

An Open Source Toolkit for R to Mitigate Discrimination and Bias in Machine Learning Models

Saishruthi Swaminathan

Agenda



Responsible AI



AI Fairness 360



Demo

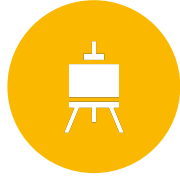


Resources



Responsible AI

- AI Opportunities
 - Increased Revenue
 - Efficiencies
- AI Risks
 - Harm to Users
 - Harm to Business
- A Solution
 - Regulation
 - Ethical & Moral Practices



DESIGN

- Human Centric
- Optimization Metrics



DATA

- Representative
- Protected



MODEL

- Interpretable
- Fair



MONITORING

- Staged rollout
- Feedback loop




ACCOUNTABILITY

- Transparency
- Responsibilities

Responsible ML Pipeline



Responsible AI Benefits

- Prevent harm
 - Build an inclusive product
 - Delightful customer experiences
 - Responsible branding
- 



Trusted AI Committee

LF AI

AlF360 is being
incubated under
Linux Foundation
AI


THE LINUX FOUNDATION PROJECTS

LF AI

About Projects Events People Resources Newsroom

Twitter Search


Projects



Acumos AI
Open source framework to build, share and deploy AI applications

Acumos is an open source platform, which supports design, integration and deployment of AI models. Furthermore, it offers an AI marketplace that empowers data scientists to publish adaptive AI models, while shielding them from the need to custom develop fully integrated solutions.


[Learn More](#)



Adlik
Open source toolkit for accelerating deep learning inference

Adlik is an end-to-end optimizing framework for deep learning models. The goal of Adlik is to accelerate deep learning inference process both on cloud and embedded environments.


[Learn More](#)



Adversarial Robustness Toolbox
Open source tools to evaluate, defend, certify and verify Machine Learning models and applications against adversarial threats

Adversarial Robustness Toolbox (ART) provides tools that enable developers and researchers to evaluate, defend, certify and verify Machine Learning models and applications against the adversarial threats.


[Learn More](#)



AI Explainability 360
Open source toolkit that can help users better understand the ways that machine learning models predict labels

AI Explainability 360 is an open source toolkit that can help users better understand the ways that machine learning models predict labels using a wide variety of techniques throughout the AI application lifecycle.

[Learn More](#)



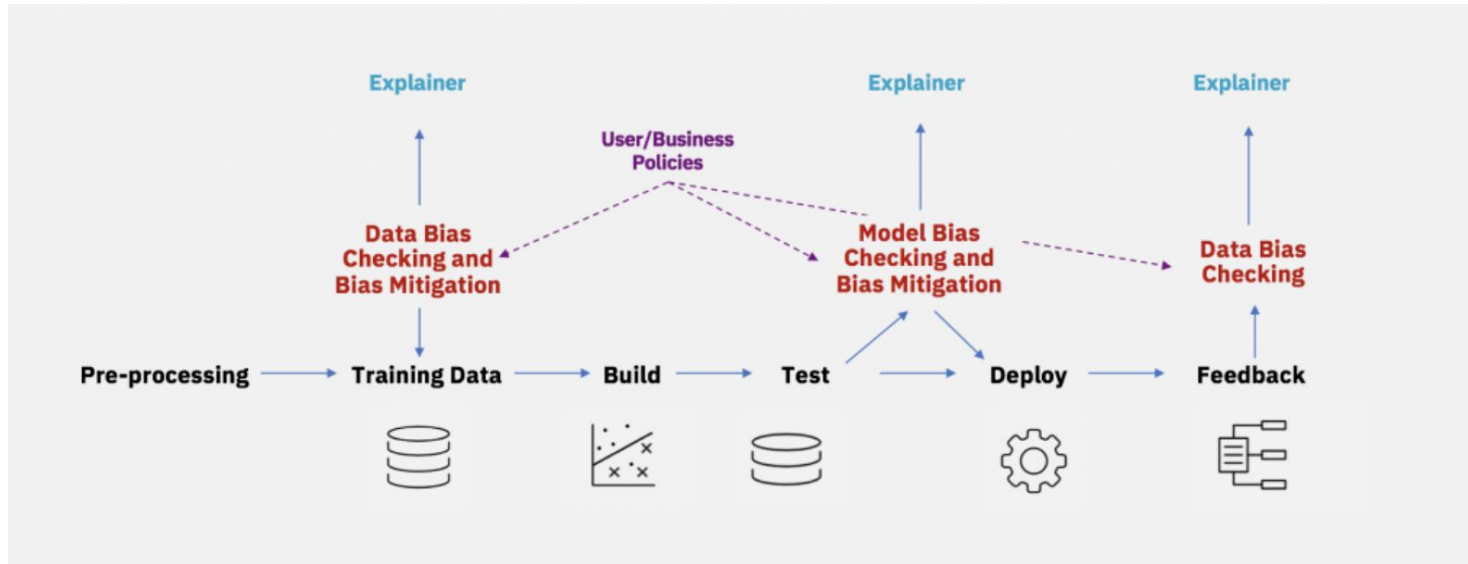
AI Fairness 360
Open source toolkit that can help users understand and mitigate bias in machine learning models throughout the AI application lifecycle

AI Fairness 360 is an extensible open source toolkit that can help users understand and mitigate bias in machine learning models throughout the AI application lifecycle.

[Learn More](#)

AIF₃₆₀

- AIF₃₆₀ toolkit is an open-source library to help detect and remove bias in machine learning models.
- AIF₃₆₀ translates algorithmic research from the lab into practice.
- Applicable domains include finance, human capital management, healthcare, and education.
- Toolbox
 - Fairness metrics
 - Fairness metric explanations
 - Bias mitigation algorithms



Mitigating bias throughout the AI application lifecycle

Metrics

A quantification of unwanted bias in training data or models.

- Individual vs. Group Fairness, or Both
Equal treatment under protected attributes
- Group Fairness: Data vs Model
Measure at different points in ML pipeline: pre-,in-,post-processing
- Group Fairness: We're All Equal vs What You See is What You Get
 - WAE: Predicted future performance is influenced by bias in measurement.
 - WISYWIG: Predicted future performance correlates only with raw score.
- Group Fairness: Ratios vs Differences

Algorithms

- Bias mitigation algorithms attempt to improve the fairness metrics by modifying the training data, the learning algorithm, or the predictions.
- These algorithm categories are known as pre-processing, in-processing, and post-processing, respectively.

Algorithms

Optimized Pre-processing

Use to mitigate bias in training data. Modifies training data features and labels.



Reweighting

Use to mitigate bias in training data. Modifies the weights of different training examples.

Adversarial Debiasing

Use to mitigate bias in classifiers. Uses adversarial techniques to maximize accuracy and reduce evidence of protected attributes in predictions.

Reject Option Classification

Use to mitigate bias in predictions. Changes predictions from a classifier to make them fairer.

Disparate Impact Remover

Use to mitigate bias in training data. Edits feature values to improve group fairness.

Learning Fair Representations

Use to mitigate bias in training data. Learns fair representations by obfuscating information about protected attributes.



Prejudice Remover

Use to mitigate bias in classifiers. Adds a discrimination-aware regularization term to the learning objective.

Calibrated Equalized Odds Post-processing

Use to mitigate bias in predictions. Optimizes over calibrated classifier score outputs that lead to fair output labels.



Equalized Odds Post-processing

Use to mitigate bias in predictions. Modifies the predicted labels using an optimization scheme to make predictions fairer.



Meta Fair Classifier

Use to mitigate bias in classifier. Meta algorithm that takes the fairness metric as part of the input and returns a classifier optimized for that metric.





Using AlF_36o in R

R Package Installation

You can install the **aif360** R package in your machine

Or you can use **Docker** for example and install the package



Using Rstudio
with Docker



docker

1) Install docker: <https://docs.docker.com/get-docker/>

2) Go to terminal and run:

```
docker run -e PASSWORD=yourpassword --rm -p 8787:8787 rocker/rstudio
```

3) Open your browser and type: localhost:8787



Username: rstudio
Password: (the one you defined
above)

A screenshot of the RStudio sign-in interface. It has a title 'Sign in to RStudio'. Below it are two input fields: 'Username:' and 'Password:'. Under the password field is a note: 'You will automatically be signed out after 60 minutes of inactivity.' Below that is a checkbox labeled 'Stay signed in when browser closes'. At the bottom is a blue button with the text 'Sign In'.

Example

```
## 3) Calculate the mean difference
metric_train <- binary_label_dataset_metric(data_aif_train,
                                             privileged_groups = privileged_groups,
                                             unprivileged_groups = unprivileged_groups)

metric_train$mean_difference()
# [1] -0.1932321
# The difference between the proportion of positive outcomes for the unprivileged vs
# the privileged group
#  $P(Y=1|D=unprivileged) - P(Y=1|D=privileged)$ 

## 4) Apply Adversarial debiasing is an in-processing technique that learns a classifier
## to maximize predictor accuracy and simultaneously reduce an adversary's ability to determine
## the protected attribute from the predictions
sess <- tfcompatv1$session()

debiased_model <- adversarial_debiasing(privileged_groups = privileged_groups,
                                       unprivileged_groups = unprivileged_groups,
                                       scope_name = 'debiased_classifier',
                                       debias = TRUE,
                                       sess = sess)

debiased_model$fit(data_aif_train)
# predictions
data_aif_train$debiased <- debiated_model$predict(data_aif_train)

# Right now we are just caring about fairness
metric_preds <- binary_label_dataset_metric(data_aif_train$debiased,
                                             privileged_groups = privileged_groups,
                                             unprivileged_groups = unprivileged_groups)

metric_preds$mean_difference()
# [1] -0.00583602 after
# [1] -0.1932321 before
```

```
adult_dataset.R
1 ## Load the library
2 library(aif360)
3 load_aif360_lib()
4
5 ## Load the data
6 original_data <- readr::read_csv(
7   "https://www.dropbox.com/s/gd8tr1g1j7nrgk/adult_data_preprocessed.csv?dl=1"
8 )
9 original_data <- original_data[, -1]
10 head(original_data)
11 str(original_data)
12
13 # Predict whether income exceeds $50K/yr based on census data.
14 # Variables:
15 # sex: 1 male, 0 female
16 # income binary: 1 >= 50k, 0 <= 50k
17
18 privileged_groups <- list("sex", 1)
19 unprivileged_groups <- list("sex", 0)
20
21 ## 1) Convert the dataframe into the aif360 format -----
22 data_aif <- aif_dataset(data_path = original_data,
23   favor_label = 1,
24   unfavor_label = 0,
25   privileged_protected_attribute = 1,
26   unprivileged_protected_attribute = 0,
27   target_column = "Income Binary",
28   protected_attribute = "sex")
29
30 ## 2) Let's split in train and test -----
31 # train should be 70%
32 # test should be 30%
33 set.seed(1234)
34 data_aif$split <- data_aif$split$num.on.size.splits = list(0.70)
35 data_aif_train <- data_aif$split[[1]]
36 data_aif_test <- data_aif$split[[2]]
37
```


Want to be a part of LFAI Trusted AI Discussion?

<https://wiki.lfai.foundation/display/DL/Trusted+AI+Committee>

Trusted AI Committee

Created by Jacqueline Serafin, last modified by Animesh Singh about 4 hours ago

Overview

Below is an overview of the current discussion topics within the Trusted AI Committee. Further updates will follow as the committee work develops.

- Focus of the committee is on policies, guidelines, tooling and use cases by industry
- Survey and contact current open source Trusted AI related projects to join LF AI efforts
- Create a badging or certification process for open source projects that meet the Trusted AI policies/guidelines defined by LF AI
- Create a document that describes the basic concepts and definitions in relation to Trusted AI and also aims to standardize the vocabulary/terminology

Mail List

Please self subscribe to the mail list here at <https://lists.lfai.foundation/g/trustedai-committee>.

Or email trustedai-committee@lists.lfai.foundation for more information.

Meetings

<https://lists.lfai.foundation/g/trustedai-committee/calendar>

Trusted AI Committee North America Monthly Meeting - 4th Thursday of the month, 10 PM Shenzhen China, 4 PM Paris, 10 AM ET, 7 AM PT USA (updated for daylight savings time as needed)

Zoom info : <https://zoom.us/j/7659717866>

We will starting ONE Monthly Call for Europe/Asia Time Zone, every 2nd Thursday of the month (Starting in November - time to be determined)

Join Principles Working Group

<https://wiki.lfai.foundation/display/DL/Principles+Working+Group>

Principles Working Group

Created by Susan Malaika, last modified on Jul 09, 2020

The working group meets every other Wednesday at 11am US Eastern - Please contact Susan Malaika malaika@us.ibm.com if you would like to join

The working group is made up of:

- Souad Ouali (Chair)
- Jeff Cao
- Francois Jezequel
- Sarah Luger
- Susan Malaika
- Alka Roy
- Alejandro Saucedo
- Marta Ziosi

Slack

AIF36o

https://join.slack.com/t/aif36o/shared_invite/zt-5hfvuafo-Xo~g6tgJQ~7tIAT~S294TQ

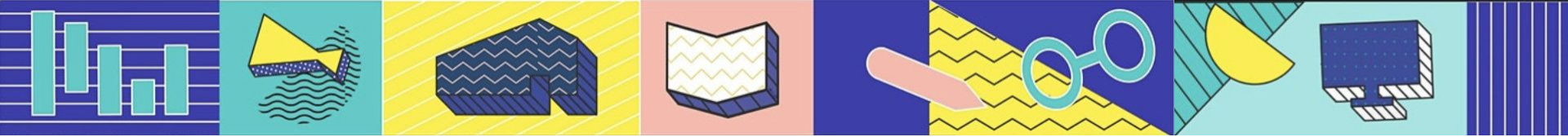
AIX36o

https://join.slack.com/t/aix36o/shared_invite/enQtNzEyOTAwOTk1NzY2LTM1ZTMwM2M4OWQzNjhmNGRiZjg3MmJiYTAzNDU1MTRiYTlyMjFhZTl4ZDUwM2M1MGYyODkwNzQ2OWQzMThlN2Q

ART

https://join.slack.com/t/ibm-art/shared_invite/enQtMzkyOTkyODE4NzM4LTA4NGQ1OTMxMzFmY2Q1MzE1NWlzMmEzN2FjNGNjOGVlODVhZDE0MjA1NTA4OGVhMjVhNmQ4MTY1NmMyOGM5YTg

- Trusted AI Wiki - <https://wiki.lfai.foundation/display/DL/Trusted+AI+Committee>
- LFAI – <https://www.linuxfoundation.org/projects/>
- LFAI Trusted AI - <https://lfai.foundation/projects/trusted-ai/>
- Trusted AI Announce – <https://lists.lfai.foundation/g/trusted-ai-360-announce>
- Trusted AI Technical Discussions: <https://lists.lfai.foundation/g/trusted-ai-360-technical-discuss>
- Trusted AI Technical Steering Committee: <https://lists.lfai.foundation/g/trusted-ai-360-tsc>



Thank You!



[linkedin.com/in/saishruthi-swaminathan](https://www.linkedin.com/in/saishruthi-swaminathan)