NTNU | Norwegian University of Science and Technology

ETHICAL AI

THE ACCURACY, FAIRNESS AND LIMITS OF PREDICTING RECIDIVISM

MA8701 ARTICLE PRESENTATION

Florian Beiser, Helene Minge Olsen, Yaolin Ge

08 March 2021

Introduction - Ethical AI

AI Based Risk Assessment

- Recidivism Risk = Risk of Reoffending
- Predictions used in all levels of a prosecution
- Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)

Introduction - Predicting Recidivism

COMPAS

- Widely spread criminal risk assessment tool
- Created by Northpointe in the US
- No public information of its actual implementation

Input

- 137 features: age, gender, criminal history etc.
- No dynamic factors or variables directly linked to race

Output

- Scores from 1-10 defining the defendants risk level in general
- Categories high/low of recidivism risk

Introduction - Recidivism assessment

There are different paradigms to measure the quality of a model

Accuracy

- Overall accuracy equality
- AUC-ROC
- Sensitivity and bias

Fairness

Check for racial bias

- Accuracy per group
- false-positive per group

Introduction - Fairness

Growing discussion around COMPAS being biased against black defendants

Discoveries

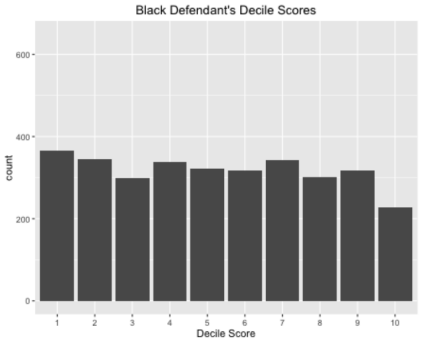
- Predictions in favor of white defendants
- Other aspects of the data correlated to race

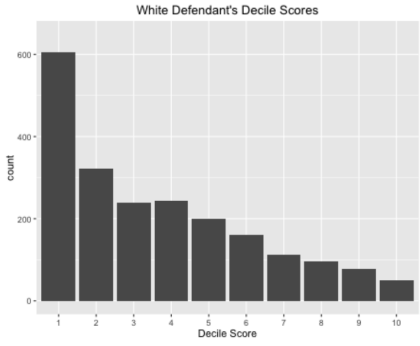
Who claimed this?

ProPublica

- Analyzed predictions on 7000 individuals
- Predictions are unreliable and racially biased

Introduction - Fairness. An alarming result

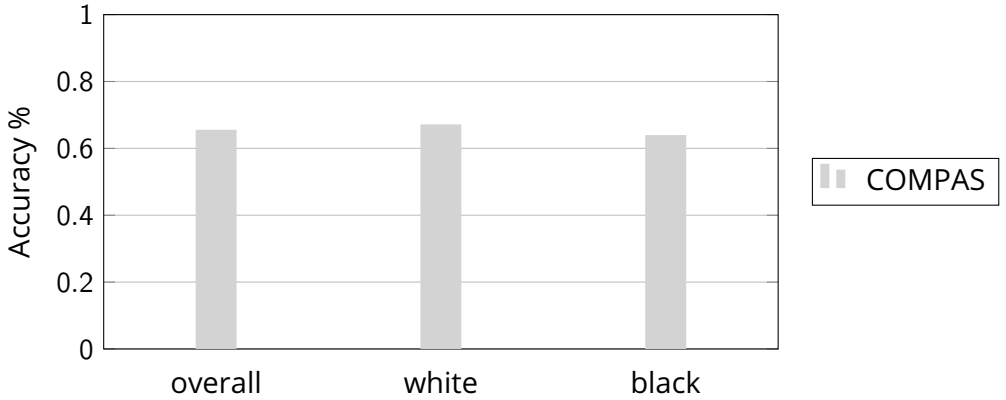




Introduction - Fairness. Not Racially biased?

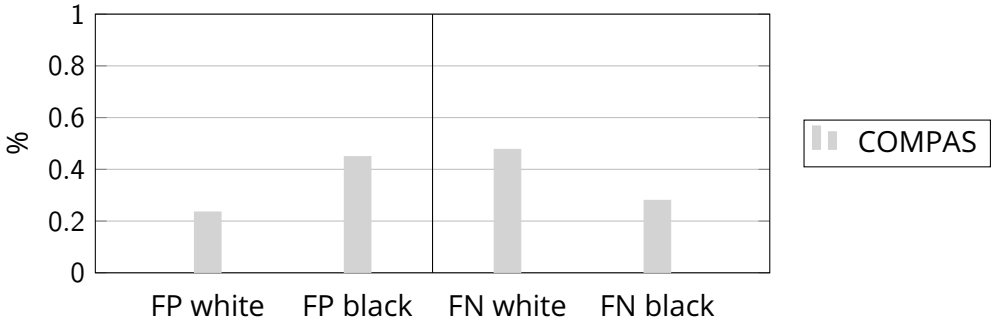
Pro Publica

Check for racial bias - Accuracy per group of defendants



Introduction - Fairness. Racially Biased

False Positives (FP): Labeled higher risk, but did not re-offend
False Negatives (FN): Labeled lower risk, yet did re-offend



Introduction - Contribution

Article addresses first a fundamental question.

Contribution I

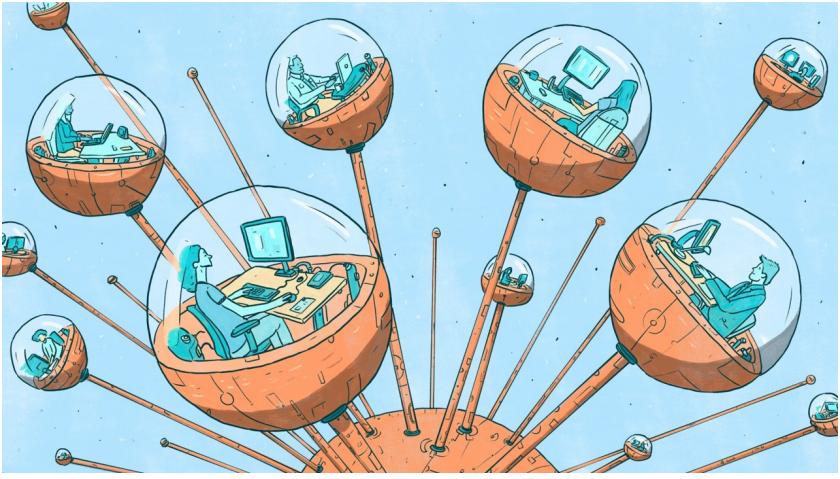
- Comparison of COMPAS assessment to human non-expert assessment

Article continues with model assessment of two algorithms:

Contribution II

- Comparison of COMPAS assessment to simple interpretable algorithmic assessment

Human Assessment - AMT(Amazon Mechanical Turk)



Human Assessment - Test setup

Test response without race

- Participants saw a short description of a defendant that included the defendant's sex, age, and previous criminal history, but **not** their race.
- 20 groups(each has 20 participants), 50 questions each (462 recruited discard 62)

Test response with race

- Participants saw a short description of a defendant that included the defendant's sex, age, and previous criminal history, and **include** their race.
- 20 groups(each has 20 participants), 50 questions each (449 recruited discard 49)

Human Assessment - Procedure

Step 1: read a paragraph of the description

The defendant is a [**SEX**] aged [**AGE**]. They have been charged with: [**CRIME CHARGE**]. This crime is classified as a [**CRIMINAL DEGREE**]. They have been convicted of [**NON-JUVENILE PRIOR COUNT**] prior crimes. They have [**JUVENILE-FELONY COUNT**] juvenile felony charges and [**JUVENILE-MISDEMEANOR COUNT**] juvenile misdemeanor charges on their record.

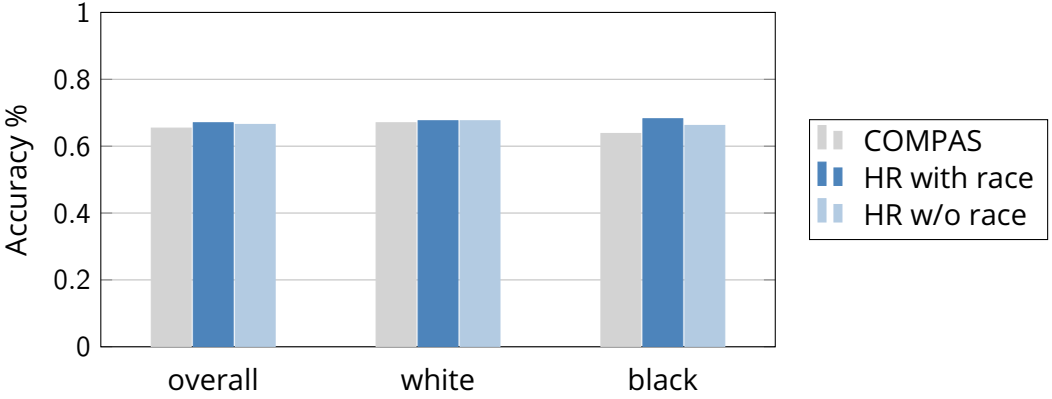
Step 2: respond to “yes” or “no”

Do you think this person will commit another crime within 2 years?

Notice: Feedback after each answer

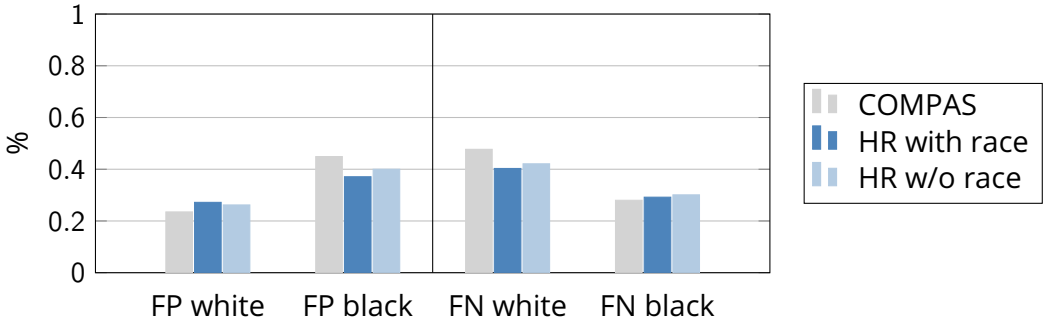
Human Assessment - Accuracy

Human vs COMPAS



Human Assessment - Fairness

Human vs COMPAS



Human Assessment - Conclusion

Discussion

- Not enough white participants in the participants
- People who answer may not represent all walks of life
- Expert may even increase more of the accuracy

Outlook

- Will weighting of features improve assessment quality?
- Will dynamic risk factors improve assessment quality?
- Will expert experience improve assessment quality?

Algorithmic Assessment - Learning Approaches

Research Question

Can the accuracy of COMPAS be achieved by an easier classifier model?

Methods

- Logistic regression (LR)
- Non-linear support vector machine (NL-SVM)

Dataset

- 7214 cases
- Validation by bootstrapping

Algorithmic Assessment - Logistic Regression

Method (Reminder)

Linear ansatz for log-odds. Log-odds relate to probability via logistic function.

Feature Selection

- 7 features

```
sex
age
crime charge
criminal degree
non-juvenile prior count
juvenile-felony count
juvenile-misdemeanor count
```

- 2 features

```
age
convicts prior count
```

Algorithmic Assessment - Support Vector Machines

Linear SVM

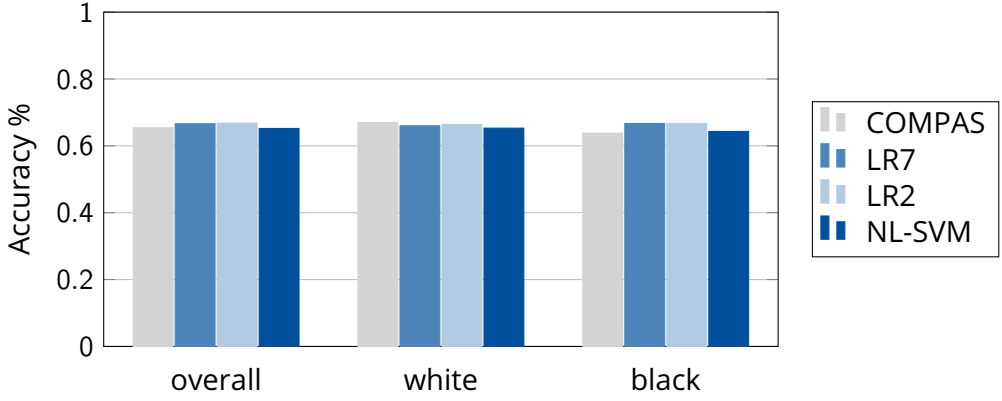
- Supervised learning technique by Vapnik et al.
- Separates data into 2 classes
- Constructs hyperplane s.t. margin between classes maximised

Non-linear SVM

- Transforming Cartesian covariate-space using a kernel function
- Radial basis kernel $k(x, y) = \exp(-\theta \|x - y\|^2)$ with $\theta > 0$
- Allowing non-linear separators

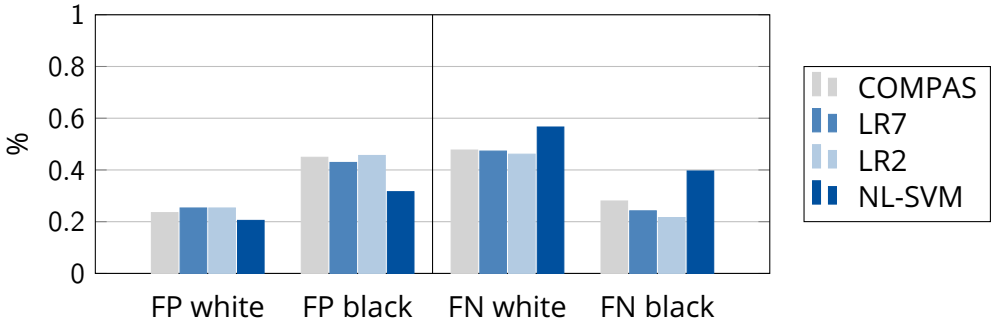
Algorithmic Assessment - Accuracy

Averaged over 1000 bootstrap samples on 80/20 splits



Algorithmic Assessment - Fairness

Averaged over 1000 bootstrap samples on 80/20 splits



Algorithmic Assessment - Conclusion

- Simple and interpretable models perform with same accuracy as COMPAS
- Non-linear methods perform similar to linear methods

Discussion - Comments to Results

COMPAS (and other commercial recidivism software) is not any more reliable than non-expert or simple-model assessments

- Uses a particular measure of fairness...
  There are many measures of fairness and it is impossible to satisfy some combination of these simultaneously
- Uses human involving data....
  Human biases and unfairness leak into the data used to train ML models
- What are the cases when the different methods disagree?
- With the accuracy unable to overgo 60% for all methods addressed here, is the data simply inseparable?

Discussion - Comments of Recidivism Assessment

- To what extent do the predictions affect the judges decision?
- What is the best measure to compare fairness?
- Are Machine Learning Algorithms just not suitable to be used on the grounds of social ethics and norms?

Discussion - References

- Julia Dressel and Hany Farid
  The accuracy, fairness, and limits of predicting recidivism
  Science advances, 2018
- Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner
  Machine Bias
  ProPublica, 2016
- Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin
  How We Analyzed the COMPAS Recidivism Algorithm
  ProPublica, 2016

Thank you for your attention