

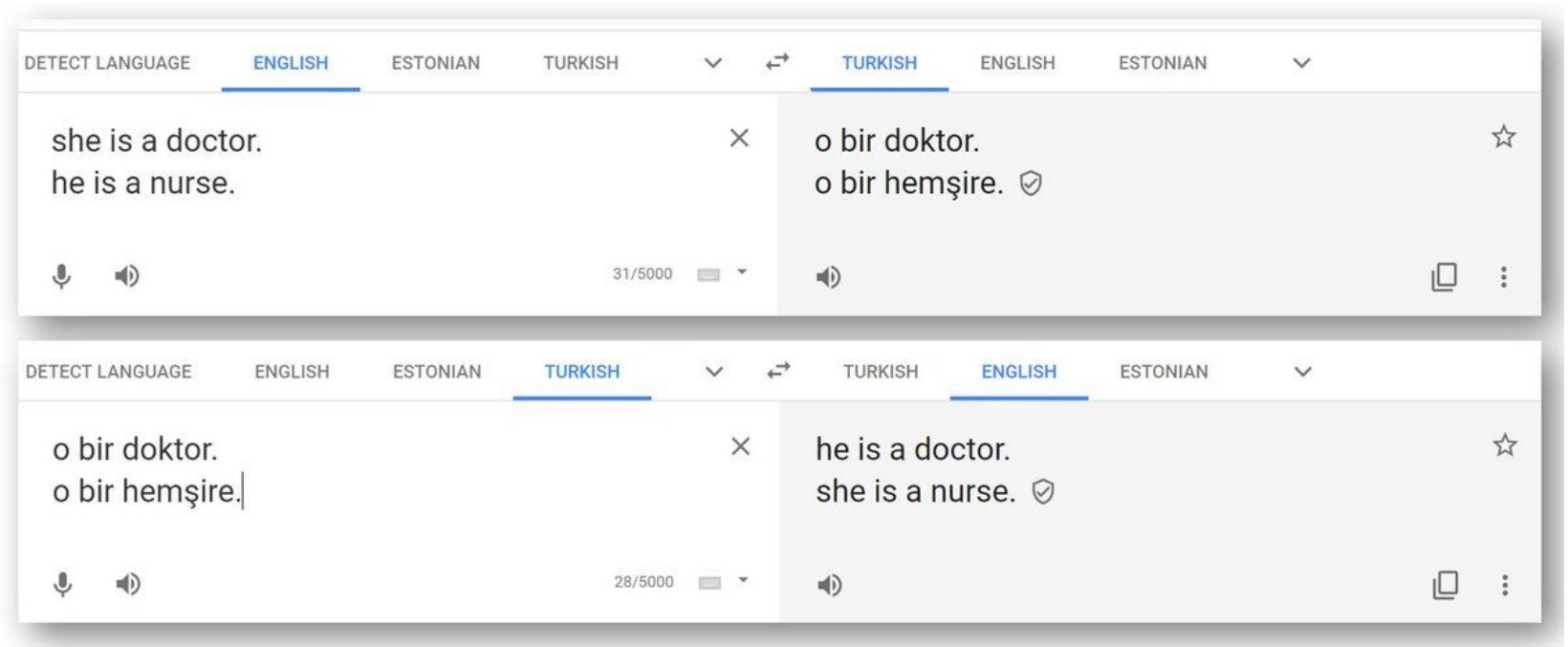
On Algorithmic Fairness and Bias Mitigation in Recidivism Prediction

An econometric view on observed tradeoffs between conflicting definitions of fairness, and the applicability of post-processing methods for bias correction of criminal sentencing algorithms

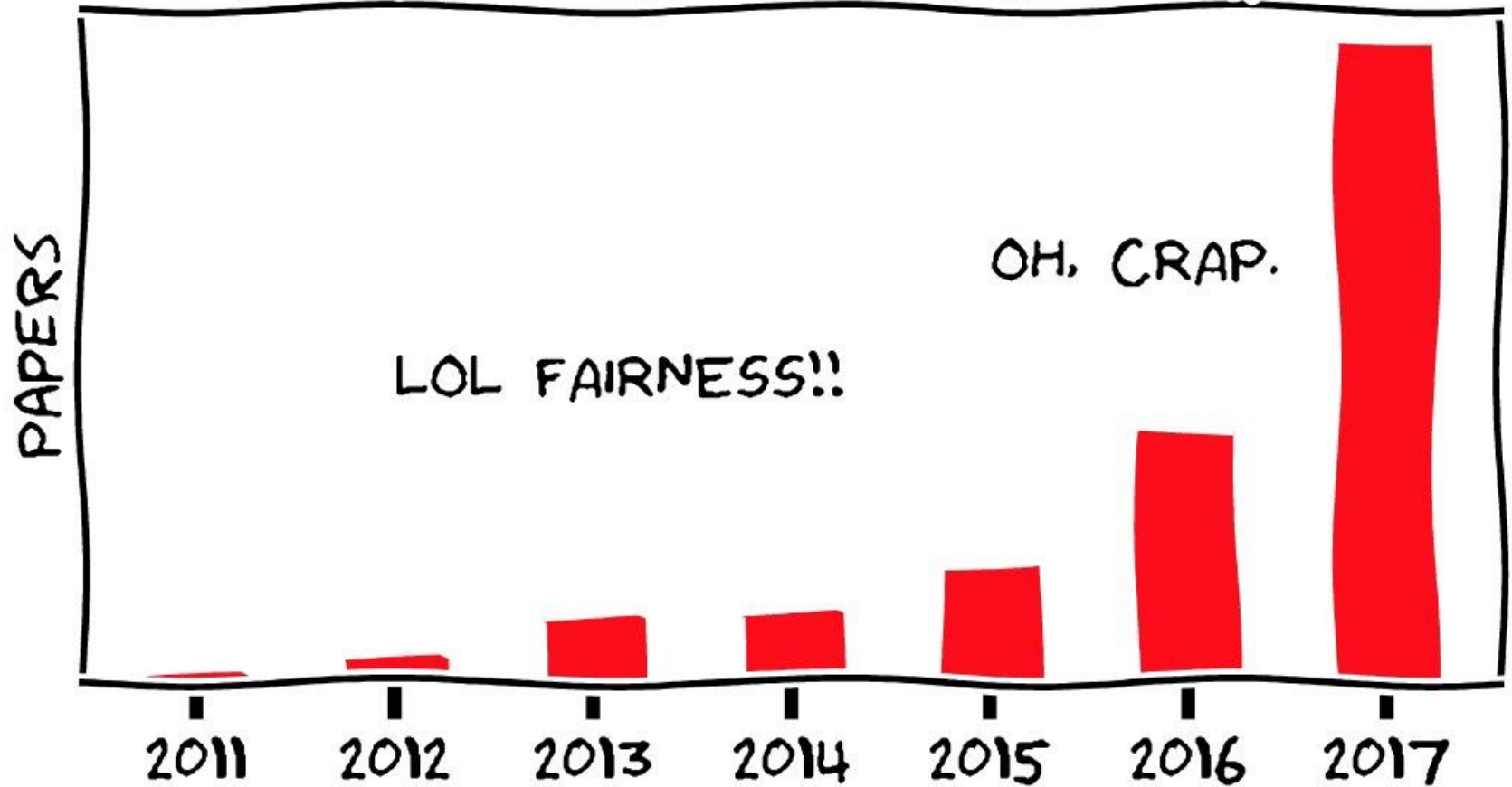
By Lennert Jansen, under the supervision of
dr. Paul Bouman (EUR) and Benjamin
Timmermans (IBM)

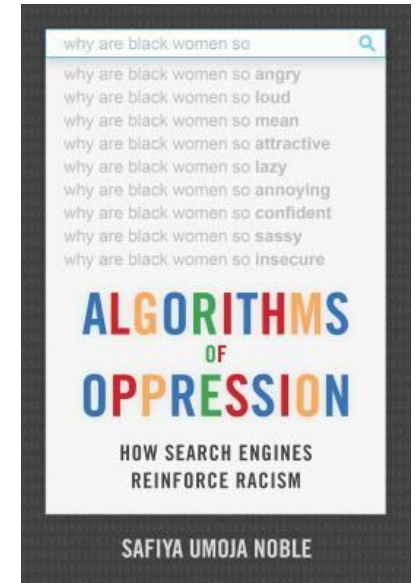
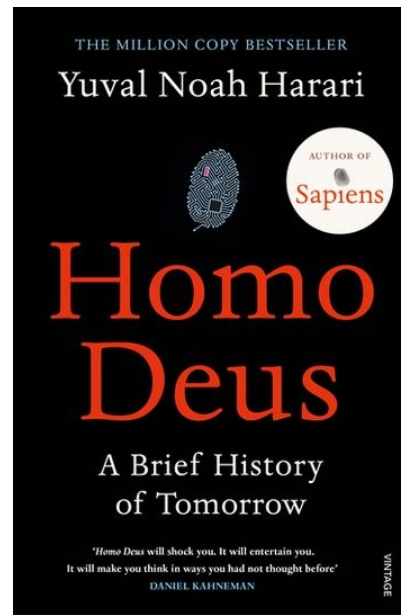
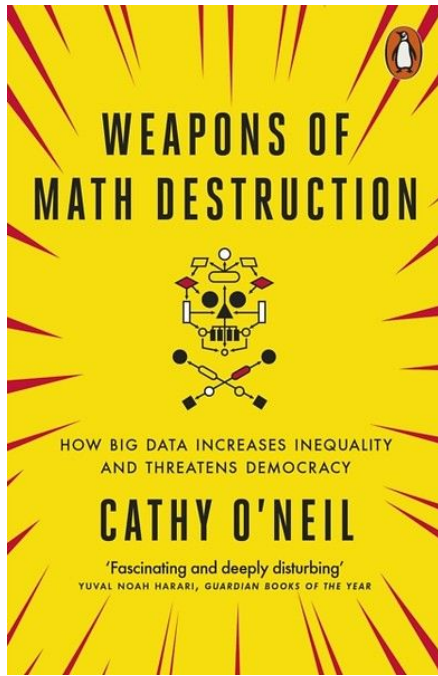
Acknowledgements

To my supervisors, family, and friends.



BRIEF HISTORY OF FAIRNESS IN ML





Fairness and machine learning

Limitations and Opportunities

Solon Barocas, Moritz Hardt, Arvind Narayanan

This online textbook is an incomplete work in progress. Essential chapters are still missing. In the spirit of open review, we solicit broad feedback that will influence existing chapters, as well as the development of later material.

CONTENTS

[ABOUT THIS BOOK](#)

1 [INTRODUCTION](#)

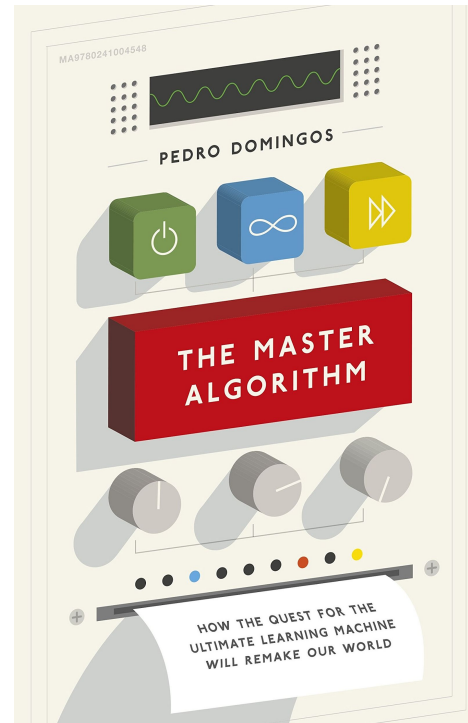
[PDF](#)

2 [CLASSIFICATION](#)

[PDF](#)

We introduce formal non-discrimination criteria, establish their relationships, and illustrate their limitations.

3 [LEGAL BACKGROUND AND NORMATIVE QUESTIONS](#)



Ezra

Theoretical background

Fairness in Machine Learning

*“In the context of decision-making, fairness is the absence of any prejudice or favouritism towards an individual or group based on their inherent or acquired characteristics.” (**)*

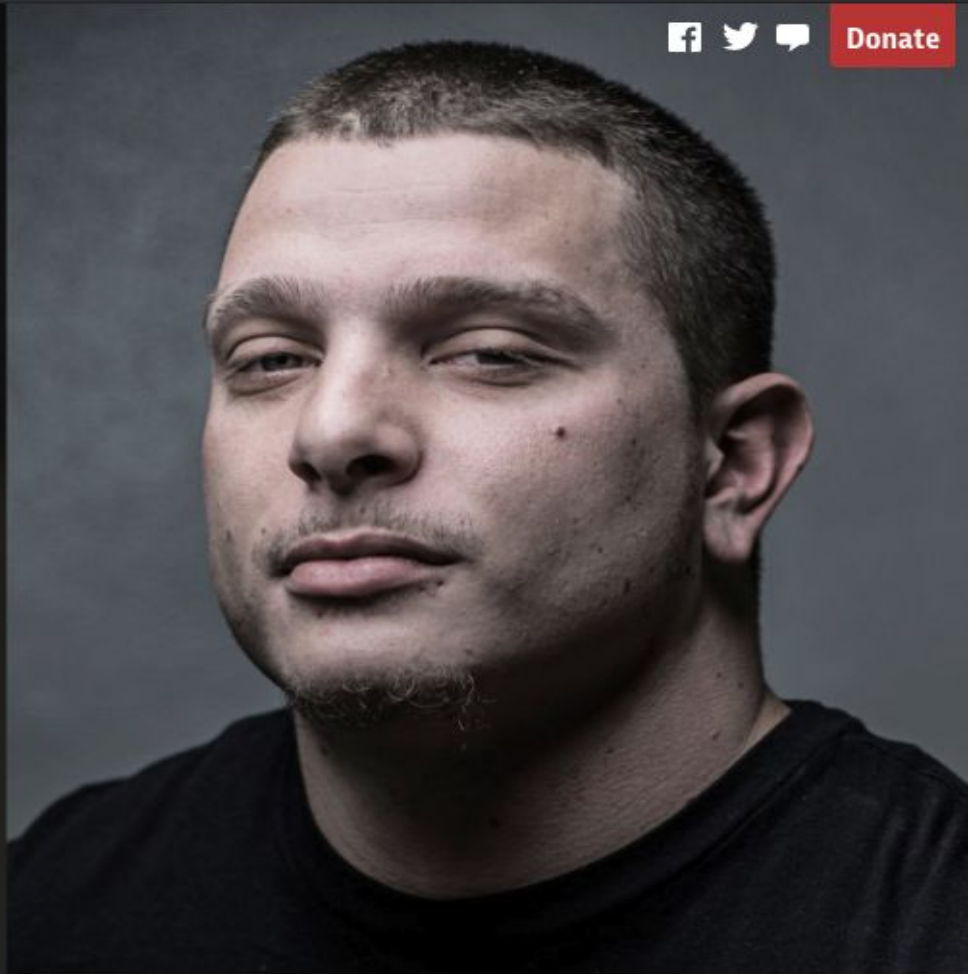
*(**)... on the basis of which discrimination is prohibited or frowned upon.*

-Mehrabi et al. (2019)

Fairness in Machine Learning

Why should we care?

The COMPAS debate



Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

The COMPAS debate

- COMPAS is a Recidivism Prediction Instrument
 - (Correctional Offender Management Profiling for Alternative Sanctions)
- ProPublica:
 - COMPAS violates error rate parities between groups → COMPAS is unfair
- COMPAS' developer:
 - COMPAS is well-calibrated by group → COMPAS is fair

The COMPAS debate

They are both right and wrong, because:

- The two definitions of fairness in question are mutually exclusive
 - (in non-trivial cases)
- *“Neither calibration nor equality of false negative rates rule out blatantly unfair practices.”*

- Corbett-Davies et al. (2017)


Research objectives

Research objectives

- I aimed to investigate...
 - ...the extent to which COMPAS can be accused of being unfair with respect to race and / or gender, using...
 - ...machine learning / econometric methods.
 - ...fairness analysis methods.
 - ...the applicability of various bias mitigation methods, using...
 - ...IBM's new fair machine learning Python library, [AIF360](#).

Methodology and data

Methodology and data

- Data
 - Criminal records of more than 6000 defendants
 - Broward County, Florida, U.S.A.
- Machine learning and econometric methods
 - Random forest
 - Logistic regression
- Fairness analysis
 - Fairness metrics
 - Bias mitigation algorithms

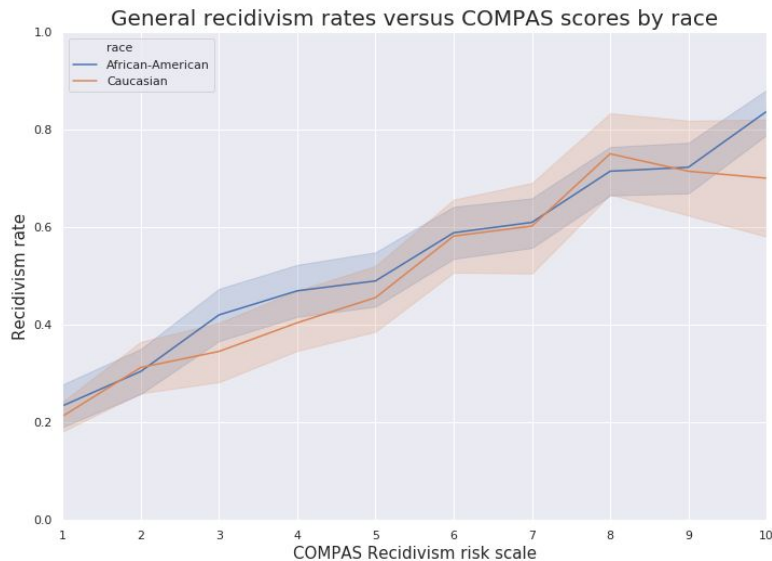
Fairness metrics

- Between-group fairness metrics
 - Is the overall outcome of group A comparable to that of group B?
- Within-group fairness metric
 - Are the outcomes of members of group A far from or close to the average outcome of group A?

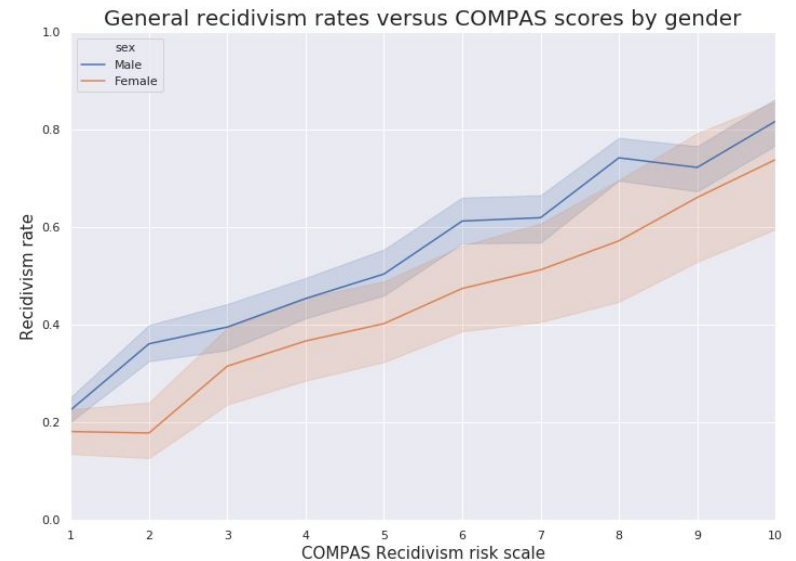
Results

Calibration by group

Well-calibrated by race

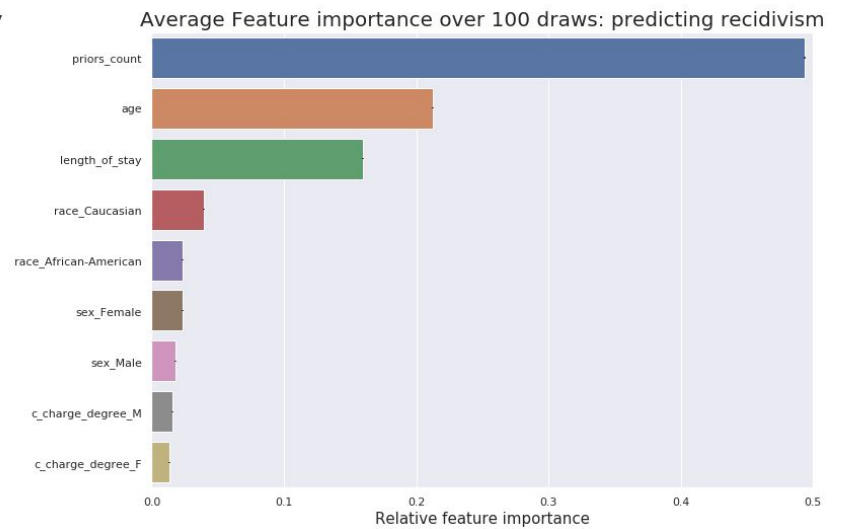
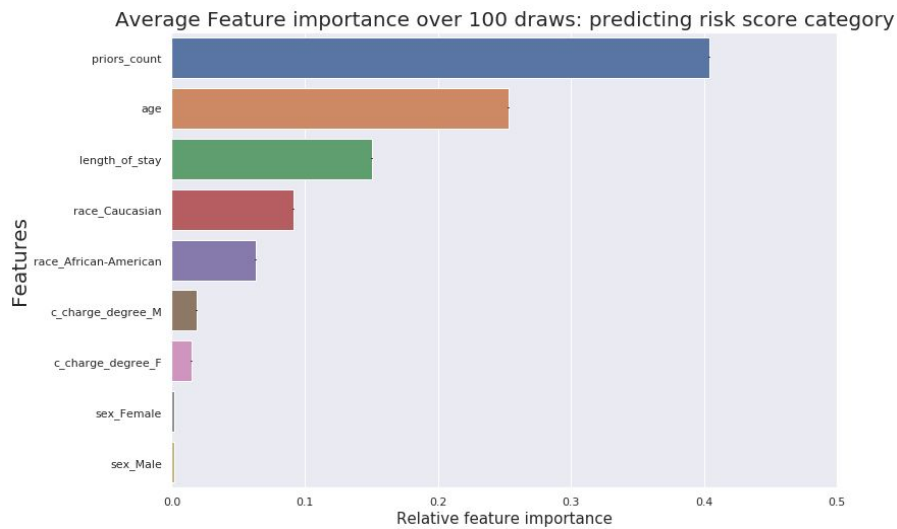


Uncalibrated by sex



Women are being held to the same standard as men, despite women recidivating significantly less

Random forest: relative feature importance



Race is about 2.5 as important for predicting score category than for predicting actual recidivism.

Logistic regression: main findings

- Score category vs. observed recidivism
 - African-Americans are 1.6 times more likely than Caucasians to receive high risk scores*
 - While no significant relationship is found between being African-American and actually recidivating**
- Error types
 - African-Americans are 1.72 times more likely to be misclassified as high-risk...
 - ...and 0.66 times as likely to be misclassified as low-risk **

*holding all other variables constant

** after controlling for relevant factors

Bias mitigation algorithms

- Equalised odds post-processing (EOPP)
- Calibrated equalised odds post-processing (CEOPP)
- Reject option based classification (RObC)

Bias mitigation algorithms

- Equalised odds post-processing (EOPP)
- Calibrated equalised odds post-processing (CEOPP)
- Reject option based classification (RObC)

Bias mitigation results

Method 1: CEOPP

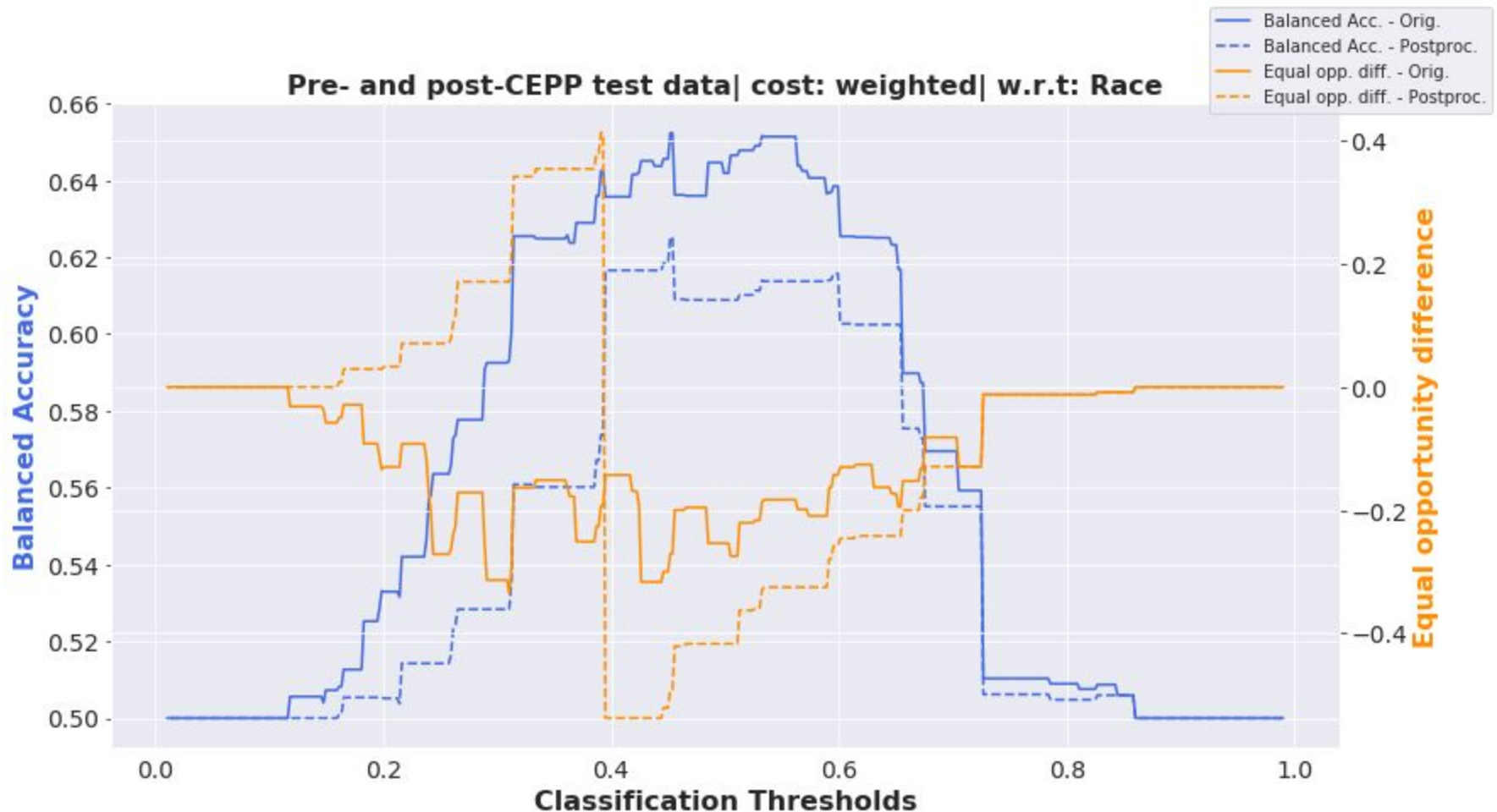
- Acceptable performance
- Satisfies calibration
- Moderately worsens accuracy
- Negligible improvement of between-group fairness
- Cases of notable improvements of within-group fairness
- Probabilistic

N.B.: both methods performed considerably worse for sex

Method 2: RObC

- Best performance by far
- Satisfies parity
- Preservation of accuracy
- Near-optimal between-group fairness
- Slight decrease in within-group fairness
- Deterministic

Bias mitigation results: observed trade-offs



Discussion, implications, and further research

Points of discussion

- Inherently limited analysis due to observed recidivism
- Calibration or parity?
 - i.e., hold different people to different standards
 - Or “arbitrarily” switch labels just to satisfy quota
 - Consult domain experts
- Both bias mitigation methods are ethically debatable
 - Affirmative action in RObC
 - Randomness in CEOPP

Further research

- Modify CEOPP and investigate whether a relaxation of calibration is compatible with error-rate parity
- Develop all-round optimised fairness metric using inequality indices (i.e., expand on the work of Speicher et al. 2018)
- Bias origin modelling
- General purpose bias mitigation algorithm using more advanced deep learning models

Conclusions

Conclusions

- COMPAS accusations confirmed
- Algorithmic fairness is an interdisciplinary problem and should be approached as such
- Achieving fairness between males and females proves difficult
 - Difference in base rates too large
- Reject option based classification achieves 'fair' outcome, while preserving utility
 - Affirmative action-style approach ethically dubious

Conclusions

- Observed trade-offs between
 - Calibration and parity
 - Within- and between-group fairness
 - Fairness and accuracy

Recommended reading, viewing, and listening material



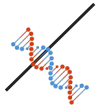
Popular science

- Domingos, P. (2015). *The master algorithm: How the quest for the ultimate learning machine will remake our world*. Basic Books.
- Harari, Y. N. (2016). *Homo Deus: A brief history of tomorrow*. Random House.
- Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.
- O'neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.



Textbooks

- Barocas, S., Hardt, M. & Narayanan, A (2019). *Fairness and Machine Learning. Limitations and Opportunities*. (INCOMPLETE WORKING DRAFT)
- Pearl, J. (2009). *Causality*. Cambridge university press.



Scientific papers

- Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. *ProPublica*, May, 23, 2016.
- Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.*, 104, 671.
- Verma, S., & Rubin, J. (2018, May). Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)* (pp. 1-7). IEEE.



Lectures & Tutorials

- [The Emerging Theory of Algorithmic Fairness. Cynthia Drowk](#)
- [Tutorial: 21 fairness definitions and their politics. by Arvind Narayanan](#)
- [Inherent trade-offs in algorithmic fairness. by Jon Kleinberg](#)
- [Tutorial on Fairness in Machine Learning. by Solon Barocas & Moritz Hardt](#)

References

1. Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2019). Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. 2016. URL [https://www. propublica. org/article/machine-bias-risk-assessments-in-criminal-sentencing](https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing).
2. Barocas, S., Hardt, M., & Narayanan, A. (2017). Fairness in machine learning. *NIPS Tutorial*.
3. Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.*, 104, 671.
4. Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... & Nagar, S. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5), 4-1.
5. 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).
6. Dieterich, W., Mendoza, C., & Brennan, T. (2016). Compas risk scales: Demonstrating accuracy equity and predictive parity. Northpoint Inc.
7. Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.
8. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.
9. Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K. P., Singla, A., Weller, A., & Zafar, M. B. (2018, July). A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2239-2248).

Slides and code available soon on GitHub

- Jupyter Notebooks & Slides soon available at
 - https://github.com/lennertjansen/msc_econometrics_thesis
- IBM's fairness analysis toolkit
 - <https://github.com/IBM/AIF360>
- Various beginner-level resources on fairness in AI
 - <https://aif360.mybluemix.net/>

Thank you for your time and attention

Questions

Appendix

TO ADD TO APPENDIX

- Results of EOPP
- Example of COMPAS survey

Logistic regression: score category

	coef	std err	z	p-value	[0.025	0.975]
const	-1.5248	0.078	-19.442	0.000	-1.679	-1.371
sex_female	0.2193	0.079	2.764	0.006	0.064	0.375
age_cat_greater_than_45	-1.3574	0.099	-13.711	0.000	-1.551	-1.163
age_cat_less_than_25	1.3063	0.076	17.231	0.000	1.158	1.455
race_african_american	0.4770	0.069	6.879	0.000	0.341	0.613
race_other2	-0.5410	0.105	-5.165	0.000	-0.746	-0.336
priors_count	0.2695	0.011	24.305	0.000	0.248	0.291
c_charge_degree_m	-0.3089	0.066	-4.646	0.000	-0.439	-0.179
two_year_recid	0.6821	0.064	10.671	0.000	0.557	0.807

Logistic regression: observed recidivism

	coef	std err	z	p-value	[0.025	0.975]
const	-0.6082	0.065	-9.430	0.000	-0.735	-0.482
sex_female	-0.3477	0.072	-4.840	0.000	-0.489	-0.207
age_cat_greater_than_45	-0.6695	0.076	-8.801	0.000	-0.819	-0.520
age_cat_less_than_25	0.7333	0.069	10.639	0.000	0.598	0.868
race_african_american	0.0959	0.063	1.529	0.126	-0.027	0.219
race_other2	-0.1780	0.088	-2.025	0.043	-0.350	-0.006
priors_count	0.1656	0.008	20.536	0.000	0.150	0.181
c_charge_degree_m	-0.2186	0.059	-3.721	0.000	-0.334	-0.103

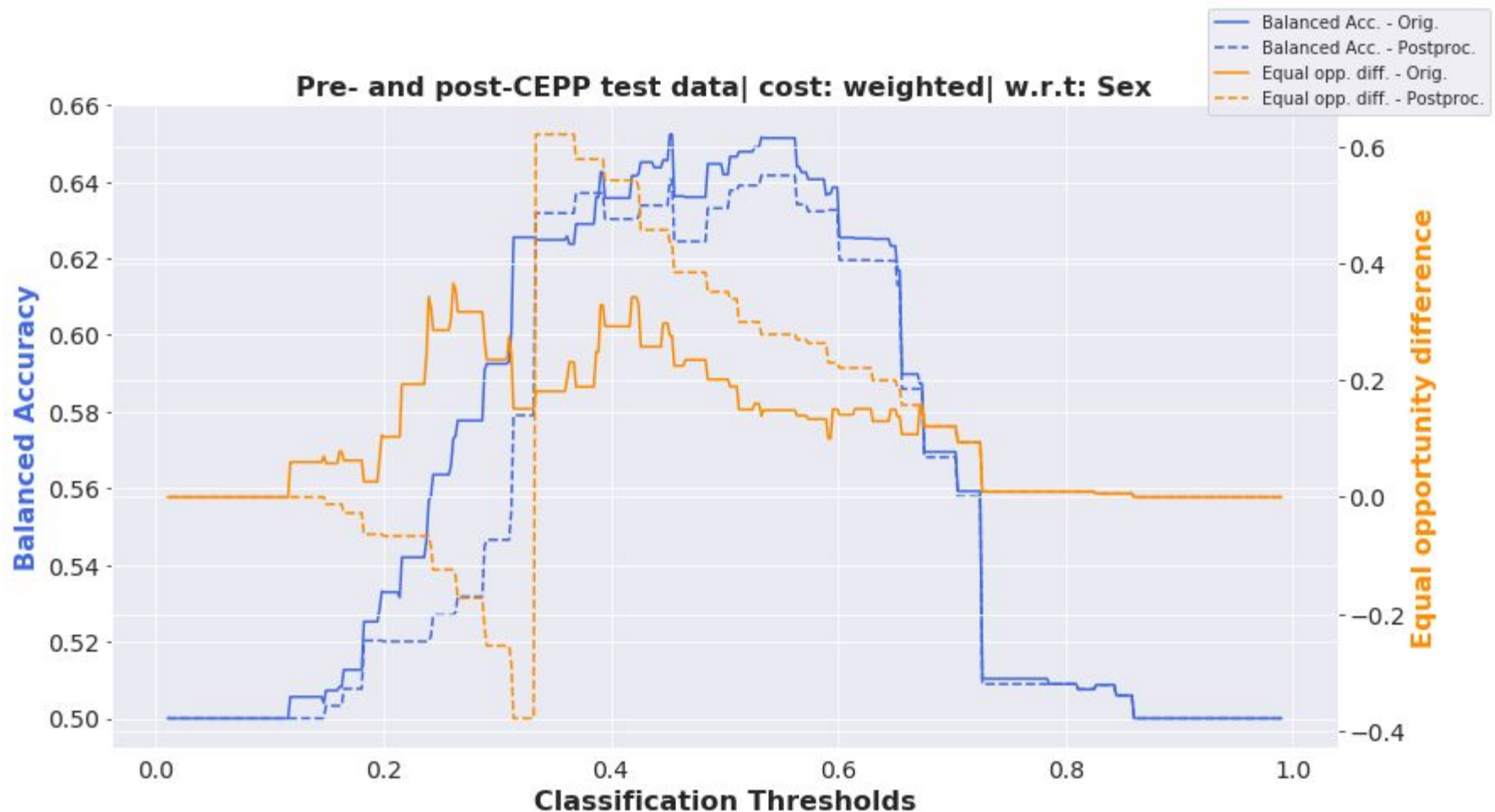
Logistic regression: false negatives

	coef	std err	z	p-value	[0.025	0.975]
const	-1.6686	0.102	-16.347	0.000	-1.869	-1.469
sex_female	0.1867	0.103	1.812	0.070	-0.015	0.389
age_cat_greater_than_45	-1.4005	0.136	-10.310	0.000	-1.667	-1.134
age_cat_less_than_25	1.3885	0.105	13.198	0.000	1.182	1.595
race_african_american	0.5431	0.096	5.669	0.000	0.355	0.731
race_other2	-0.4806	0.145	-3.309	0.001	-0.765	-0.196
priors_count	0.2884	0.017	17.150	0.000	0.255	0.321
c_charge_degree_m	-0.1601	0.091	-1.766	0.077	-0.338	0.018

Logistic regression: false positives

	coef	std err	z	p-value	[0.025	0.975]
const	0.6857	0.108	6.346	0.000	0.474	0.898
sex_female	-0.2483	0.127	-1.961	0.050	-0.496	-0.000
age_cat_greater_than_45	1.3046	0.146	8.940	0.000	1.019	1.591
age_cat_less_than_25	-1.2275	0.110	-11.189	0.000	-1.442	-1.012
race_african_american	-0.4137	0.102	-4.073	0.000	-0.613	-0.215
race_other2	0.6138	0.152	4.046	0.000	0.316	0.911
priors_count	-0.2536	0.015	-17.234	0.000	-0.282	-0.225
c_charge_degree_m	0.4759	0.097	4.883	0.000	0.285	0.667

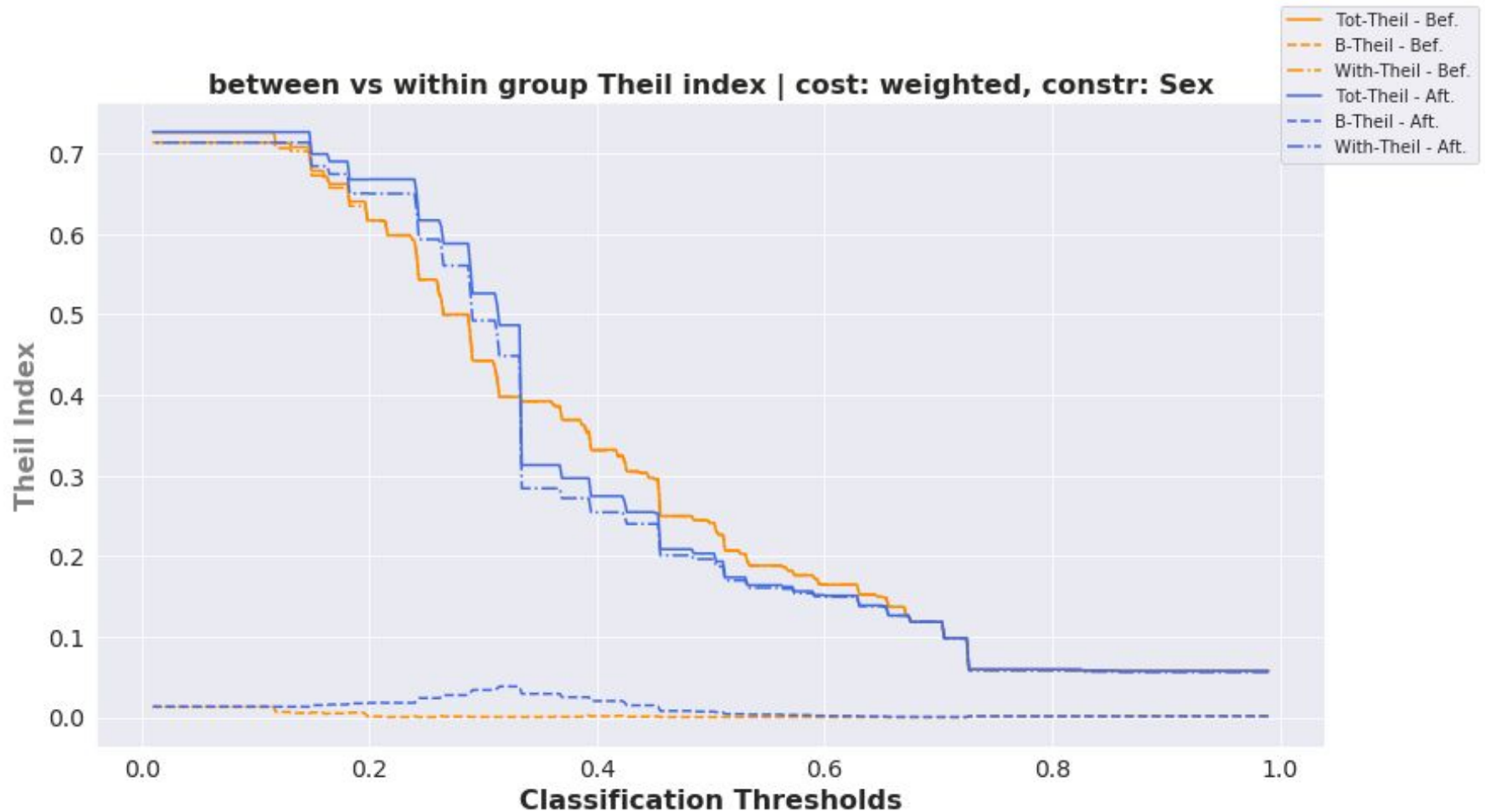
Bias mitigation results: observed trade-offs



Bias mitigation results: observed trade-offs



Bias mitigation results: observed trade-offs



Discrimination...

- ...is domain specific
- ...is feature specific

Two types of discrimination

Disparate treatment

- Formal or intentional
differential treatment
- Intention > outcome

Disparate impact

- Avoidable or unjustified
disadvantageous outcome
- Outcome > intention

Algorithms, Bias, and Criminals

Recidivism prediction instrument (RPI):

Model-based assessment of criminal risk

Commonplace in international judicial systems:

Terbeschikkingstelling (Netherlands)

COMPAS (USA)

*** THINK OF MORE EXAMPLES***

Recidivism Prediction Instruments (RPIs)

Proponents (+)

Opponents (-)

Three fundamental principles: Exceptions

Some fairness criteria outside of these three categories:

- Unawareness
- Individual fairness
- Counterfactual fairness

IBM's AIF360 Python library

Outline of presentation

1. Introduction
 - a. Research objectives and contributions
2. Theoretical framework
 - a. Fairness in Machine Learning
 - b. Defining discrimination
 - c. Algorithms, Bias, and Criminal justice
 - d. On the origin of biases
 - e. The three fundamental principles of algorithmic fairness
 - f. Algorithmic interventions: achieving fairness
3. Related work and the COMPAS-debate
4. Methodology
 - a. Fairness metrics
 - b. Econometric methods
 - c. Bias mitigation algorithms
 - d. Data
5. Experimental setup
6. Results
7. Discussion and topics for further research
8. Conclusion
9. Questions

Three fundamental principles: Independence

- Intuition
 - Prediction model and protected attribute (e.g., race) must be independent
- Examples
 - Demographic parity, statistical parity, the four-fifths rule
- Pros
 - Simple, intuitive, easily compatible with legal notions
- Cons
 - Ignores possible correlation, laziness

Three fundamental principles: Separation

- Intuition
 - Prediction model and protected attribute (e.g., race) must be independent, conditional on target variable
- Examples
 - Equalised odds, equality of opportunity
- Pros
 - Optimality compatible, penalizes laziness
- Cons
 - Requires a reliable predictor (TODO) or data

Three fundamental principles: Sufficiency

(TODO)

- Intuition
 - Target variable is independent of the sensitive attribute, given the prediction score
- Examples
 - Calibration
- Pros
 - The risk score is *sufficient* for equitable prediction
- Cons
 - (TODO)

Three fundamental principles: Trade-offs

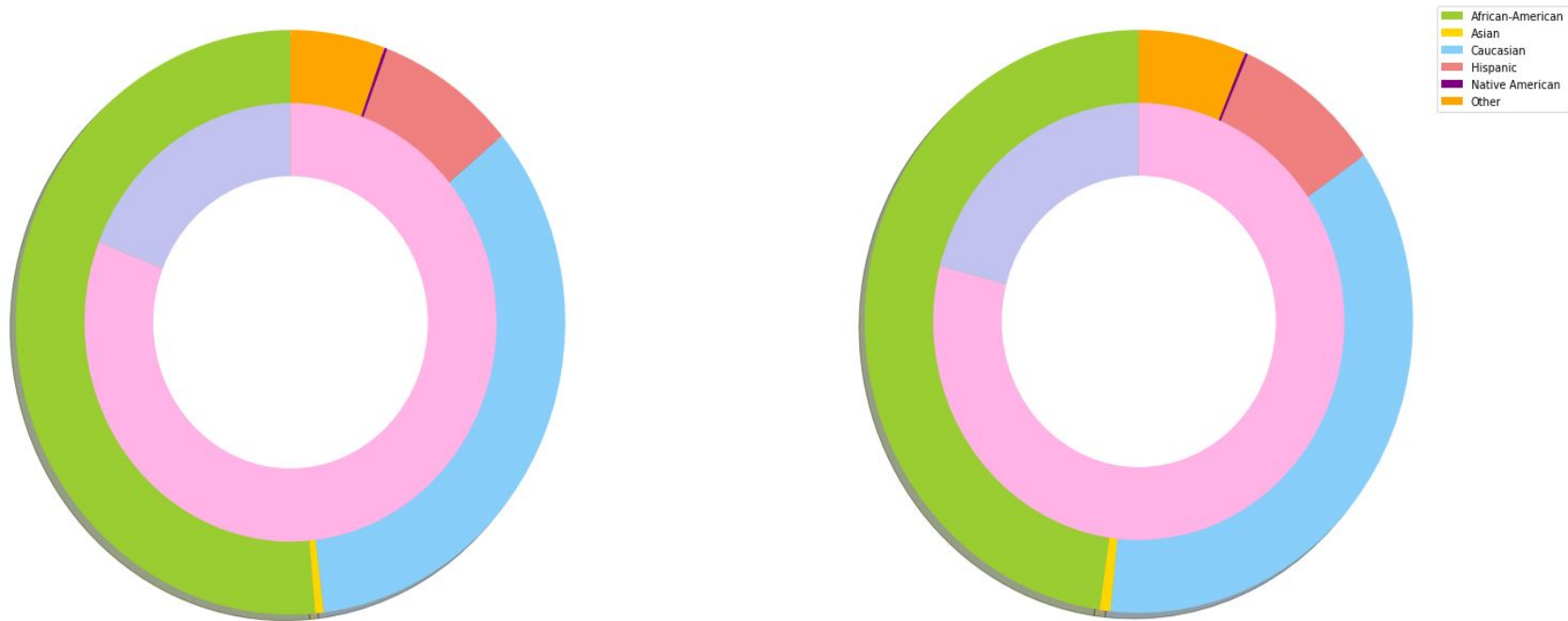
These three fairness criteria are mutually exclusive
(except in degenerate cases)

Three fundamental principles

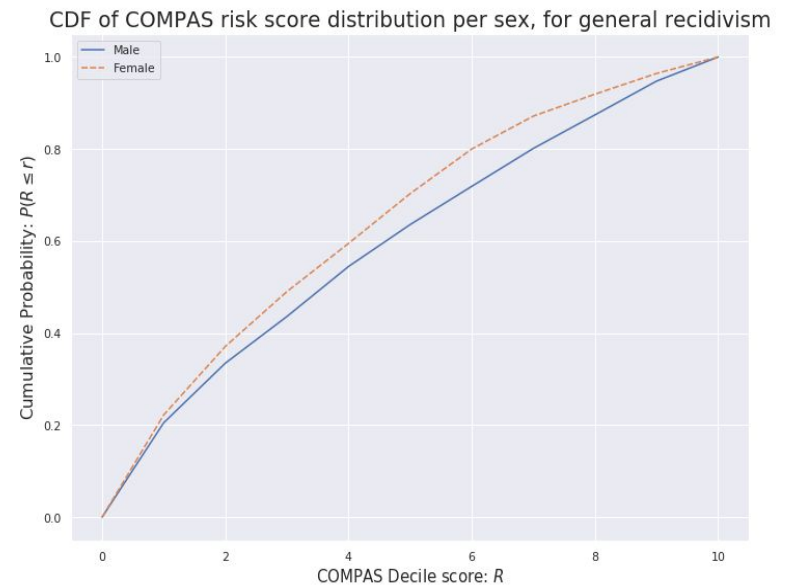
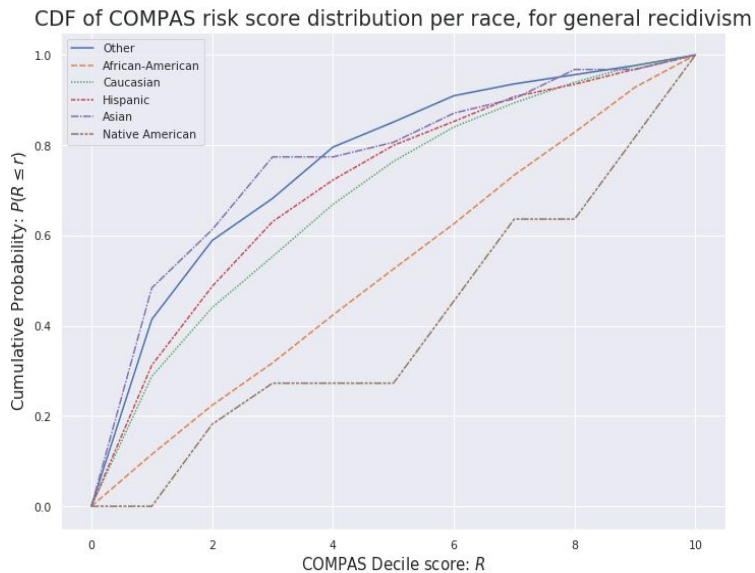
- Independence
- Separation
- Sufficiency

Demographic breakdown of dataset (TODO)

- Broward County, Florida, USA



Exploratory data analysis: distribution of scores



Bias mitigation results

- Equalised odds post-processing
 - ...
- Calibrated equalised odds post-processing
 - ...
- Reject option based classification
 - ...

Bias mitigation results

Equalised odds post-processing

- Unsatisfactory performance
 - Returned either the unchanged unfair model, or a fair random guessing classifier
- Possible explanations
 - There exist only trivial intersection points in the combined problem-space*
 - Model misspecification

Bias mitigation results

Calibrated equalised odds post-processing

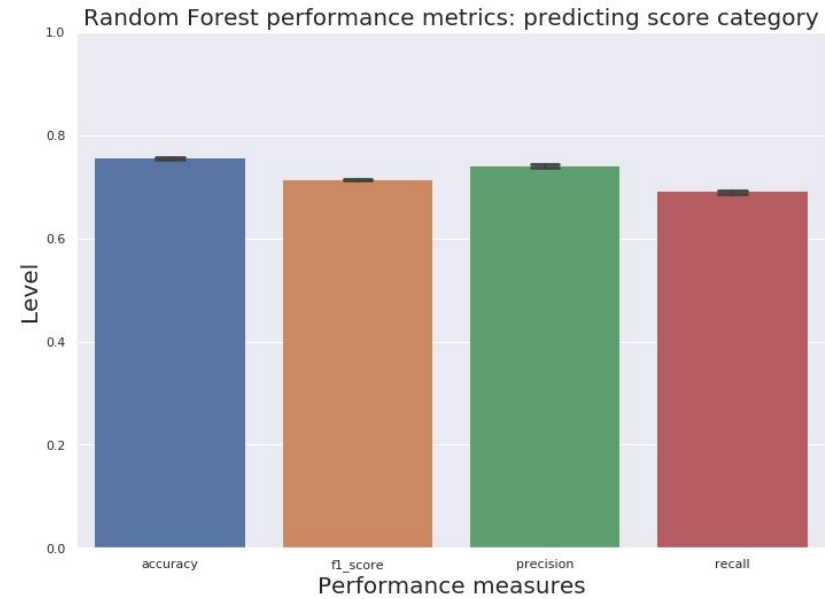
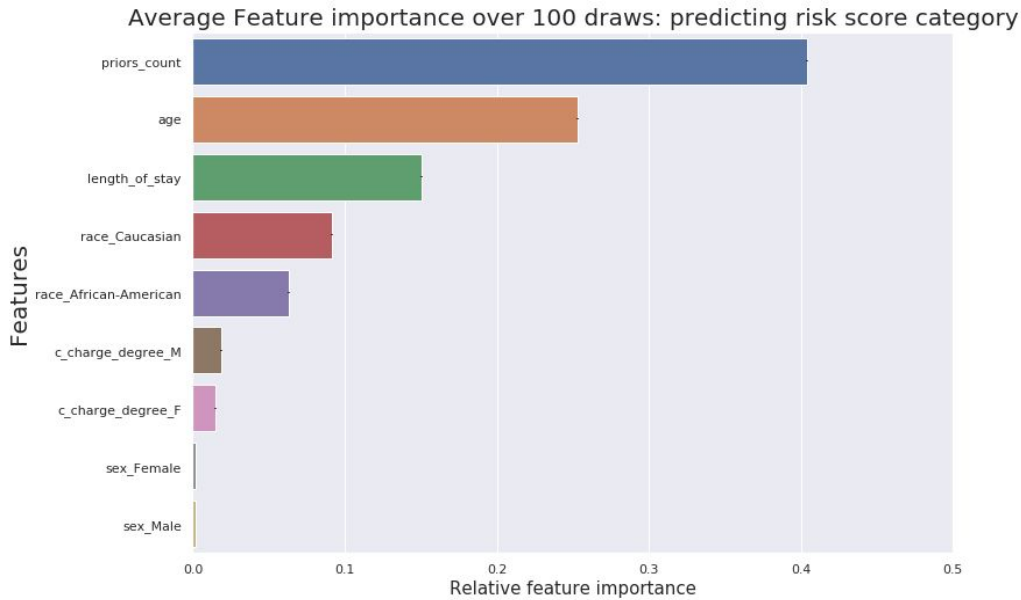
- Acceptable performance
 - Moderate decrease in prediction performance
 - Satisfies calibration
 - Fails to notably improve group-fairness
- Possible explanations
 - Incompatibility of calibration and error rate parity (Kleinberg's impossibility result)
 -

Bias mitigation results

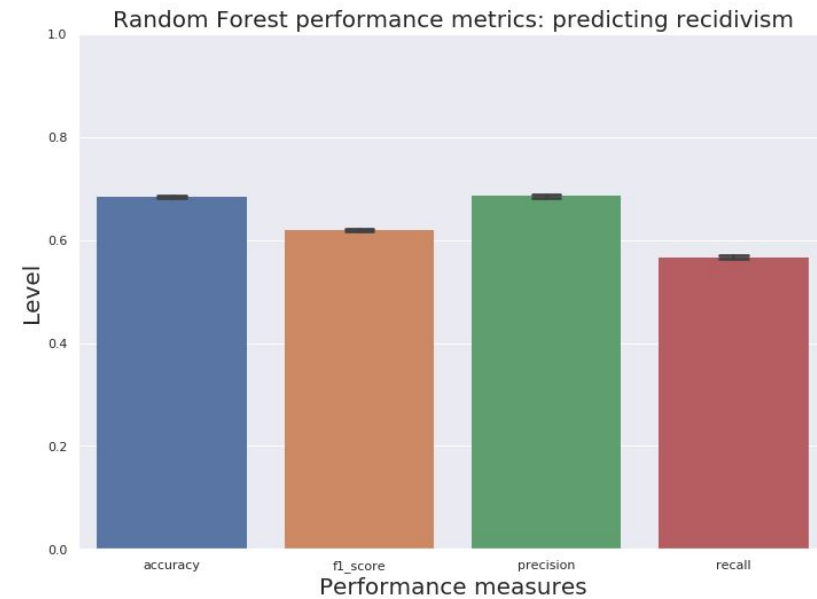
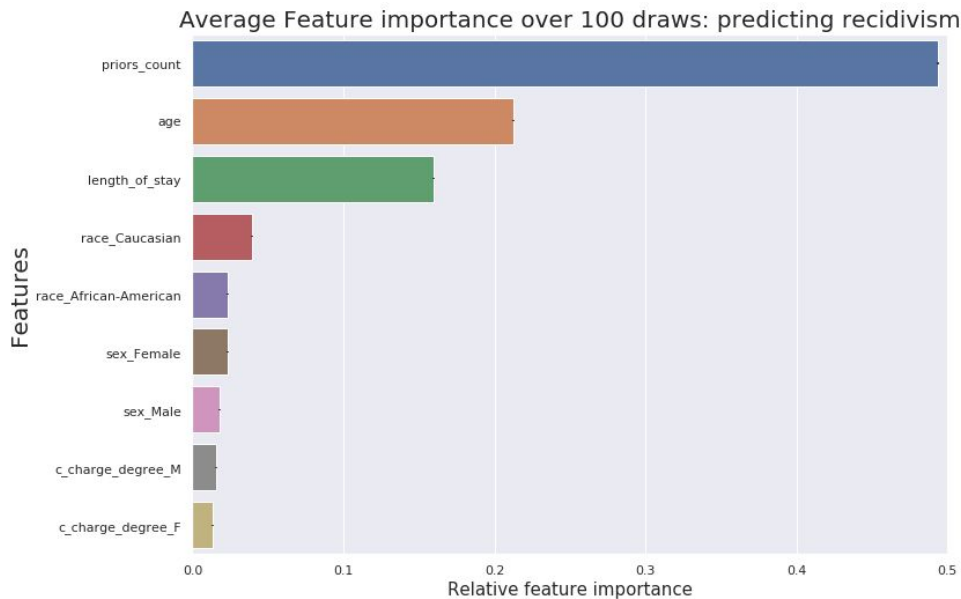
Reject option based classification

- Superior performance
 - ...
- Possible explanations
 -

Random forest: relative feature importance



Random forest: relative feature importance



Bias

“...an unwanted systematic error that places unprivileged groups at a disadvantage...”

-Bellamy et al. (2018)

Algorithmic fairness: conventions

- Positive → favourable label or outcome
 - Labelled as low risk, not committing a crime, etc.
- Negative → unfavourable label or outcome
 - Labelled as high risk, committing a crime (i.e., recidivating), etc.
- Error types
 - False positive: falsely being labelled as low risk, while re-committing a crime
 - False negative: falsely being labelled as high risk, while not becoming a recidivist

On the origin of biases

- Biased data
 - Unfair examples from an unjust past
- Features
 - Focussing on characteristics that are unfair towards a group
- Proxies
 - Using seemingly harmless variables as an approximation for protected attributes
- Sample size disparity
- Skewed sample

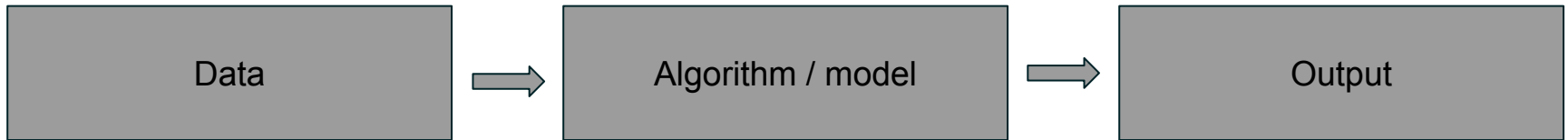
Strongly recommend the work of Solon Barocas (see references)

Achieving fairness

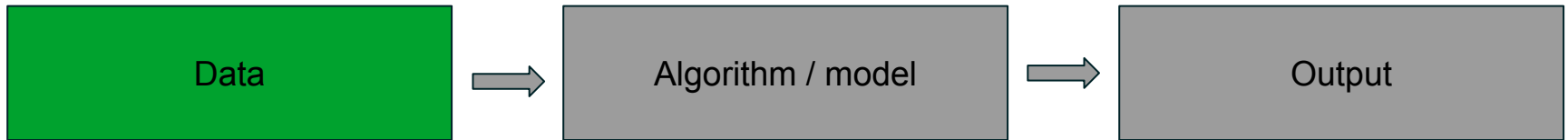
Three types of bias mitigation algorithms

- Pre-processing
- In-processing
- Post-processing

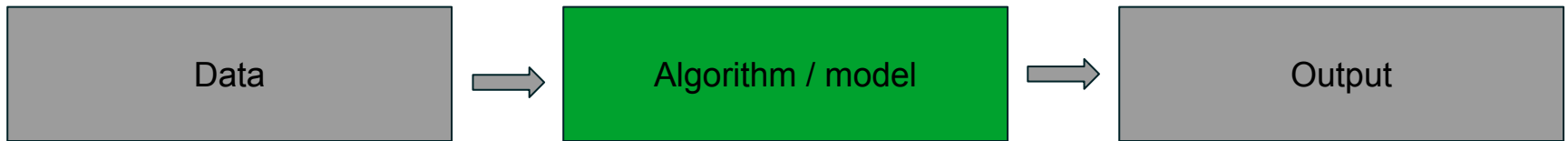
Achieving fairness: the machine learning pipeline



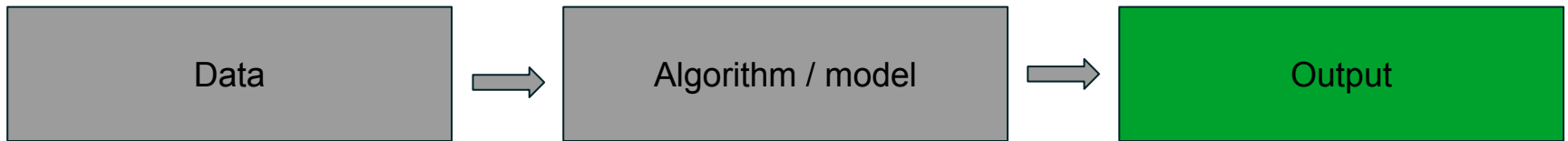
Achieving fairness: pre-processing



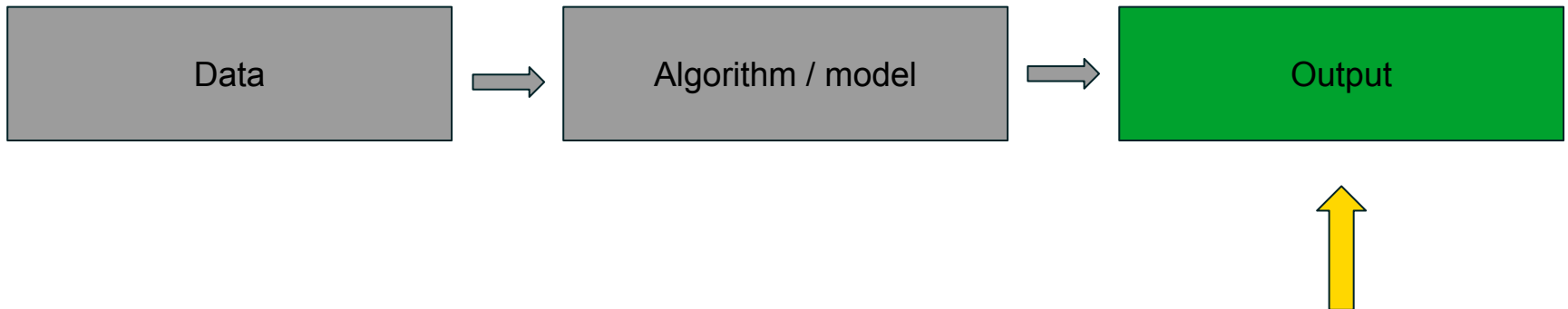
Achieving fairness: in-processing



Achieving fairness: post-processing



Achieving fairness: post-processing



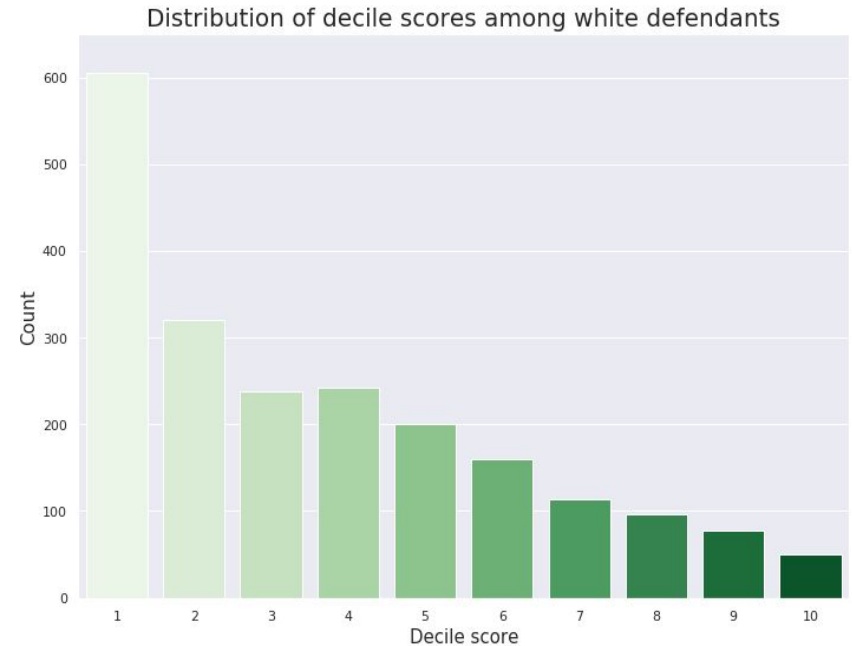
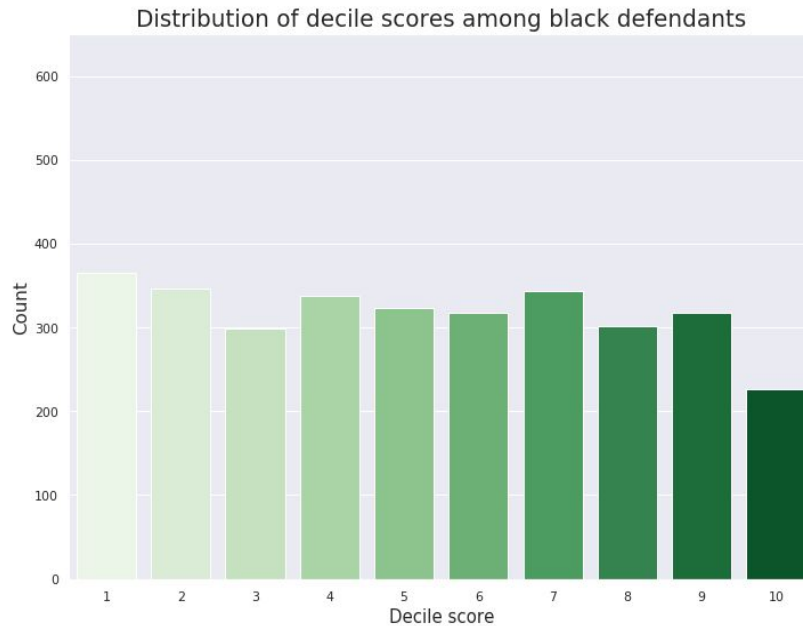
Algorithmic fairness: terminology

- Bias
 - An unwanted systematic error that systematically places certain groups at a disadvantage
- Bias mitigation algorithm
 - A set of procedures to make a decision-making model fair with respect to a certain characteristic (e.g., race, gender, age, etc.)

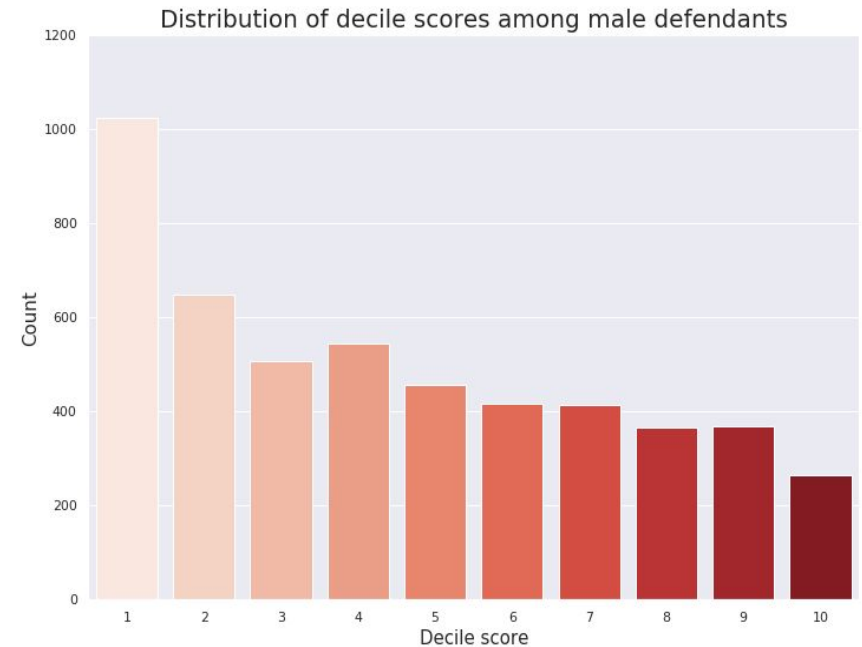
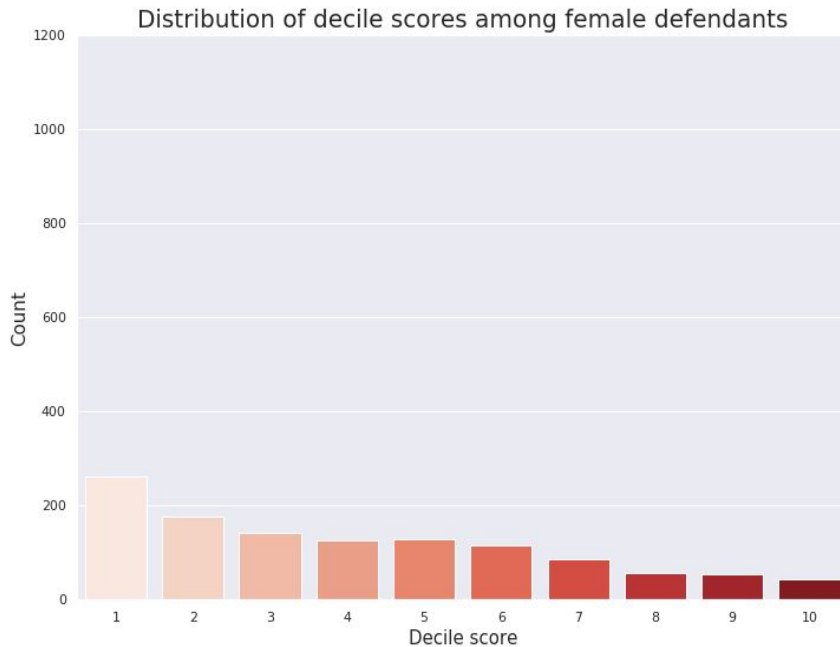
Econometric methods

- Random forest
 - Natural framework to determine relative variable importance
- Logistic regression
 - Provides insights about the direction of statistical relationships, after controlling for relevant variables

Exploratory data analysis: distribution of scores



Exploratory data analysis: distribution of scores



Exploratory data analysis: base rates

