Wharton
UNIVERSITY of PENNSYLVANIA

Generalized Shapley Additive Explanations (G-SHAP)

By Dillon Bowen and Lyle Ungar

Zoom link

SUMMARY

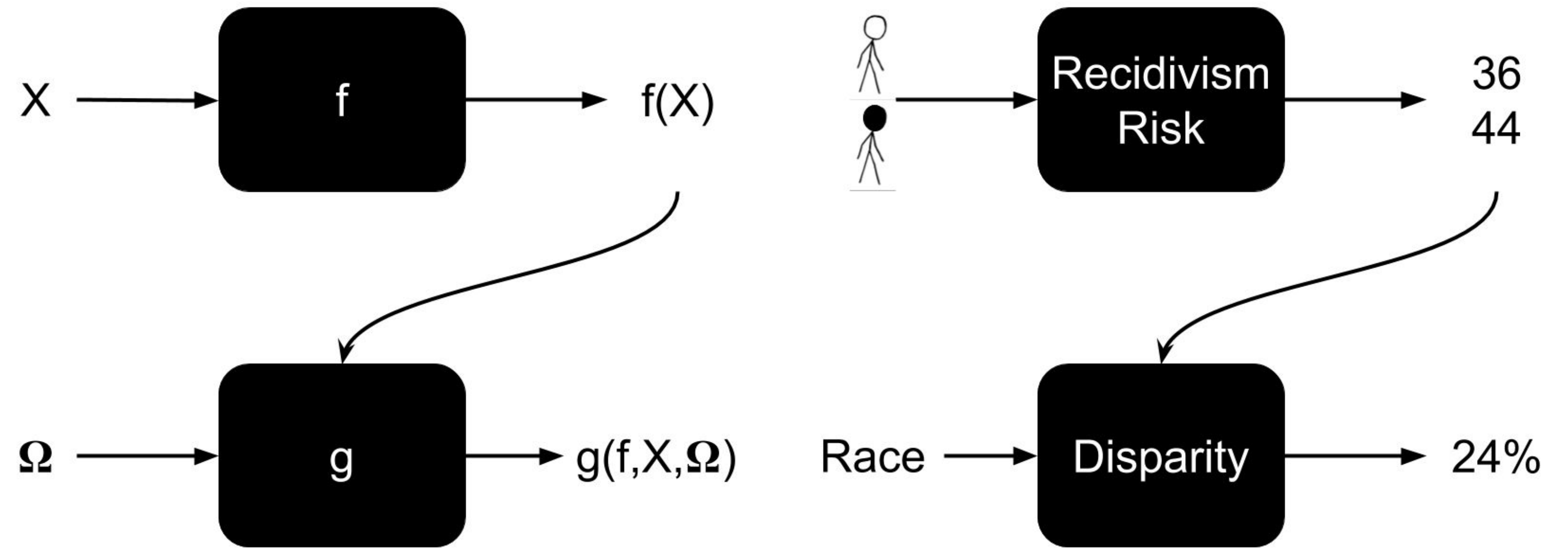
Machine learning is increasingly used both for decision support and for predicting human behavior. However, the opacity of these models is an impediment to their adoption as decision support tools and makes them unsuitable for scientists seeking to understand human behavior. This paper generalizes a popular method in explainable artificial intelligence, Shapley Additive Explanations (SHAP). Our new method, Generalized Shapley Additive Explanations (G-SHAP), broadens the set of questions we can ask of machine learning models.

We apply G-SHAP to compare how humans and machine learning models predict criminal recidivism using demographics and criminal records. Our analysis suggests the following:

1. **Our machine learning model is both more accurate and fairer** (exhibits less racial disparity) than humans.
2. **Humans over-rely on a single variable** (prior convictions), whereas our model considers a broader range of variables.
3. **Human predictions exhibit racial disparity**, driven by prior convictions, age, and race itself. **Our model eliminates racial disparity by implementing a type of affirmative action without sacrificing performance.**

G-SHAP METHOD

General formulation **Recidivism application**



X → f → f(X)

Ω → g → g(f,X,Ω)

Recidivism Risk → 36 44

Race → Disparity → 24%

1. Compute the model output for a dataset *X*.
2. Transform the output (*g*).
3. Run a Shapley decomposition on the transformed output.

1. Predict recidivism risk for criminal offenders.
2. Calculate the racial disparity.
3. Run a Shapley decomposition on the racial disparity.

DATA AND ANALYSIS

Data. Arrest records, demographics, and criminal history of offenders from Broward County, Florida (N_{White}=2859, N_{Black}=3565).

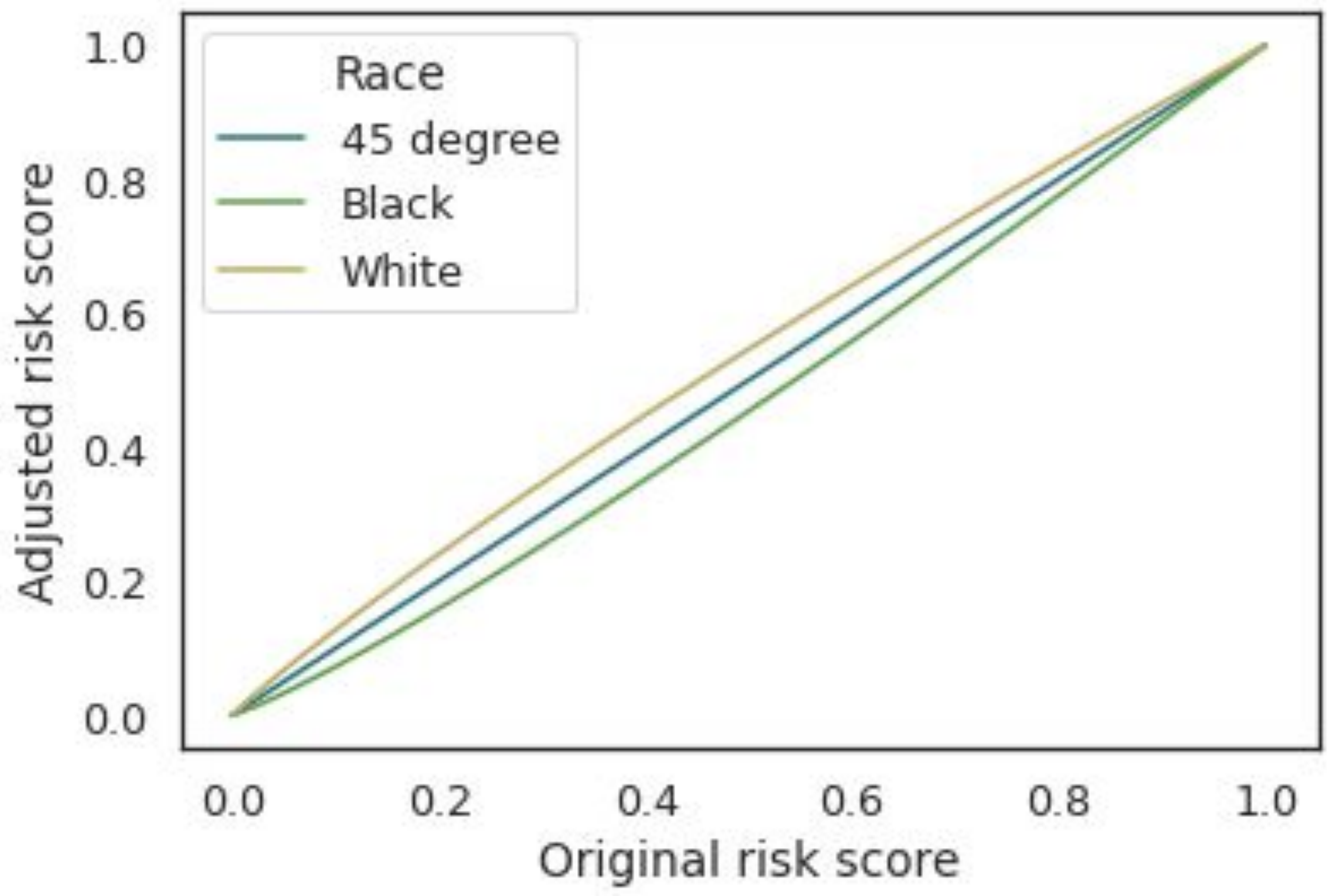
Task. Predict the probability a criminal offender is arrested for another crime within two years (recidivism risk).

Measuring racial disparity. Among offenders who do not recidivate, how much higher is the predicted recidivism risk for Black offenders than for White offenders?

Original model. Trained using an automatic machine learning package (TPOT).

Human subjects. MTurk study where participants predict recidivism risk using the same information as the model (*N*=105).

DEBIASING THE MODEL

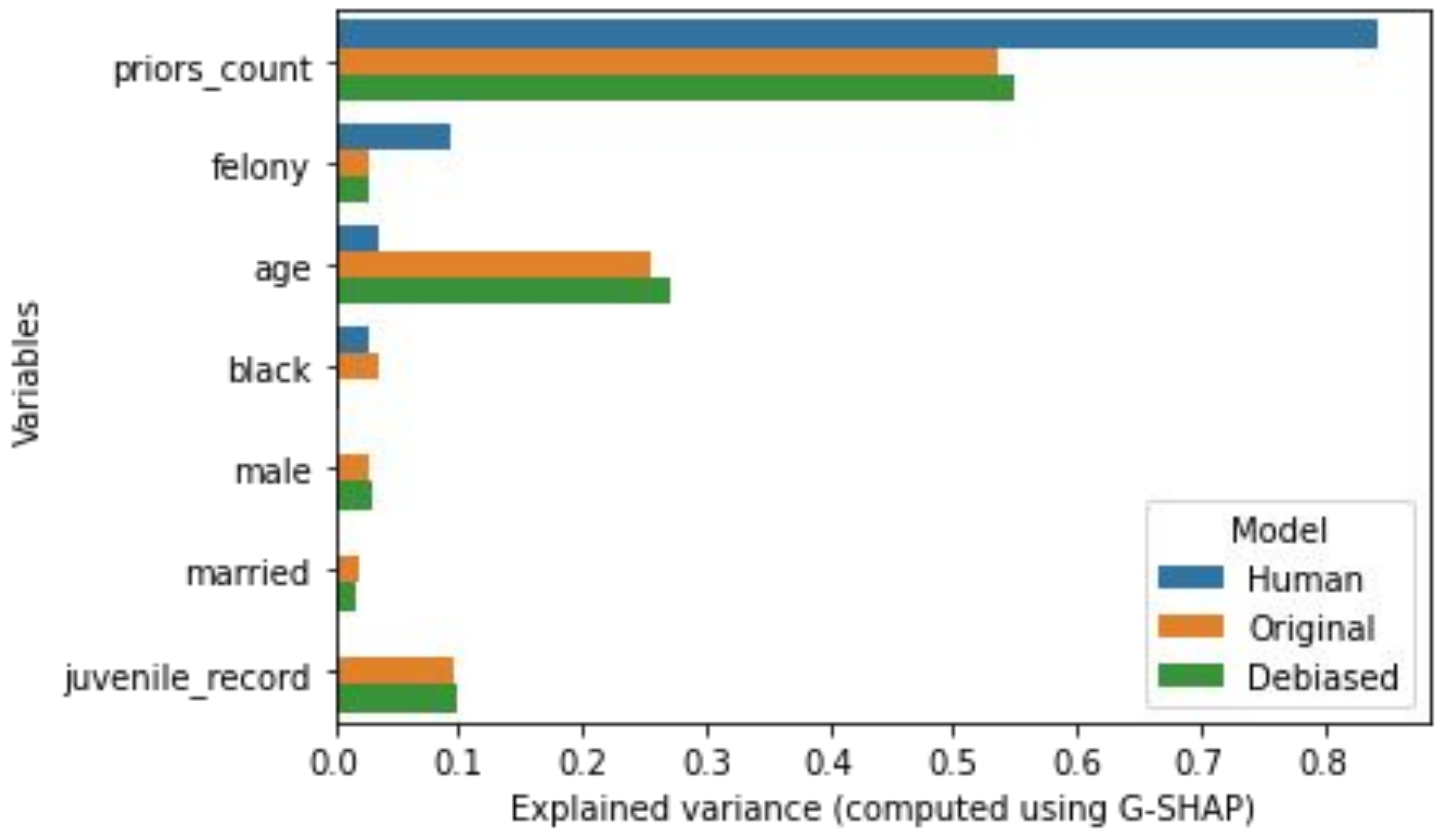


Slightly adjusting predicted risk in favor of Black offenders reduces racial disparity from 24% to 0%.

TEST PERFORMANCE

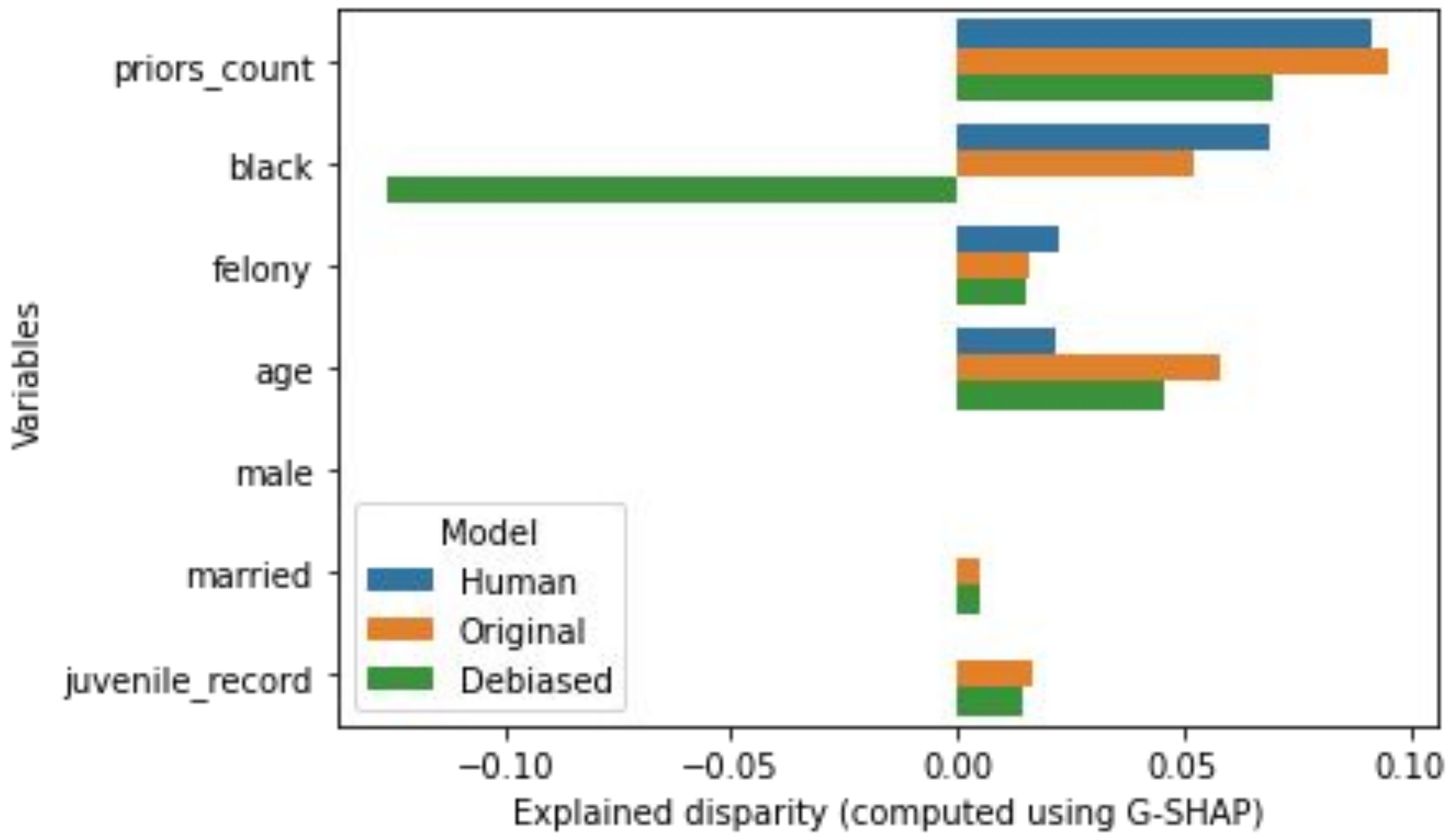
	Human	Original model	Debiased model
Accuracy	63%	67%	67%
AUC	0.61	0.73	0.73
Disparity	20%	24%	2% (ns)

EXPLAINING PREDICTIONS



Discussion. Humans over-rely on prior convictions. Our model considers a broader range of variables like age and juvenile record.

EXPLAINING DISPARITY



Discussion. Among offenders who do not recidivate, people predict the recidivism risk is 20% higher for Black offenders than for White offenders. Of this 20%, 9% is explained by prior convictions, 7% by race, etc. Our debiased model uses race to offset the discriminatory effects of other variables.

SELECTED REFERENCES

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems* (pp. 4765-4774).

Kleinberg, J., Ludwig, J., Mullainathan, S., & Rambachan, A. (2018, May). Algorithmic fairness. In *Aea papers and proceedings* (Vol. 108, pp. 22-27).

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017, August). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining* (pp. 797-806).