Can Explainable AI Techniques Explain Unfairness?

Kiana Alikhademi, Emma Drobina, Brianna Richardson

Objectives

- Define a rubric for evaluating XAI tools in terms of their use in evaluating fairness
- Apply this rubric to three case studies: COMPAS, Hate Speech, Patient Survival rates

Contribution

- Developed holistic fairness rubrics with respect to the access and capabilities of XAI
- Examine the state-of-the art fairness tools with respects to our comprehensive rubrics
- Outline the gaps withins this area

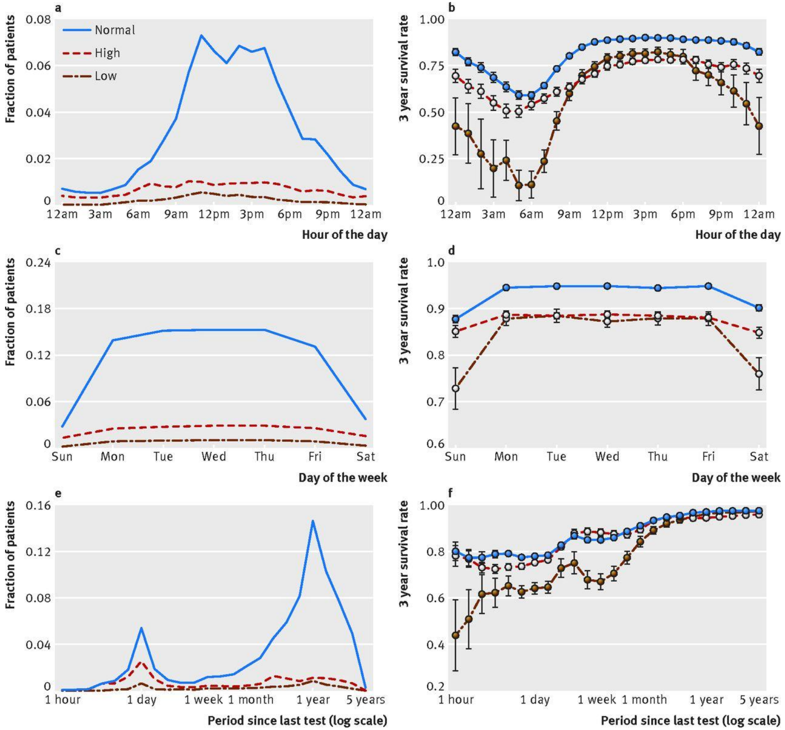
Defining Fairness

Used the literature to distinguish major recurring issues of fairness involved in machine learning.

Distinguished four major areas of need:

- Biased data & bias that are reflected in the data
- Pre-processing procedures
- Selection and optimization of ML models
- Perception of ML results

Health Data[3]



Patients had decreased survival rate when:

- Tests run early in the morning, or
- Tests ordered on the weekend, or
- Consecutive tests ordered in shorter period of time

Healthcare processes could be a better predictor than actual lab values.

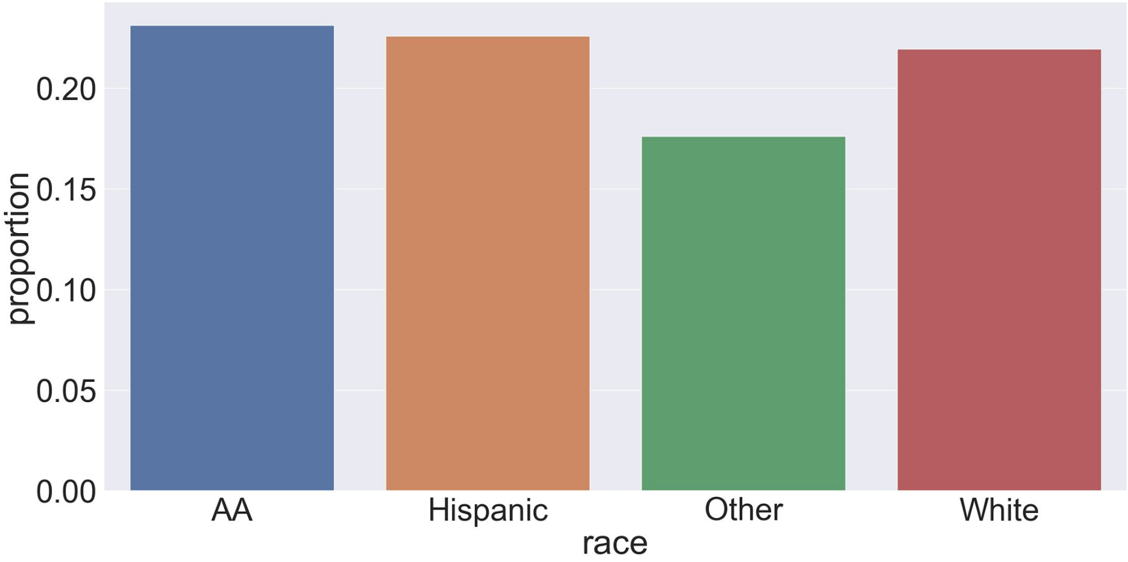
3. Agniel, D., Kohane, I. S., and Weber, G. M.Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. BMJ 361(2018).

Twitter Data[2]

Based on previous research indicating racial bias in abusive language detection

Unique because we trained on data that did not include race & tested on data that did include race

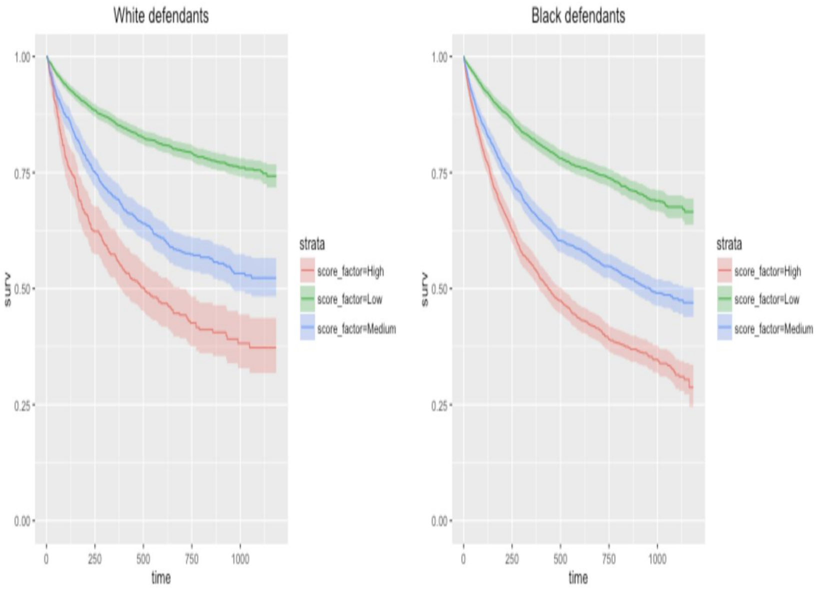
Only slight differences in how our model classified different races



Proportion of derogatory tweets by race (for random forest classifier)

2. https://dataverse.mpi-sws.org/dataset.xhtml?persistentId=doi:10.5072/FK2/ZDTEMN

COMPAS Data[1]




```
Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -2.24274    0.11326 -19.802  < 2e-16 ***
gender_factorFemale         -0.72890    0.12666  -5.755 8.66e-09 ***
age_factorGreater than 45   -1.74208    0.18415  -9.460  < 2e-16 ***
age_factorLess than 25       3.14591    0.11541  27.259  < 2e-16 ***
race_factorAfrican-American  0.65893    0.10815   6.093 1.11e-09 ***
race_factorAsian            -0.98521    0.70537  -1.397   0.1625
race_factorHispanic         -0.06416    0.19133  -0.335   0.7374
race_factorNative American   0.44793    1.03546   0.433   0.6653
race_factorOther            -0.20543    0.22464  -0.914   0.3605
priors_count                 0.13764    0.01161  11.854  < 2e-16 ***
crime_factorM               -0.16367    0.09807  -1.669   0.0951 .
two_year_recid               0.93448    0.11527   8.107 5.20e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

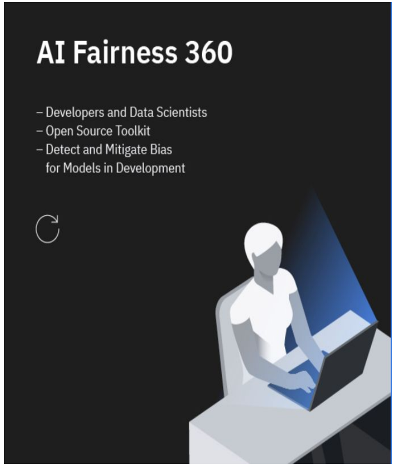
Black defendants do recidivate at higher rates according to race specific Kaplan Meier plots.

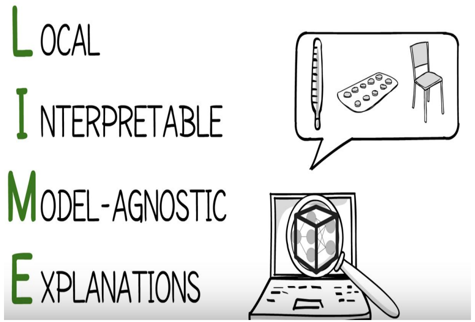
The violent score overpredicts recidivism for black defendants by 77.3% compared to white defendants.

1. https://github.com/propublica/compas-analysis/

XAI Tools

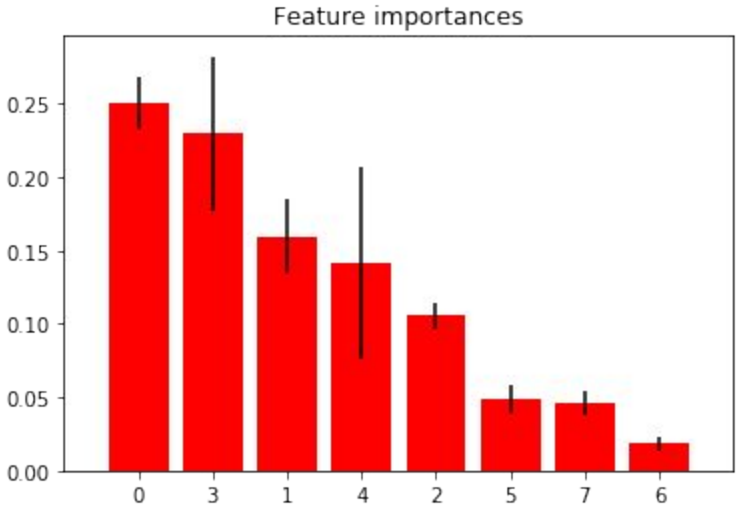
- Two explainable models
  - Random Forests
  - Logistic Regression
- Explainable AI tools:
  - LIME [2][video]
  - AI Fairness 360(AIF 360) [1][video]





XAI Model Evaluation - Random Forests

```
|--- age <= 37.50
|   |--- c_charge_degree <= 1.50
|   |   |--- score_text <= 0.50
|   |   |   |--- priors_count <= 6.50
|   |   |   |   |--- days_b_screening_arrest <= 21.00
|   |   |   |   |   |--- sex <= 0.50
|   |   |   |   |   |   |--- age <= 21.50
|   |   |   |   |   |   |   |--- days_b_screening_arrest <= -0.50
|   |   |   |   |   |   |   |   |--- days_b_screening_arrest <= -12.50
|   |   |   |   |   |   |   |   |   |--- class: 2.0
|   |   |   |   |   |   |   |   |--- days_b_screening_arrest >  -12.50
|   |   |   |   |   |   |   |   |   |--- race <= 1.00
|   |   |   |   |   |   |   |   |   |   |--- c_charge_degree <= 0.50
|   |   |   |   |   |   |   |   |   |   |   |--- truncated branch of depth 6
|   |   |   |   |   |   |   |   |   |   |--- c_charge_degree >  0.50
|   |   |   |   |   |   |   |   |   |   |   |--- class: 2.0
|   |   |   |   |   |   |   |   |   |--- race >  1.00
|   |   |   |   |   |   |   |   |   |   |--- class: 1.0
|   |   |   |   |   |   |   |--- days_b_screening_arrest >  -0.50
|   |   |   |   |   |   |   |   |--- race <= 1.00
|   |   |   |   |   |   |   |   |   |--- class: 1.0
|   |   |   |   |   |   |   |   |--- race >  1.00
|   |   |   |   |   |   |   |   |   |--- class: 2.0
|   |   |   |   |   |   |--- age >  21.50
|   |   |   |   |   |   |   |--- race <= 1.00
|   |   |   |   |   |   |   |   |--- age <= 22.50
|   |   |   |   |   |   |   |   |   |--- priors_count <= 3.50
|   |   |   |   |   |   |   |   |   |   |--- class: 1.0
|   |   |   |   |   |   |   |   |   |--- priors_count >  3.50
|   |   |   |   |   |   |   |   |   |   |--- days_b_screening_arrest <= -6.50
|   |   |   |   |   |   |   |   |   |   |   |--- class: 1.0
|   |   |   |   |   |   |   |   |   |   |--- days_b_screening_arrest >  -6.50
|   |   |   |   |   |   |   |   |   |   |   |--- class: 2.0
|   |   |   |   |   |   |   |   |--- age >  22.50
|   |   |   |   |   |   |   |   |   |--- priors_count <= 3.50
|   |   |   |   |   |   |   |   |   |   |--- age <= 24.50
|   |   |   |   |   |   |   |   |   |   |   |--- truncated branch of depth 2
|   |   |   |   |   |   |   |   |   |   |--- age >  24.50
|   |   |   |   |   |   |   |   |   |   |   |--- truncated branch of depth 5
|   |   |   |   |   |   |   |   |   |--- priors_count >  3.50
|   |   |   |   |   |   |   |   |   |   |--- days_b_screening_arrest <= -0.50
|   |   |   |   |   |   |   |   |   |   |   |--- truncated branch of depth 5
|   |   |   |   |   |   |   |   |   |   |--- days_b_screening_arrest >  -0.50
```



Easy to implement

Needs additional processing to come up with presentation of explanations and feature importance

XAI Model Evaluation - LIME

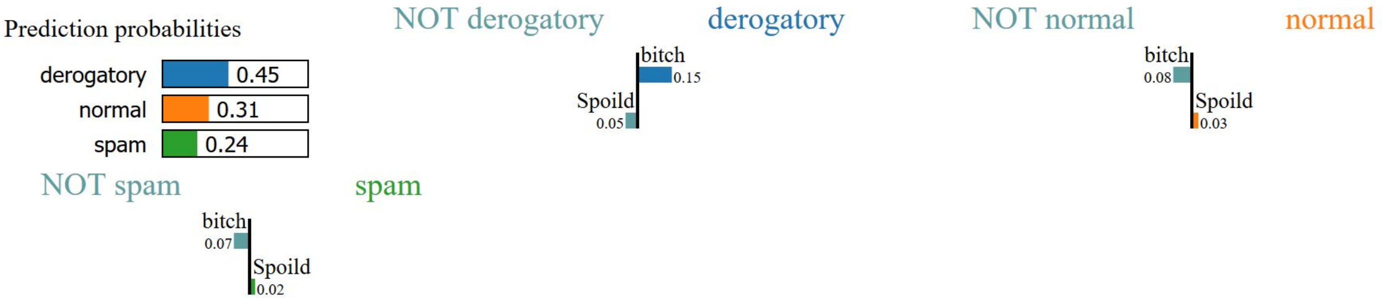
```
Explanation for class derogatory
('bitch', 0.14928979336061113)
('Spoild', -0.05243819052375206)

Explanation for class normal
('bitch', -0.08101678414794018)
('Spoild', 0.028457227162992704)

Explanation for class spam
('bitch', -0.06827297618255586)
('Spoild', 0.023980951758601272)
```

Explains one sample at a time

To extract relevant fairness information, more processing needed!

Its lack of global awareness is a weakness, but it gives detailed information on why a sample belongs to a given class (output is similar to logistic regression)

```
In [212]:  1  exp.show_in_notebook(text=False)
```

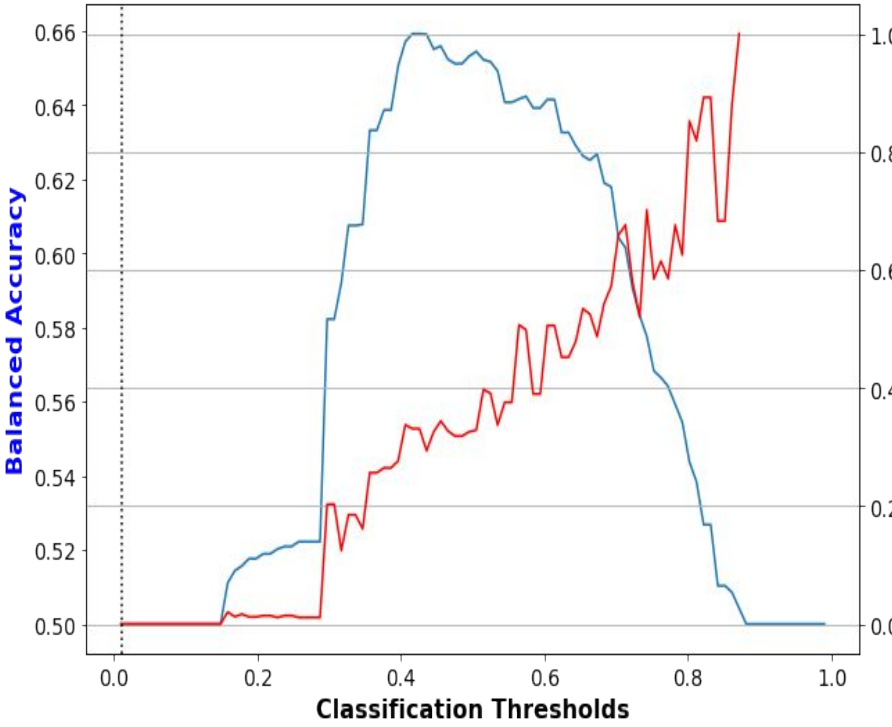


XAI Model Evaluation - AIF360

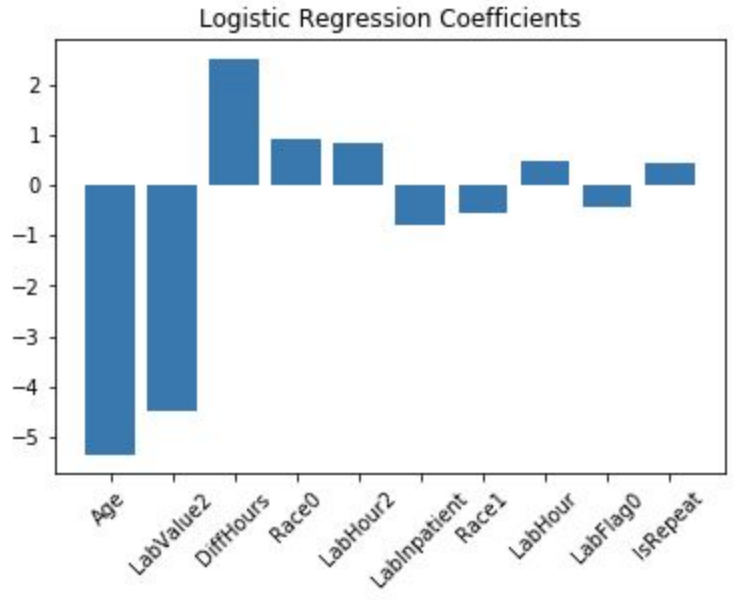
Considers the fairness and parity between user-determined unprivileged and privileged groups.

Looking into explanations across different subgroups.

Detects imbalanced data and bias



XAI Model Evaluation - Logistic Regression



Easy to understand

For fairness, requires additional processing!

- What scores in each category associated to survival rates?
- Are there disproportional representations of data?

Completed Rubric

	Random Forest	LIME	Fairness 360	Ad-hoc explainability
Model used	Random Forest	Deep learning		Logistic Regression
Issues with Biased Data				
Imbalanced data	0	0	1	0
Influential variable identification	1	1	0	1
Preprocessing issues	0	0	1	0
Sensitive attributes	0	0	1	0
Issues involved in Machine Learning Models				
Model-Specific influences	0	0	0	0
Accuracy equity	0	1	1	0
Issues involved with XAI results				
Target audience	0	0	0	0
Presentation of explanations	1	1	0	1

Conclusion/Future Work

- Conclusion:
  - While current XAI tools have the data and the model, they still are lacking when it comes to a thorough investigation of issues involved in the results.
- Future Work:
  - Incorporate WHATIF/AI Explainability 360
  - Develop a more detailed way to quantify the level of explainability each of these tools yield
  - Consider more sensitive attributes and how we can provide global explanations across all of them.

Citations

1. Bellamy, R.K., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A. and Nagar, S., 2018. AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv preprint arXiv:1810.01943.

2. Tulio Ribeiro, M., Singh, S., & Guestrin, C. (2016). " Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv preprint arXiv:1602.04938*.

3. Agniel, D., Kohane, I. S., and Weber, G. M.Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. BMJ 361(2018).