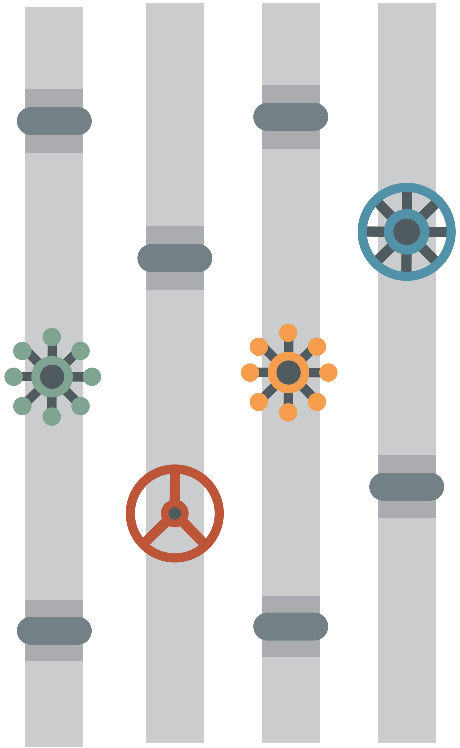
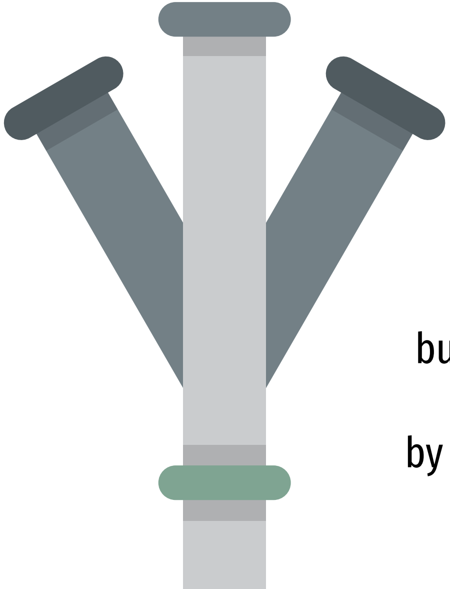
Improving the DualFair Pipeline to Evaluate Alternate World Index

by Nayna Pashilkar & Abigail Starr



Research Question



Using the novel, **intersectional** fairness metric **Alternate World Index** (AWI), how does balancing and debiasing 2022 HMDA datasets with the **DualFair pipeline** influence the accuracy and fairness of loan classifiers?

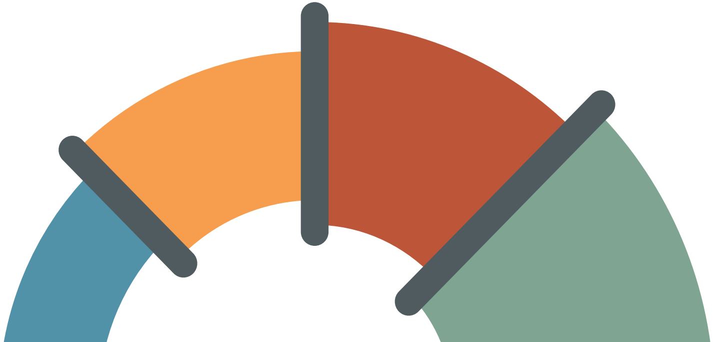
built off ideas from **"Developing a Novel Fair-Loan Classifier through a Multi-Sensitive Debiasing Pipeline: DualFair"**
by Arashdeep Singh, Jashandeep Singh, Ariba Khan, & Amir Gupta (2022)

What is AWI?

- **normalized count** of biased points
- a point is **“unbiased”** if model prediction is **consistent** across all counterfactual worlds
- a **counterfactual world** is a copy of the dataset where sensitive parameters (i.e. race, sex) have been toggled and are identical for all points

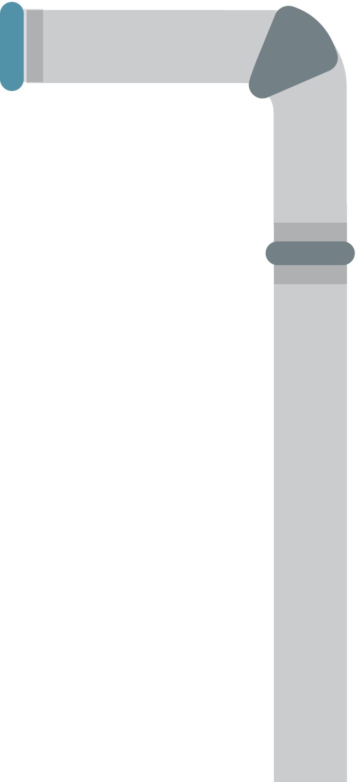
Why Do We Care?

modern fairness metrics rarely account for **intersectionality**. If successful, the DualFair model could help **train fairer models**.

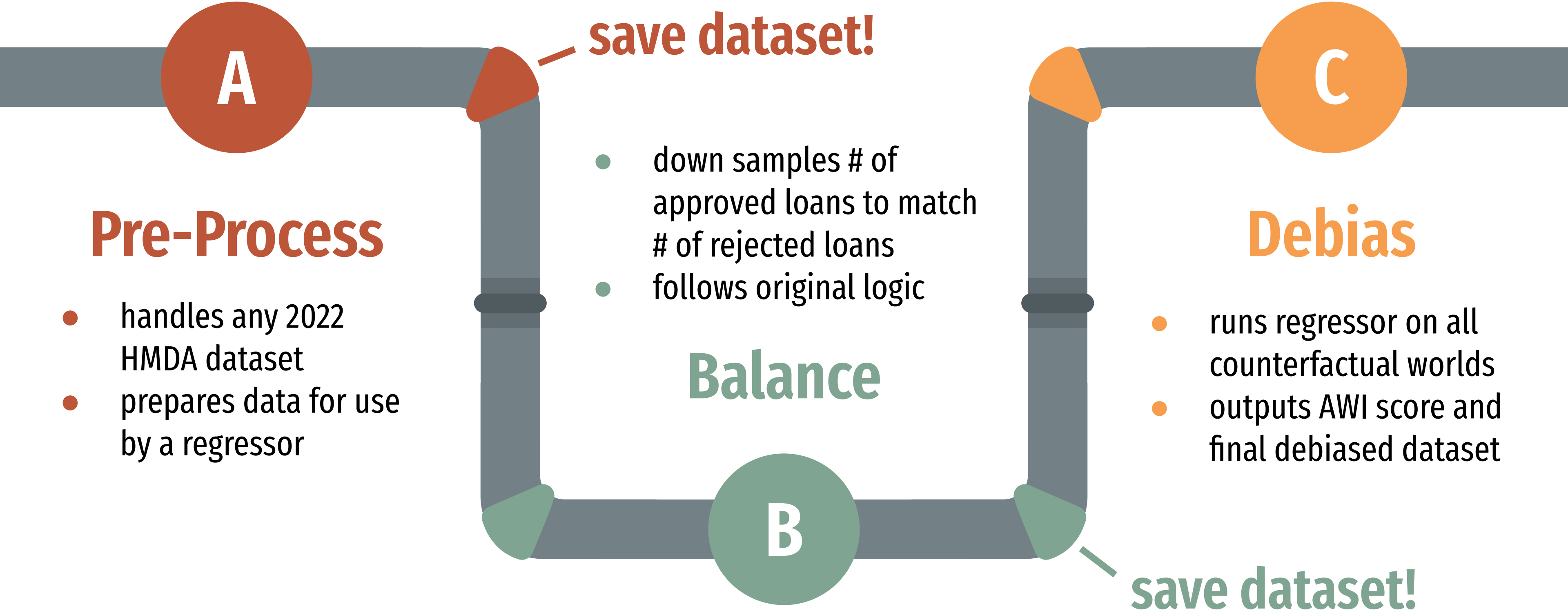


Pitfalls

of the original DualFair model



- **confusing** implementation of AWI
- consistently **failed to remove** any points from datasets
- unusable **output** & does not return an AWI score



A

Pre-Process

- handles any 2022 HMDA dataset
- prepares data for use by a regressor

save dataset!

- down samples # of approved loans to match # of rejected loans
- follows original logic

Balance

B

save dataset!

C

Debias

- runs regressor on all counterfactual worlds
- outputs AWI score and final debiased dataset

Our Findings

State Size	Increase in Accuracy	Increase in F1 Score	Average Increase in Proportion Given Loans for Majority Groups	Average Increase in Proportion Given Loans for Minority Groups
Small	8.5%	9.46%	20.32%	20.04%
Medium	17.45%	20.44%	26%	25.26%
Large	50.15%	61.36%	43.69%	25.9%

after comparing balanced datasets to their balanced + debiased counterparts, we determined...

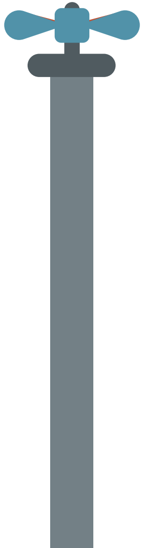
The DualFair Model does **NOT** proportionally increase the overall fairness of a classifier

State Size	Average AWI
Small	0.00115
~ 50,000 rows	0.07131
	0.10165
Medium	0.00133
~ 300,000 rows	0.04734
	0.08292
Large	0.00069
~ 500,000+ rows	0.04999
	0.10815

Limitations

script is hardcoded for 2022 HMDA data

future researchers could not automatically apply our pipeline to another dataset



we consider only 8 counterfactual worlds

future researchers should account for more demographics, although this exponentially increases runtime