Created by Murat Durmus (CEO AISOMA)
LinkedIn: https://www.linkedin.com/in/ceosaisoma/

An overview of some useful *Fairness Frameworks* & *Packages*

Content

1. The LinkedIn Fairness Toolkit (LiFT)

The LinkedIn Fairness Toolkit (LiFT) is a Scala/Spark library that enables the measurement of fairness in large scale machine learning workflows. The library can be deployed in training and scoring workflows to measure biases in training data, evaluate fairness metrics for ML models, and detect statistically significant

differences in their performance across different subgroups. It can also be used for ad-hoc fairness analysis.

This library was created by **Sriram Vasudevan** and **Krishnaram Kenthapadi** (work done while at LinkedIn).

More info: **https://github.com/linkedin/LiFT**

2. Fairness-indicators: Tensorflow's Fairness Evaluation and Visualization Toolkit (Google)

Fairness Indicators is designed to support teams in evaluating, improving, and comparing models for fairness concerns in partnership with the broader Tensorflow toolkit.

The tool is currently actively used internally by many of our products. We would love to partner with you to understand where Fairness Indicators is most useful, and where added functionality would be valuable. Please reach out at **tfx@tensorflow.org**.

More info: **https://github.com/tensorflow/fairness-indicators**

3. AI Fairness 360 (IBM)

The AI Fairness 360 toolkit is an extensible open-source library containg techniques developed by the research community to help detect and mitigate bias in machine learning models throughout the AI application lifecycle. AI Fairness 360 package is available in both Python and R.

The AI Fairness 360 package includes

· a comprehensive set of metrics for datasets and models to test for biases,

· explanations for these metrics, and

· algorithms to mitigate bias in datasets and models. It is designed to translate algorithmic research from the lab into the actual practice of domains as wide-ranging as finance, human capital management, healthcare, and education. We invite you to use it and improve it.

More info: **https://github.com/Trusted-AI/AIF360**

4. Fairlearn: Fairness in machine learning mitigation algorithms (Microsoft)

Fairlearn is a Python package that empowers developers of artificial intelligence (AI) systems to assess their system's fairness and mitigate any observed unfairness issues. Fairlearn contains mitigation algorithms as well as a Jupyter widget for model assessment. Besides the source code, this repository also contains Jupyter notebooks with examples of Fairlearn usage.

More Info: **https://github.com/fairlearn/fairlearn**

5. Algofairness

BlackBoxAuditing

This repository contains a sample implementation of Gradient Feature Auditing (GFA) meant to be generalizable to most datasets. For more information on the repair process, see our paper on Certifying and Removing Disparate Impact. For information on the full auditing process, see our paper on Auditing Black-box Models for Indirect Influence.

More info: **https://github.com/algofairness/BlackBoxAuditing**

fairness-comparison

This repository is meant to facilitate the benchmarking of fairness aware machine learning algorithms.

The associated paper is:

A comparative study of fairness-enhancing interventions in machine learning by Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek
Roth. **https://arxiv.org/abs/1802.04422**

More info: **https://github.com/algofairness/fairness-comparison**

fatconference-2019-toolkit-tutorial

More info: **https://github.com/algofairness/fatconference-2019-toolkit-tutorial**

fatconference-2018-auditing-tutorial

More info: **https://github.com/algofairness/fatconference-2018-auditing-tutorial**

runaway-feedback-loops-src

More info: **https://github.com/algofairness/runaway-feedback-loops-src**

Knight

More info: **https://github.com/algofairness/knight**

6. FairSight: Visual Analytics for Fairness in Decision Making

FairSight is a viable fair decision-making system to assist decision makers in achieving fair decision making through the machine learning workflow.

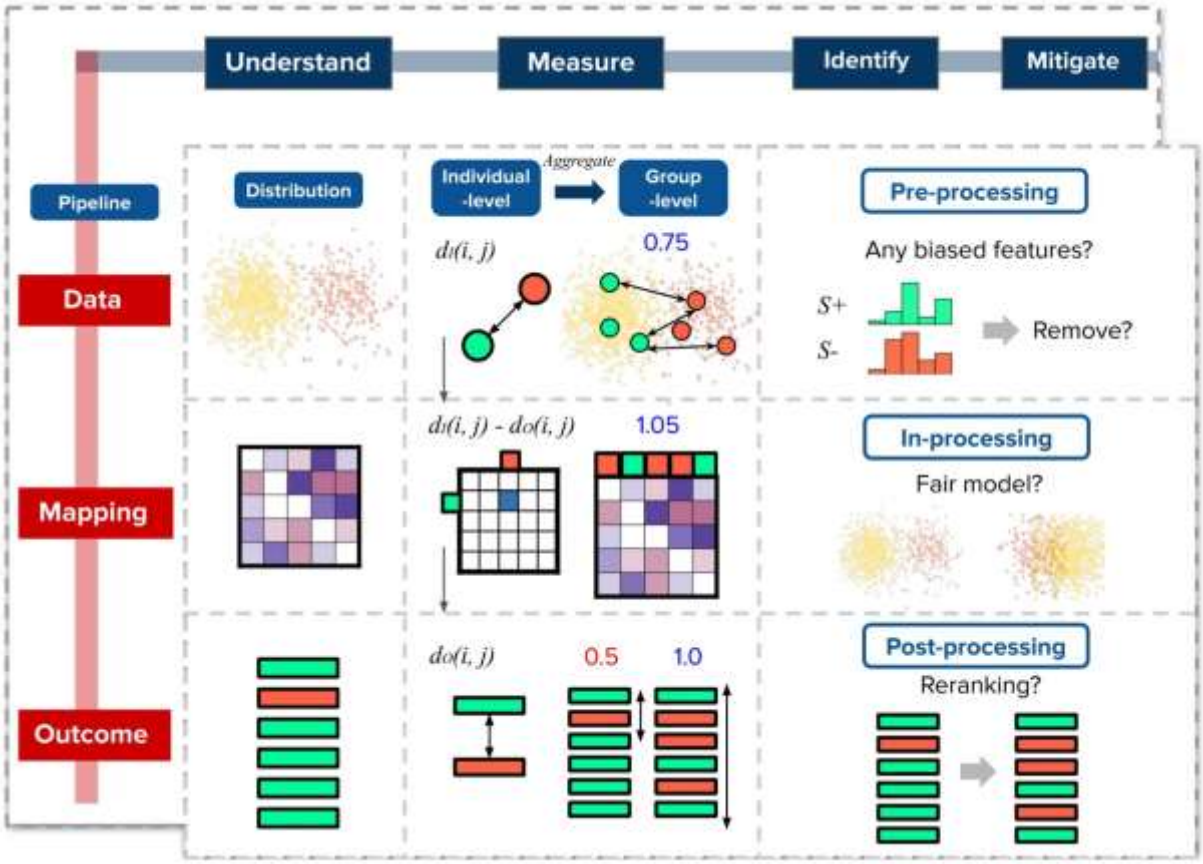
Specification

React: Frontend framework for rendering and communicating with data

django: Python-based backend framework for serving API of data and running machine learning work

scss: The stylesheet grammar for more flexible structure

d3.js: Javascript-based visualization library



More info: **https://github.com/ayong8/FairSight**

7. Aequitas: Bias and Fairness Audit Toolkit

Aequitas is an open-source bias audit toolkit for data scientists, machine learning researchers, and policymakers to audit machine learning models for discrimination and bias, and to make informed and equitable decisions around developing and deploying predictive tools.

More info: **https://github.com/dssg/aequitas**

8. CERTIFAI: A Common Framework to Provide Explanations and Analyse the Fairness and Robustness of Black-box Models

Concerns within the machine learning community and external pressures from regulators over the vulnerabilities of machine learning algorithms have spurred on the fields of explainability, robustness, and fairness. Often, issues in explainability, robustness, and fairness are confined to their specific sub-fields and few tools exist for model developers to use to simultaneously build their modeling pipelines in a transparent, accountable, and fair way. This can lead to a bottleneck on the model developer's side as they must juggle multiple methods to evaluate their algorithms. In this paper, we present a single framework for analyzing the robustness, fairness, and explainability of a classifier. The framework, which is based on the generation of counterfactual explanations1 through a custom genetic algorithm, is flexible, model-agnostic, and does not require access to model internals. The framework allows the user to calculate robustness and fairness scores for individual models and generate explanations for individual predictions which provide a means for actionable recourse (changes to an input to help get a desired outcome). This is the first time that a unified tool has been developed to address three key issues pertaining towards building a responsible artificial intelligence system.

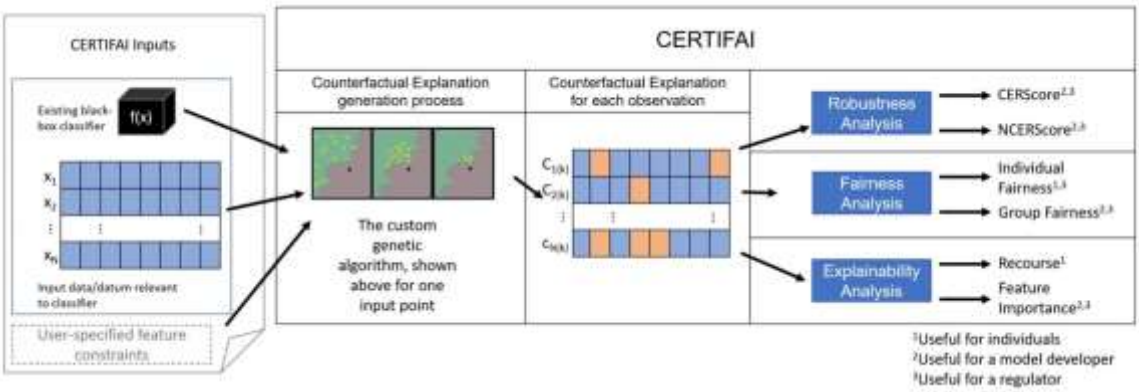


Figure 2: The CERTIFAI framework. Given a black-box ML model and input data along with optional user-specified feature constraints (such as feature type, range, etc.), the method generates counterfactual explanations using a genetic algorithm. The explanations can then be used for three purposes: explainability, fairness and robustness. k represents the number of explanations per input which can be set by the user for recourse purposes and is set to 1 for the feature importance, fairness and robustness analysis. On the right, we show how each of CERTIFAI's attributes is useful for different stakeholders using the tool

More info: **CERTIFAI: A Common Framework to Provide Explanations and Analyse the Fairness and Robustness of Black-box Models**

9. ML-fairness-gym: Google's implementation based on OpenAI's Gym

ML-fairness-gym is a set of components for building simple simulations that explore the potential long-run impacts of deploying machine learning-based decision systems in social environments. As the importance of machine learning fairness has become increasingly apparent, recent research has focused on potentially surprising long term behaviors of enforcing measures of fairness that were originally defined in a static setting. Key findings have shown that under specific assumptions in simplified dynamic simulations, long term effects may in fact counteract the desired goals.

More info: **https://github.com/google/ml-fairness-gym**

10. scikit-fairness

The goal of this project is to attempt to consolidate fairness related metrics, transformers and models into a package that (hopefully) will become a contribution project to scikit-learn.

Fairness, in data science, is a complex unsolved problem for which many tactics are proposed - each with their own advantage and disadvantages. This packages aims to make these tactics readily available, therefore enabling users to try and evaluate different fairness techniques.

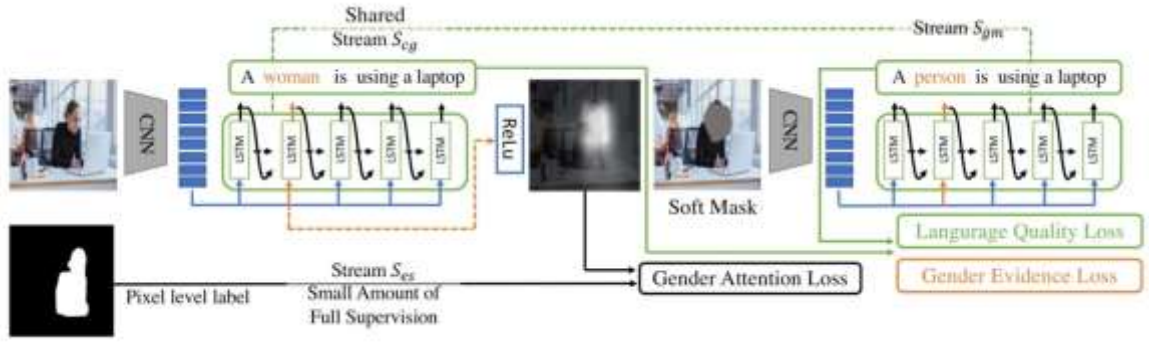
More info: **https://github.com/koaning/scikit-fairness**

11. Mitigating Gender Bias In Captioning System

This is the pytorch implemention for paper “Mitigating Gender Bias In Captioning system”. Recent studies have shown that captioning datasets, such as the COCO dataset, may contain severe social bias which could potentially lead to unintentional discrimination in learning models. In this work, we specifically focus on the gender bias problem.

Image Captioning Model with Guided Attention

We propose a novel Guided Attention Image Captioning model (GAIC) to mitigate gender bias by self-supervising on model’s visual attention. GAIC has two complementary streams to encourage the model to explore correct gender features. The training pipeline can seamlessly add extra supervision to accelerate the self-exploration process. Besides, GAIC is model-agnostic and can be easily applied to various captioning models.



More info:

https://github.com/CaptionGenderBias2020/Mitigating_Gender_Bias_In_Captioning_System

This might be also of interest: **A collection of useful Slides & Quotes on AI-Ethics and XAI**

aisoma.de

Contact

Murat Durmus

CEO & Founder @ AISOMA AG

Frankfurt am Main, Hessen, Deutschland

https://www.linkedin.com/in/ceosaisoma/

murat.durmus@aisoma.de

https://www.aisoma.de