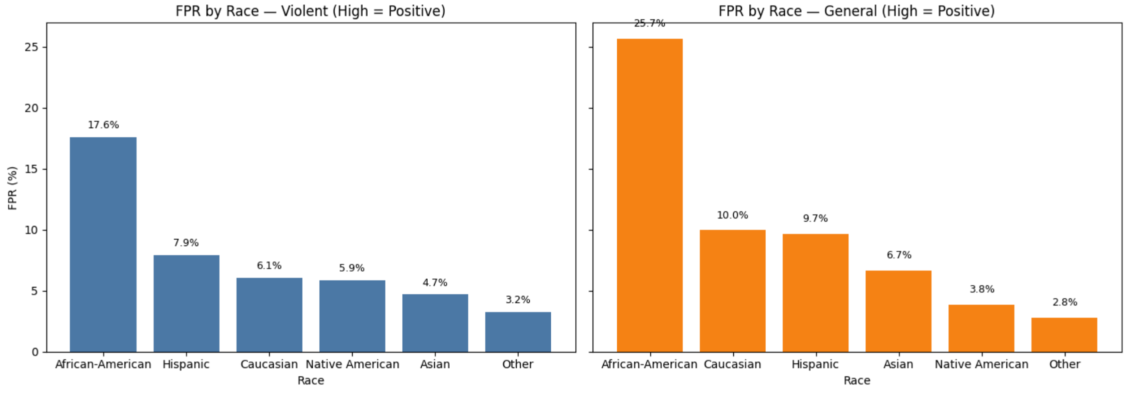
- Representation imbalance:
    - African-American and Caucasian individuals dominate the dataset.
    - Asian and Native American groups are very small, making their metrics less stable.
- Gender imbalance:
    - ~80% of individuals are male.
- Base rate differences:
    - Overall recidivism is roughly balanced (~50% yes, ~50% no).
    - Recidivism rates are higher among males than females and vary significantly by race.

These imbalances indicate that certain groups may carry more influence in model training and that fairness analysis should account for group size differences.

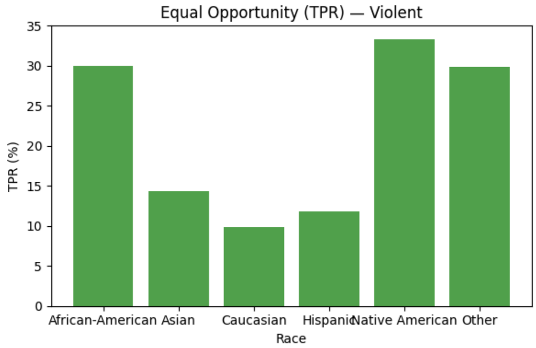
Bias Identification

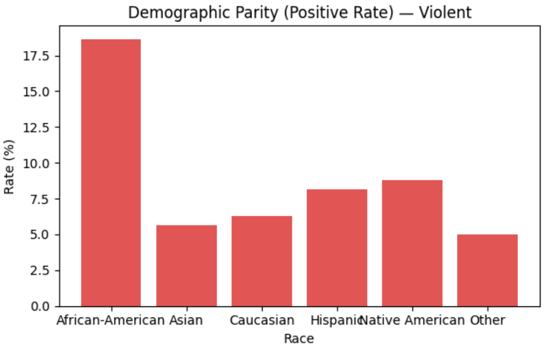
An initial fairness audit of the raw scores and targets highlights multiple forms of disparity across racial groups.

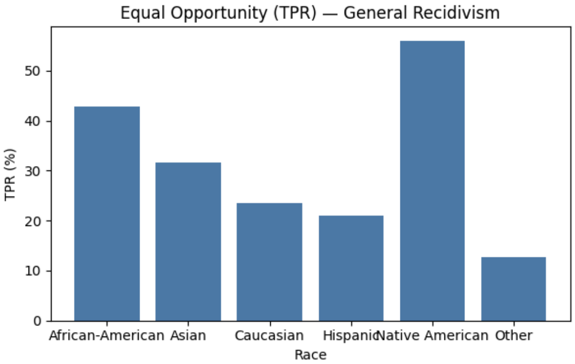
- False Positive Rate (FPR):
    - *General recidivism:* African-American individuals show the highest FPR, more than double that of some other groups.
    - *Violent recidivism:* The same trend holds, with African-American individuals predicted as high risk incorrectly more often.

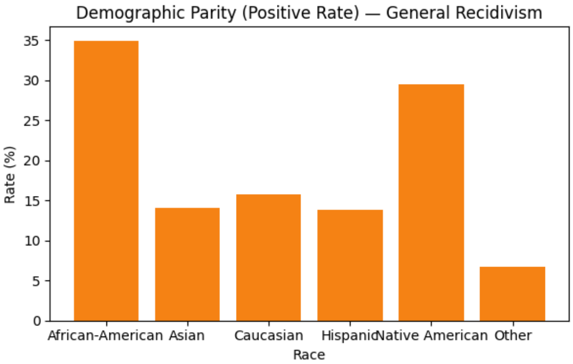


- Equal Opportunity (True Positive Rate, TPR):
  - *General recidivism:* Native American and African-American groups have the highest TPRs, while Caucasian and Hispanic groups are lower.
  - *Violent recidivism:* African-American and Native American groups again lead in TPR, with Caucasian groups lagging.
- Demographic Parity (Positive Prediction Rate, PPR):
  - African-American individuals are labeled high-risk far more frequently than other groups, in both general and violent predictions.









Initial Model Training

A logistic regression model was trained with age, priors_count, decile_score, race, and sex as features. Class balancing was applied to reduce bias from the slight target imbalance.

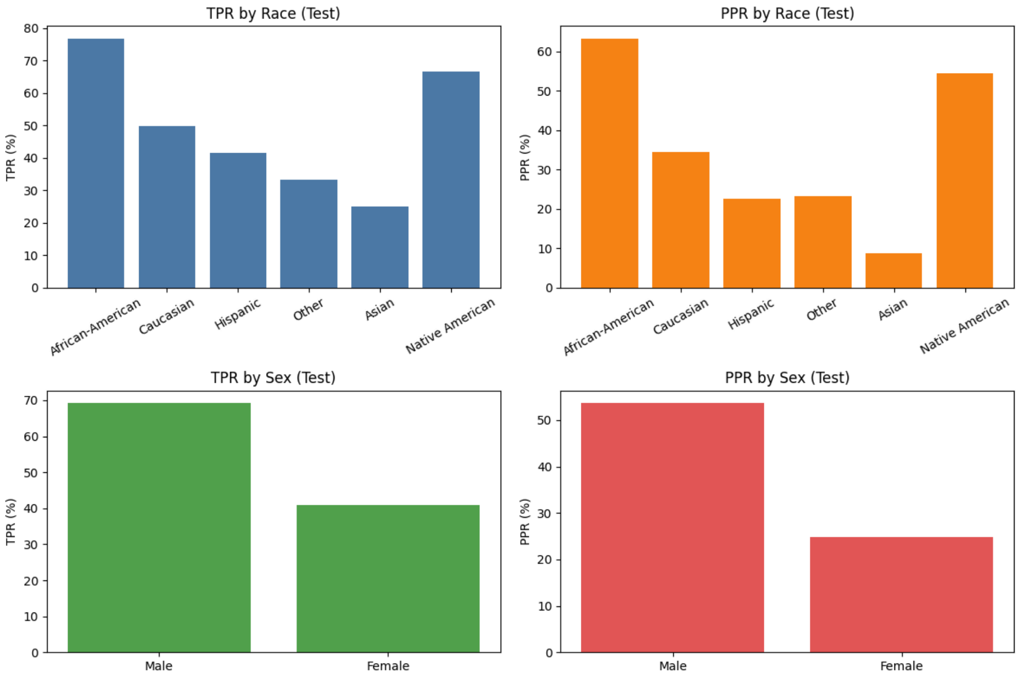
- Overall performance:
  - Accuracy: ~0.66
  - Precision: ~0.65
  - Recall: ~0.65
  - ROC-AUC: ~0.71

This baseline model performs moderately well overall but does not address fairness concerns.

Group-Level Evaluation

The model was evaluated by race and sex using fairness metrics such as Positive Prediction Rate (PPR) and True Positive Rate (TPR).

- By Race:
  - African-American and Native American groups show higher PPR and TPR, meaning they are flagged as high risk more often.
  - Caucasian, Asian, and Hispanic groups see lower TPR, indicating under-identification of risk.
- By Sex:
  - Males: Higher PPR (~54%) and TPR (~69%).
  - Females: Lower PPR (~25%) and TPR (~41%).



These disparities confirm that the model mirrors and amplifies the imbalances in the data.

Proposed Mitigation Strategies

To address the disparities:

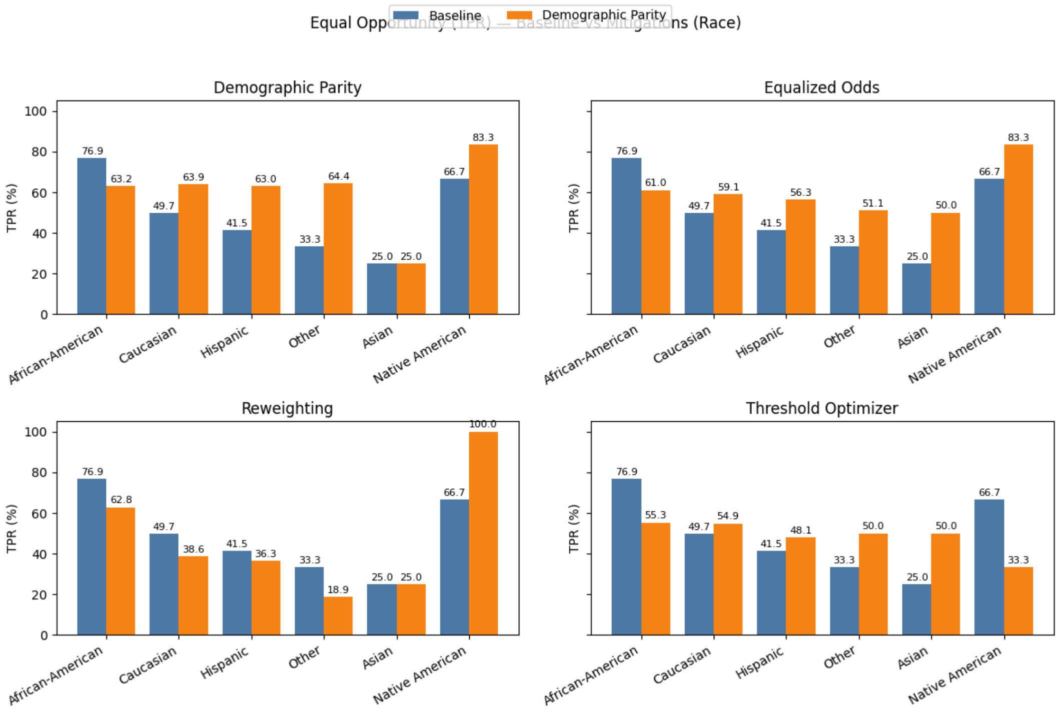
- Data-level actions:
  - Collect more samples for underrepresented groups (Asian, Native American, female).
  - Oversample minority groups during training.

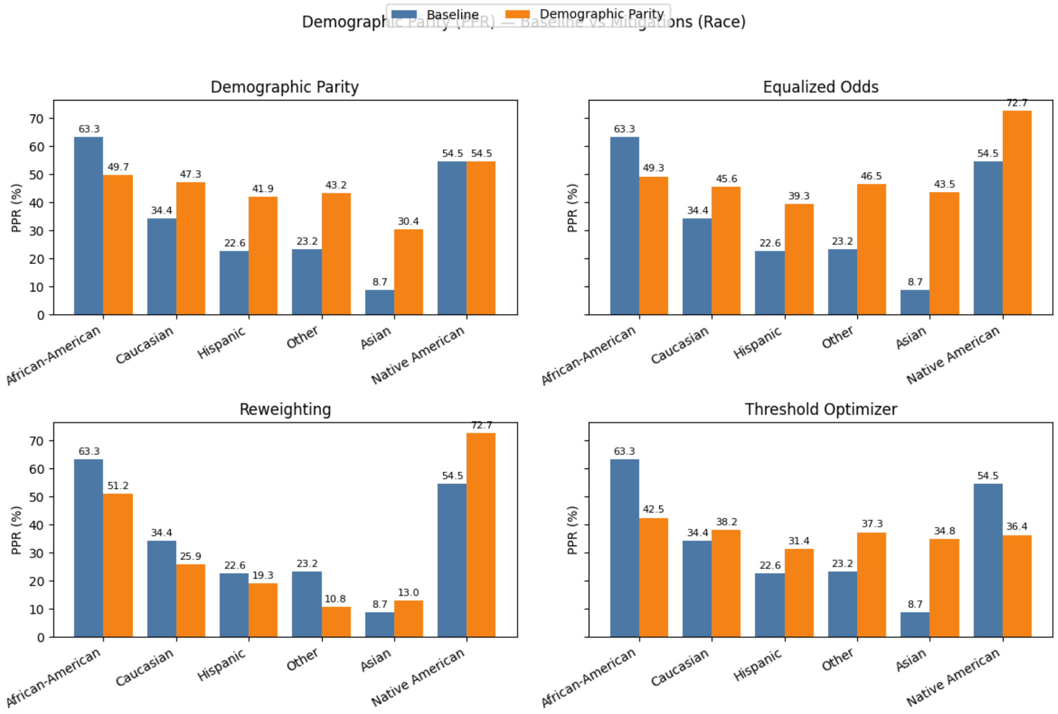
- Algorithm-level actions:
  - Apply fairness-constrained models (e.g., with Fairlearn or AIF360).
  - Use reweighting techniques to adjust the influence of sensitive groups.

- Post-processing actions:
  - Adjust decision thresholds per group to balance error rates.

Comparison: Baseline vs. Mitigation

- Equal Opportunity (TPR):
  Mitigation methods like Demographic Parity and Equalized Odds reduced gaps between groups, especially narrowing the advantage African-American and Native American groups had in TPR. However, small group sizes still caused volatility for some races (e.g., Native American showing extreme increases).
- Demographic Parity (PPR):
  The overall positive prediction rates became more balanced across groups after mitigation. However, the improvement came with trade-offs: some groups experienced reduced accuracy, which highlights the tension between fairness and performance.
- Reweighting and Threshold Optimizer:
  These methods balanced group metrics better but introduced instability in smaller groups, confirming that techniques must be applied carefully, especially where sample sizes are limited.





Conclusions and Reflections

- Bias persists even after mitigation.
  Improvements were achieved, but disparities remain, especially in small demographic groups where noise skews results.

- Fairness comes with trade-offs.
  While metrics like TPR and PPR became more equitable, accuracy for some groups decreased, showing the importance of balancing fairness objectives with predictive reliability.

- Responsibility in AI use:
  - Transparency about biases and limitations should always accompany model outputs.
  - Models should not be used blindly in high-stakes scenarios like parole decisions without human oversight.
  - Periodic re-evaluation and inclusion of diverse data sources are essential to reduce structural bias over time.

- Key recommendations:
  - Collect more data for underrepresented groups.
  - Use fairness metrics in every retraining cycle.
  - Involve stakeholders in reviewing fairness-performance trade-offs to ensure ethical deployment.