Ethics of Artificial Intelligence in Crime

UC San Diego
HALICIOĞLU DATA SCIENCE INSTITUTE

Kavya Sriram kasriram@ucsd.edu
Catherine Back cback@ucsd.edu
Aj Falak pfalak@ucsd.edu
Yuancheng Cao yuc094@ucsd.edu
Mentor: Emily Ramond eramond@deloitte.com
Mentor: Greg Thein gthein@deloitte.com

Background

Predictive models in pretrial risk assessment influence judicial decisions but often inherit racial biases from historical criminal justice data. This work examines racial bias in these models and applies bias mitigation techniques to improve fairness.

Pretrial Risk-Assessment Algorithms

- Predicts a defendant's risk of failing to appear or reoffending, influencing bail and detention decisions
- Aims to reduce subjectivity but often reinforces systemic biases

Bias in Risk Assessments

- COMPAS analysis (ProPublica, 2016) found Black defendants were twice as likely as White defendants to be falsely labeled high-risk
- Bias originates from policing practices and socioeconomic disparities embedded in arrest data

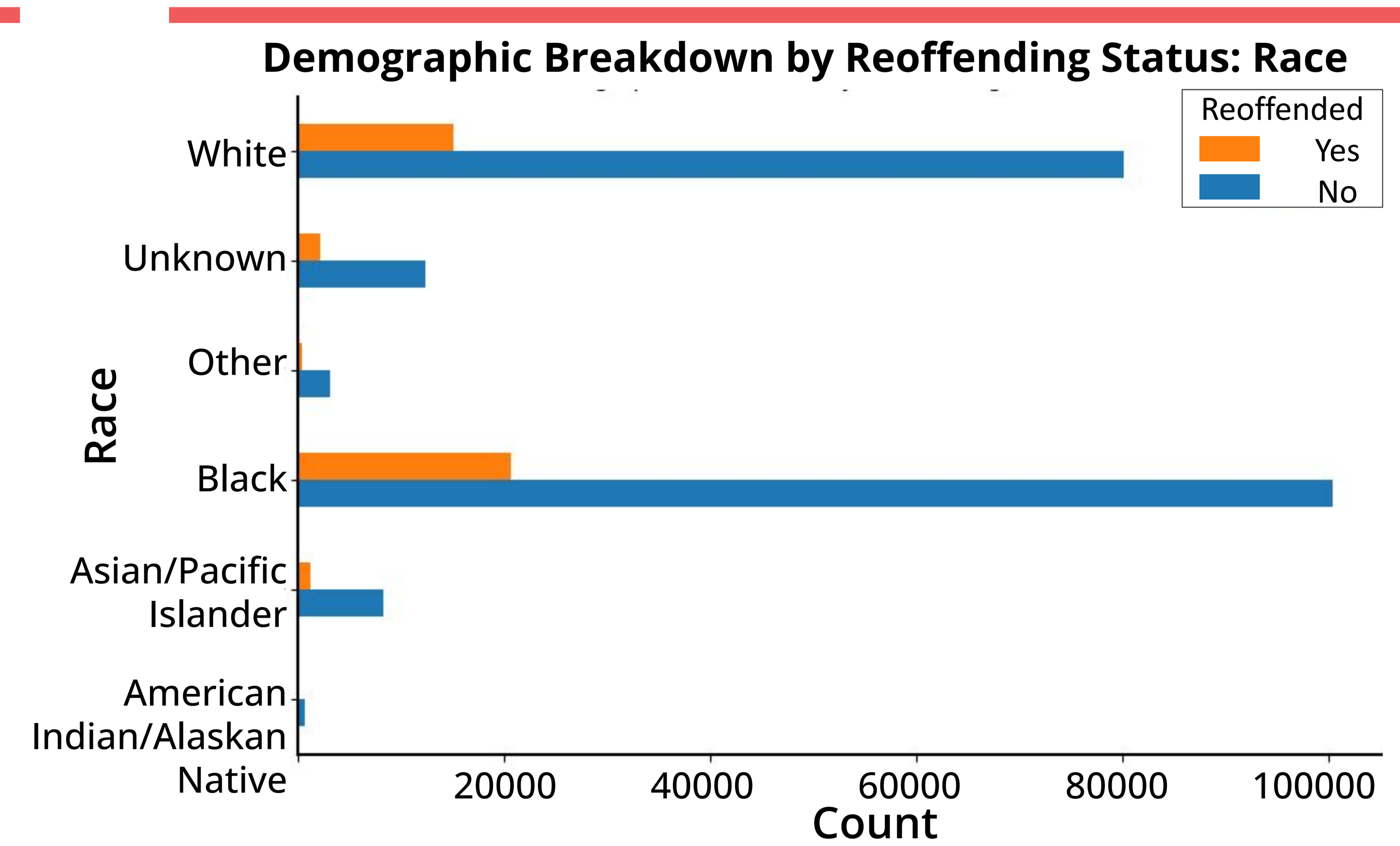
Data

- Created with the Division of Criminal Justice Services (DCJS)
- Contains 244,271 records from 2023, including protected attributes like race

Key Features

- Demographics: Race, ethnicity, gender, age at arrest/crime
- Pretrial Decisions: Bail set/posted, release type, supervision type
- Outcomes: Failure to Appear (FTA), reoffended, release decision, rearrest
- Financial Factors: Bail amount, bond type

Exploratory Data Analysis



- The majority of individuals (83.7%) had no arrests during their pretrial period
- Black individuals accounted for 49.5% of cases, followed by White individuals at 38.9%, Hispanic population represented 24.8% of cases, while Non-Hispanic individuals accounted for 65.2%
- Black defendants showed a higher rate of pretrial rearrest (19.3%) compared to White defendants (14.7%) and Asian/Pacific Islander defendants (11.2%)

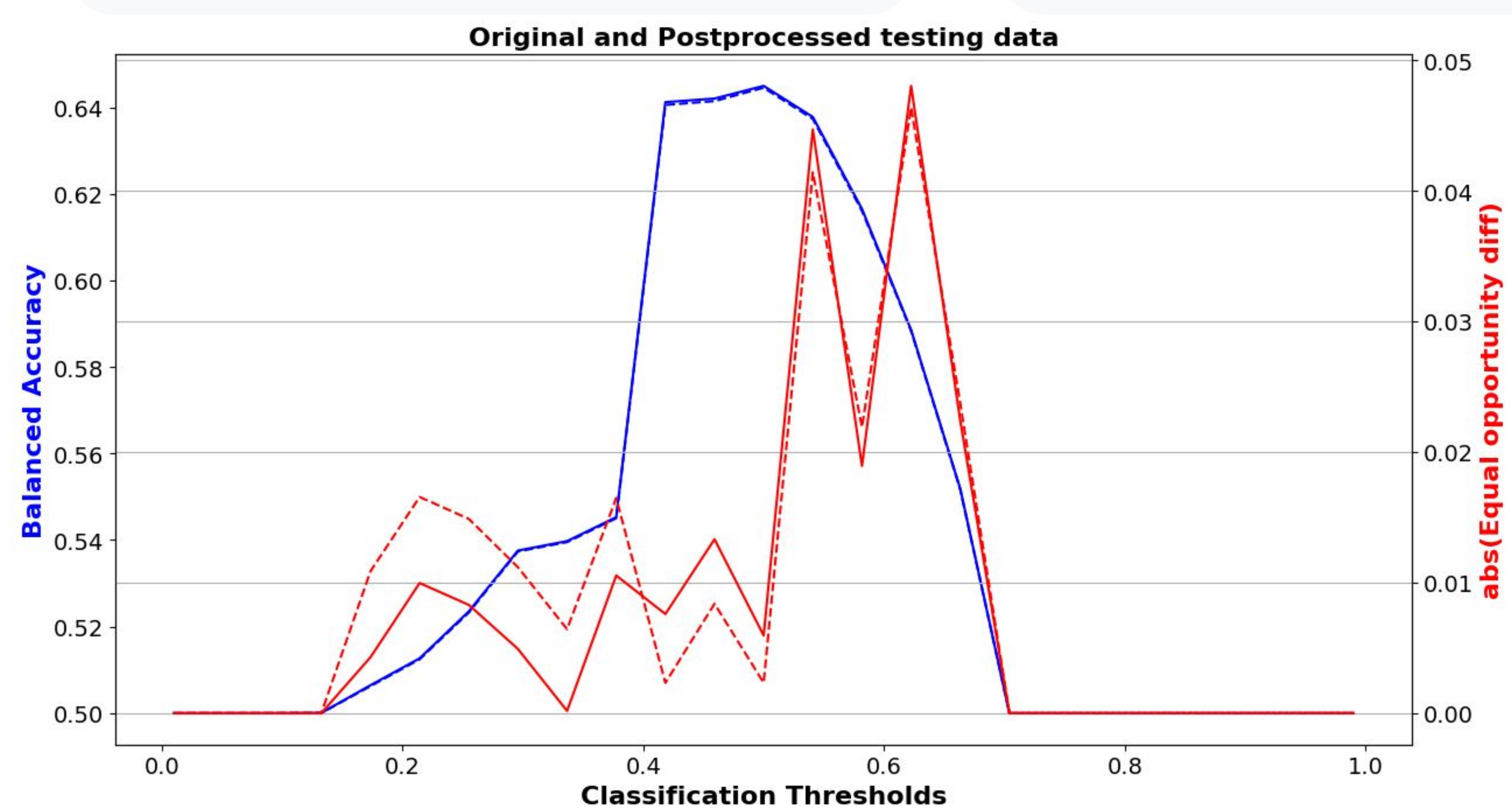
Model Development

- Built a Random Forest model using prior offences, pending charges, and crime severity, with binary indicators, 100 estimators, and a fixed random state for reproducibility
- Integrated SMOTE (Synthetic Minority Over-sampling Technique) with a Random Forest Classifier and setting class_weight='balanced' to adjust for class distribution
- Applied Stratified K-Fold Cross-Validation and performed Grid Search to optimize parameters

Bias Mitigation

Reweighing (Pre-processing)
Assigns different weights to groups to correct imbalances to prevent bias toward the majority groups

Calibrated Equalized Odds (Post-processing)
Modifies final model predictions to balance false positive and false negative rates across groups, ensuring fair classification



The postprocessing approach is a better option for models deployed in production, where maintaining the integrity of the original data is important. If fairness constraints must be met without modifying the training data, postprocessing is preferable.

scikit learn
AI Fairness 360
jupyter

Results

- The original Random Forest Classifier achieved 83.3% accuracy but had a balanced accuracy of 50.7% indicating bias toward non-offenders
- After applying oversampling and fine-tuning, balanced accuracy improved to 64.66%

	Before Fine-Tuning	After Fine-Tuning
Recall (Class 1)	0.50	0.66
Recall (Class 0)	0.73	0.63
Precision (Class 1)	0.26	0.26
Precision (Class 0)	0.88	0.90

- Statistical analysis revealed racial bias in the model, with Black individuals having the highest predicted reoffender and false positive rates
- After Reweighing → Improved fairness, but a slight drop in accuracy due to balancing efforts
- Postprocessing did not significantly impact balanced accuracy, meaning the model maintained its overall predictive performance

Limitations & Next Steps

- Limited access to existing criminal justice models
- Enhance model accuracy while mitigating bias and expand to diverse datasets for broader applicability and fairness

Reference

[1] Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks." ProPublica.

[2] Bellamy, Rachel K. E., et al. *AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias.*

Learn More Here 👉