Fairness-Aware Candidate Pre-Screening in Hiring

Using Equal Opportunity Post-Processing to Reduce Bias in AI Hiring Tools

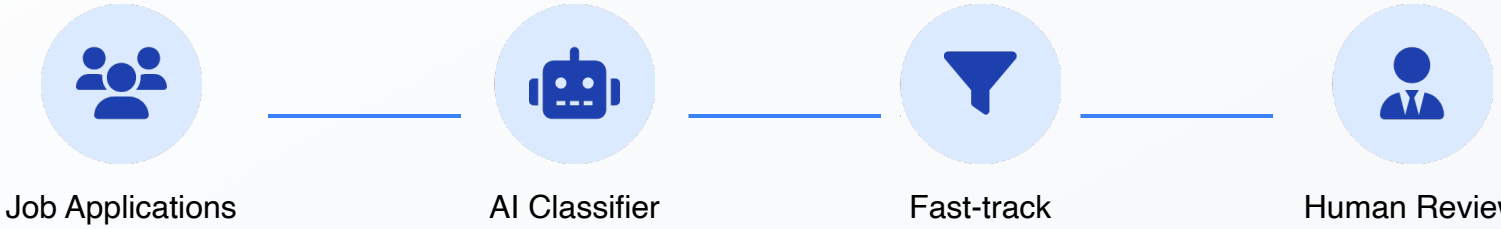
University Project using Adult (Census Income) Dataset

Focus on Sex as Protected Attribute

Presented by: Charles Santhakumar & Roberto Villafuerte
University of Helsinki

Trustworthy Machine Learning

Problem Setup: AI in Hiring Decisions



Classifier Function

- Predicts candidate income **>50K** (proxy for experience)
- **>50K** → fast-tracked to interviews
- **≤50K** → deprioritized, may not reach recruiters

Fairness Concern

- Model may underestimate women more often than men
- Reduces interview opportunities for women
- Reinforces inequality in who reaches better-paid positions

Focus: Analyzing fairness in candidate pre-screening using the Adult (Census Income) dataset with **sex** as the protected attribute.

Dataset and Methodology Overview

Adult Census Income Dataset

Size: ~50,000 rows

Purpose: Predicting income level based on census data

Target: Binary classification (>50K vs ≤50K)

Dataset Structure

Feature Columns	Categorical & Continuous
Target Column	Binary (>50K/≤50K)
Protected Attribute	sex (Male/Female)

Analysis Methodology

Protected Attribute

Focus on **sex** as the protected attribute (Male/Female)

Race is not analyzed in this presentation

Fairness Analysis

- Demographic Parity (DP)
- Statistical Parity Difference (SPD)
- Disparate Impact (DI)
- True Positive Rate by sex

Note: This project analyzes fairness in candidate pre-screening using the Adult (Census Income) dataset with **sex** as the protected attribute

Step 1: Baseline Models Methodology

Three baseline models were tested without fairness mitigation to establish performance and fairness metrics.

Logistic Regression

- Linear model for binary classification
- Regularized to prevent overfitting

Random Forest

- Ensemble of decision trees
- Handles mixed data types well

XGBoost

- Gradient boosting framework
- Optimized for performance

Evaluation Metrics

Performance Metrics

- Accuracy
- Precision
- Recall
- F1-score

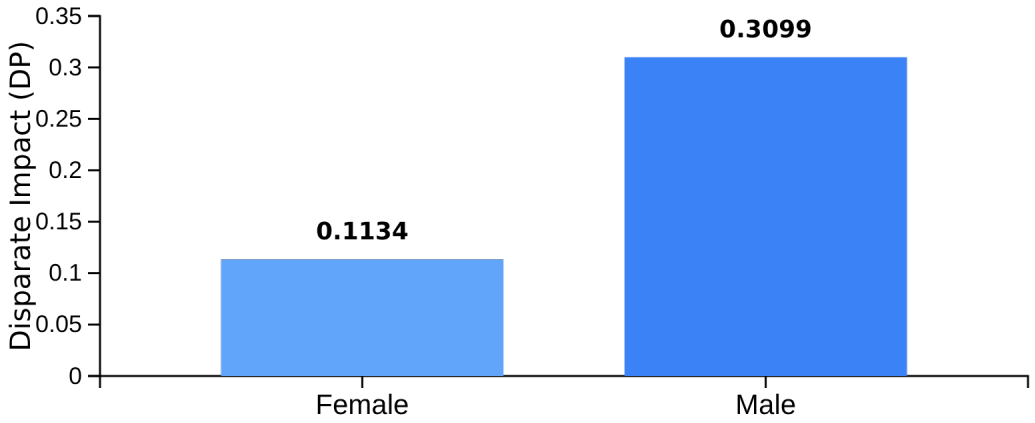
Fairness Metrics (by sex)

- Demographic Parity (DP)
- SPD (Male – Female)
- DI (Female / Male)
- TPR gap (M–F)

Methodology Note: All models were trained on the Adult (Census Income) dataset with sex as the protected attribute, without fairness mitigation techniques.

Raw Data Fairness Analysis by Sex

Disparate Impact (DP) by Sex



Female DP: 0.113

Male DP: 0.310

Fairness Metrics

Metric	Value	Interpretation
DP Female	0.1134	Proportion of females with positive outcome
DP Male	0.3099	Proportion of males with positive outcome
SPD	0.1965	Difference in positive prediction rates
DI	0.3659	Ratio of positive prediction rates

Why This Matters

The raw data shows significant gender disparity before any model is applied. This indicates a need for fairness-aware algorithms to prevent reinforcing existing inequalities in the hiring process.

Disparate Impact (DI)

Ratio of positive outcomes between groups. Values close to 1 indicate parity.

SPD

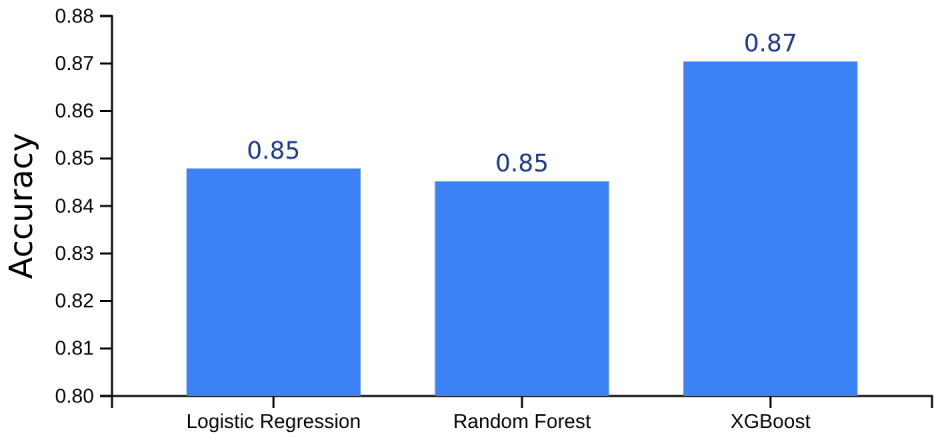
Signed difference between positive prediction rates. Measures group disparity.

DP

Proportion of positive outcomes in each group. Direct measure of group representation.

Baseline Model Performance Results

Model Accuracy Comparison



All models achieve high accuracy (84-87%)

Detailed Performance Metrics

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.8479	0.7327	0.6000	0.6597
Random Forest	0.8452	0.7120	0.6214	0.6636
XGBoost	0.8704	0.7802	0.6581	0.7140

Gender Bias in Predictions

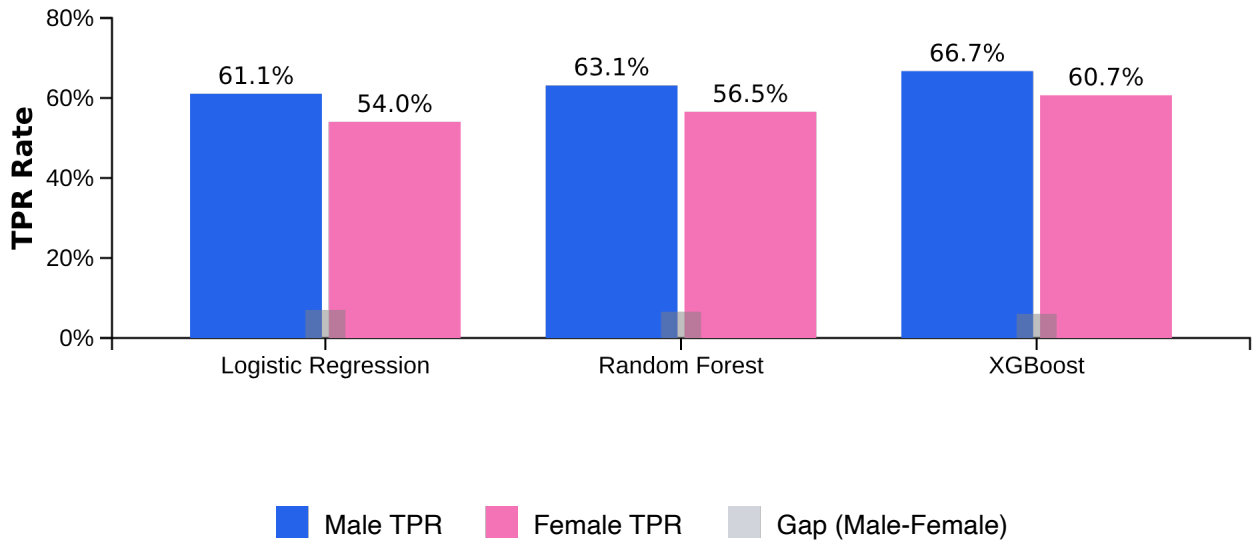
All models show **TPR gaps by sex**, indicating women have lower positive prediction rates:

Logistic Regression:	0.0702	Random Forest:	0.0657	XGBoost:	0.0604

Baseline Fairness Metrics by Sex

Key Finding: All baseline models demonstrate lower positive prediction rates and TPR for women, creating hiring disadvantages.

TPR Gap by Model and Sex



Fairness Metrics by Model

Model	SPD	DI	TPR Gap
Logistic Regression	0.1754	0.3213	0.0702
Random Forest	0.1806	0.3396	0.0657
XGBoost	0.1753	0.3371	0.0604

DP by Sex (Baseline)

DP Female:	0.083 - 0.093
DP Male:	0.259 - 0.273

All models show lower positive prediction rates for women, meaning women are fast-tracked less often in the hiring process.

Why TPR Gaps Matter in Hiring

Impact of Unequal TPRs

True Positive Rate (TPR) represents the probability that a candidate with actual income >50K is correctly fast-tracked to interviews. When TPR gaps exist between groups, it creates a systemic disadvantage.

Logistic Regression

- Male: **61.06%**
- Female: **54.04%**

7.02%

Random Forest

- Male: **63.12%**
- Female: **56.55%**

6.57%

XGBoost

- Male: **66.72%**
- Female: **60.68%**

6.04%

Real-World Consequences

Reduced Interview Opportunities

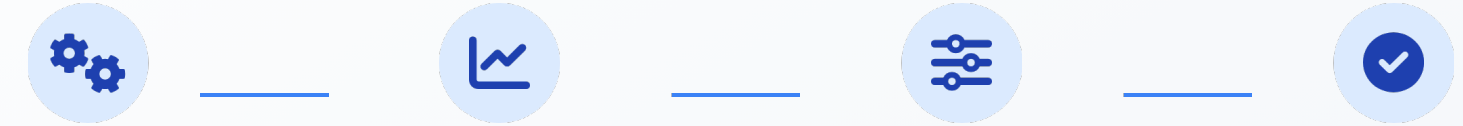
Qualified women with income >50K are less likely to be fast-tracked, reducing their interview chances.

Reinforces Inequality

Systematic disadvantage accumulates, reinforcing existing inequalities in who reaches better-paid positions.

Step 2: Equal Opportunity Post-Processing

Equal Opportunity post-processing adjusts decision thresholds for each group (Male/Female) after model training to reduce TPR gaps without retraining.



How It Works

- Adjusts decision thresholds for Male vs Female groups
- Equalizes TPRs across groups
- Preserves model predictions for each group

Equal Opportunity Goal

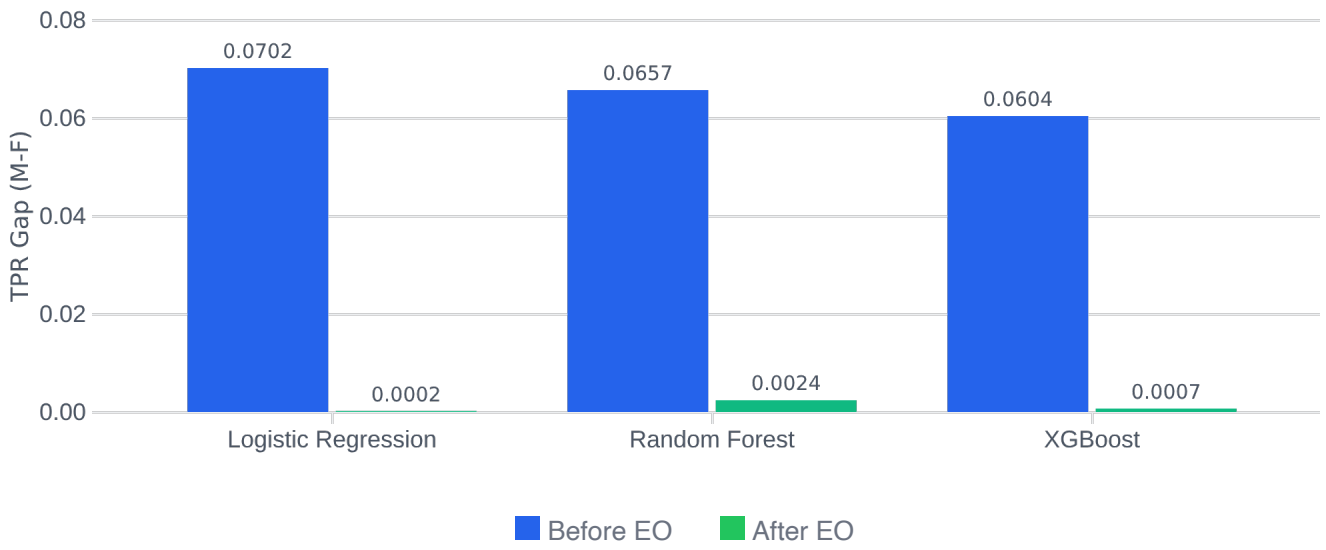
- Minimizes TPR gap (M–F) = 0
- Maintains overall accuracy
- Reduces fairness metrics: SPD, DI

Key Difference from Baseline: Post-processing adjusts the model output after training, rather than modifying the training process. This allows us to correct for fairness issues without retraining the model.

Equal Opportunity Results Comparison

Equal Opportunity post-processing reduces TPR gaps from 0.06-0.07 to nearly zero across all three models.

TPR Gap Comparison



Key Metrics Comparison

MODEL	TPR GAP (BEFORE)	TPR GAP (AFTER)
Logistic Regression	0.0702	≈ 0.0002
Random Forest	0.0657	≈ -0.0024
XGBoost	0.0604	≈ -0.0007

All models achieve near-equal TPR after EO processing.

TPR Gap Reduction

EO reduces TPR gaps from 0.06-0.07 down to ≈ 0 across all models

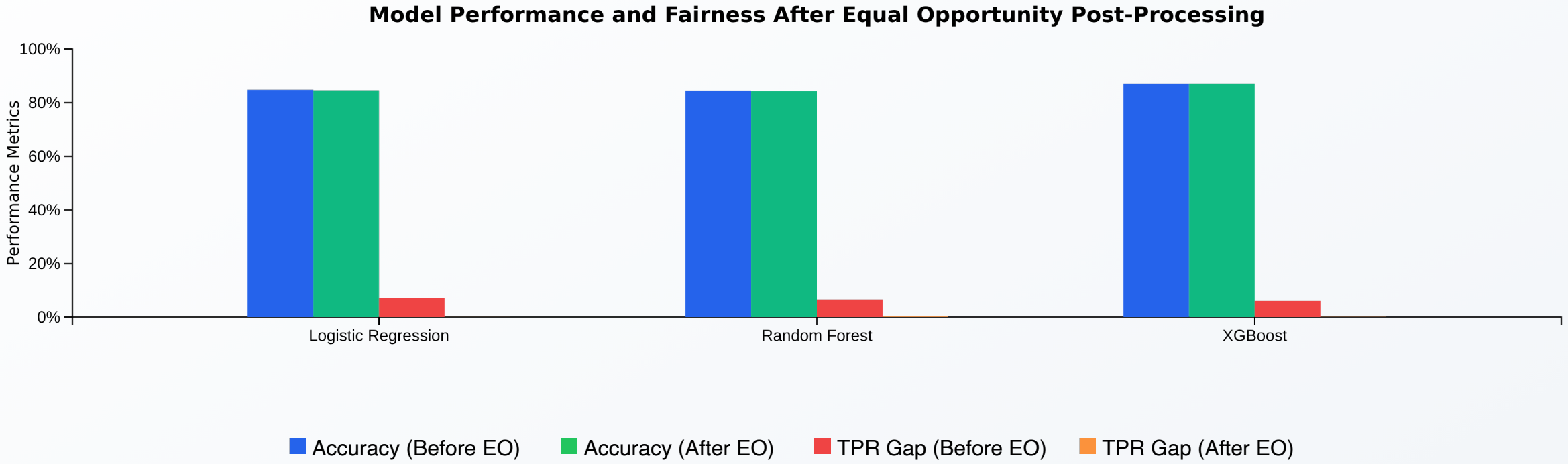
Accuracy Impact

LR and RF show small accuracy decreases, while XGBoost maintains almost the same accuracy

SPD & DI Improvement

SPD and DI move slightly toward parity (still not perfect, but better than baseline)

Trade-offs: Fairness vs Performance



Logistic Regression

- Accuracy decrease: **0.2%**
- TPR gap reduced to nearly equal
- Small performance cost for fairness

Random Forest

- Accuracy decrease: **0.2%**
- TPR gap reduced to nearly equal
- Small performance cost for fairness

XGBoost

- Accuracy slightly improved: **0.1%**
- TPR gap reduced to nearly equal
- Balances performance and fairness

Key Insight: All models achieved near-equal TPR after EO, with XGBoost maintaining performance while LR and RF showed small accuracy decreases.

Conclusions and Limitations

Key Conclusions

EO greatly reduces TPR gaps by sex
From ~0.06-0.07 down to ≈ 0

Accuracy cost is minimal for LR and RF
XGBoost maintains almost the same accuracy

SPD and DI move towards parity
Still not perfect, but better than baseline

Study Limitations

Focus on sex only
Ignoring race and other protected attributes

Dataset limitations
Adult dataset is not a perfect proxy for real hiring

Evaluation scope
Models evaluated only on bias reduction, not other factors

Future Directions

Expand to multiple protected attributes
Include race and other demographic factors

Real-world deployment
Test in actual hiring workflows

Broader evaluation metrics
Include more aspects of fairness and performance

"Fairness-aware algorithms can reduce bias with minimal performance cost, but careful evaluation is needed before deployment."