Trusted AI: Robust, Unbiased and Reproducible AI through Open Source

Dr. Margriet Groenendijk
Data & AI Developer Advocate

IBM **Developer**

IBM

AI is now used in many high-stakes decision making applications

Credit

Employment

Admission

Sentencing

Healthcare

What does it take to trust a decision made by a machine?

(Other than that it is 99% accurate)?

Did anyone tamper with it?

Alt Text

How would you describe this object and its context to someone who is blind?

(1-2 sentences recommended)

A group of brown and white dog

Description automatically generated

Mark as decorative

Generate a description for me

(a) Husky classified as wolf

(b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

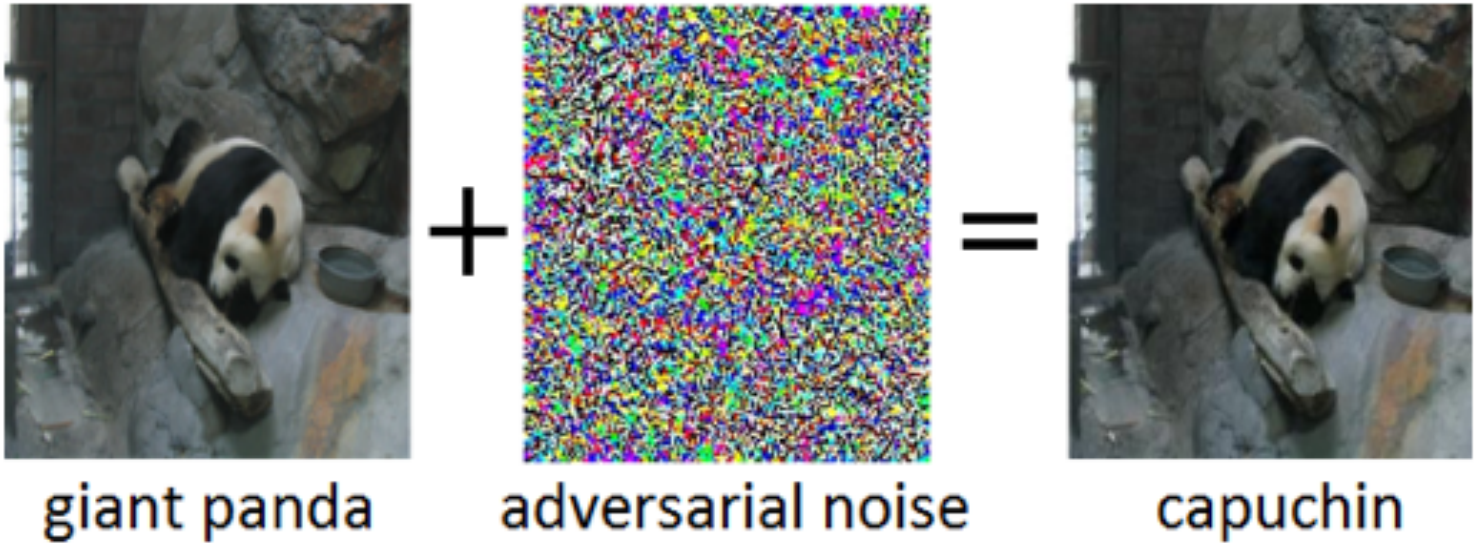
	Before	After
Trusted the bad model	10 out of 27	3 out of 27
Snow as a potential feature	12 out of 27	25 out of 27

https://hackernoon.com/dogs-wolves-data-science-and-why-machines-must-learn-like-humans-do-41c43bc7f982

Robust AI Example: Misclassification

Adversarial machine learning

Adversarial machine learning can be used to “trick” machine learning models into providing incorrect predictions



https://www.ibm.com/blogs/research/2018/04/ai-adversarial-robustness-toolbox/

https://arxiv.org/pdf/1707.08945.pdf

What does it take to trust a decision made by a machine?

(Other than that it is 99% accurate)?

Did anyone tamper with it?

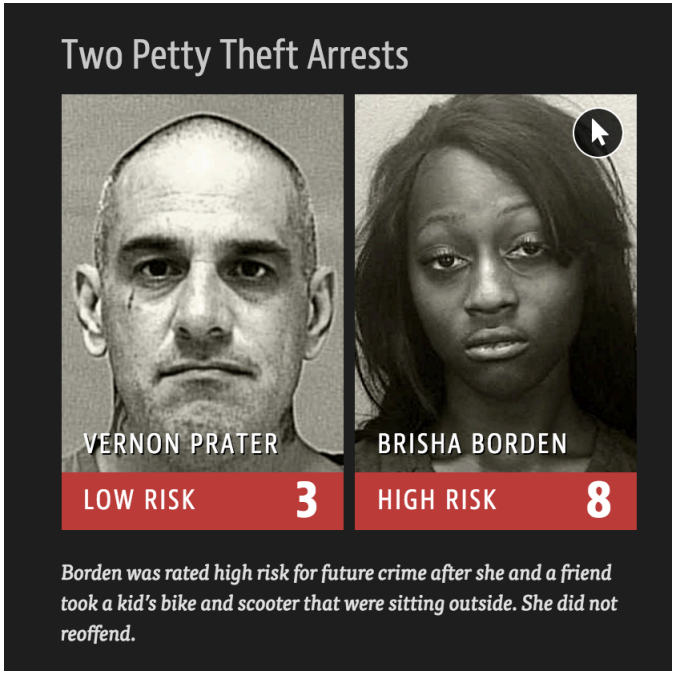
Is it fair?

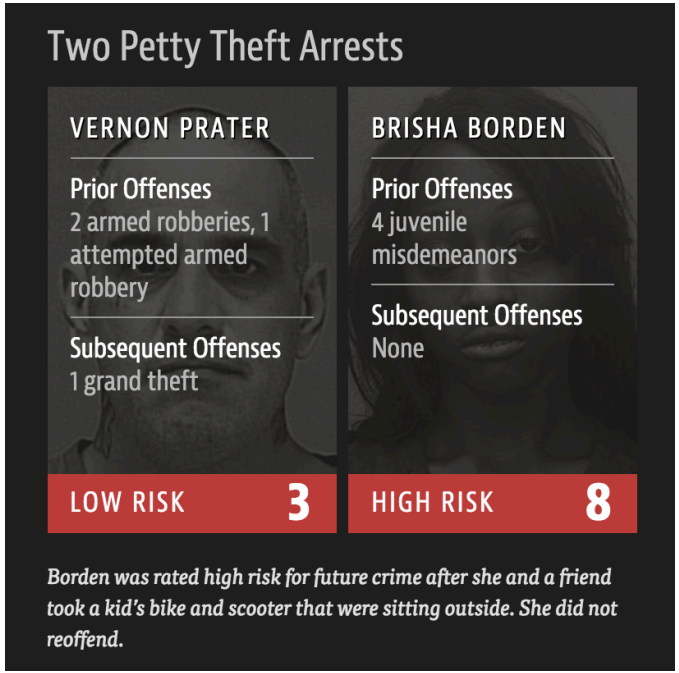
Bias in AI Example: Criminal Justice System

Since 2008, nearly every arrestee in Broward County, Florida has been assigned a risk score using Northpointe's COMPAS algorithm.

Defendants with low risk scores are released on bail.

It falsely flagged black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants





Machine Bias — ProPublica <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.

What does it take to trust a decision made by a machine?

(Other than that it is 99% accurate)?

Did anyone tamper with it?

Is it fair?

Is it easy to understand?

What does it take to trust a decision made by a machine?

(Other than that it is 99% accurate)?

Did anyone tamper with it?

Is it fair?

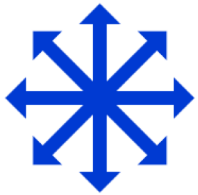
Is it easy to understand?

Is it accountable?

Trusted AI Lifecycle through Open Source

Pillars of trust, woven into the lifecycle of an AI application

Did anyone tamper with it?



ROBUSTNESS

Adversarial Robustness 360

↳ (ART)

github.com/IBM/adversarial-robustness-toolbox

art-demo.mybluemix.net

Is it fair?

FAIRNESS

AI Fairness 360

↳ (AIF360)

github.com/IBM/AIF360

aif360.mybluemix.net

Is it easy to understand?

EXPLAINABILITY

AI Explainability 360

↳ (AIX360)

github.com/IBM/AIX360

aix360.mybluemix.net

Is it accountable?

LINEAGE

In the works!

IBM also has a long history in the open source ecosystem

and

We are leveraging this to bring Trust and Transparency into AI though Open Source..

AI Fairness 360
↳ (AIF360)

https://github.com/IBM/AIF360

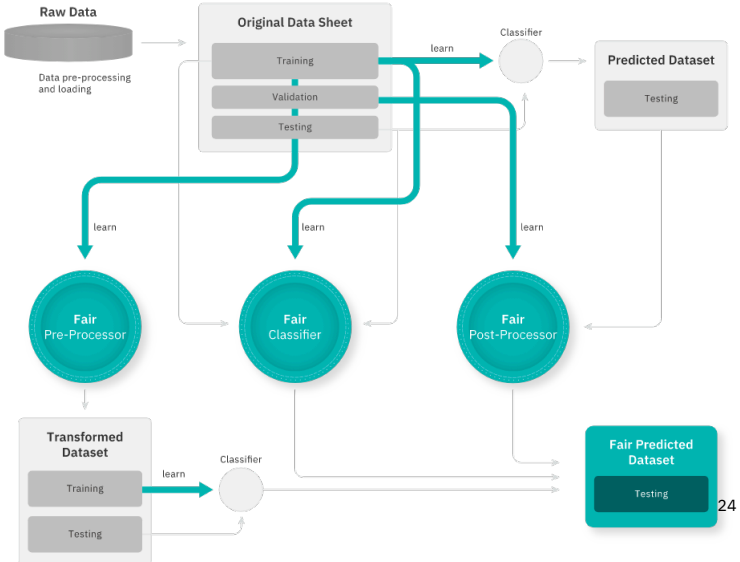
AIF360 toolkit is an open-source library to help detect and remove bias in machine learning models. **AIF360 translates algorithmic research from the lab into practice**. Applicable domains include finance, human capital management, healthcare, and education.

The **AI Fairness 360** Python 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.

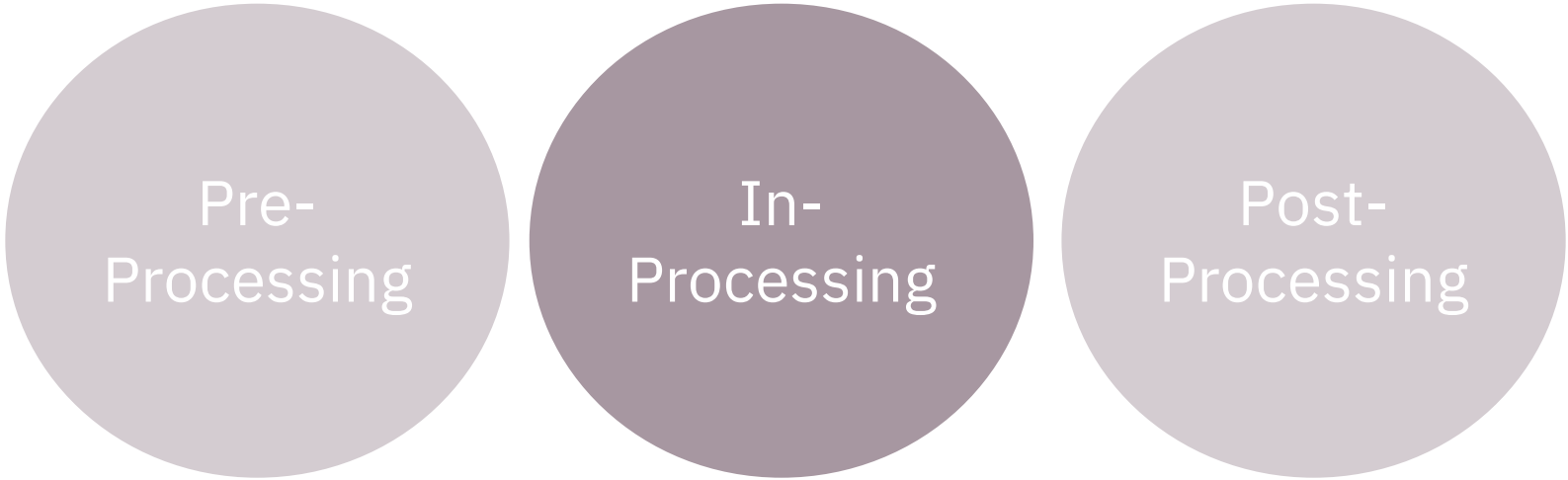
Toolbox
Fairness metrics (70+)
Fairness metric explanations
Bias mitigation algorithms (10+)

http://aif360.mybluemix.net/

AIF360



Machine Learning Pipeline



Modifying the training data.

Modifying the learning algorithm.

Modifying the predictions (or outcomes.)

http://aif360.mybluemix.net/

IBM Research Trusted AI | Home | **Demo** | Resources | Events | Videos | Community

AI Fairness 360 - Demo

Data — Check — Mitigate — Compare

Next

1. Choose sample data set

Bias occurs in data used to train a model. We have provided three sample datasets that you can use to explore bias checking and mitigation. Each dataset contains attributes that should be protected to avoid bias.

◉ **Compas (ProPublica recidivism)**

Predict a criminal defendant's likelihood of reoffending.
Protected Attributes:
- **Sex**, privileged: ***Female***, unprivileged: ***Male***
- **Race**, privileged: ***Caucasian***, unprivileged: ***Not Caucasian***

Learn more

○ **German credit scoring**

Predict an individual's credit risk.
Protected Attributes:
- **Sex**, privileged: ***Male***, unprivileged: ***Female***
- **Age**, privileged: ***Old***, unprivileged: ***Young***

Learn more

○ **Adult census income**

Predict whether income exceeds $50K/yr based on census data.
Protected Attributes:
- **Race**, privileged: ***White***, unprivileged: ***Non-white***

http://aif360.mybluemix.net/

IBM Research Trusted AI | Home | **Demo** | Resources | Events | Videos | Community

AI Fairness 360 - Demo

Data — Check — Mitigate — Compare

Back Next

3. Choose bias mitigation algorithm

A variety of algorithms can be used to mitigate bias. The choice of which to use depends on whether you want to fix the data (pre-process), the classifier (in-process), or the predictions (post-process). Learn more about how to choose.

◉ **Reweighing**

Weights the examples in each (group, label) combination differently to ensure fairness before classification.

○ **Optimized Pre-Processing**

Learns a probabilistic transformation that can modify the features and the labels in the training data.

http://aif360.mybluemix.net/

IBM Research Trusted AI | Home | **Demo** | Resources | Events | Videos | Community

AI Fairness 360 - Demo

Data — Check — Mitigate — Compare

Back

4. Compare original vs. mitigated results

Dataset: Compas (ProPublica recidivism)

Mitigation: **Reweighing algorithm applied**

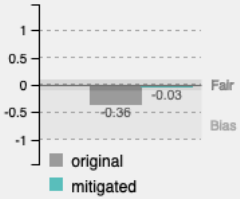
Protected Attribute: Sex

Privileged Group: ***Female***, Unprivileged Group: ***Male***

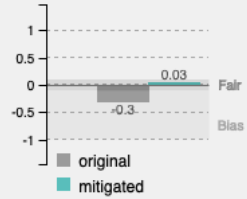
Accuracy after mitigation changed from 66% to 65%

Bias against unprivileged group was reduced to acceptable levels* for 4 of 4 previously biased metrics (0 of 5 metrics still indicate bias for unprivileged group)

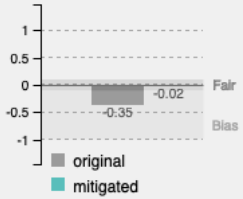
***Statistical Parity Difference**



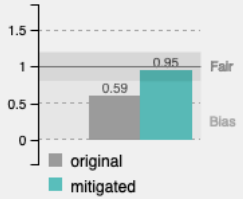
***Equal Opportunity Difference**



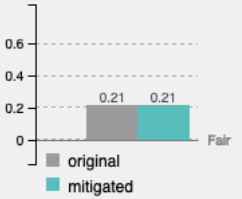
***Average Odds Difference**



***Disparate Impact**



Theil Index



Cookie Preferences

AI Explainability 360
↳ (AIX360)

https://github.com/IBM/AIX360

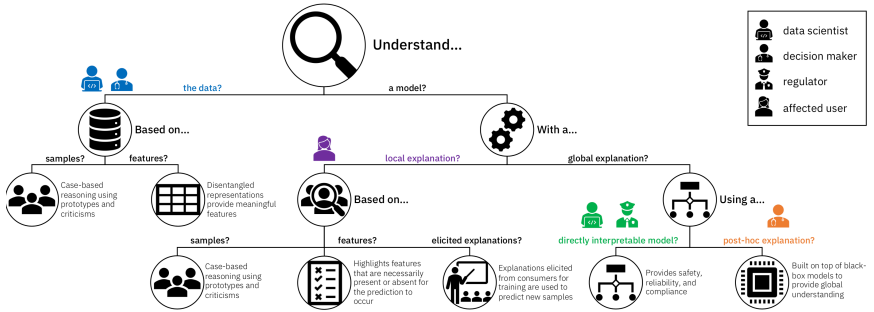
AIX360

AIX360 toolkit is an open-source library to help explain AI and machine learning models and their predictions. This includes three classes of algorithms: local post-hoc, global post-hoc, and directly interpretable explainers for models that use image, text, and structured/tabular data.

The **AI Explainability360** Python package includes a comprehensive set of explainers, both at global and local level.

Toolbox
Local post-hoc
Global post-hoc
Directly interpretable

http://aix360.mybluemix.net



AIX360: Different Ways to explain

One explanation does not fit all

Different stakeholders require explanations for different purposes and with different objectives, and explanations will have to be tailored to their needs.

End users/customers (trust)

Doctors: Why did you recommend this treatment?

Customers: Why was my loan denied?

Teachers: Why was my teaching evaluated in this way?

Gov’t/regulators (compliance, safety)

Prove to me that you didn't discriminate.

Developers (quality, “debuggability”)

Is our system performing well?

How can we improve it?

http://aix360.mybluemix.net/

IBM Research Trusted AI | Home | **Demo** | Resources | Events | Community

AI Explainability 360 - Demo

Data — Consumer — Explanation

Next

Data: FICO Explainable Machine Learning Challenge

Machine learning models are used to support an increasing number of important decisions. These decisions are consumed by various users, who may have different needs and require different kinds of explanations. For this reason, AI Explainability 360 offers a collection of algorithms that provide diverse ways of explaining decisions generated by machine learning models.

To explore these different types of algorithmic explanations, we consider an AI-powered credit approval system using the FICO Explainable Machine Learning Challenge dataset and probe into it from the perspective of different users. We illustrate how different users – a data scientist, a loan officer, and a bank consumer – require different explanations.

FICO, a credit scoring company, released an anonymized dataset of Home Equity Line of Credit (HELOC) applications made by real homeowners. A HELOC is a line of credit typically offered by a bank as a percentage of home equity (the difference between the current market value of a home and the outstanding balance of all liens, e.g., mortgages). The customers in this dataset have requested a credit line in the range of $5,000 - $150,000. The fundamental task is to use the information about the applicant in their credit report to predict whether they will make timely payments over a two-year period. This is the machine learning task that we focus on. The machine learning prediction is then used by loan officers to decide whether the homeowner qualifies for a line of credit and, if so, how much credit should be extended. Learn more about the dataset.

http://aix360.mybluemix.net/

IBM Research Trusted AI | Home | **Demo** | Resources | Events | Community

AI Explainability 360 - Demo

Data — Consumer — Explanation

Back Next

Choose a consumer type

Data Scientist
must ensure the model works appropriately before deployment

Loan Officer
needs to assess the model's prediction and make the final judgement

Bank Customer
wants to understand the reason for the application result

http://aix360.mybluemix.net/

IBM Research Trusted AI | Home | **Demo** | Resources | Events | Community

AI Explainability 360 - Demo

Data — Consumer — Explanation

Back

A Bank Customer wants to understand:

Why was my application rejected?
What can I improve to increase the likelihood my application is accepted?

Providing Contrastive Explanationsfor Insight into Loan Application Outcomes

The Bank Customer wants to know how and why the decision was made to accept or reject their loan application. The explanation given will help them understand if they've been treated fairly, and also provide insight into what – if their application was rejected – they can improve in order to increase the likelihood it will be accepted in the future. To help provide that insight and suggest avenues for improvement, we will use the Contrastive Explanations Method (CEM) algorithm available in AI Explainability 360. This algorithm sits on top of an existing predictive model and helps detect both the features that a bank customer could improve (e.g., amount of time since last credit inquiry, average age of accounts), and also further detects the features that will increase the likelihood of approval and those that are within reach for the customer. See examples below.

http://aix360.mybluemix.net/

IBM Research Trusted AI | Home | **Demo** | Resources | Events | Community

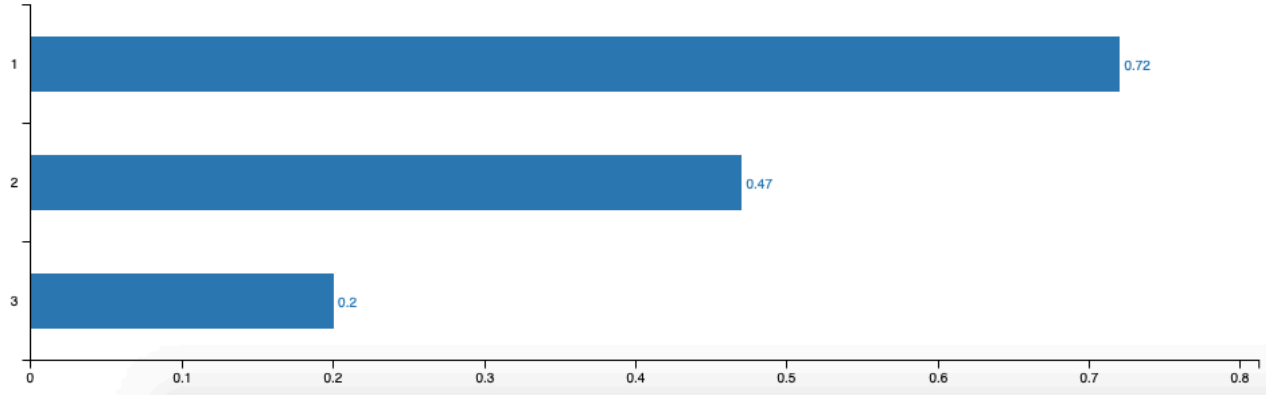
Several features in Jason's application fall outside the acceptable range. All would need to improve before acceptance was recommended.

Factors contributing to Jason's application denial

1. The value of **Consolidated risk markers** is **65**. It needs to be around **72** for the application to be approved.
2. The value of **Average age of accounts in months** is **52**. It needs to be around **68** for the application to be approved.
3. The value of **Months since most recent credit inquiry not within the last 7 days** is **2**. It needs to be around **3** for the application to be approved.

Relative importance of factors contributing to denial

While all three factors need to improve as indicated above, the most important to improve first is the Consolidated risk markers.
Jason now has insight into what he can do to improve his likelihood of being accepted.



We are also making these capabilities around Trusted AI available to businesses through

Watson OpenScale

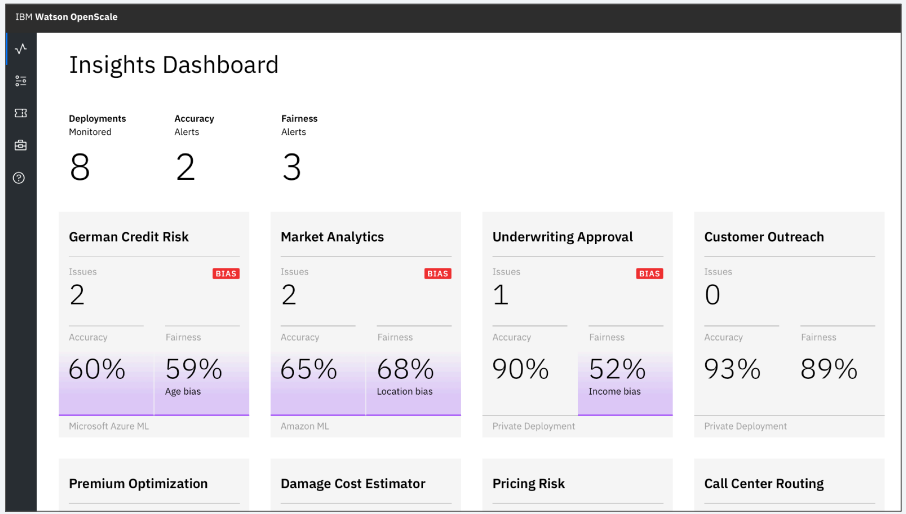
Watson OpenScale **tracks and measures trusted AI outcomes across its lifecycle**, and adapts and governs AI to changing business situations — for models built and running anywhere.

Measure and track AI outcomes

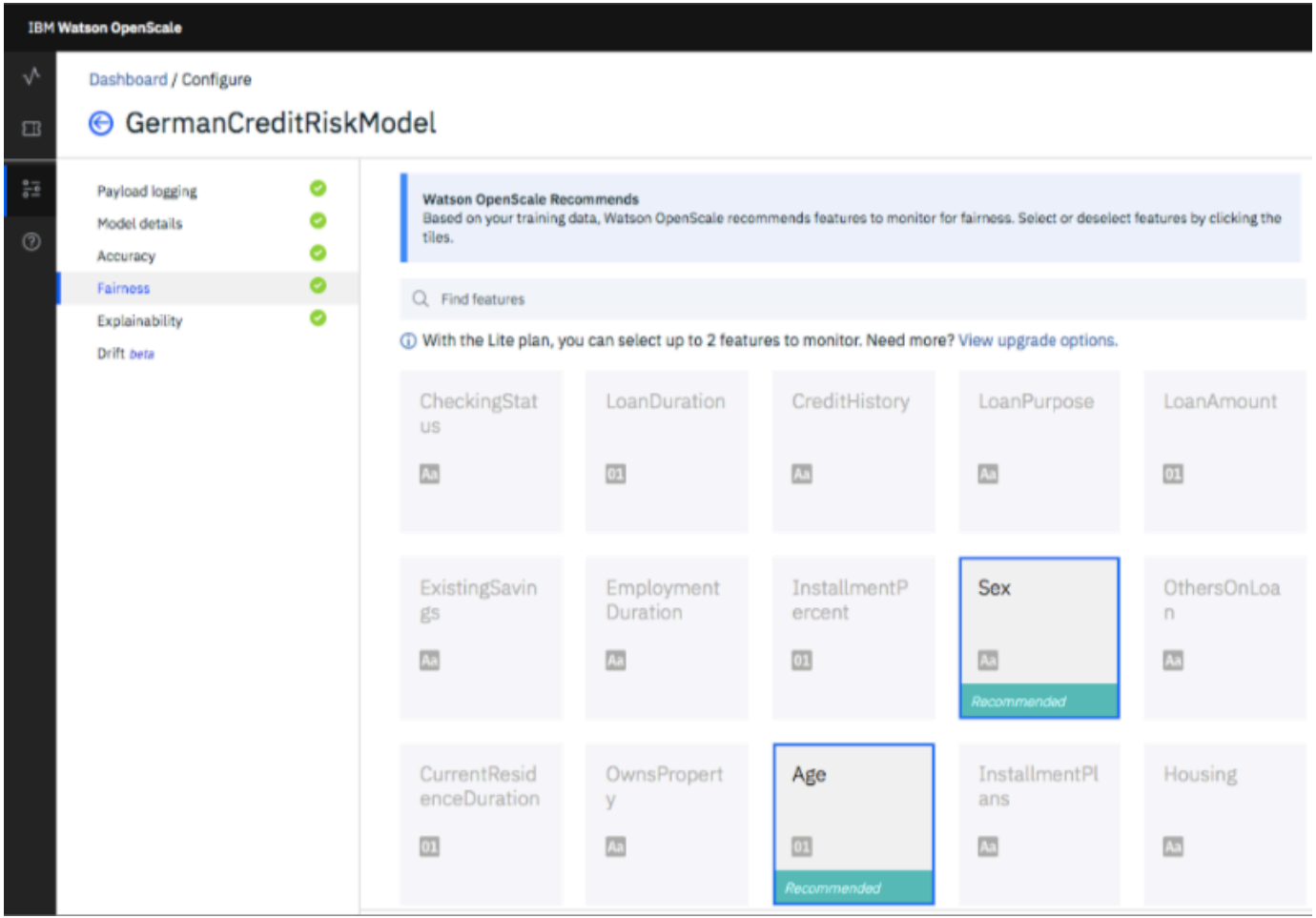
Track performance of production AI and its impact on business goals, with actionable metrics in a single console.

Govern, detect bias and explain AI

Maintain **regulatory compliance** by **tracing and explaining AI decisions** across workflows, and intelligently **detect and correct bias** to improve outcomes.



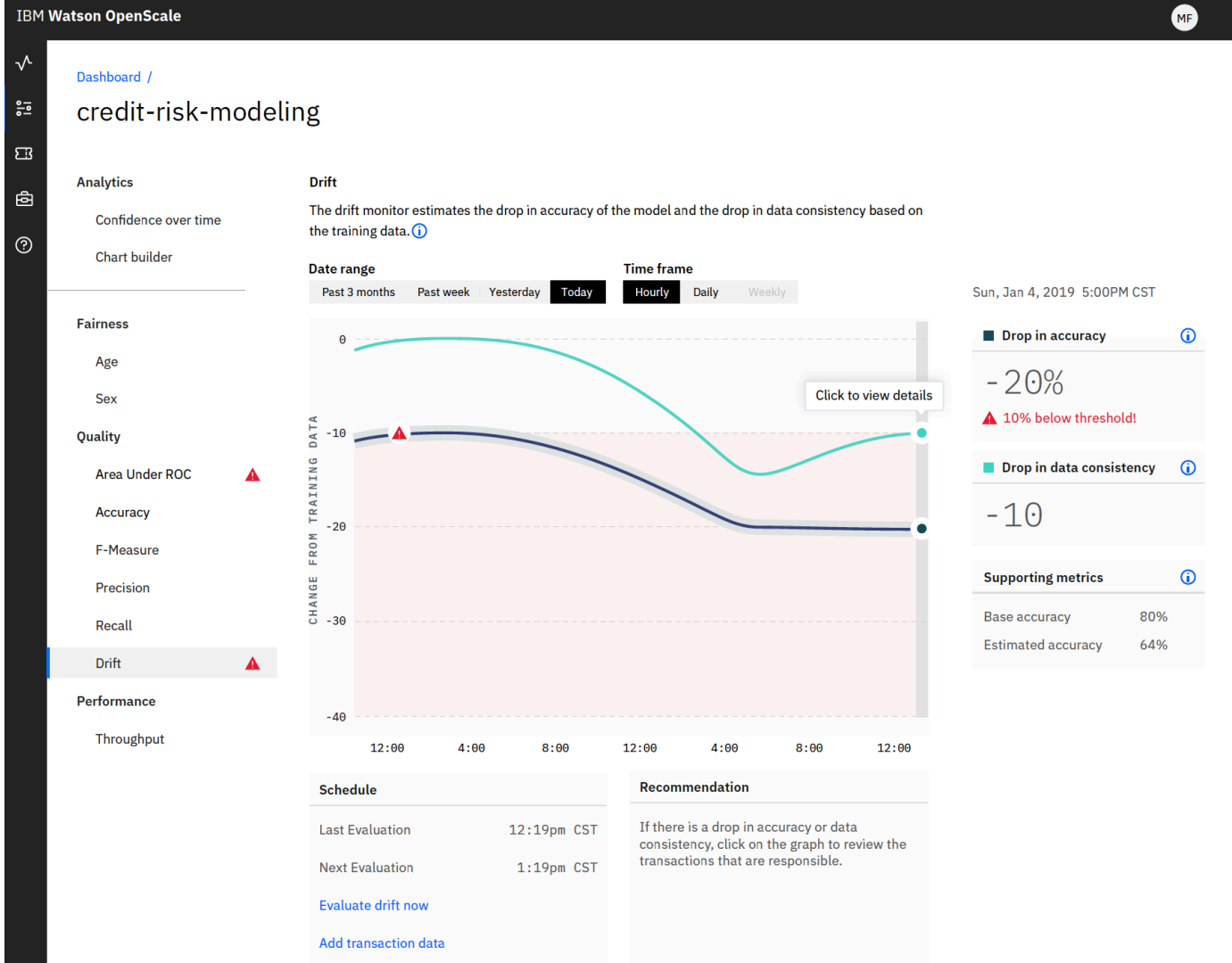
Biases from correlated attributes in a model, so that user does not miss unknown biases in the model



Fairness


GermanCreditRiskModel
Model ID: 34be7106-2f0c-4118-b46d-77e11abba8f6
Created date: 6/19/2019
Configure monitors
Fairness
Age
Sex
Quality
Area under ROC
Area under PR
Accuracy
True positive rate (TPR)
False positive rate (FPR)
Recall
Precision
F1-Measure
Logarithmic loss
Performance
Throughput
Analytics
Predictions by Confidence
Chart Builder
Fairness for Sex
The models propensity to deliver favorable outcomes to one group over another. Learn more.
Time frame
Hourly
Daily
Weekly
Date range
Past 3 months
Past week
Yesterday
Today
100%
75%
50%
25%
0%
FAIRNESS SCORE
Jun 12, 10:00 AM
Jun 14, 4:00 AM
Jun 15, 5:00 PM
Jun 17, 6:00 AM
TIME [PAST WEEK]
Fairness Score for Sex
99%
1% above threshold
Wed, Jun 19, 2019, 2:00 AM EDT
Threshold 98%
Monitored Groups
Average 99%
female 99%
Schedule
Last Evaluation 10:18 AM EDT
Next Evaluation 11:18 AM EDT
Check fairness now
Make a scoring request

Drift

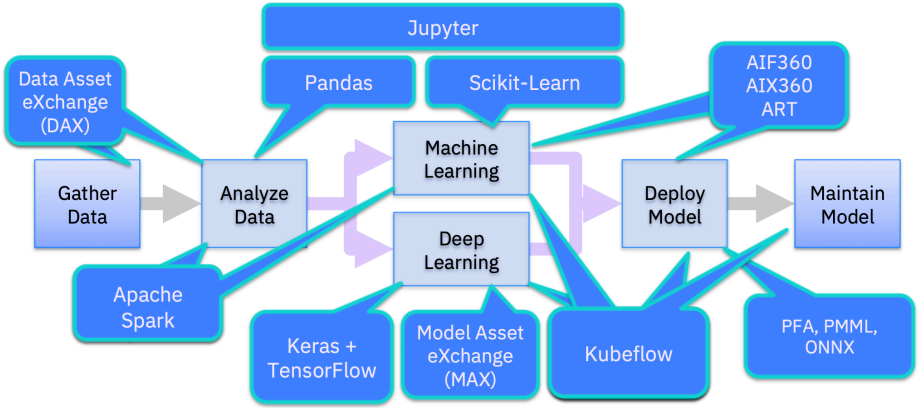

IBM Watson OpenScale
MF
Dashboard /
credit-risk-modeling
Analytics
Confidence over time
Chart builder
Fairness
Age
Sex
Quality
Area Under ROC
Accuracy
F-Measure
Precision
Recall
Drift
Performance
Throughput
Drift
The drift monitor estimates the drop in accuracy of the model and the drop in data consistency based on the training data.
Date range
Past 3 months
Past week
Yesterday
Today
Time frame
Hourly
Daily
Weekly
Sun, Jan 4, 2019 5:00PM CST
0
-10
-20
-30
-40
CHANGE FROM TRAINING DATA
Click to view details
12:00
4:00
8:00
12:00
4:00
8:00
12:00
Drop in accuracy
-20%
10% below threshold!
Drop in data consistency
-10
Supporting metrics
Base accuracy 80%
Estimated accuracy 64%
Schedule
Last Evaluation 12:19pm CST
Next Evaluation 1:19pm CST
Evaluate drift now
Add transaction data
Recommendation
If there is a drop in accuracy or data consistency, click on the graph to review the transactions that are responsible.

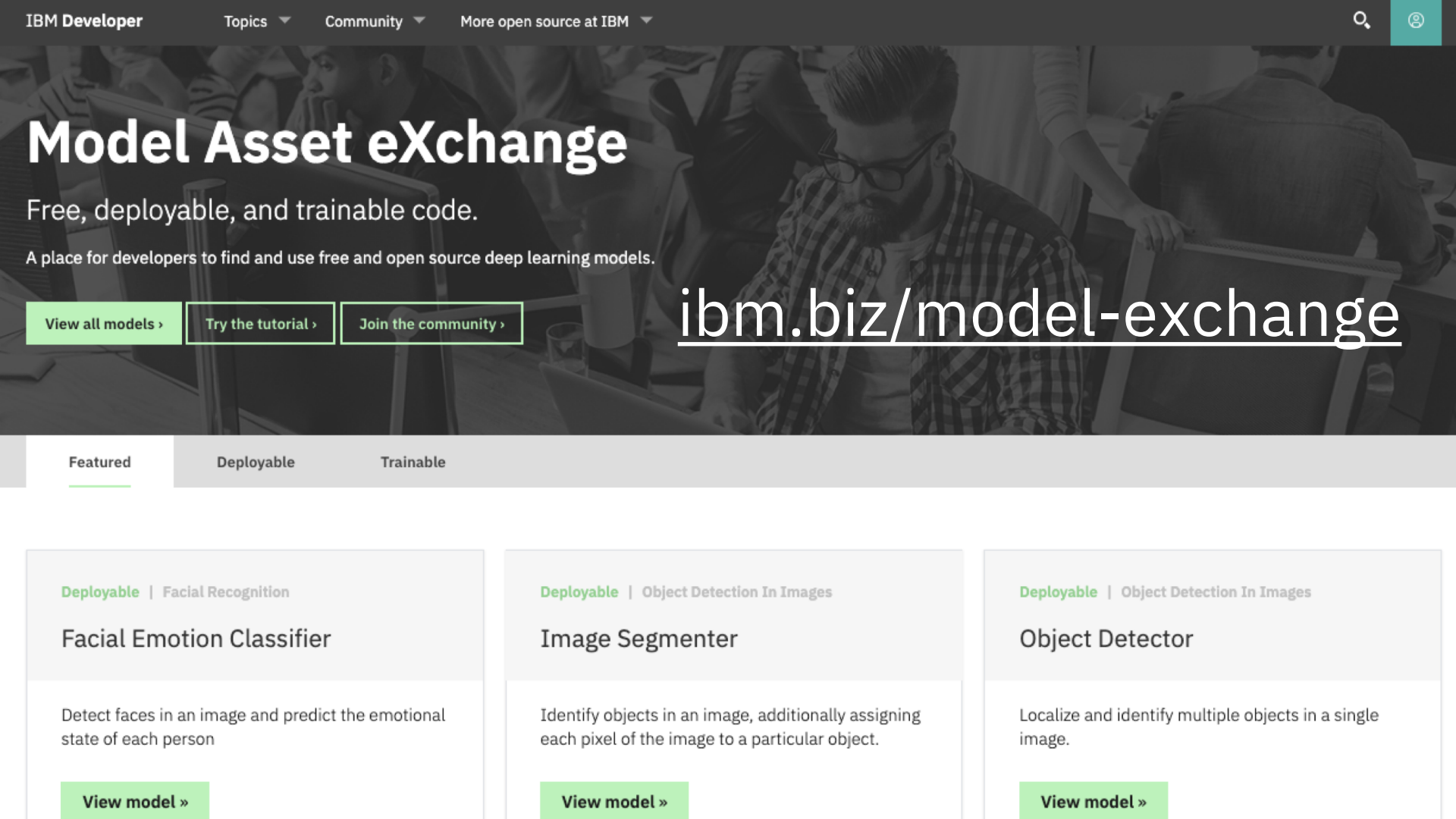
Center for Open Source Data and AI Technologies

CODAIT aims to make AI solutions dramatically easier to create, deploy, and manage in the enterprise.

CODAIT

Center for Open Source Data and AI Technologies




IBM Developer
Topics
Community
More open source at IBM
Model Asset eXchange
Free, deployable, and trainable code.
A place for developers to find and use free and open source deep learning models.
View all models ›
Try the tutorial ›
Join the community ›
ibm.biz/model-exchange
Featured
Deployable
Trainable
Deployable | Facial Recognition
Facial Emotion Classifier
Detect faces in an image and predict the emotional state of each person
View model »
Deployable | Object Detection In Images
Image Segmenter
Identify objects in an image, additionally assigning each pixel of the image to a particular object.
View model »
Deployable | Object Detection In Images
Object Detector
Localize and identify multiple objects in a single image.
View model »

MIT | CSV

Fashion-MNIST

A dataset of standardized images of fashion items from 10 classes

View dataset »

CDLA-Sharing | CoNLL-U

Contracts Proposition Bank

Text from approximately 1000 English compliance sentences obtained from IBM's publicly available contracts, annotated with a layer of "universal" semantic role labels.

CDLA-Sharing | CSV

NOAA Weather Data – JFK Airport

Local climatological data originally collected by JFK airport.

View dataset »

Trust and Transparency into AI though Open Source

We would like to partner with community to build Trusted and Transparent AI

To collaborate, look at the corresponding projects here

codait.org

or

https://github.com/topics/trusted-ai

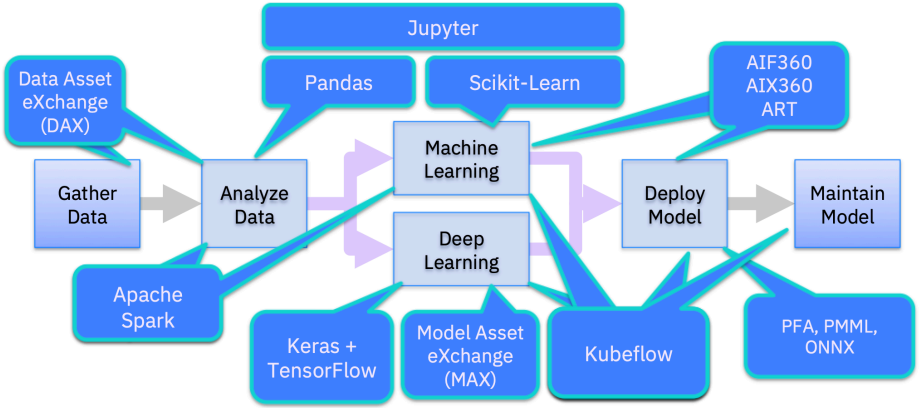
and reach out via github or send an email to

mgroenen@uk.ibm.com

@MargrietGr

CODAIT

Center for Open Source Data and AI Technologies




IBM