Fairness Monitoring

SWP AIML

FU Berlin: SWP 2022
Examiner: Eirini Ntoutsi
Presented by: Jonas Schäfer, Marius Wawerek, Tolga Yurtseven

Outline

Intro

Key methods

Datasets

Results

Conclusions

Demonstration

Intro

Why fairness monitoring is important?

- AI and ML are widely used today
- \+ ability to analyze large amount of data with high accuracy
- \- discriminative impact on individuals and groups
- handling bias and fairness brings technical challenges

Example

Does fairness impact real lifes?

- (COMPAS) software used by US courts to value recidivism risk
- black defendants recidivism risk was higher predicted than their actual risk compared to white defendants

How can this happen?

- protected attributes don’t guarantee that sensitive information is used
- proxy attributes (zip code/race, credit rating/safe driving)
- methods for handling proxy attributes -> reduce utility of data

Main functionality

What is fairness?

-> Measure of **“Discrimination”** against certain individuals or groups of ppl

How to make a model fair?

-> Make it **aware** of possible **bias (Fairness Awareness)**

-> Actively work against this bias

Main functionality - Our Focus

Evaluate the different impact of the **context** that a model is trained on

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

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

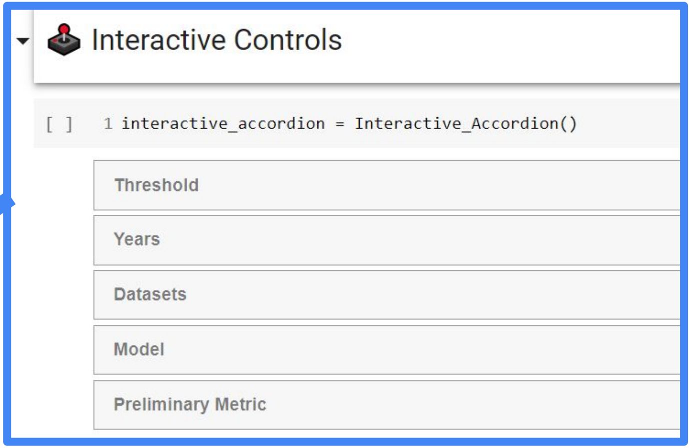
➔We want to test different settings of context to help make a model fair

Functionality

- an training framework
- an plotting function for our datasets
- an visualization framework for evaluating different fairness metrics
- Training Dashboard
- Visualization Dashboard

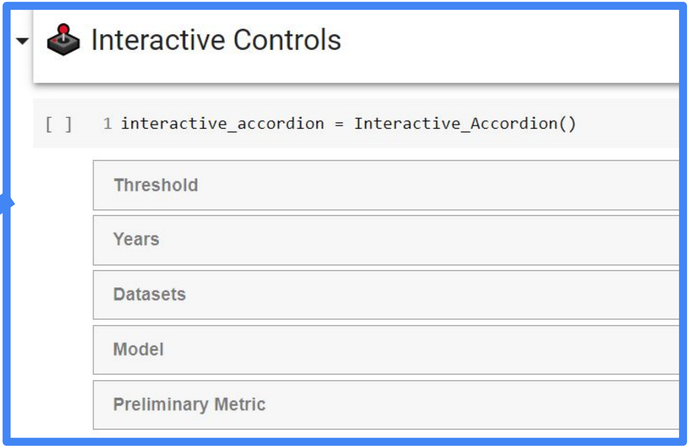
Functionality

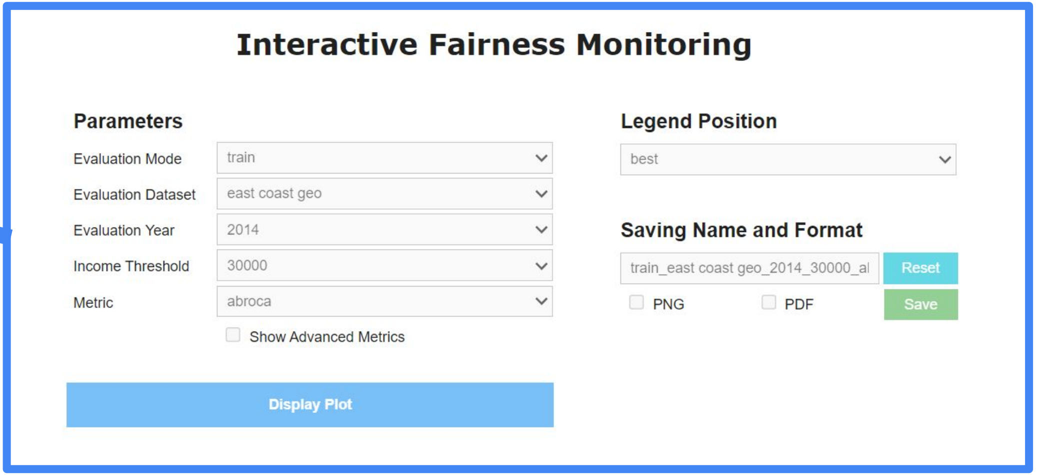
- an training framework
- an plotting function for our datasets
- an visualization framework for evaluating different fairness metrics
- Training Dashboard
- Visualization Dashboard



Functionality

- an training framework
- an plotting function for our datasets
- an visualization framework for evaluating different fairness metrics
- Training Dashboard
- Visualization Dashboard





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

General Dataset

ACI Income (also known as “Adult Dataset”)

-> Based on US Census (≈1-2% of USA

We can choose any subset

Time scale: 2014 -2018

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

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

General Dataset

ACI Income (also known as “Adult Dataset”)

-> Based on US Census (≈1-2% of USA

We can choose any subset

Time scale: 2014 -2018

We can adjust this threshold

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

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

Datasets

-Northern States

-Southern States

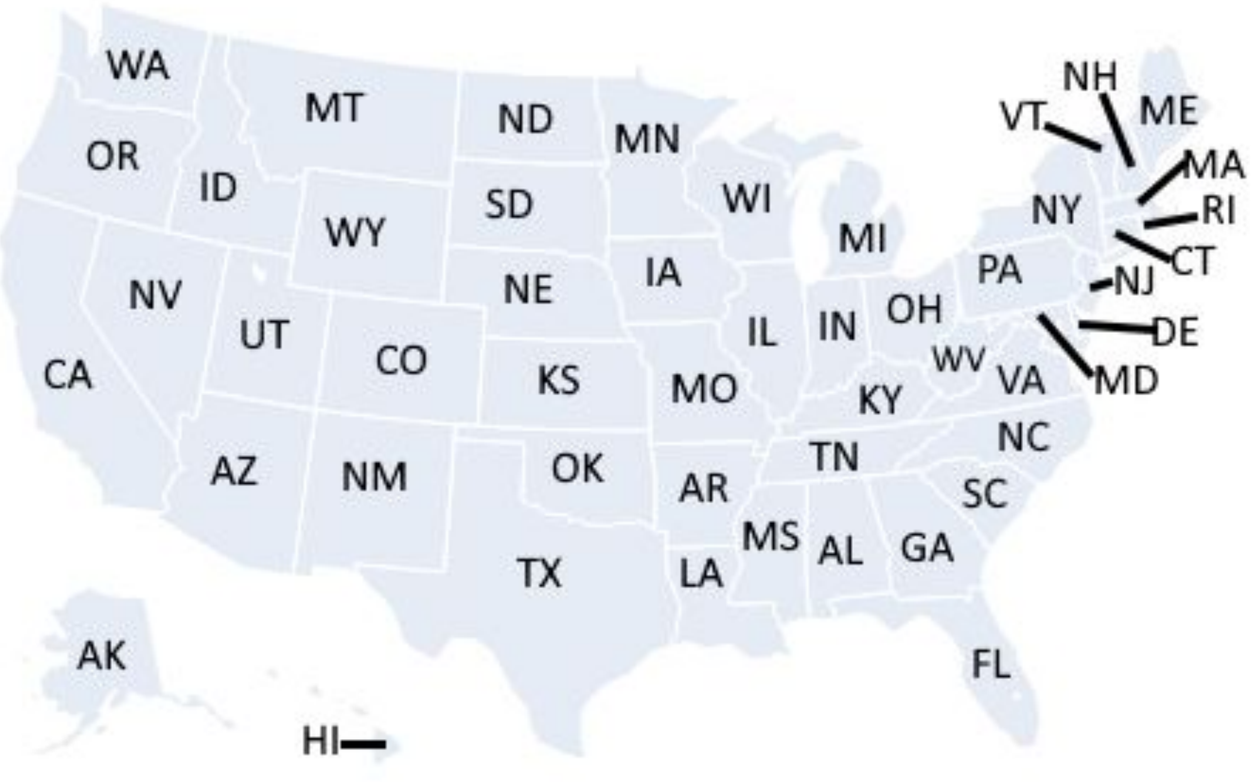
-East Coast

-West Coast

-None Coast

-Urban States

-Rural States



Sample Size: 1.6 Mio (2016)

mean class: 34%

Datasets

-Northern States

-Southern States

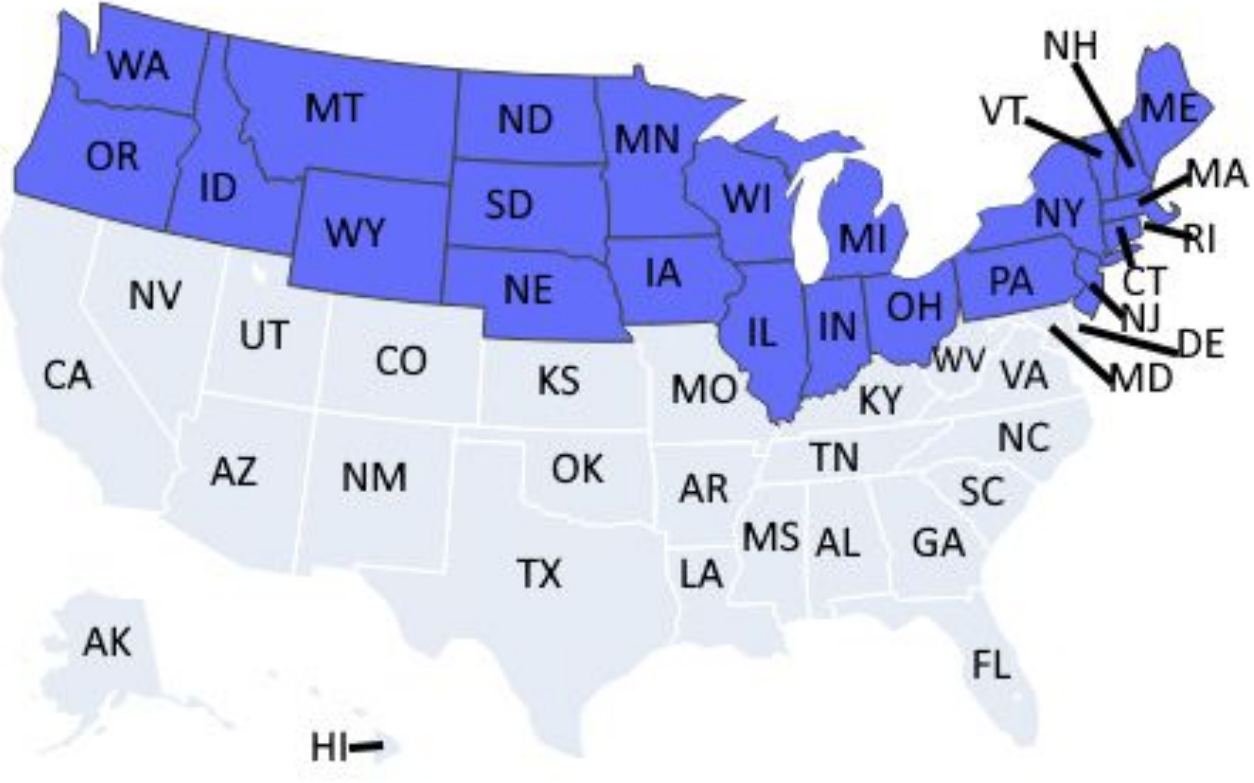
-East Coast

-West Coast

-None Coast

-Urban States

-Rural States



Sample Size: 676k (2016)

mean class: 35%

Datasets

-Northern States

-Southern States

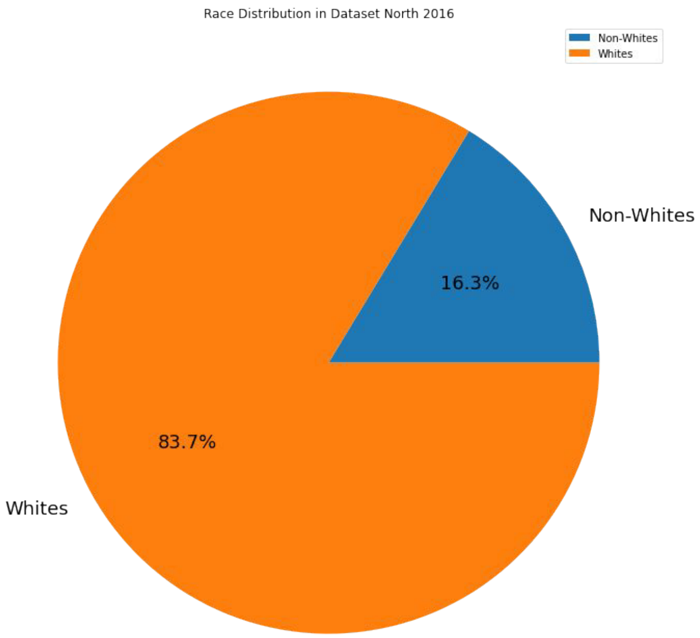
-East Coast

-West Coast

-None Coast

-Urban States

-Rural States



Sample Size: 676k (2016)

mean class: 35%

Datasets

-Northern States

-**Southern States**

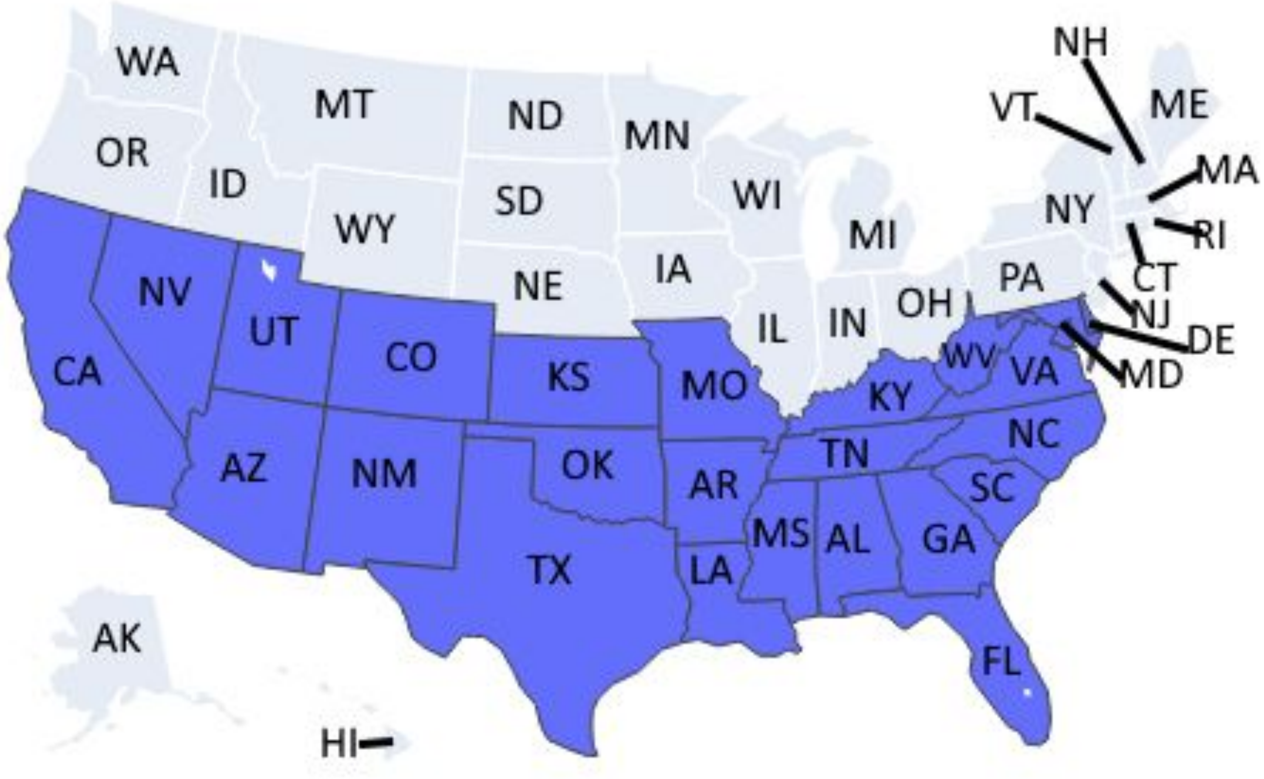
-East Coast

-West Coast

-None Coast

-Urban States

-Rural States



Sample Size: 915k (2016)

mean class: 33%

Datasets

-Northern States

-Southern States

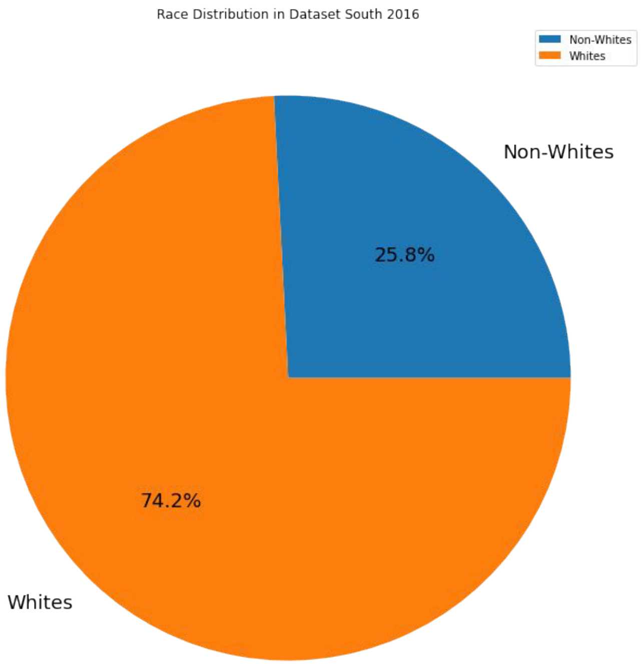
-East Coast

-West Coast

-None Coast

-Urban States

-Rural States



Sample Size: 915k (2016)

mean class: 33%

Datasets

-Northern States

-Southern States

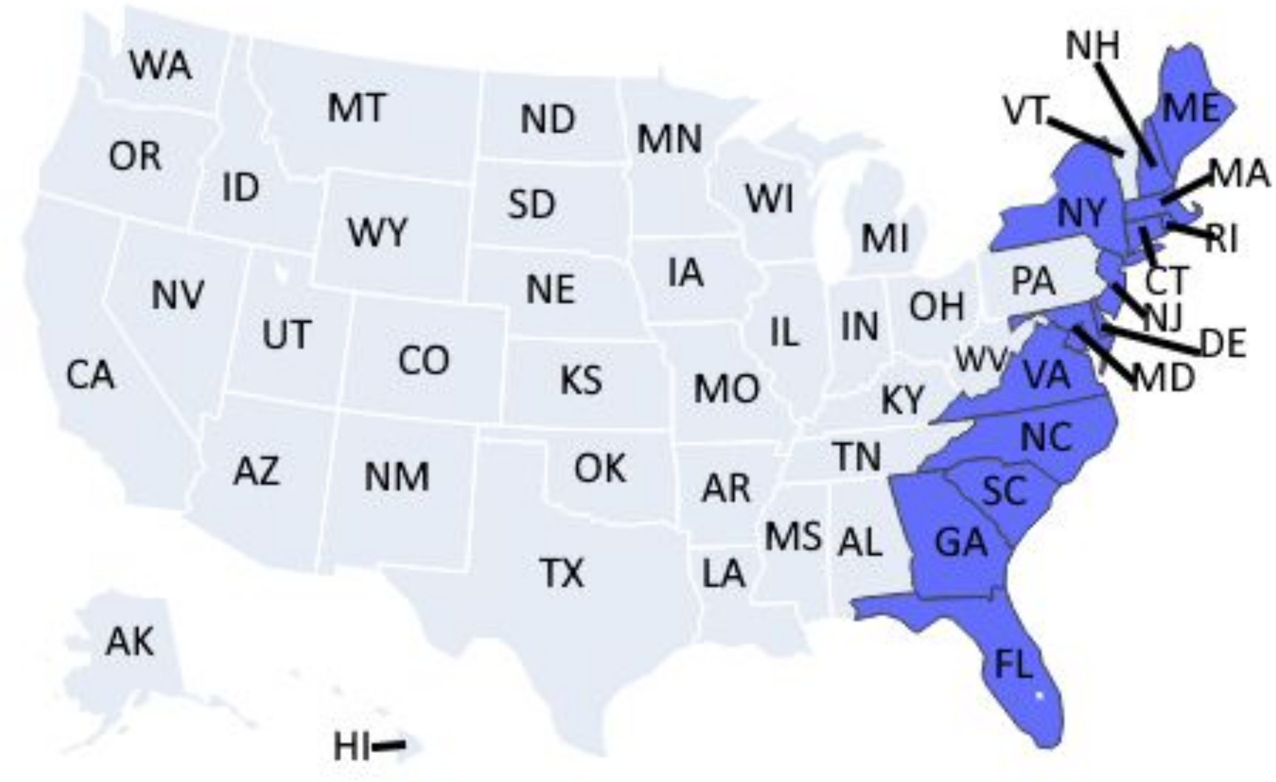
-East Coast

-West Coast

-None Coast

-Urban States

-Rural States



Sample Size: 526k (2016)

mean class: 37%

Datasets

-Northern States

-Southern States

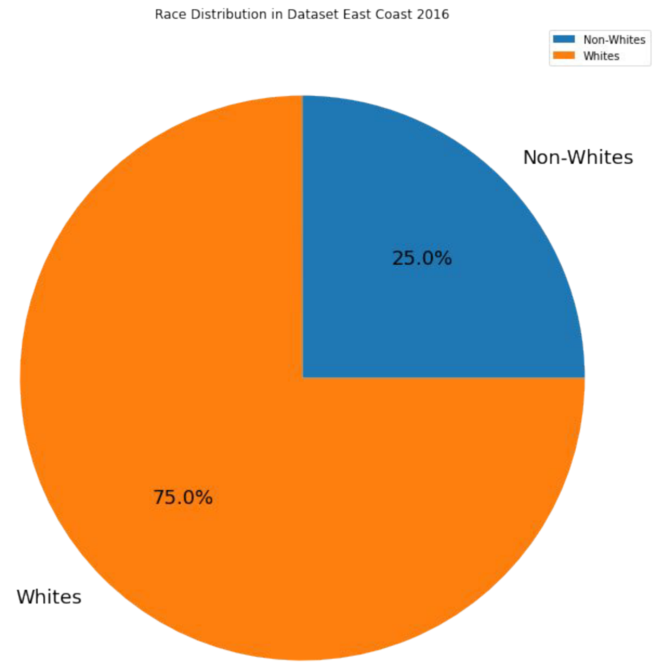
-**East Coast**

-West Coast

-None Coast

-Urban States

-Rural States



Sample Size: 526k (2016)

mean class: 37%

Datasets

-Northern States

-Southern States

-East Coast

-West Coast

-None Coast

-Urban States

-Rural States

Sample Size: 260k (2016)

mean class: 38%

Datasets

-Northern States

-Southern States

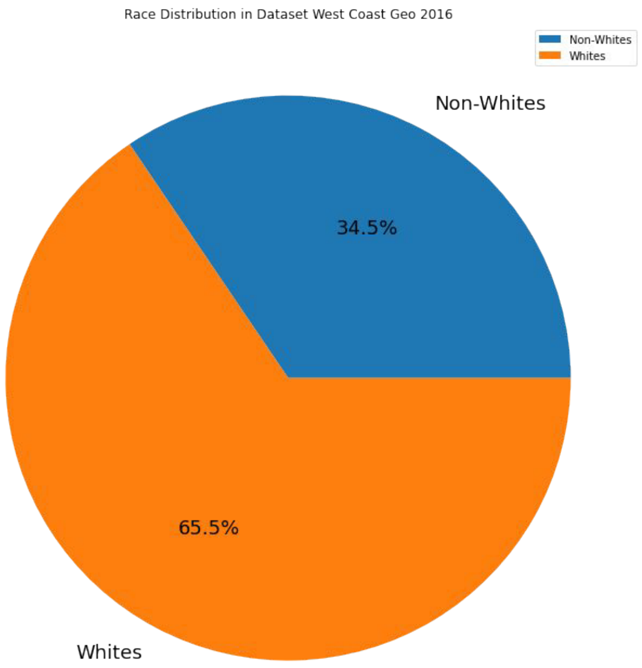
-East Coast

-**West Coast**

-None Coast

-Urban States

-Rural States



Sample Size: 260k (2016)

mean class: 38%

Datasets

-Northern States

-Southern States

-East Coast

-West Coast

-None Coast

-Urban States

-Rural States

Sample Size: 831k (2016)

mean class: 31%

Datasets

-Northern States

-Southern States

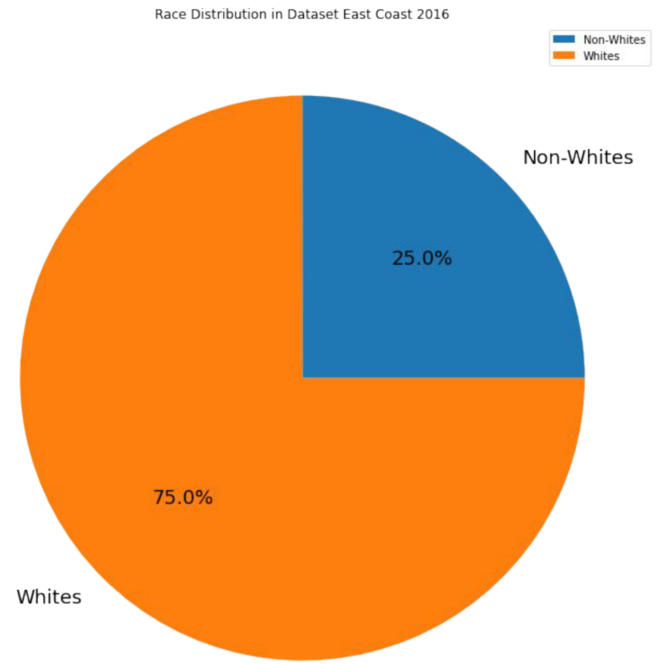
-East Coast

-West Coast

-None Coast

-Urban States

-Rural States



Sample Size: 831k (2016)

mean class: 31%

Datasets

-Northern States

-Southern States

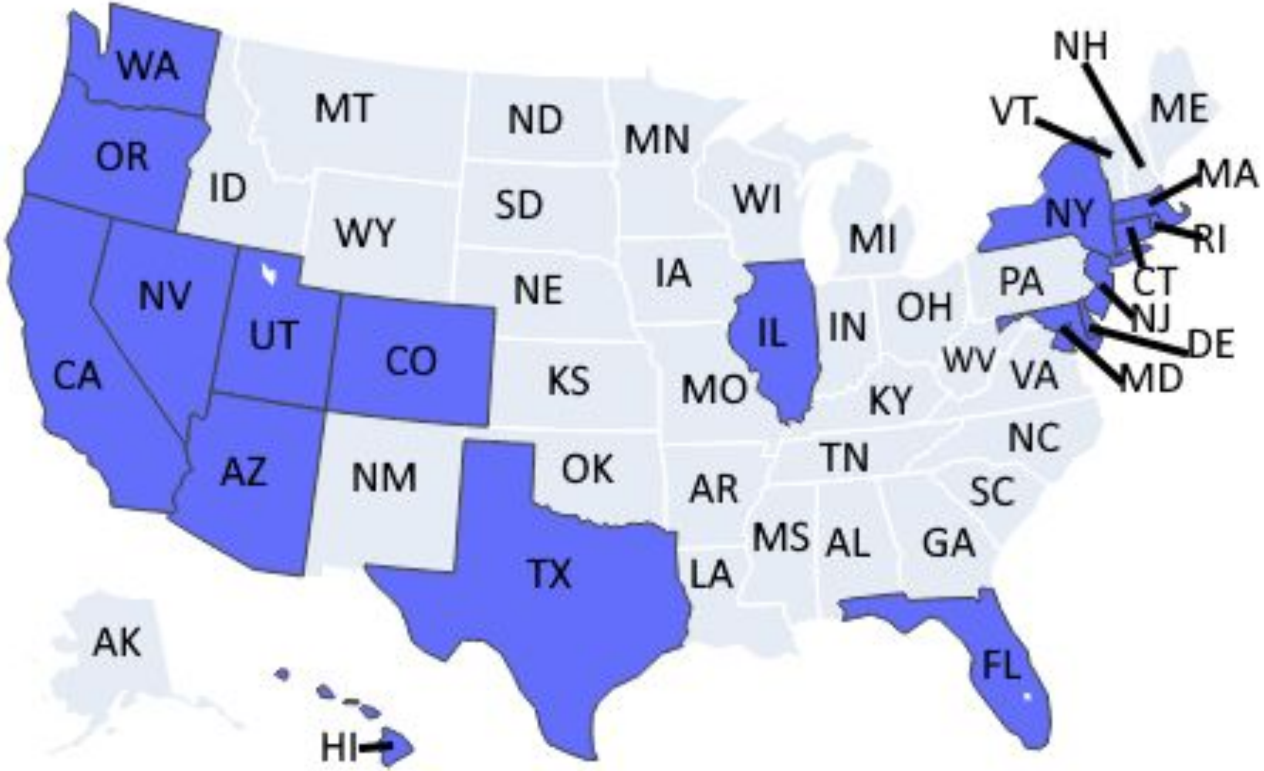
-East Coast

-West Coast

-None Coast

-Urban States

-Rural States



Sample Size: 898k (2016)

mean class: 37%

Datasets

-Northern States

-Southern States

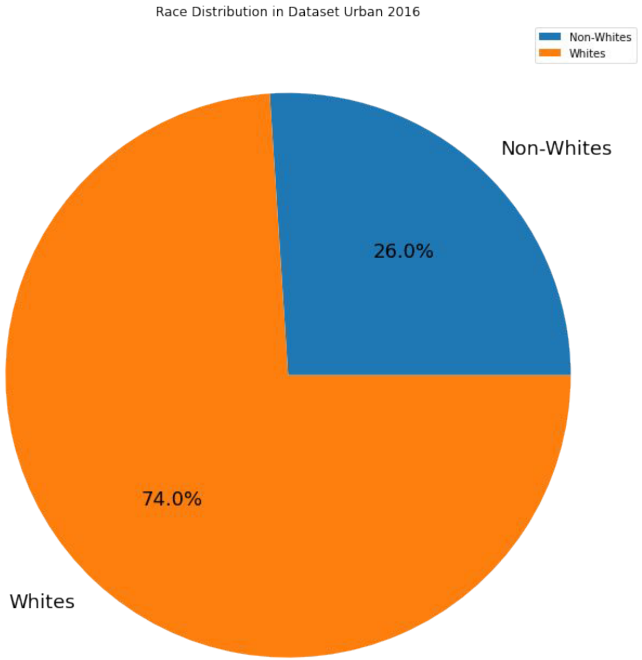
-East Coast

-West Coast

-None Coast

-Urban States

-Rural States



Sample Size: 898k (2016)

mean class: 37%

Datasets

-Northern States

-Southern States

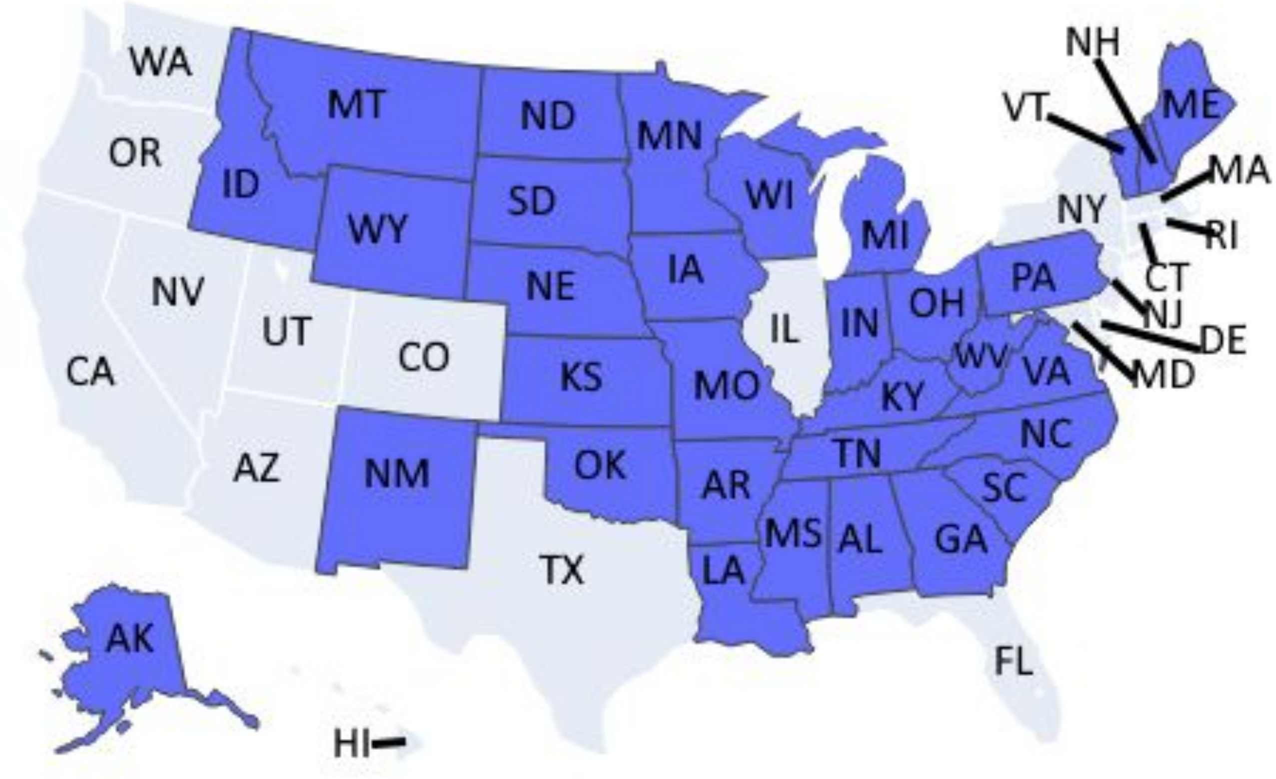
-East Coast

-West Coast

-None Coast

-Urban States

-Rural States



Sample Size: 719k (2016)

mean class: 31%

Datasets

-Northern States

-Southern States

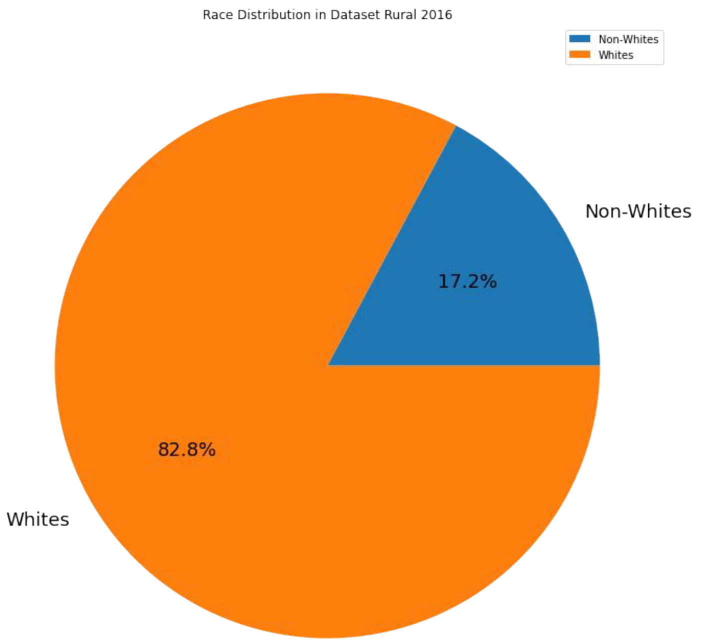
-East Coast

-West Coast

-None Coast

-Urban States

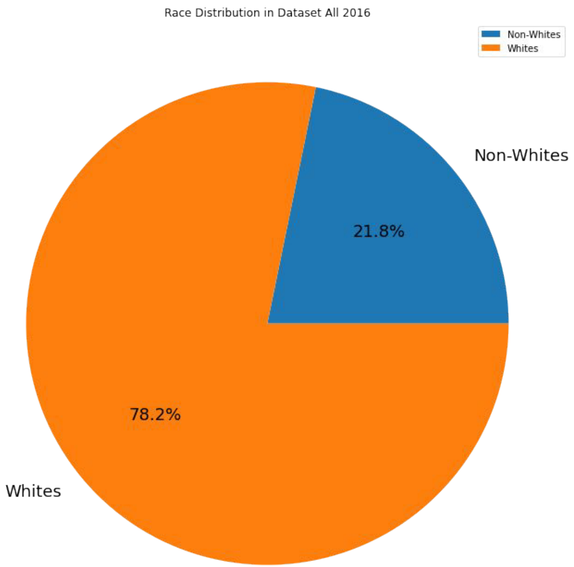
-**Rural States**



Sample Size: 719k (2016)

mean class: 31%

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

Experiments

Main metrics:

Classification Performance: Accuracy, F1-Score,

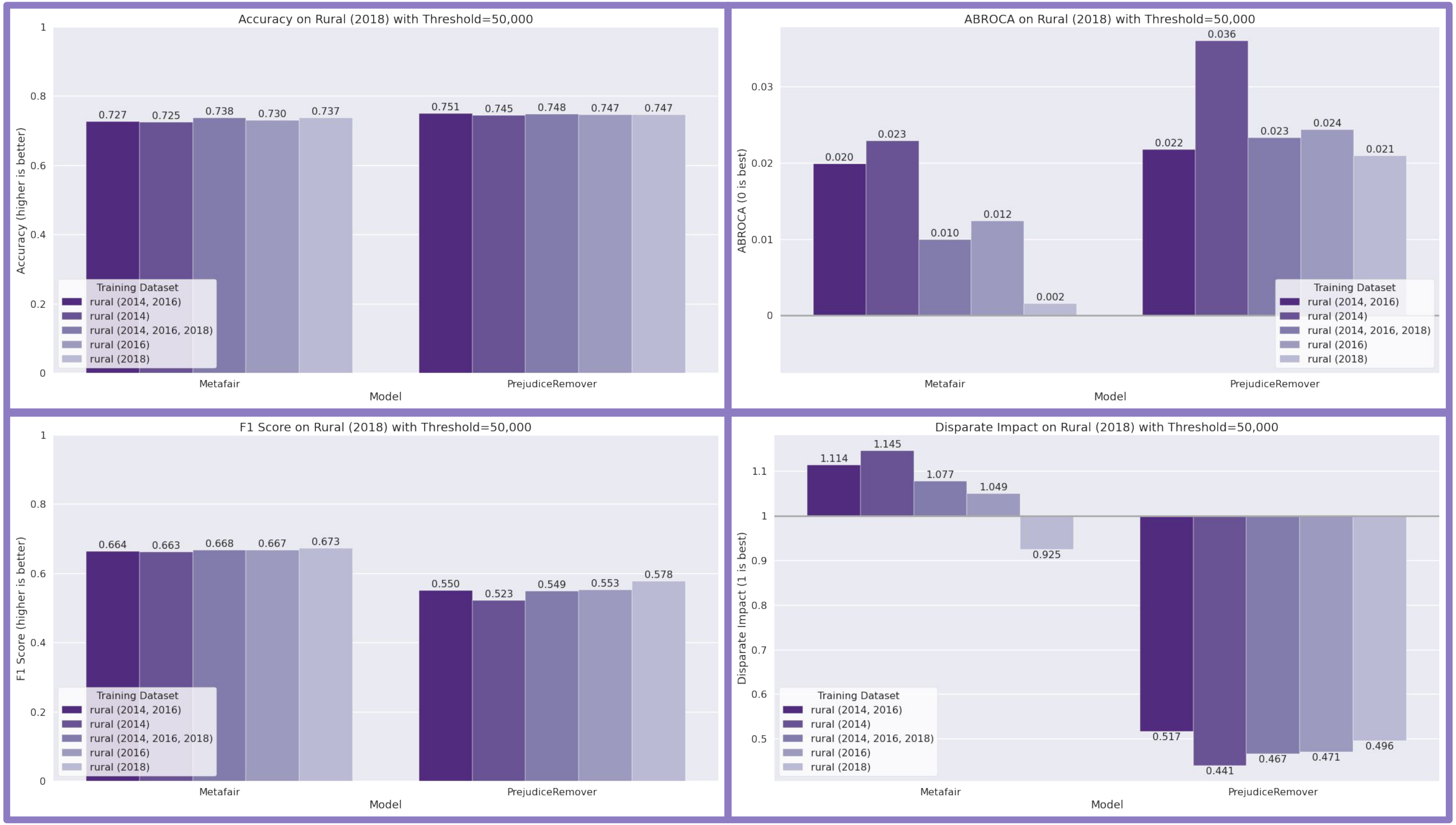
Classification Fairness: ABROCA, Disparate Impact

- Impact of **temporal** context shifts
- Impact of **spatial** context shifts
- Impact of both **temporal and spatial** context shifts
- Impact of the **method of data binarization** (threshold choice)

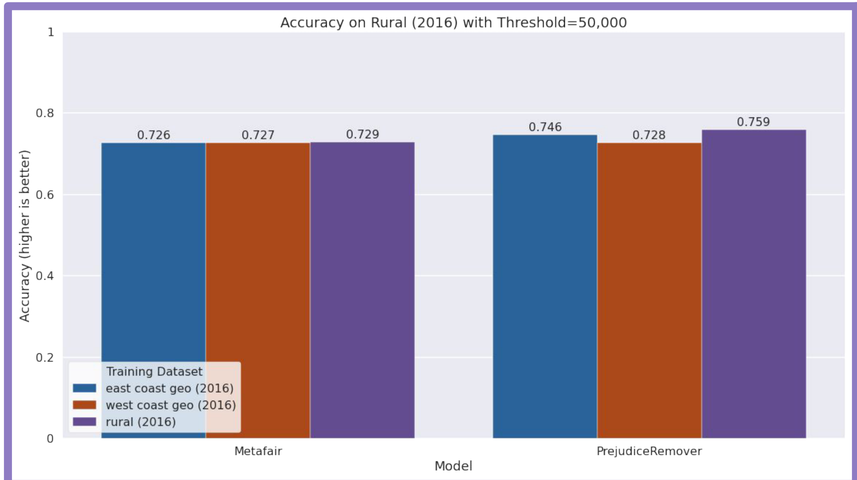
Results - How to interpret

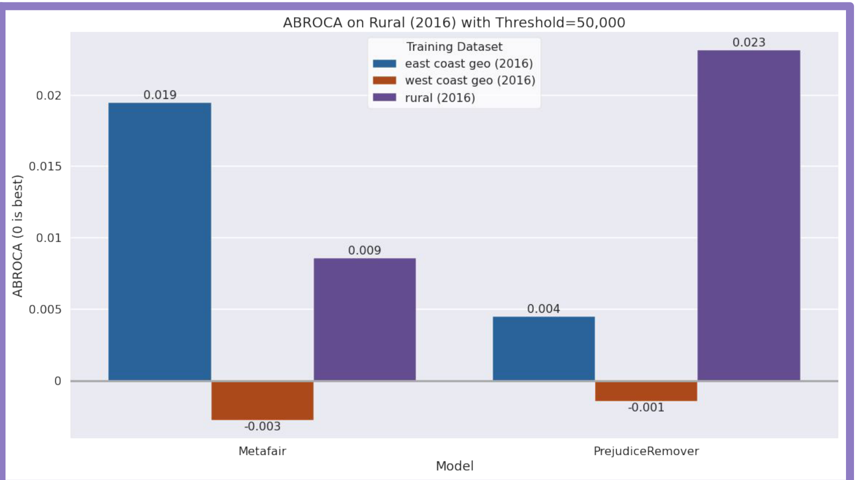
Performance: Accuracy Range: [0, 1] ⬆	Fairness: ABROCA ⬍ 0 is best
Performance: F1 Score Range: [0, 1] ⬆	Fairness: Disparate Impact ⬍ 1 is best

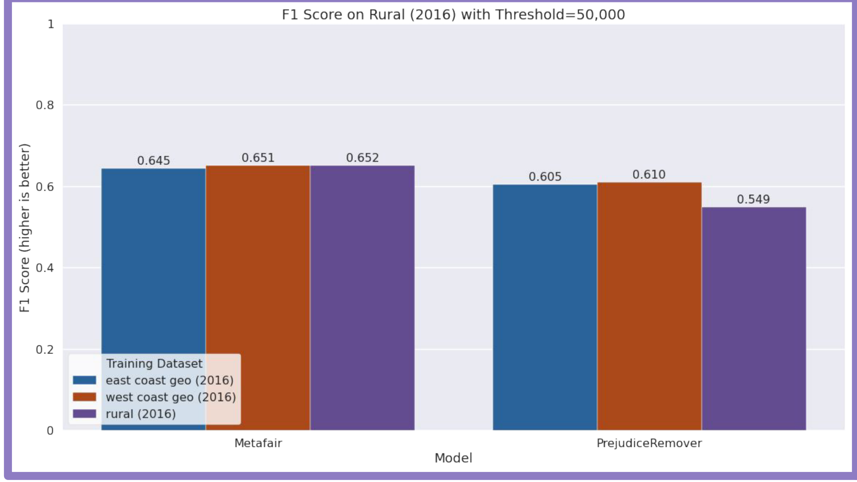
Results - Temporal Context

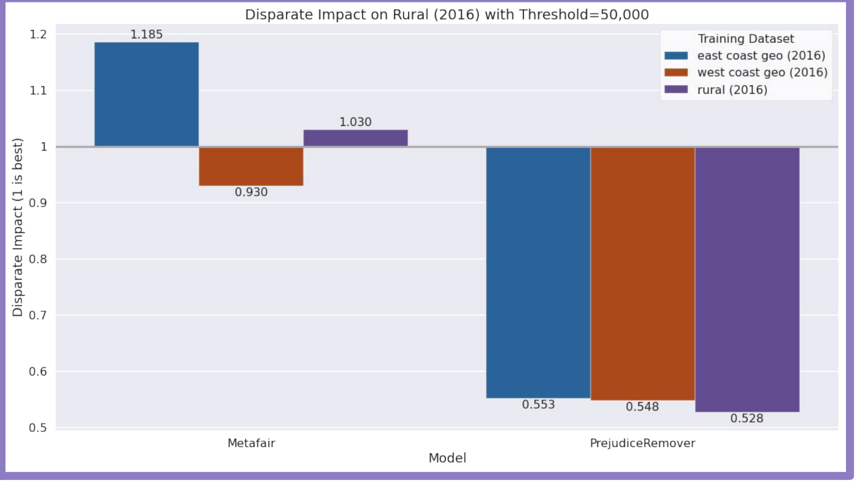


Results - Spatial Context

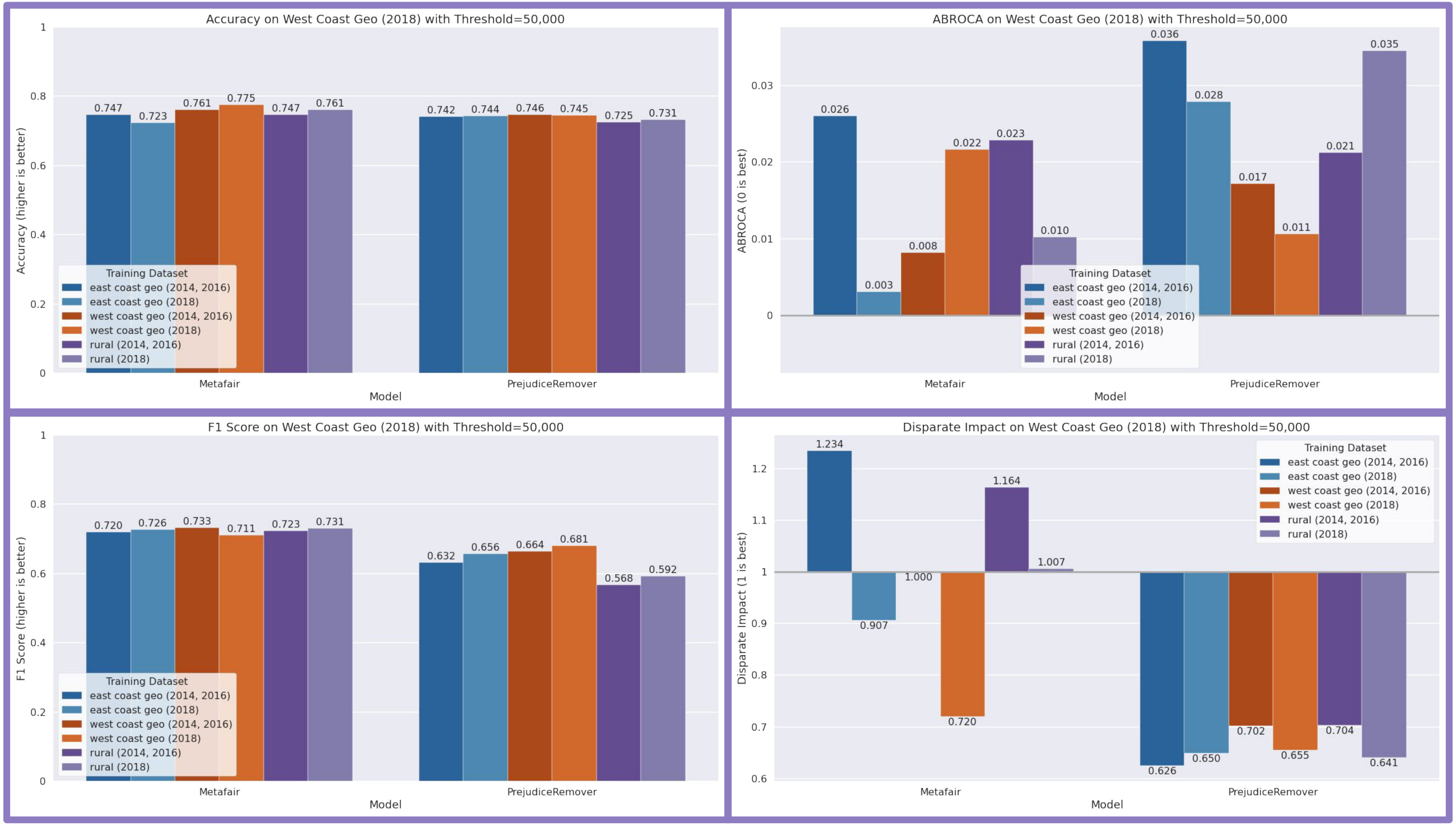




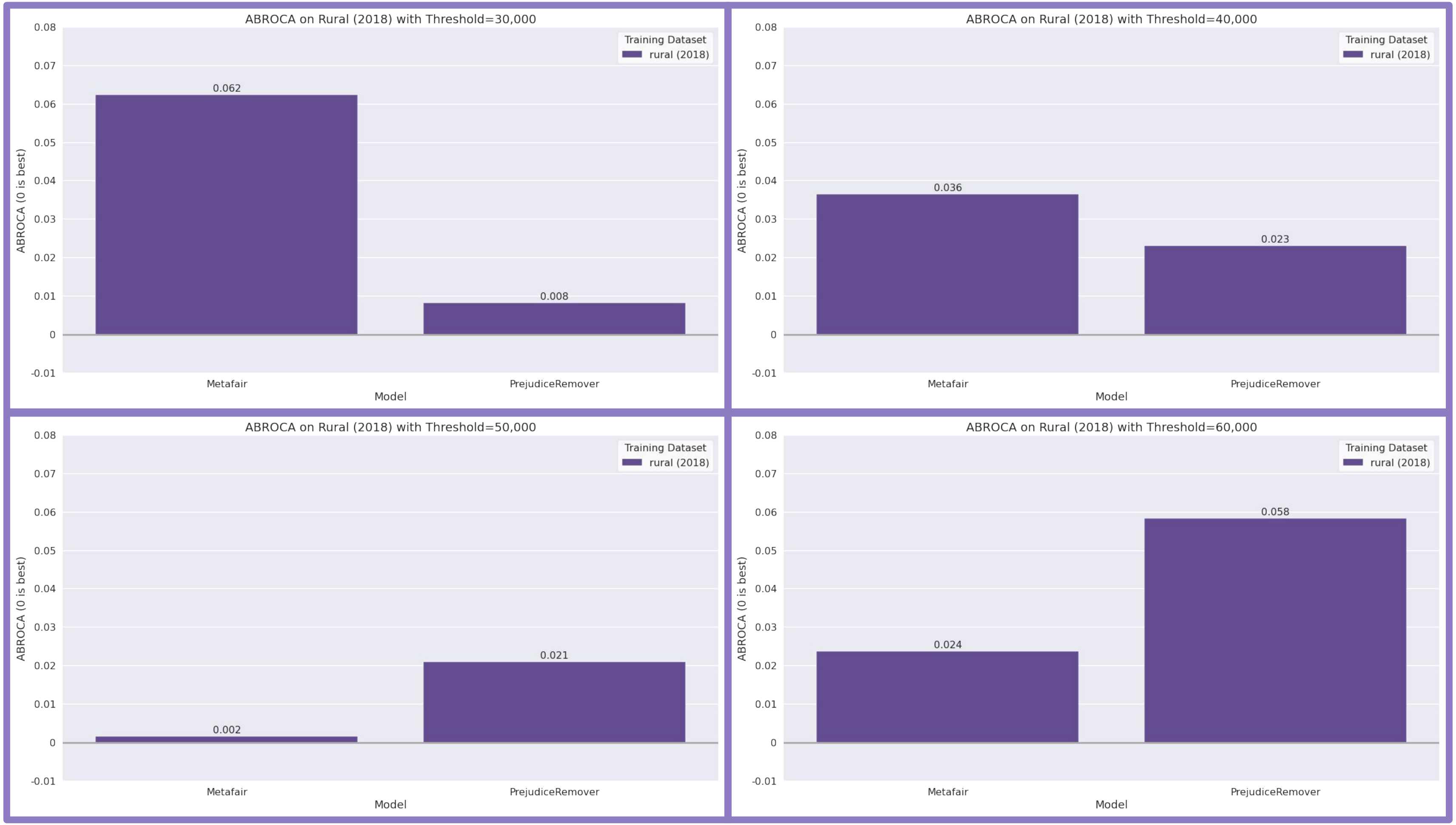




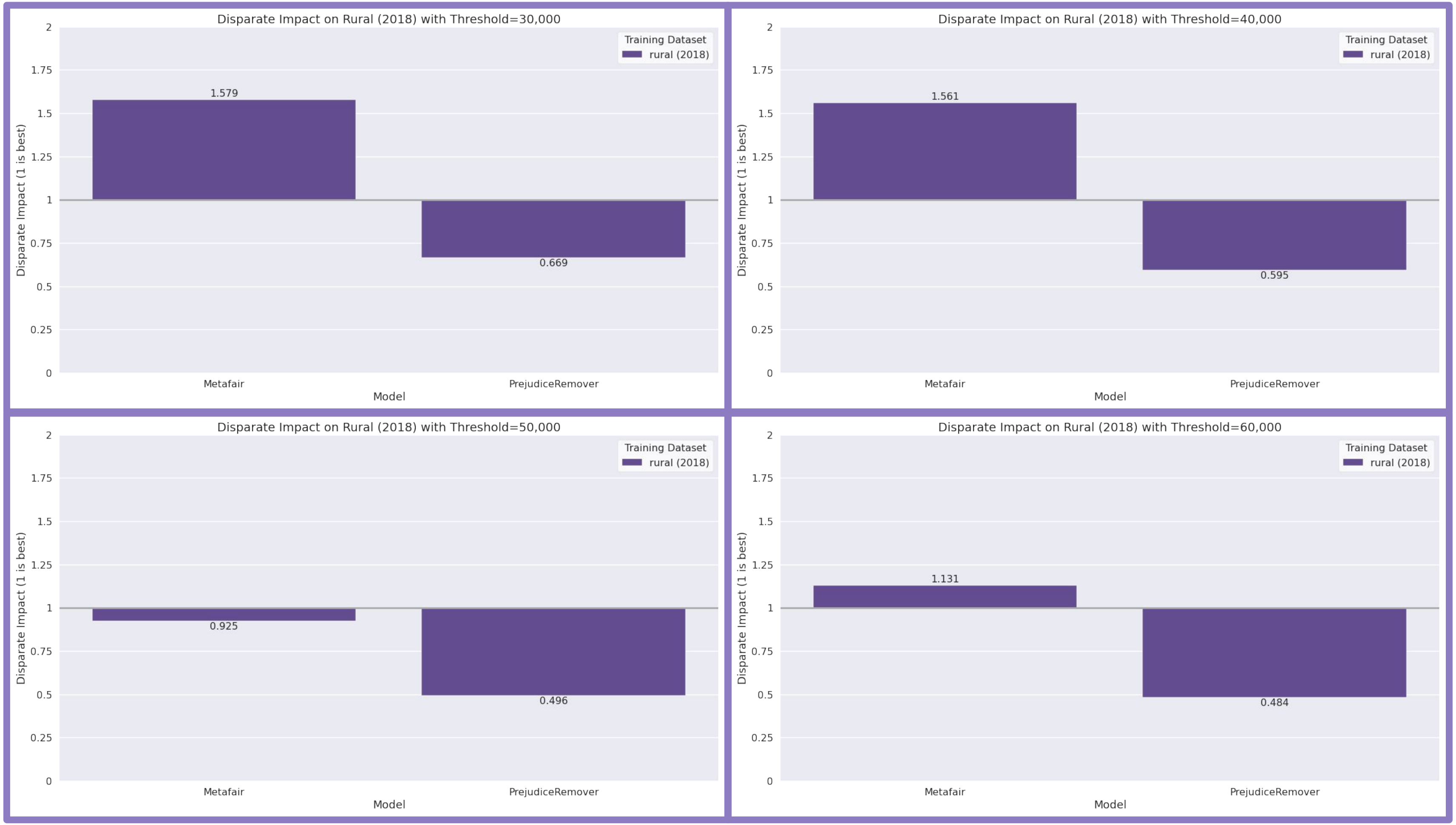
Results - Temporal and Spatial Context



Results - Threshold - ABROCA



Results - Threshold - Disparate Impact



Conclusions:

Most experiments show that the fairness of models **decreases** after both spatial and temporal **distribution shifts**.

If the spatial and/or temporal context **does not change**, the models **mostly retain their fairness**.

The **method of data labelling** can have **strong effects** on model fairness.

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.

Sources

E. Ntoutsi et al., "Bias in data-driven artificial intelligence systems—An introductory survey," WIREs Data Mining and Knowledge Discovery, vol. 10, no. 3, p. e1356, 2020, doi: 10.1002/widm.1356.

S. Ghodsi, H. Alani, and E. Ntoutsi, "Context matters for fairness – a case study on the effect of spatial distribution shifts." arXiv, 2022. doi: 10.48550/ARXIV.2206.11436.

T. L. Quy, A. Roy, V. Iosifidis, W. Zhang, and E. Ntoutsi, "A survey on datasets for fairness-aware machine learning," WIREs Data Mining and Knowledge Discovery, vol. 12, no. 3, Mar. 2022, doi: 10.1002/widm.1452.

F. Ding, M. Hardt, J. Miller, and L. Schmidt, "Retiring Adult: New Datasets for Fair Machine Learning," in Advances in Neural Information Processing Systems, 2021, vol. 34, pp. 6478–6490.

A. Castelnovo, R. Crupi, G. Greco, and D. Regoli, "The zoo of Fairness metrics in Machine Learning," CoRR, vol. abs/2106.00467, 2021, [Online]. Available: https://arxiv.org/abs/2106.00467

A. Fabris, S. Messina, G. Silvello, and G. A. Susto, "Algorithmic Fairness Datasets: the Story so Far," ArXiv, vol. abs/2202.01711, 2022.

J. Gardner, C. Brooks, and R. Baker, "Evaluating the Fairness of Predictive Student Models Through Slicing Analysis," Feb. 2019. doi: 10.1145/3303772.3303791.

Thank you for listening!

But wait… there’s more!

Demonstration

Interactive Notebooks

-> Interactive Gridtraining

-> Plots/Outputs

-> Interactive Metric Visualization