MSc Data Analytics 2019

Department of Computer Science and Information Systems Birkbeck, University of London

Assessing fairness/bias in binary classification machine learning models on consumers

Benedict Faria

Assessing fairness/bias in binary classification machine learning models on consumers

Project Objectives

1. Measure data set and classification model bias using selected metrics from the AIF360 toolkit.
2. Apply selected bias-mitigation algorithms to the data sets.
3. Measure the effect of each bias-mitigation algorithm on model performance
4. Assess the business implications of bias-mitigation algorithms.

What is Bias in Machine Learning

Definition

In law, bias or discrimination refers to unfair treatment of individuals because of their membership of a certain group. An algorithm is biased if it produces results that are unfair to individuals or groups with respect to the population it is being used to analyse.

Why is it important to address

- Looming legislation
- Some examples:
  - Street Bump app in Boston.
  - XING job platform found to rank less qualified male candidates higher than more qualified female candidates
  - Face recognition services from Microsoft, Face++, and IBM respectively, achieving lower accuracy on darker-skinned females

Where does this bias come from?

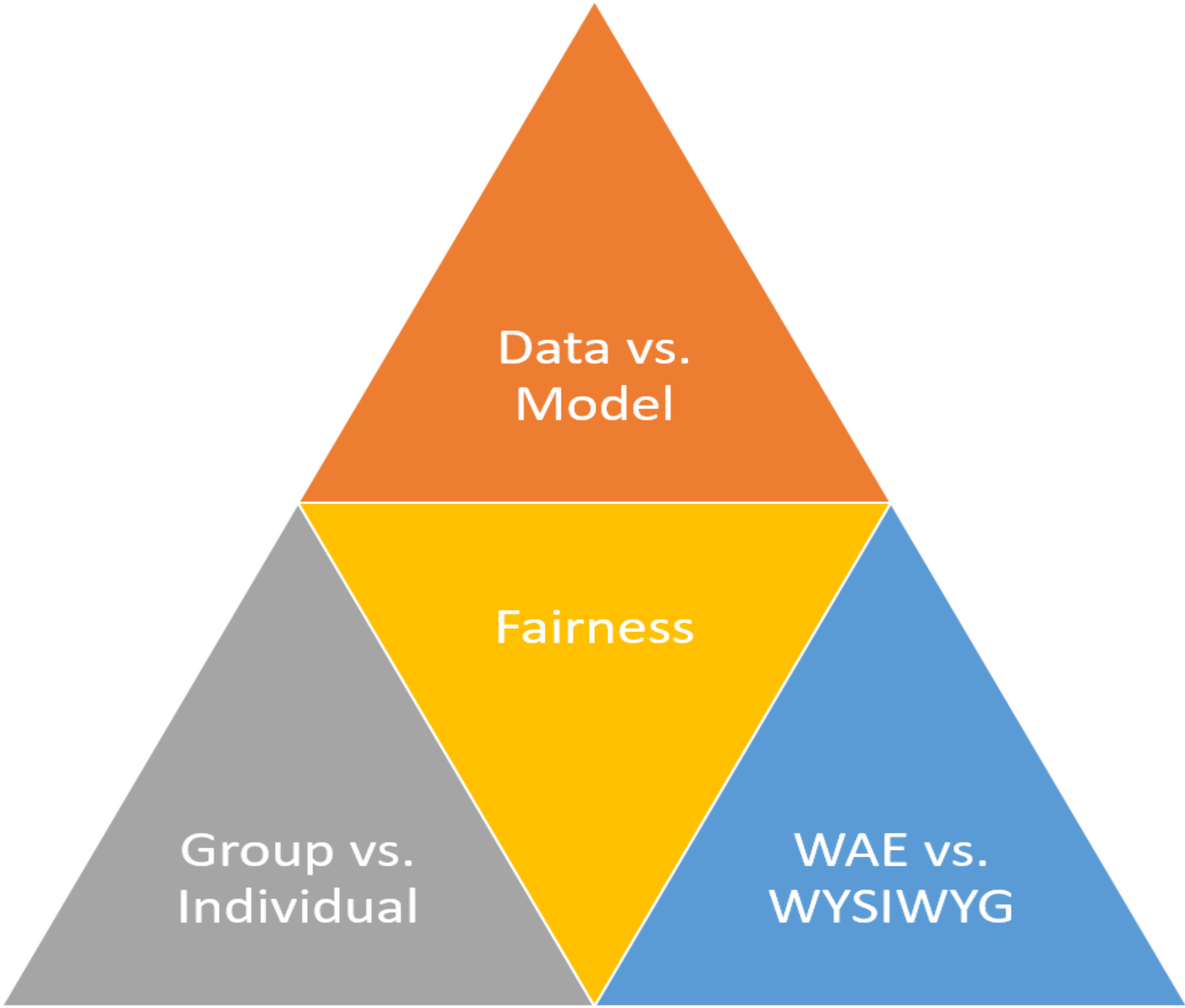
Training datasets

- Historic bias
- Over / Under representation - Skewed sampling or sample size disparity
- Incorrect labelling – perceptions
- Feature Selection

Measuring bias in Machine Learning

Three perspectives

- Data v/s Model – where do we mitigate bias in a ML pipeline?
- Group v/s Individual – what is the classification use case?
- WAE v/s WYSIWYG – what feature engineering is needed on the training dataset?

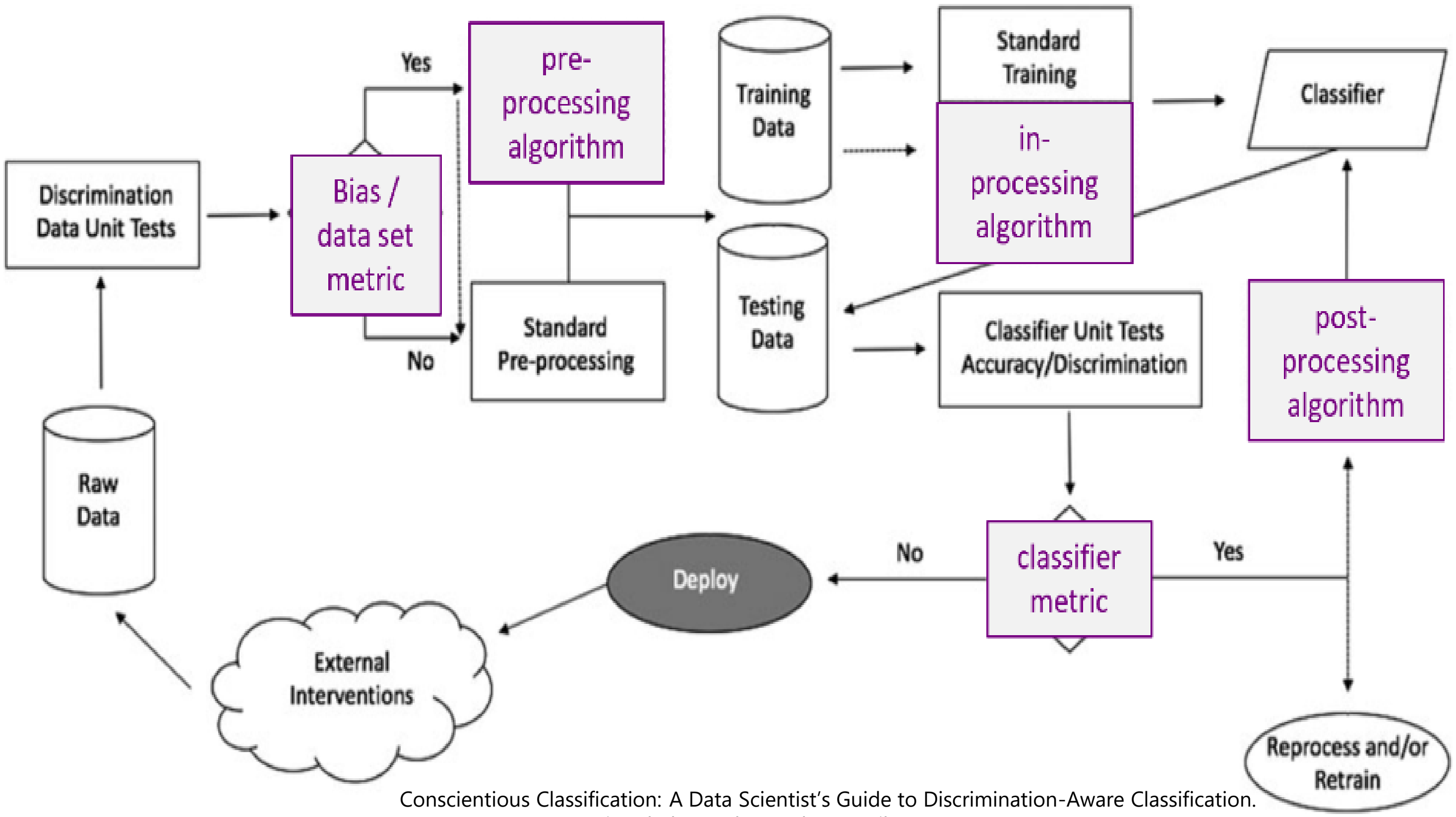


What's New in Deep Learning Research: Reducing Bias and Discrimination in Machine Learning Models with AI Fairness 360
Jesus Rodriguez 2018

Measuring & Mitigating bias in Machine Learning

The AIF360 toolkit

- Dataset class
  - Structured dataset
  - Binary Labelled dataset
- Metrics
  - Dataset metrics
  - Binary label dataset metrics
  - Classification Metrics
- Algorithms
  - Pre-processing
  - In-Processing
  - Post-Processing



Conscientious Classification: A Data Scientist's Guide to Discrimination-Aware Classification.
Brian d'Alessandro, Cathy O'Neil, Tom LaGatta 2017

Project Scope

Datasets

• **Taiwan – Gender**	30,000 entries with 24 attributes
• **Taiwan – Marriage**	30,000 entries with 24 attributes
• **German**	1000 entries with 20 attributes
• **Adult**	48,842 entries with 14 attributes

Classification Models

- **Logistic Regression**
- **Random Forest Classifier**

Project Scope

Metrics

- **Group fairness**
  - **Statistical Parity:** probability of achieving a favourable outcome is the same.
  - **Disparate Impact:** ratio between the probabilities less than 80% or some other legal threshold.

- **Individual Fairness**
  - **Consistency:** similar individuals from privileged and unprivileged groups treated similarly
  - **Equalised odds**: probability of an individual with an actual un/favourable outcome to be correctly assigned a un/favourable outcome

- **Sensitive Attribute test – balanced accuracy**
  - Test for Sensitive attribute obfuscation in the transformed dataset

- **Model Performance – balanced accuracy**
  - Test for classification model performance before and after mitigation

Project Scope

Algorithms

- **Pre-processing**
  - **Learning Fair Representations** – Group and Individual Fairness
  - **Disparate Impact** – Group fairness
  - **Re-weighing** – Group Fairness
- **In-processing**
  - **Adversarial de-biasing** – Group and Individual fairness
- **Post-processing**
  - **Reject Option Classification (ROC)** – Group fairness

Project Approach

Data Discovery / Feature Engineering

- Sensitive attribute and label distribution, ratios between groups
- New features (Taiwan dataset)

Baseline measures

- For group and individual fairness
- Model performance to predict label
- Model performance to predict Sensitive attribute

Applying Mitigation algorithms

- Using all algorithms in scope

Post Mitigation measures

- Comparing metrics with baseline measures

Findings

Taiwan-Gender	LFR	DI	Re-Weighing	Adversarial Debiasing	ROC
Statistical Parity (Before)	0.0298339	0.0332706	0.0354055	0.0310865	0.0519715
Statistical Parity (After)	0.0092356	0.0666004	2.220e-16	0.0308645	0.0399190 LR
					0.0167206 RF
Disparate Impact (Before)	1.0391294	1.0437451	1.0467924	1.0409955	1.0691643
Disparate Impact (After)	1.0093853	1.1722253	1.0000000	0.9662387	1.0504995 LR
					1.0232405 RF
Consistency (Before)	0.7854190	0.7844583	0.7836833	0.7812333	0.7388666
Consistency (After)	0.9996666	0.9187666	0.7836833	0.9738083	0.7840000 LR
					0.7476000 RF
Equality of Odds (Before)	0.0303846	0.0416668	0.0111595	0.0247457	0.059800
Equality of Odds (After)	0.0059917	0.0549178	0.0252449	0.0162113	0.008900
Logistic Regression - Sensitive attribute model performance					
Balanced Accuracy (Before)	0.5373862	0.5391934	0.5379510	0.5303988	0.5008655
Balanced Accuracy (After)	0.4857890	0.5390558	0.5379510	0.5556714	0.5008655
Random Forest - Sensitive attribute model performance					
Balanced Accuracy (Before)	0.5205672	0.5121004	0.5251818	0.5112115	0.5102468
Balanced Accuracy (After)	0.4983500	0.5004226	0.5251818	0.5089027	0.5102468
Logistic Regression					
Balanced Accuracy (Before)	0.6432656	0.6481056	0.5941323	0.6470951	0.6007241
Balanced Accuracy (After)	0.5055595	0.6196062	0.5948680	0.8460778	0.6889000
Random Forest					
Balanced Accuracy (Before)	0.5963679	0.5845750	0.5982413	0.6002443	0.6180156
Balanced Accuracy (After)	0.5044285	0.6037297	0.6004866	0.7485771	0.7264000
Label changes					Test dataset L
Before Label=0	4611	5265	6636	5352	663
After Label = 0	218	5265	6636	2505	565
Before Label=1	16389	18735	23364	18648	2337
After Label = 1	20782	18735	23364	21495	2435

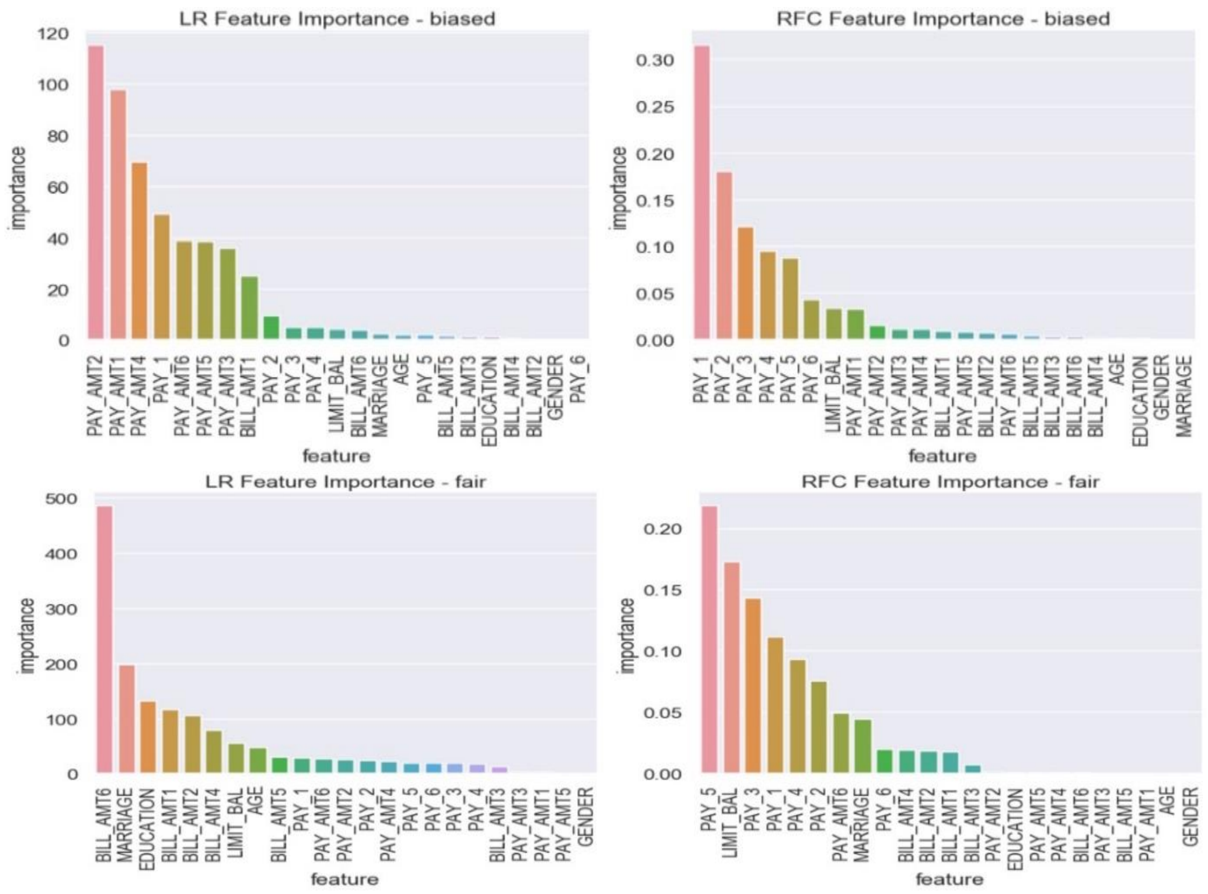
Findings

Adult	LFR	DI	Re-Weighing	Adversarial Debiasing	ROC
Statistical Parity (Before)	0.1929391	0.1929391	0.1906723	0.1942230	0.2003601
Statistical Parity (After)	0.0085174	0.0864186	8.326e-17	0.0343408	0.0350745 LR
					0.0370729 RF
Disparate Impact (Before)	0.3620832	0.3620832	0.3685749	0.3616879	0.3605318
Disparate Impact (After)	0.1198747	0.7838038	1.0000000	0.7974075	0.6281866 LR
					0.5079544 RF
Consistency (Before)	0.8326312	0.8326312	0.8323445	0.8316945	0.8245240
Consistency (After)	0.9997287	0.9120073	0.8323445	0.9591533	0.9578710 LR
					0.9882906 RF
Equality of Odds (Before)	0.1724093	0.1724093	0.1716719	0.0904799	0.327600
Equality of Odds (After)	0.0028328	0.0407425	0.0170516	0.0904670	0.030200
Logistic Regression - Sensitive attribute model performance					
Balanced Accuracy (Before)	0.6952920	0.8003889	0.7019187	0.6893517	0.6957168
Balanced Accuracy (After)	0.4790547	0.7480806	0.7019187	0.6898084	0.6957168
Random Forest - Sensitive attribute model performance					
Balanced Accuracy (Before)	0.7569811	0.7569811	0.7566385	0.7554315	0.7531003
Balanced Accuracy (After)	0.4362819	0.6635963	0.7566385	0.7525725	0.7531003
Logistic Regression					
Balanced Accuracy (Before)	0.6914758	0.6914758	0.6861426	0.6818849	0.6783533
Balanced Accuracy (After)	0.5004230	0.7330637	0.6616281	0.7480738	0.6824996
Random Forest					
Balanced Accuracy (Before)	0.6781526	0.6781526	0.6664366	0.6725080	0.6709555
Balanced Accuracy (After)	0.5002115	0.7092362	0.6469660	0.7403673	0.6709555
Label changes					Test dataset L
Before Label=0	29750	29750	37115	29713	3679
After Label = 0	38805	29750	37115	32897	4481
Before Label=1	9322	9323	11687	9360	1206
After Label = 1	268	9323	11687	6176	404

Findings

Datasets

- The Taiwan dataset showed minimal response to mitigation treatment
- The Adult dataset responded best to bias mitigation treatment
- The German dataset responded worst



Findings

Pre-processing Algorithms

- **Learning Fair Representations**
  - Had the best test results in terms of meeting its all success criteria.
  - Aggressive in modifying the original label and attribute values

- **Disparate Impact**
  - Did not show improvements in mitigating Disparate Impact (DI) across all datasets
  - Inconsistent results across all datasets.

- **Re-weighing**
  - Achieved good results for group fairness across all datasets
  - Less well on individual fairness
  - Showed no improvement in decreasing a classifier’s ability to predict the sensitive attribute – as expected

Findings

In-processing Algorithms

- **Adversarial De-biasing**
    - Had good test results in terms of meeting its all success criteria.
    - Showed little improvement in decreasing a classifier's ability to predict the sensitive attribute.
    - Aggressive in modifying the original label values.

Post-processing Algorithms

- **Reject Option Classification**
    - Had the best test results for the Adult dataset
    - Showed no improvement in decreasing a classifier's ability to predict the sensitive attribute – as expected
    - Aggressive in modifying the original label values.

Personal Perspectives

Different Approach

- Preparation for data discovery and feature engineering
- Common baseline metrics for all datasets
- Common approach in using training / test / validation datasets
- Could have attempted all algorithms and metrics
- Business impacts

Learning

- Bias in machine learning
- Python skills

Future development

- Keep current with the subject in MS / Police work

Future work

Datasets

- Datasheets for datasets, to establish provenance and change log of public datasets.

AIF360 Tools

- To provide the before and after-mitigation falls or gains in classifier accuracy in percentage terms and establish thresholds for acceptable falls in accuracy.
- To measure the impact of label or attribute value changes or allow thresholds to be changed in the mitigation algorithms to limit the number of label changes.
- To automate the findings in this report, and return the various measures of discrimination, and an optimal debiased model accuracy / discrimination trade-off threshold.

Questions?