Fairness Project

Creating a Tool

Presented by: Jonas Schäfer, Marius Wawerek

FU Berlin
Supervisor:
Eirini Ntoutsi, Siamak Ghodsi

Main functionality - Our Focus

Most important point:

-> **evaluation of adaptability of algorithms to distribution shifts/context shifts**

to make it interesting:

-> out of distribution testing (**new algorithms**)

Scope of the project

We do this project in “part-time”

We are not sure if we can polish this paper up to “conference standard”

Should we ‘only’ focus on one key aspect? -> No jack of all trades

How likely are we to attend a conference? -> What happens if we do not make it?

(paper about context shifts?)

Scope of the project

We do this project in “part-time”

We are not sure if we can polish this paper up to “conference standard”

Should we ‘only’ focus on one key aspect? -> No jack of all trades

How likely are we to attend a conference? -> What happens if we do not make it?

(paper about context shifts?)

Scope of the project

We do this project in “part-time”

We are not sure if we can polish this paper up to “conference standard”

Should we ‘only’ focus on one key aspect? -> No jack of all trades

How likely are we to attend a conference? -> What happens if we do not make it?

(paper about context shifts?)

Scope of the project

We do this project in “part-time”

We are not sure if we can polish this paper up to “conference standard”

Should we ‘only’ focus on one key aspect? -> No jack of all trades

How likely are we to attend a conference? -> What happens if we do not make it?

(paper about context shifts?)

Main functionality - Our Focus

Main Questions:

-Should we allow custom datasets?

-Which Datasets can support such shifts?

-Should we allow different evaluation methods/metrics?

Main functionality - Our Focus

What is a distribution shift precisely?

-> Spatial -> “From **where** is the data?”

-> Temporal -> “From **when** is the data?”

-> Other Shifts?

Keywords: **covariate shift**, concept shift

Functionality

- Standalone, Notebook or ‘Library’?
- Step by Step selection of Parameters -> AIF360 Demo
- Evaluation/Quantification of context shifts (KL-Divergence)
- How much was the model affected by the shifts? How to evaluate?
    - Absolute and relative difference in **which** metrics (e.g. 90% acc vs 70% acc)
    - Runtime?
- Should we have prepared/precomputed models to compare against?

Functionality

- Standalone, Notebook or ‘Library’?
- Step by Step selection of Parameters -> AIF360 Demo
- Evaluation/Quantification of context shifts (KL-Divergence)
- How much was the model affected by the shifts? How to evaluate?
    - Absolute and relative difference in **which** metrics (e.g. 90% acc vs 70% acc)
    - Runtime?
- Should we have prepared/precomputed models to compare against?

Functionality

- Standalone, Notebook or ‘Library’?
- Step by Step selection of Parameters -> AIF360 Demo
- Evaluation/Quantification of context shifts (KL-Divergence)
- How much was the model affected by the shifts? How to evaluate?
  - Absolute and relative difference in **which** metrics (e.g. 90% acc vs 70% acc)
  - Runtime?
- Should we have prepared/precomputed models to compare against?

Functionality

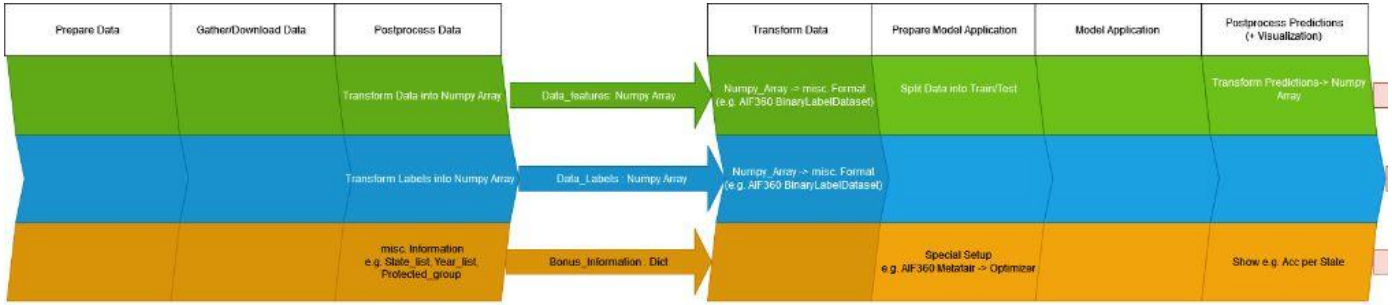
- Standalone, Notebook or ‘Library’?
- Step by Step selection of Parameters -> AIF360 Demo
- Evaluation/Quantification of context shifts (KL-Divergence)
- How much was the model affected by the shifts? How to evaluate?
    - Absolute and relative difference in **which** metrics (e.g. 90% acc vs 70% acc)
    - Runtime?
- Should we have prepared/precomputed models to compare against?

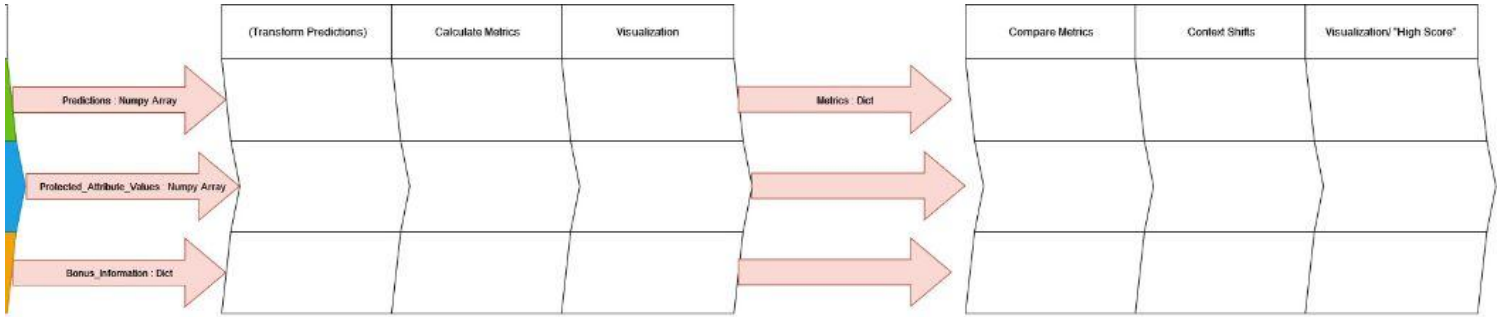
Functionality

- Standalone, Notebook or ‘Library’?
- Step by Step selection of Parameters -> AIF360 Demo
- Evaluation/Quantification of context shifts (KL-Divergence)
- How much was the model affected by the shifts? How to evaluate?
    - Absolute and relative difference in **which** metrics (e.g. 90% acc vs 70% acc)
    - Runtime?
- Should we have prepared/precomputed models to compare against?

Progress so far

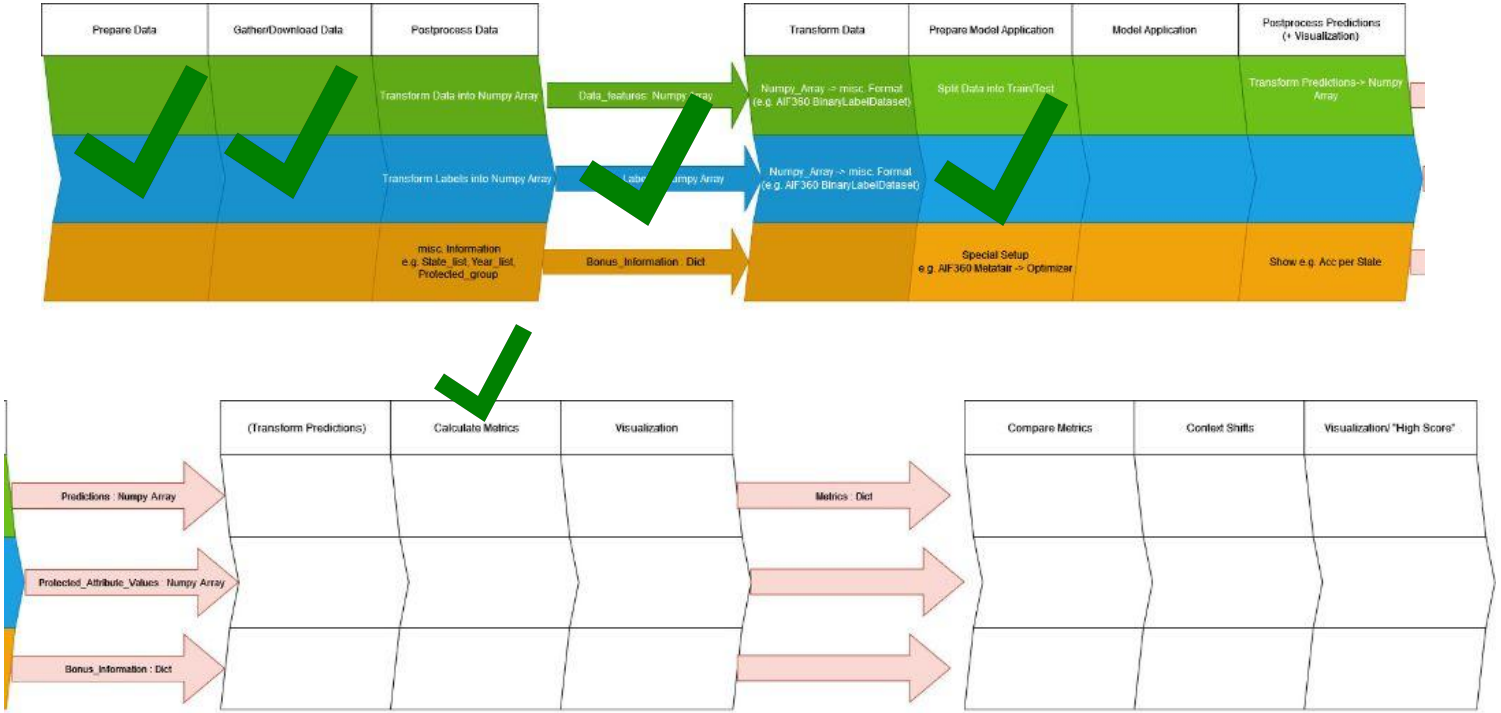
- Pipeline:





Progress so far

- Pipeline:



Algorithms for developing fair models

Preprocessing

- Reweighing
- Disparate Impact Removing
- “Learning fair representations”
- “Optimized preprocessing”

Inprocessing

- **Adversarial Debiasing**
- **“Meta Fair Classifier”**
- **Prejudice Removing**
- “Gerry Fair Classifier”

Postprocessing

- Equalized Odds Postprocessing
- Reject Option Classification

Source: https://aif360.readthedocs.io/en/stable/modules/algorithms.html

Metrics for evaluating fairness

Three types of fairness:

Group

- (Conditional) Demographic Parity
- Error Parity
  - Equal Accuracy
  - **ABROCA**
  - Equality of Odds
  - **Disparate Impact**
  - Predictive Parity

Individual

- FTU/Blindness
- Fairness Through Awareness

Causality-based

Observational

In total, we collected around **70 different metrics** for assessing classification performance and fairness.

Source: https://arxiv.org/pdf/2106.00467.pdf

General Dataset

ACI Income (also known as New “Adult Dataset”)

-> Based on US Census (≈1-2% of USA Pop.)

Time scale: 2014 -2018

Features: 10 (e.g. Occupation, Worktime per Week, Race, ...)

Prediction goal: earn more than 50k? -> Yes / No

Datasets

-Northern States

-Southern States

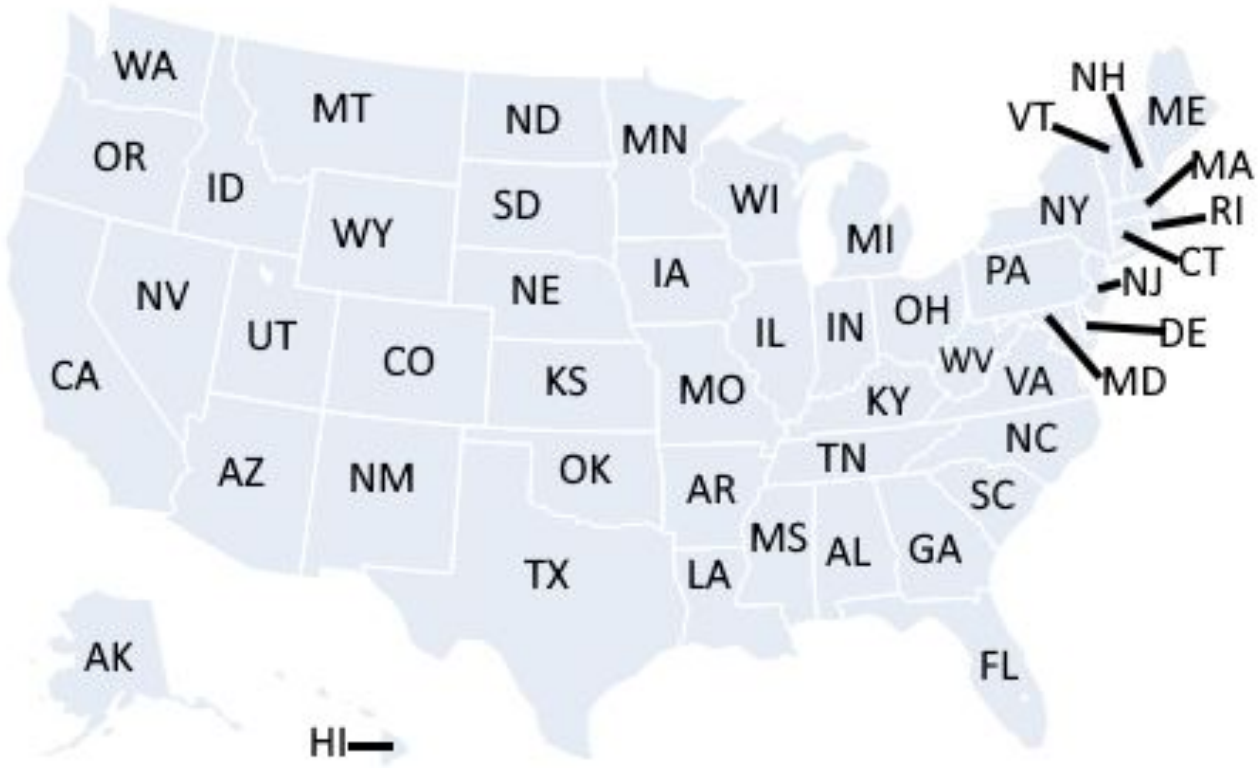
-East Coast

-West Coast

-None Coast

-Urban States

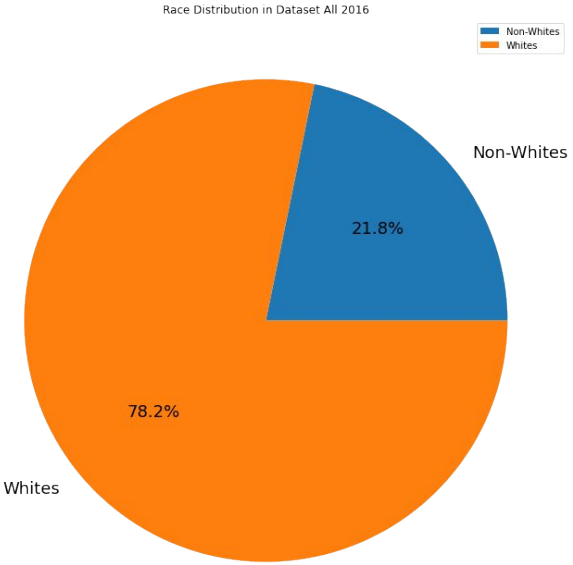
-Rural States



Sample Size: 1.6 Mio (2016)

mean class: 34%

Data Overview:

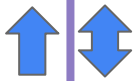


Datasets	Size (in k)	Mean Label (in %)
All	1600	34
Northern	676	35
Southern	915	33
East Coast	526	37
West Coast	260	38
None Coast	831	31
Urban States	898	37
Rural States	719	31

Results - How to interpret

Performance:
Accuracy

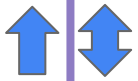
Range: [0, 1]



0 is best

Fairness:
ABROCA

Range: [0, 1]

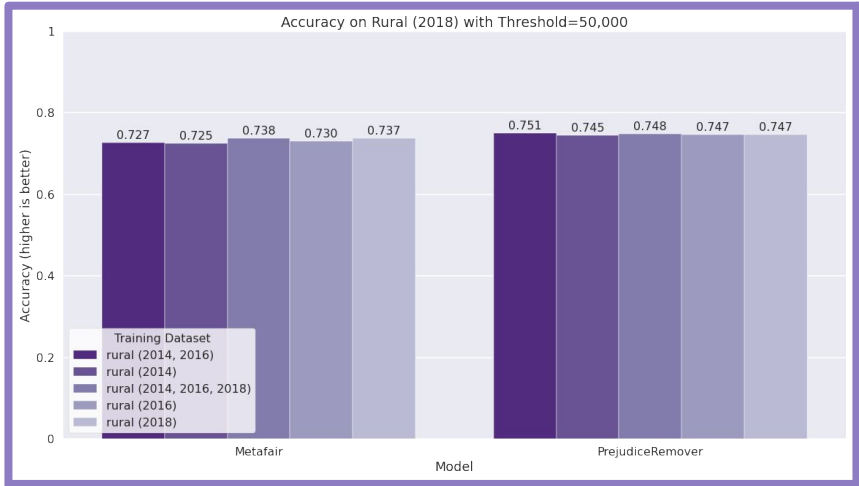


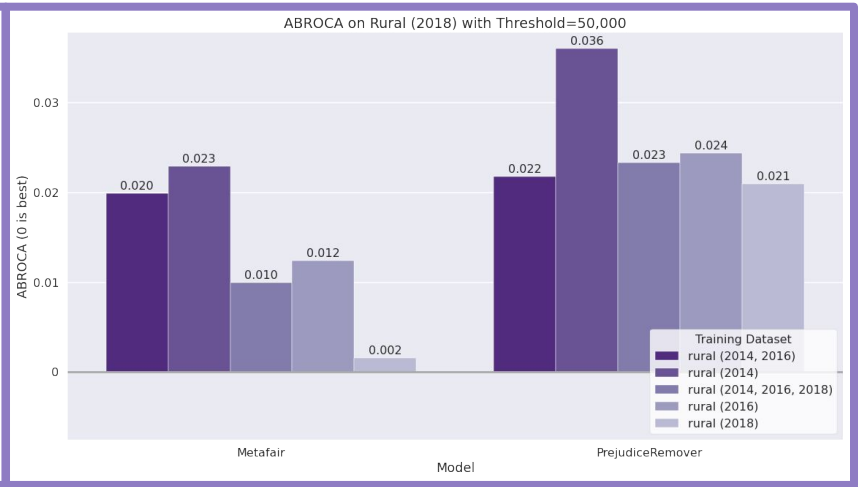
1 is best

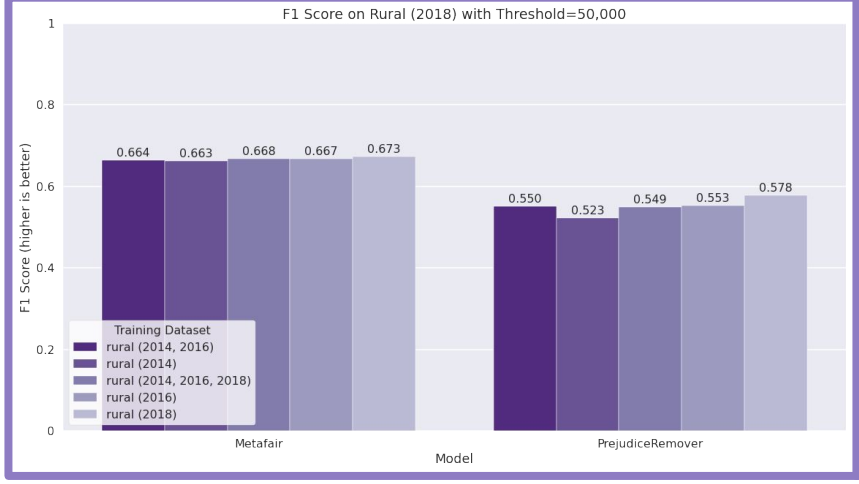
Performance:
F1 Score

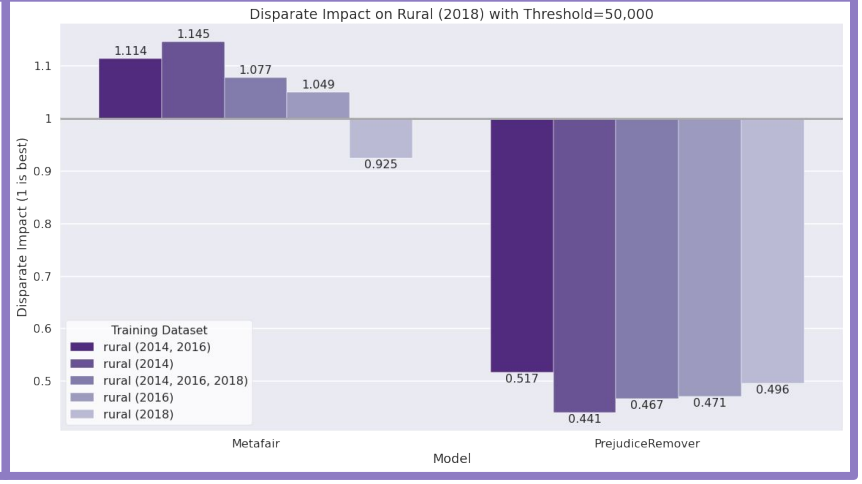
Fairness:
Disparate Impact

Results - Temporal Context

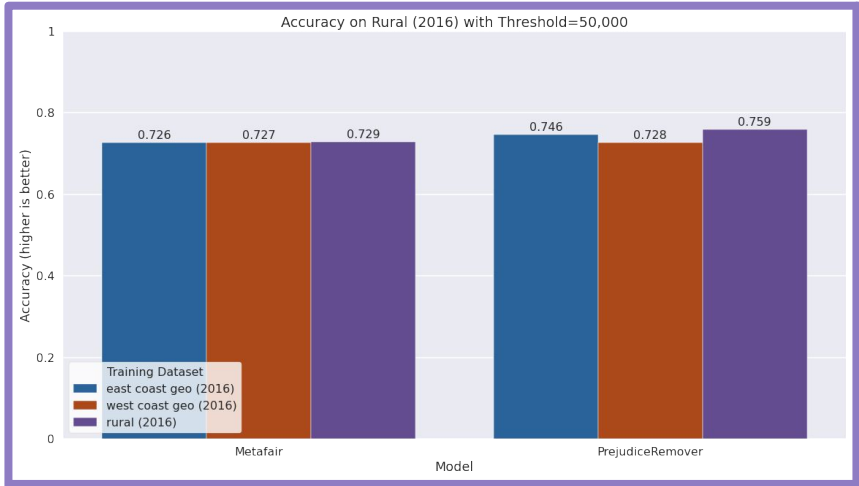


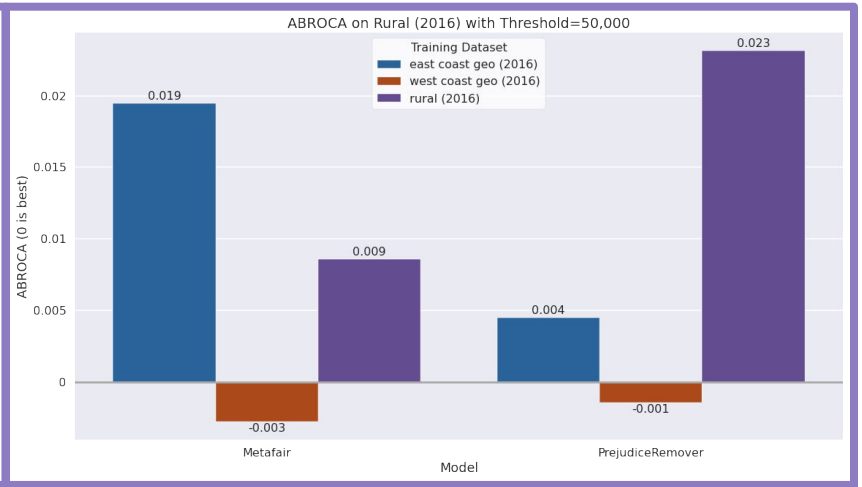


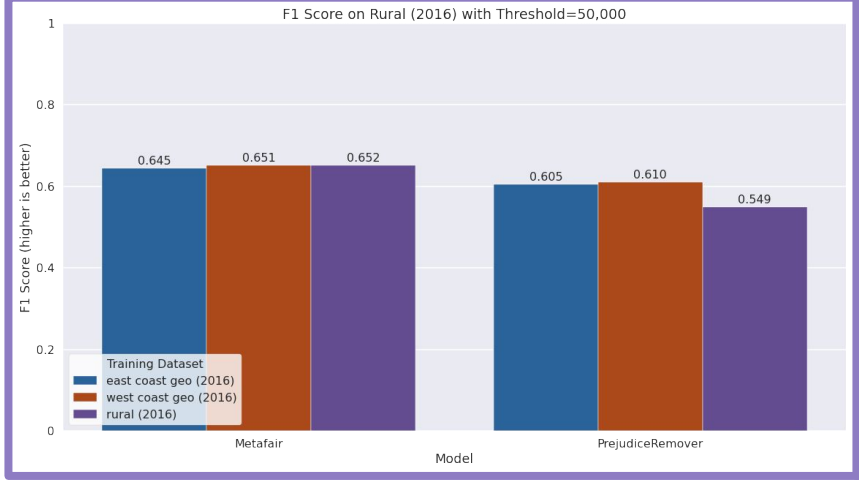


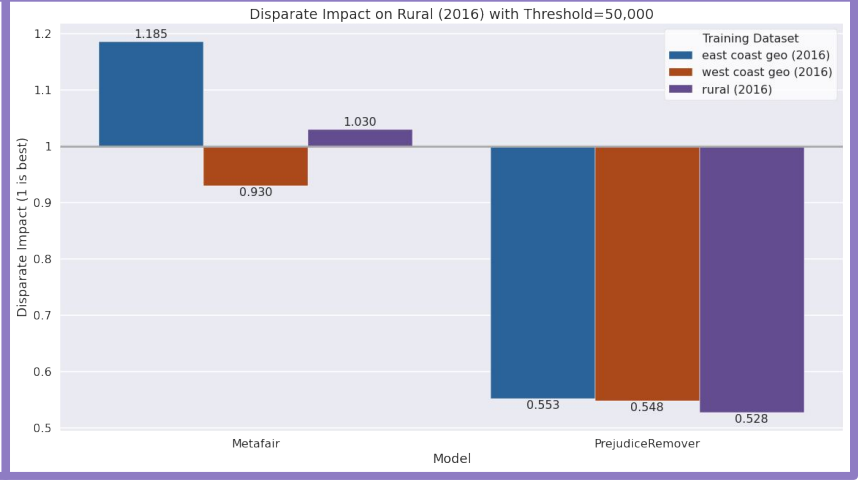


Results - Spatial Context

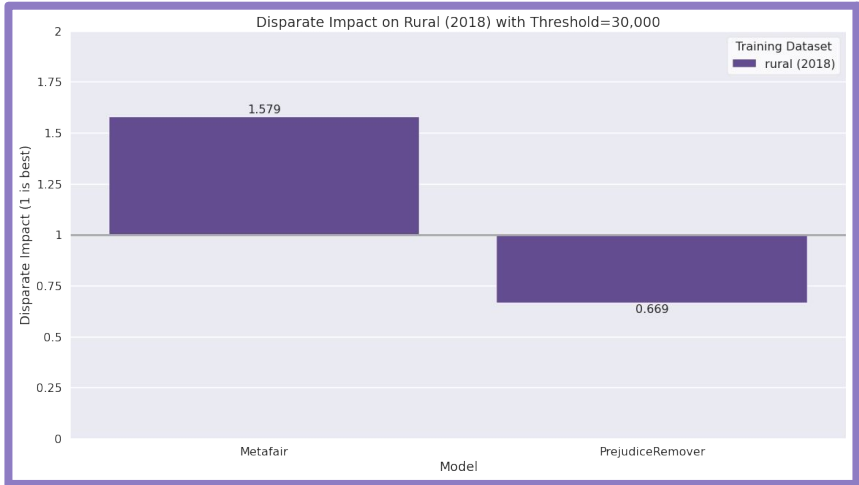


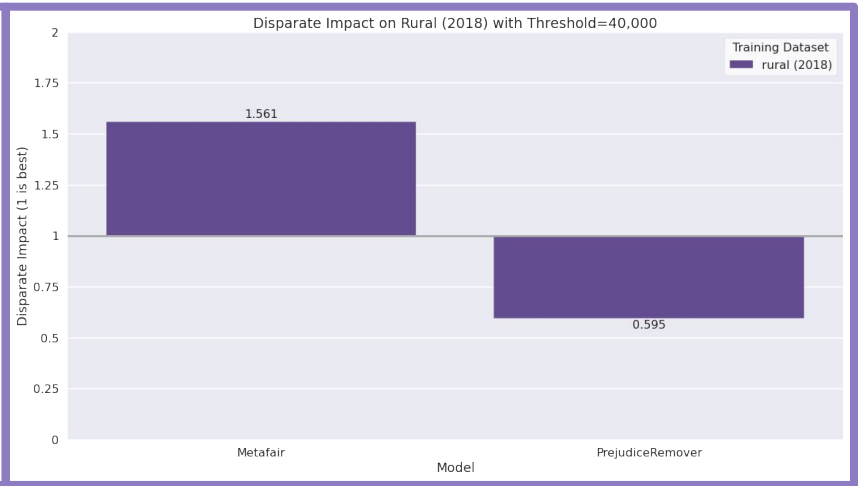


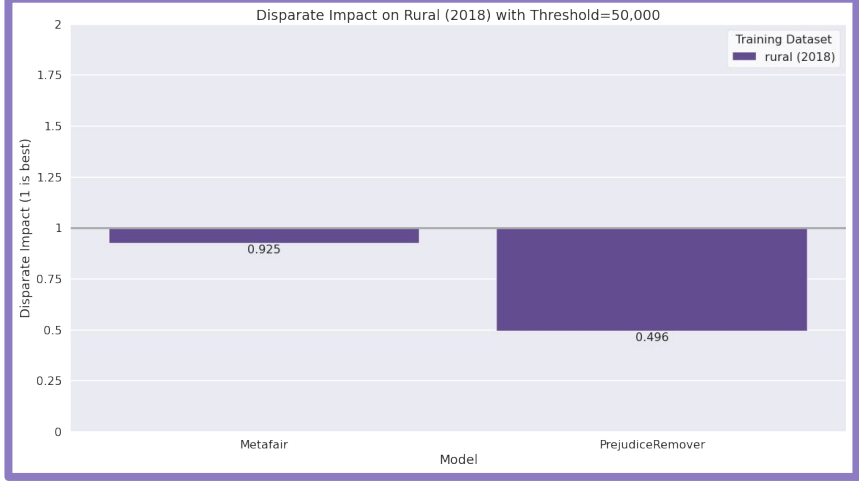


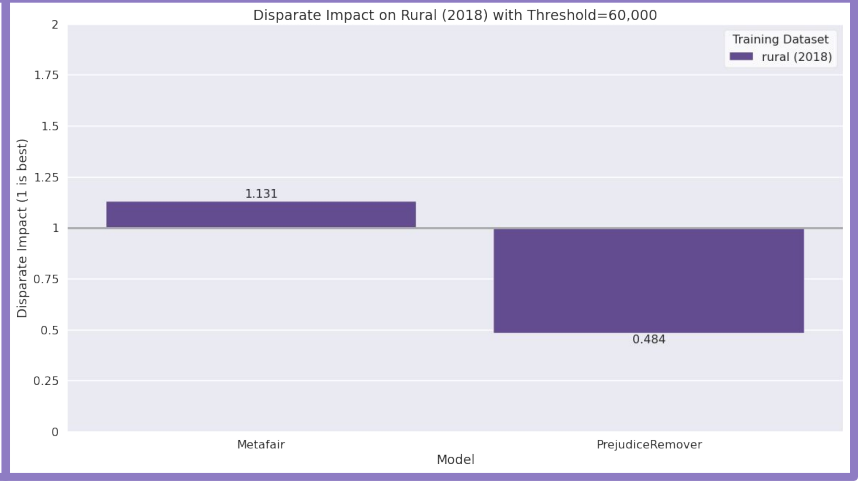


Results - Threshold - Disparate Impact









Limits

- Not every model could run on every data set
- Regional Context limited -> Only US States, not international
- Only limited comparisons with non-fairness aware models
- Only group fairness metrics, no individual fairness metric
- no/ few data set metrics

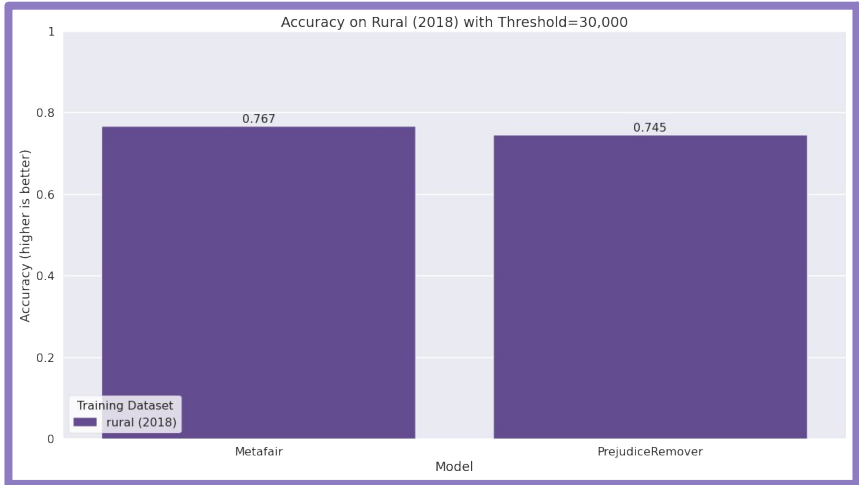
Possible Extensions

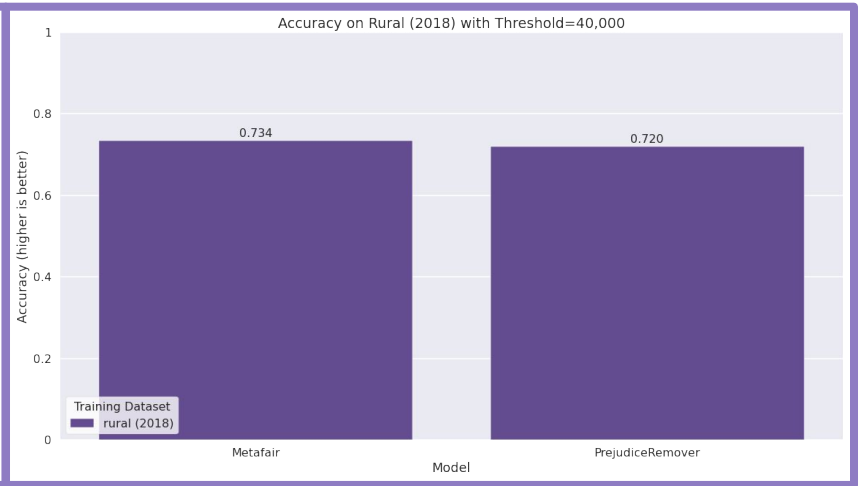
- Include more models
- Calculate more metrics (e.g. individual fairness)
- Use more data -> More datasets
- Use more data sources -> e.g. international data
- Run more experiments -> threshold, temporal/spatial context etc.

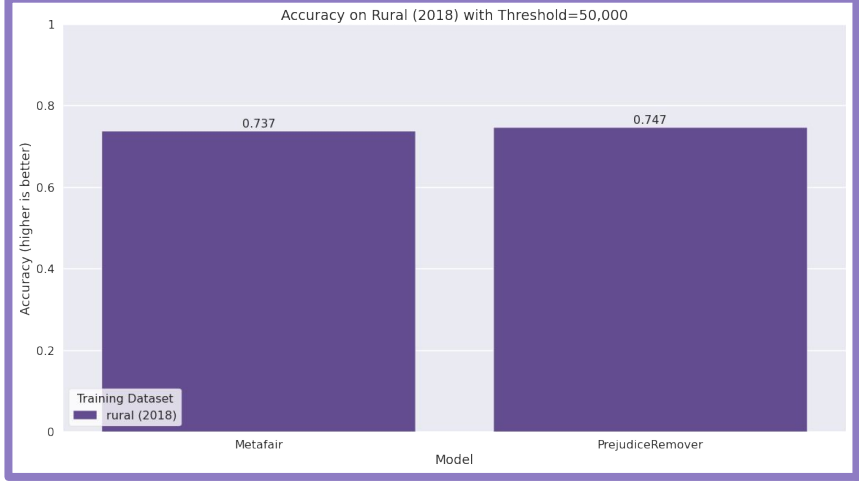
Thank you for listening!

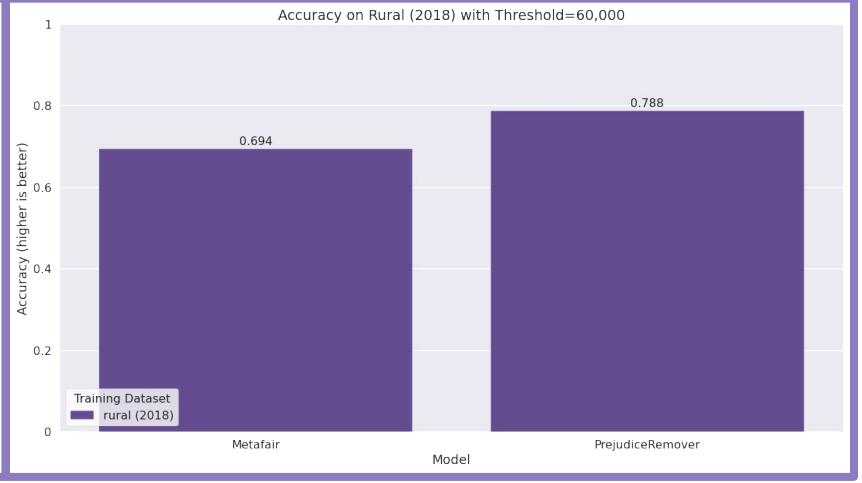
But wait… there's more!

Results - Threshold - Accuracy









Results - Threshold - F1 Score