Technical Report: Racial Disparities in COMPAS Risk Assessment

Analysis of Recidivism Prediction for Black and White Women

Summary

This study examines racial bias in the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) algorithm's assessment of recidivism risk among women. Analyzing a subset of the ProPublica COMPAS dataset focusing on Black and White women, it was found that despite nearly identical 2-year recidivism rates (37.0% vs. 35.3%), Black women receive systematically higher risk scores. Kaplan-Meier survival analysis confirms no statistically significant difference in time to recidivism ($p = 0.45$), yet score distributions show clear racial disparities. The predictive models demonstrate that age and prior criminal history are the strongest legitimate predictors of recidivism, suggesting that racial disparities in COMPAS scores may reflect algorithmic bias rather than actual risk differences.

---

1. Introduction

1.1 Background

The COMPAS algorithm is widely used in the United States criminal justice system to assess the risk of recidivism. Judges and parole boards rely on these risk scores for critical decisions regarding bail, sentencing, and parole. In 2016, ProPublica's investigation "Machine Bias" revealed that COMPAS exhibited racial bias, with Black defendants receiving higher risk scores than White defendants with similar profiles.

1.2 Research Questions

This study focuses specifically on women and investigates:

1. Do Black and White women receive different COMPAS risk scores?
2. Are actual recidivism rates different between these groups?
3. What factors legitimately predict recidivism?
4. Can we build predictive models that achieve similar accuracy without racial bias?

1.3 Significance

Women represent a growing but understudied population in the criminal justice system. Understanding how algorithmic bias affects women specifically is crucial for: - Ensuring fair treatment in judicial decisions - Identifying gender-specific patterns in recidivism - Informing policy on algorithmic accountability - Contributing to criminal justice reform

---

2. Data and Methodology

2.1 Dataset

Source: ProPublica COMPAS Analysis (2016)
Original Size: ~11,000 records
Filtered Sample: 700-800 Black and White women
Time Period: 2-year recidivism tracking
Data Quality: $< 2\%$ missing values

2.2 Feature Selection

Demographic Features: - Age (continuous and categorical) - Race (African-American vs. Caucasian) - Sex (Female only in this analysis)

Criminal History: - Prior convictions count - Juvenile felony, misdemeanor, and other charges - Current charge degree (Felony vs. Misdemeanor)

COMPAS Assessment: - Decile score (1-10) - Risk category (Low, Medium, High)

Outcome Variable: - Two-year recidivism (binary) - Time to recidivism (days)

2.3 Analytical Approach

Phase 1: Exploratory Data Analysis

- Correlation analysis
- Demographic comparisons
- COMPAS score distribution analysis
- Recidivism rate calculations

Phase 2: Bias Quantification

- False Positive Rate (FPR): P(High Risk | No Recidivism)
- False Negative Rate (FNR): P(Low Risk | Recidivism)
- Recidivism rates by risk category
- Age-stratified analysis

Phase 3: Dimensionality Reduction

- Principal Component Analysis (PCA)
- Variance explained analysis
- Feature space visualization by race

Phase 4: Predictive Modeling **Models**: 1. Logistic Regression 2. Random Forest Classifier

Evaluation Metrics: - Area Under ROC Curve (AUC) - Confusion Matrix - Feature Importance - Accuracy, Precision, Recall

Phase 5: Survival Analysis

- Kaplan-Meier survival curves
- Log-rank statistical test
- Stratification by age and COMPAS risk
- Cumulative hazard functions

---

3. Results

3.1 Demographic Comparison

Age Distribution

- **Black Women**: Mean age ~32 years, similar distribution to White women
- **White Women**: Mean age ~31 years
- **Finding**: No significant age difference between groups

Prior Convictions

- **Black Women**: Mean = 2.53 priors
- **White Women**: Mean = 1.81 priors
- **Difference**: 40% higher for Black women
- **Implication**: This disparity partially explains COMPAS score differences but raises questions about upstream bias in arrests/convictions

Charge Severity

- **Black Women**: 67% face felony charges
- **White Women**: 52% face felony charges
- **Finding**: Black women disproportionately charged with more serious offenses

3.2 COMPAS Score Analysis

Score Distribution

- **Black Women**: Concentrated in scores 6-8 (higher risk)
- **White Women**: Concentrated in scores 1-3 (lower risk)
- **Statistical Test**: Distributions are significantly different (Kolmogorov-Smirnov test, $p < 0.001$)

Risk Category Assignment

Risk Category	Black Women	White Women
Low (1-4)	~30%	~55%
Medium (5-7)	~45%	~30%
High (8-10)	~25%	~15%

Finding: Black women are disproportionately assigned to higher-risk categories.

3.3 Actual Recidivism Outcomes

Overall 2-Year Recidivism

- **Black Women**: 37.0%
- **White Women**: 35.3%
- **Difference**: 1.7 percentage points (not statistically significant)

Recidivism by Age Group

Age Group	Black Women	White Women
< 25	~37%	~41%
25-45	~25%	~24%
> 45	~44%	~31%

Finding: Age patterns are generally consistent across races, with younger individuals showing higher recidivism.

Recidivism by COMPAS Risk Category

Risk Category	Black Women	White Women
Low	~23%	~26%
Medium	~45%	~46%
High	~65%	~66%

Finding: Within each risk category, recidivism rates are nearly identical, suggesting COMPAS can distinguish risk levels but assigns groups to categories differently.

3.4 Bias Metrics

False Positive Rate (Labeled High Risk but Did NOT Recidivate)

- **Black Women**: 35.7%
- **White Women**: 34.0%
- **Difference**: 1.7 percentage points

False Negative Rate (Labeled Low Risk but DID Recidivate)

- **Black Women**: 22.4%
- **White Women**: 25.5%
- **Difference**: 3.1 percentage points (White women more likely to be underestimated)

Interpretation: While differences are modest, they suggest asymmetric treatment - Black women slightly more likely to be overestimated, White women more likely to be underestimated.

3.5 Principal Component Analysis

Variance Explained

- **PC1**: 32.5% of variance
- **PC2**: 25.5% of variance
- **PC1 + PC2**: 58% cumulative variance
- **First 6 components**: 90% cumulative variance

Feature Loadings PC1 (Risk/Criminal History Axis): - High positive loadings: decile_score, risk_encoded, priors_count - Interpretation: Captures overall risk profile

PC2 (Age Axis): - High loadings: age-related features, age_priors_interaction - Interpretation: Captures age-related risk patterns

Biplot Insights

- Substantial overlap between Black and White women
- Separation primarily driven by risk-related features
- Race × prior interaction appears prominently
- Suggests risk assessment differences are not based on fundamentally different characteristics

3.6 Predictive Modeling

Model Performance

Model	AUC	Accuracy	Precision	Recall
Logistic Regression	0.696	69.4%	66.7%	32.3%
Random Forest	0.676	66.7%	55.9%	40.9%

Finding: Both models perform reasonably well (AUC ~0.68-0.70), suggesting recidivism is moderately predictable using legitimate factors.

Feature Importance **Logistic Regression (Top 5)**: 1. priors_count (coef = 0.65) 2. age_priors_interaction (coef = 0.35) 3. decile_score (coef = 0.30) 4. risk_encoded (coef = 0.20) 5. days_b_screening_arrest (coef = 0.18)

Random Forest (Top 5): 1. age (importance = 0.18) 2. age_priors_interaction (importance = 0.16) 3. decile_score (importance = 0.14) 4. priors_count (importance = 0.12) 5. days_b_screening_arrest (importance = 0.10)

Key Finding: Both models identify age and criminal history as most important. Race alone has low importance, but race × prior interactions show moderate importance.

Confusion Matrix Analysis **Logistic Regression**: - True Negatives: 147, False Positives: 15 - False Negatives: 63, True Positives: 30 - Tends to under-predict recidivism (more FN than FP)

Random Forest: - True Negatives: 132, False Positives: 30 - False Negatives: 55, True Positives: 38 - More balanced but still struggles with FN

Interpretation: Our models show different error patterns than COMPAS, which tends to over-predict risk for Black women.

3.7 Survival Analysis

Kaplan-Meier Curves: Race Comparison **Log-Rank Test**: p-value = 0.4476

Interpretation: No statistically significant difference in survival curves between Black and White women. Both groups show: - Similar initial recidivism rates - Parallel curve trajectories - Overlapping confidence intervals - Convergence around 35-37% recidivism by day 700

Conclusion: Time to recidivism is statistically identical between groups.

Age-Stratified Survival Analysis **Pattern for Both Races**: - **< 25 years**: Steepest decline, ~60-65% survival by day 700 - **25-45 years**: Moderate decline, ~65-70% survival - **> 45 years**: Flattest curve, ~75-80% survival

Key Finding: Age is a strong predictor of recidivism risk across both racial groups, with consistent patterns.

COMPAS Risk-Stratified Analysis **Pattern for Both Races**: - **Low Risk**: ~75-80% survival (correctly identified low risk) - **Medium Risk**: ~55-60% survival - **High Risk**: ~32-35% survival (correctly identified high risk)

Key Finding: COMPAS successfully distinguishes risk levels within both racial groups. The problem is differential assignment to risk categories, not prediction accuracy within categories.

Cumulative Hazard Analysis Both Black and White women show: - Rapid hazard accumulation in first 200 days - Gradual increase from days 200-500 - Plateau around days 500-700 - Final cumulative probability ~35-37%

Conclusion: Risk accumulation patterns are identical.

---

4. Discussion

4.1 Evidence of Algorithmic Bias

Our analysis provides multiple lines of evidence for algorithmic bias in COMPAS:

1. **Score Disparity**: Black women receive significantly higher scores despite similar outcomes
2. **Outcome Similarity**: Recidivism rates differ by only 1.7 percentage points (not statistically significant)
3. **Temporal Equivalence**: No difference in time to recidivism (p = 0.45)
4. **Within-Category Accuracy**: COMPAS predicts equally well for both races, yet assigns them differently

4.2 The Prior Convictions Complication

Black women in our dataset have 40% more prior convictions (2.53 vs. 1.81). This presents a complex question:

Interpretation 1: Legitimate Risk Factor - Prior convictions predict future behavior - Higher priors justify higher scores - Disparity reflects real criminogenic risk

Interpretation 2: Upstream Bias - Higher priors may reflect biased policing/prosecution - Black individuals arrested/convicted more for same behaviors - COMPAS amplifies pre-existing bias

Our Finding: Even controlling for priors in our models, actual recidivism remains identical between groups. This suggests either: - Priors are being overweighted by COMPAS - The priors themselves reflect systemic bias - Some combination of both

4.3 Legitimate Predictors vs. Proxies

Our models reveal: - **Strong Predictors**: Age, prior convictions, interaction terms - **Weak Predictors**: Race alone, current charge degree - **Moderate Predictors**: Race × prior interactions

This suggests COMPAS may be using race as a proxy or amplifying its effects through interaction terms rather than as a direct predictor.

4.4 Model Performance Without Explicit Race

Our Logistic Regression and Random Forest models achieve: - AUC ~0.68-0.70 (comparable to COMPAS) - Similar accuracy levels - Without explicit racial weighting

Implication: Fair risk assessment is possible without relying on race or racial proxies.

4.5 Policy Implications

1. **Algorithmic Auditing**: Risk assessment tools should be regularly audited for racial disparities
2. **Transparency**: Scoring methodologies should be transparent and explainable
3. **Human Oversight**: Algorithms should inform, not replace, human judgment
4. **Bias Correction**: Models should be adjusted when outcome parity exists despite score disparity
5. **Upstream Reform**: Address bias in arrests and prosecutions that create disparate priors

4.6 Limitations

Sample Size: Analysis limited to ~700-800 women; larger samples would increase statistical power

Scope: Focus on Black vs. White women; does not examine other races, men, or intersectional effects

Causality: Observational study cannot definitively prove causation

Ground Truth: Recidivism based on arrests, which may themselves reflect bias

External Validity: COMPAS implementation may vary by jurisdiction

Feature Availability: Analysis limited to features in ProPublica dataset; COMPAS may use additional proprietary features

---

5. Conclusions

This study provides empirical evidence of racial disparities in COMPAS risk assessment for women. Despite nearly identical recidivism rates (37.0% vs. 35.3%) and no statistical difference in time to recidivism ($p = 0.45$), Black women receive systematically higher risk scores. While some disparity is explained by higher average prior convictions among Black women, this raises fundamental questions about whether those priors reflect actual risk or upstream systemic bias.

Our predictive models demonstrate that age and criminal history are the most important legitimate predictors of recidivism, and that comparable accuracy can be achieved without explicit racial weighting. The finding that COMPAS predicts equally well within risk categories for both races, yet assigns groups to categories differently, suggests the bias occurs in risk categorization rather than in the algorithm's ability to predict.

These findings have significant implications for criminal justice policy. If algorithmic risk assessment tools are to be used in high-stakes decisions about bail, sentencing, and parole, they must be: 1. Regularly audited for racial bias 2. Transparent in their methodology 3. Subject to human oversight 4. Adjusted when disparities cannot be justified by actual outcome differences

Future research should examine: - Intersectional effects (race $\times$ gender $\times$ age) - Longitudinal bias patterns (does bias persist over time?) - Jurisdictional variations (does COMPAS behave differently across locations?) - Alternative risk assessment approaches that minimize bias

Ultimately, this analysis contributes to the growing body of evidence that algorithmic decision-making in criminal justice requires careful scrutiny to ensure fairness and equal treatment under the law.