Trusted AI:
Addressing Bias, Explainability and Robustness in Machine Learning Models

January 28, 2021

The session starts at 12:00pm.

Trusted AI

Addressing Bias, Explainability and Robustness in Machine Learning Models

Topics

- **What is Machine Learning?**
- Model Fairness
  - AIF360 Toolkit
- Model Explainability
  - AIX360 Toolkit
- Model Robustness
  - Adversarial Robustness Toolbox
- Watson OpenScale
- Lab Overview

What is Machine Learning?

"Computer that learn without being explicitly programmed"

data data data data data data

Historical Data + Outcome

Training

Algorithm

Trained model

1 A machine learning model is trained to recognize patterns in historical data

data data data data data data

New Data

Trained model

prediction or classification

2 The model is then shown new data and asked to predict or classify it. If the patterns In the new data match the training data then the model makes accurate predictions.

Machine Learning is used in many high-stakes decision-making applications

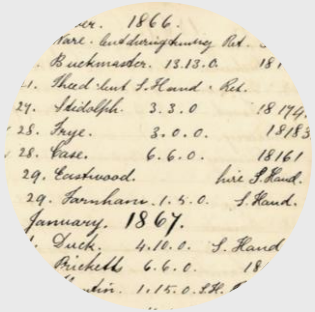
Credit

Employment

Healthcare

Self-Driving Cars

Our vision for trusted Machine Learning Models

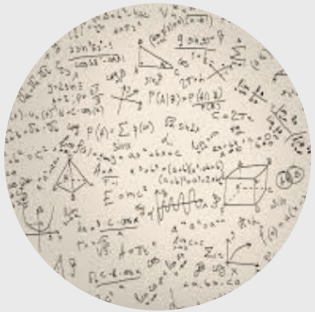


Is it accurate?

- Watson OpenScale

Is it fair?

- AIF360
- Watson OpenScale



Is it easy to understand?

- AIX360
- Watson OpenScale

Did anyone tamper with it?

- ART

Topics

- What is Machine Learning?
- **Model Fairness**
  - AIF360 Toolkit
- Model Explainability
  - AIX360 Toolkit
- Model Robustness
  - Adversarial Robustness Toolbox
- Watson OpenScale
- Lab Overview

Model Fairness

Machine Learning is a form of statistical discrimination by nature

Sometimes this discrimination becomes objectionable or possibly illegal

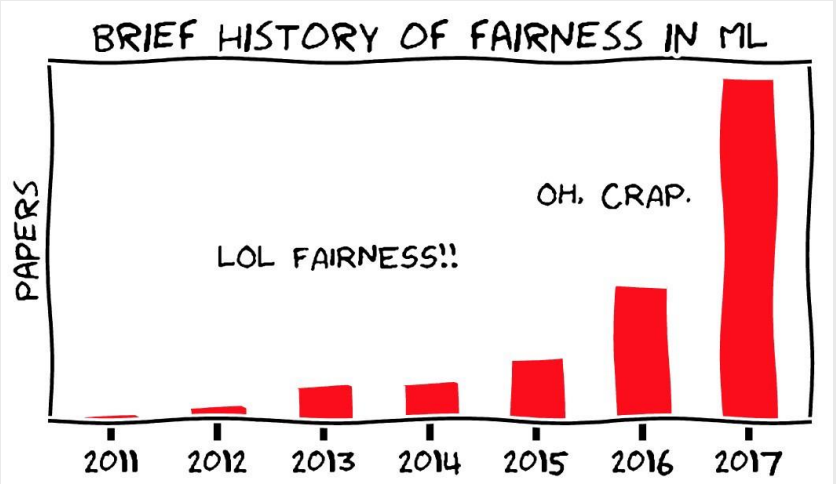
— Systematically favoring privileged groups like Caucasian male

Mitigating bias can be done on:

— training data*: pre-processing*

— the learned model: *in-processing*

— the model outcomes: post-processing



Source: A Tutorial on Fairness in Machine Learning

Example - Hiring

Source: A Tutorial on Fairness in Machine Learning

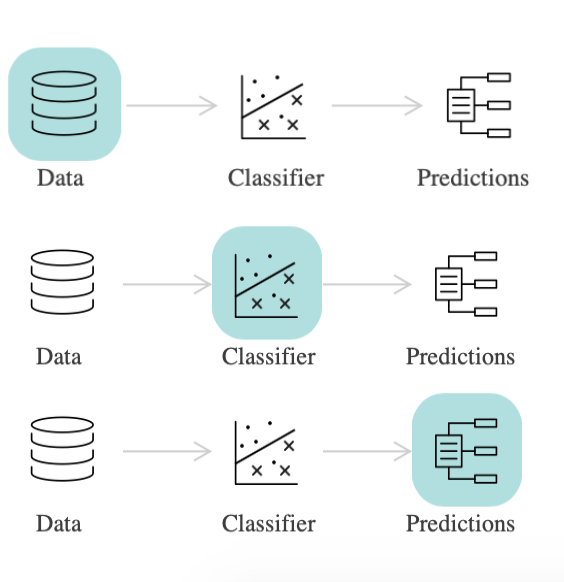
XING, a job platform similar to Linked-in, was found to rank less qualified male candidates higher than more qualified female candidates

Search query	Work experience	Education experience	Profile views	Candidate	Xing ranking
Brand Strategist	146	57	12992	male	1
Brand Strategist	327	0	4715	female	2
Brand Strategist	502	74	6978	male	3
Brand Strategist	444	56	1504	female	4
Brand Strategist	139	25	63	male	5
Brand Strategist	110	65	3479	female	6
Brand Strategist	12	73	846	male	7
Brand Strategist	99	41	3019	male	8
Brand Strategist	42	51	1359	female	9
Brand Strategist	220	102	17186	female	10

TABLE II: Top k results on www.xing.com (Jan 2017) for the job serach query “Brand Strategist”.

AIF360 Toolkit

- Provides fairness metrics to examine each stage of ML pipeline
- Provides bias mitigation in all stages



Pre-Processing: improve training data to remove bias from it

In-Processing: modify ML algorithm. Often in form of adding extra regularizations

Post-Processing: modify the prediction. Treats the ML model as a black box

For more information ...

AI Fairness 360

This extensible open source toolkit can help you examine, report, and mitigate discrimination and bias in machine learning models throughout the AI application lifecycle. We invite you to use and improve it.

Python API Docs ↗ | Get Python Code ↗ | Get R Code ↗

Not sure what to do first? Start here!

Read More

Learn more about fairness and bias mitigation concepts, terminology, and tools before you begin.

Try a Web Demo

Step through the process of checking and remediating bias in an interactive web demo that shows a sample of capabilities available in this toolkit.

Watch Videos

Watch videos to learn more about AI Fairness 360.

Read a paper

Read a paper describing how we designed AI Fairness 360.

Use Tutorials

Step through a set of in-depth examples that introduces developers to code that checks and mitigates bias in different industry and application

http://aif360.mybluemix.net/

https://github.com/IBM/AIF360

Topics

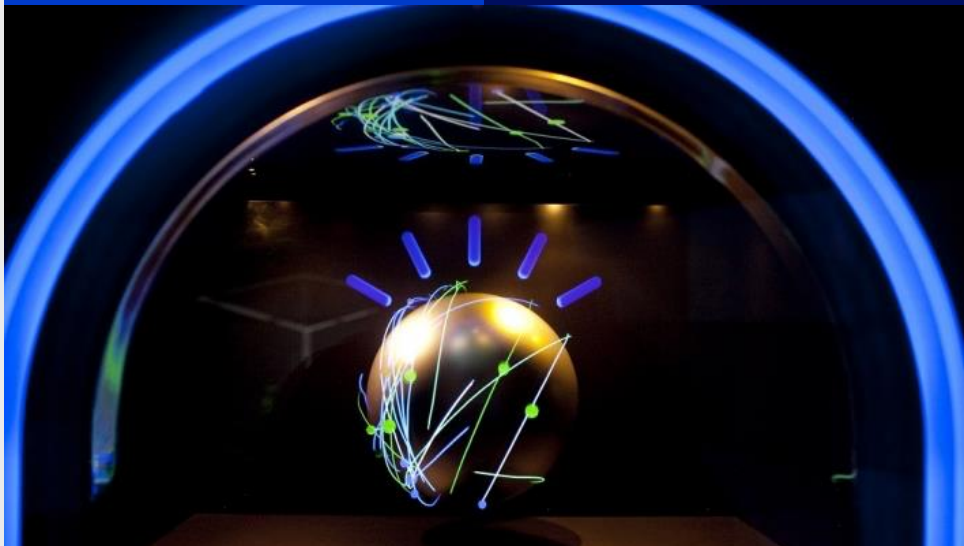
- What is Machine Learning?
- Model Fairness
  - AIF360 Toolkit
- **Model Explainability**
  - AIX360 Toolkit
- Model Robustness
  - Adversarial Robustness Toolbox
- Watson OpenScale
- Lab Overview

Model Explainability

- To explanation our decisions in daily life, we use expressive vocabulary and several examples
- Must do the same with algorithmic decisions

Produce more explainable models, while maintaining a high level of learning performance (prediction accuracy)

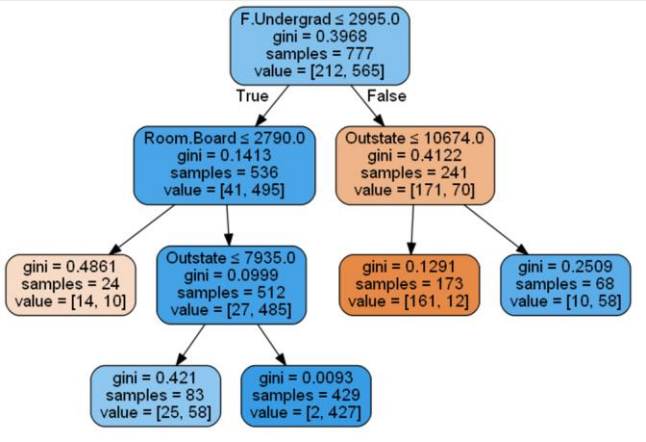
Enable human users to understand, trust, and effectively manage the emerging generation of artificially intelligent partners.

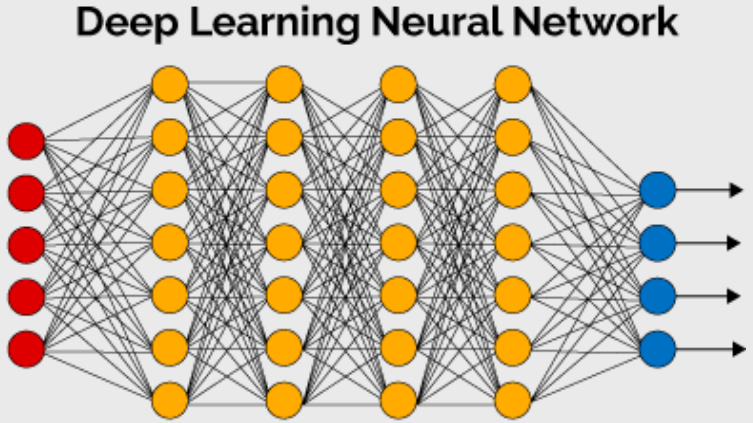


What is going on in The Black Box?

Source: If applicable, describe source origin

- Black box ML models cannot be understood by people
- The more powerful a model gets, the harder it is to understand





Which model's decision is easier to explain?

Example – Healthcare

AI System Diagnoses a patient with heart disease

Explain to the Doctor

What symptoms contributed to such diagnosis? What was similar/ different between this patient and the ones who where diagnosed before?

Explain to the Patient

What could have she done to prevent the illness?

AIX360 Toolkit

- Comprehensive toolkit of state-of-the-art algorithms
- Support the interpretability and explainability of machine learning models
- Open-source library written in python

Data Scientists

Interested in technical details of why a model works the way it does.

Decision Makers (e.g. Doctors)

Interested in understanding the entire decision-making process and ensure its safety, reliability, or compliance.

Affected Users (e.g. Patients)

Need to understand the decision about them in simple terms.

For more information ....

AI Explainability 360

This extensible open source toolkit can help you comprehend how machine learning models predict labels by various means throughout the AI application lifecycle. We invite you to use it and improve it.

API Docs ↗ Get Code ↗

Not sure what to do first? Start here!

Read More

Learn more about explainability concepts, terminology, and tools before you begin.

Try a Web Demo

Step through the process of explaining models to consumers with different personas in an interactive web demo that shows a

Watch Videos

Watch videos to learn more about AI Explainability 360 toolkit.

Read a Paper

Read a paper describing how we designed AI Explainability 360 toolkit.

Use Tutorials

Step through a set of in-depth examples that introduce developers to code that explains data and models in different industry

http://aix360.mybluemix.net/#

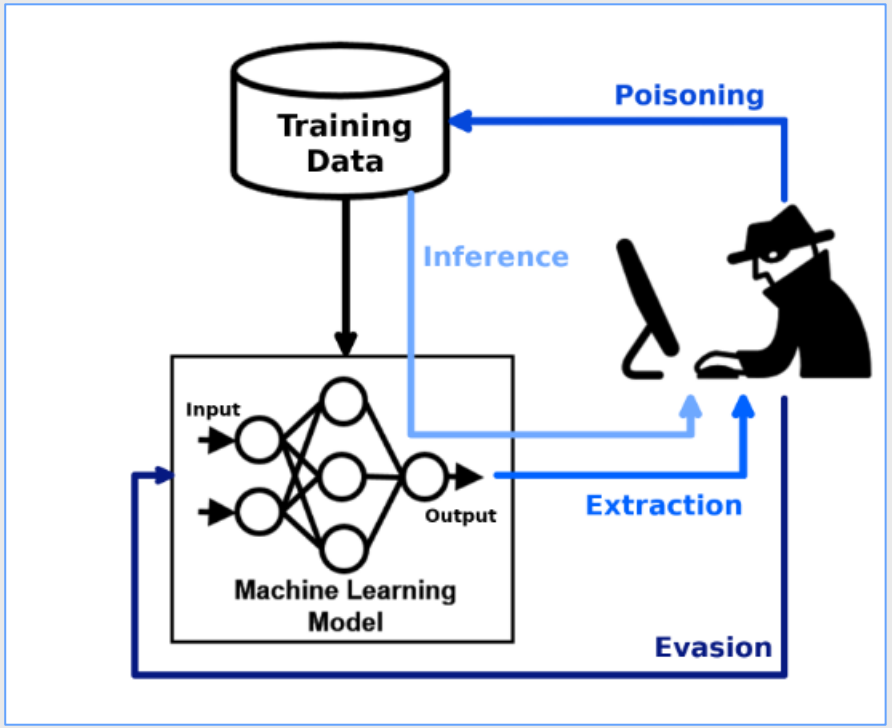
https://github.com/IBM/AIX360

Topics

- What is Machine Learning?
- Model Fairness
  - AIF360 Toolkit
- Model Explainability
  - AIX360 Toolkit
- **Model Robustness**
  - Adversarial Robustness Toolbox
- Watson OpenScale
- Lab Overview

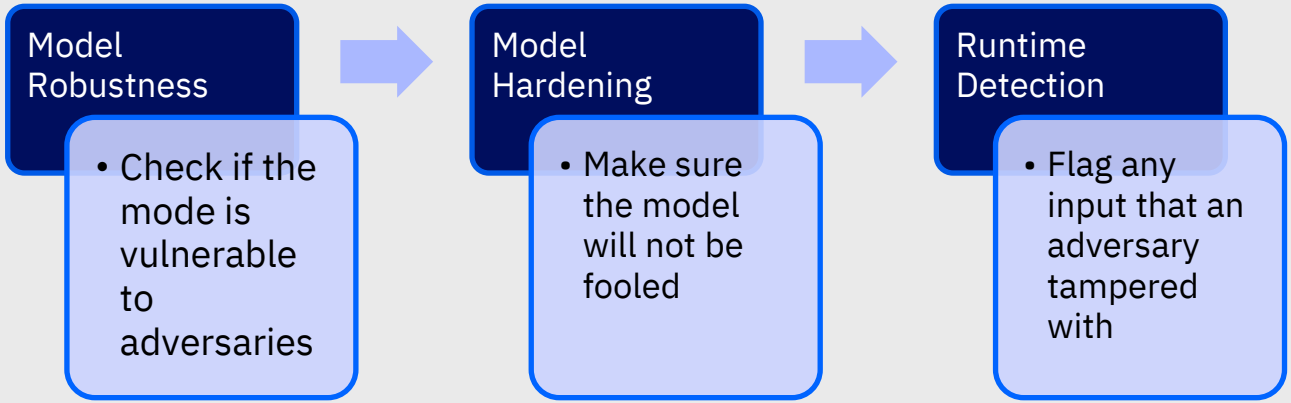
Adversarial Robustness Toolbox (ART)

- Provides tools to defend and evaluate Machine Learning models against the adversarial threats of Evasion, Poisoning, Extraction, and Inference.
- Supports all popular machine learning frameworks (TensorFlow, Keras, PyTorch, MXNet, scikit-learn, XGBoost, LightGBM, CatBoost, GPy, etc.)
- Supports all data types (images, tables, audio, video, etc.)
- Supports machine learning tasks (classification, object detection, speech recognition, generation, etc.).



Adversarial Robustness Toolbox (ART)

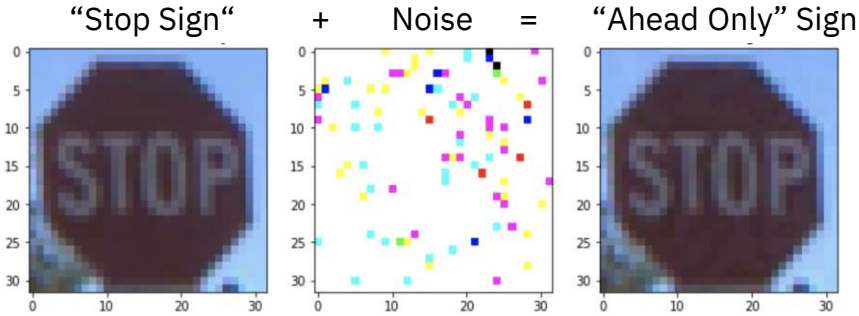
- IBM Research team in Ireland developed the toolbox to help defend Machine Learning models against adversarial attacks
- Open-source software library written in Python
- It creates adversarial examples AND provides methods for defending Machine Learning models against those.



Adversarial Images

- Adversarial examples are inputs (say, images) which have deliberately been modified to produce a desired response by a VR system.

- Often, the target of adversarial examples is misclassification or a specific incorrect prediction which would benefit an attacker.



Why are they dangerous?

- Can be crafted even if the attacker doesn't have exact knowledge of the architecture of the Machine Learning Model
- Adversarial attacks can be launched in the physical world
  - adversaries could evade face recognition systems by wearing specially designed glasses
  - defeat visual recognition systems in autonomous vehicles by sticking patches to traffic signs

* Pictures from paper: Kevin Eykholt, et al. "Robust Physical-World Attacks on Deep Learning Visual Classification"

For more information …

https://art-demo.mybluemix.net/

https://github.com/IBM/AIF360

Topics

- What is Machine Learning?
- Model Fairness
  - AIF360 Toolkit
- Model Explainability
  - AIX360 Toolkit
- Model Robustness
  - Adversarial Robustness Toolbox
- **Watson OpenScale**
- Lab Overview

Watson OpenScale

Trust and Transparency

- Intelligently delivers bias mitigation help
- Provides traceability & auditability of AI predictions made in production applications
- Tracks AI accuracy in applications
- Explains an outcome in business terms
- Provides drift detection

Automation

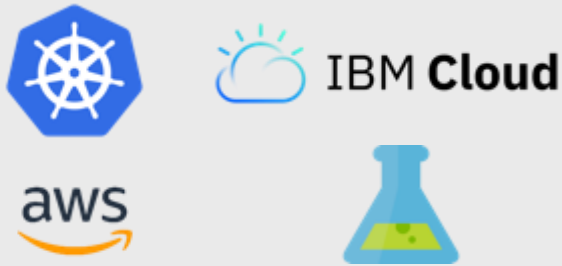
- Automatically detects and mitigates bias in model output, without affecting currently deployed model or outcomes

Open by Design

- Monitor models deployed on third party mode server engines
- Deploy behind enterprise firewall or on IaaS provider.

Model build / train frameworks

Model serving environments



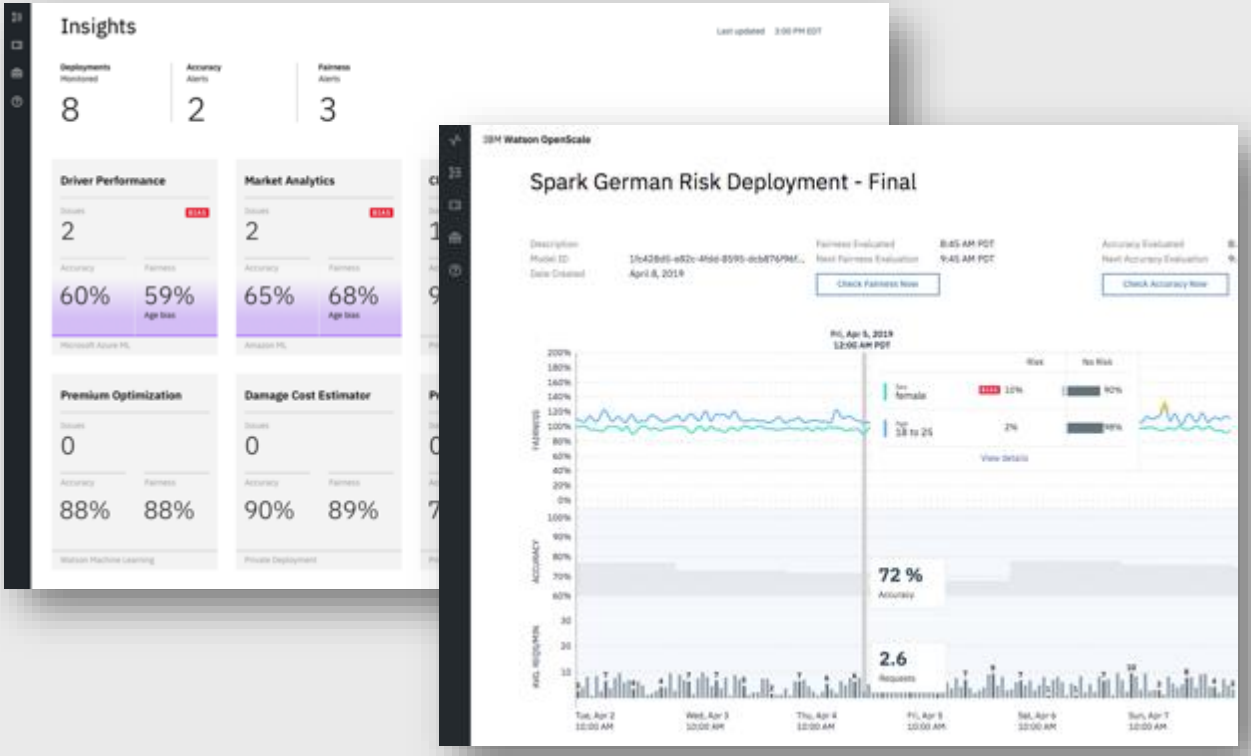
OpenScale Operations Dashboard

Description:

Monitor deployed models in a single dashboard that can be filtered by deployment making it easy to manage AI in apps

Value:

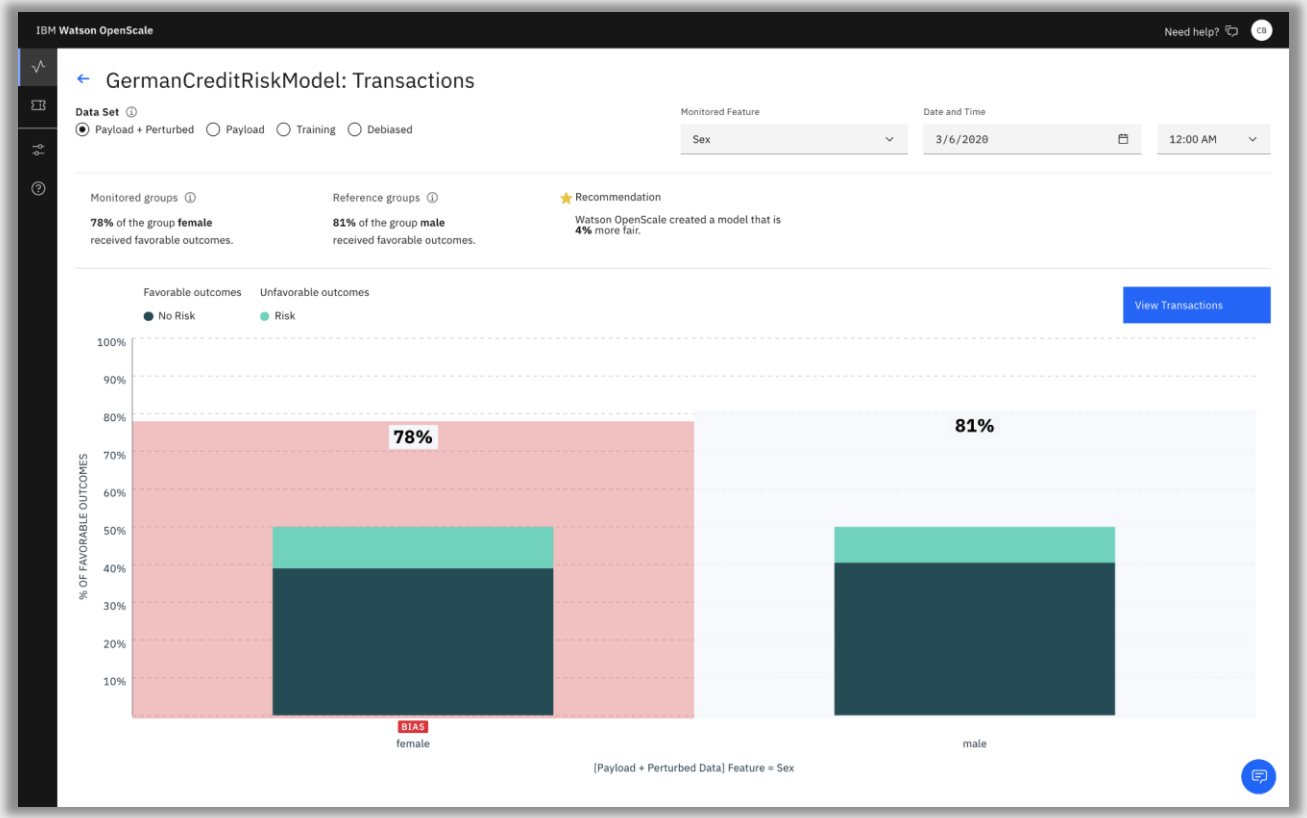
- Configure alerts or actions to be triggered when KPIs exceed threshold, ensuring model quality for improve business outcomes
- Measure model accuracy as it pertains to it's ability to deliver outcomes more accurate than knowledge workers
- Provides "continuous evolution" for your models



Bias Mitigation – Original Model Output

Credit Risk Example Model

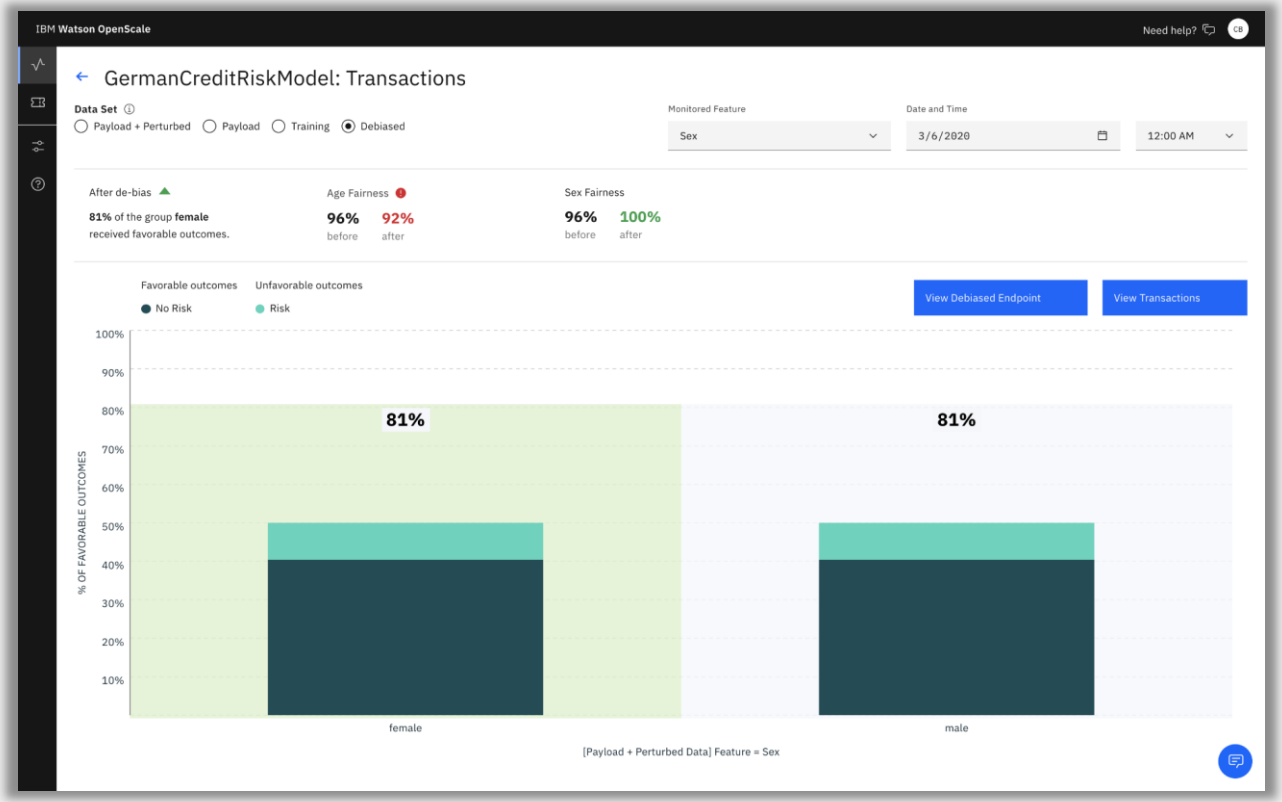
- 78% of the protected group (female) have a favorable output
- 81% of the reference group (male) get a favorable output
- Disparate impact Value: 96%



Bias Mitigation – De-biased Model Output

After De-biasing algorithms was applied

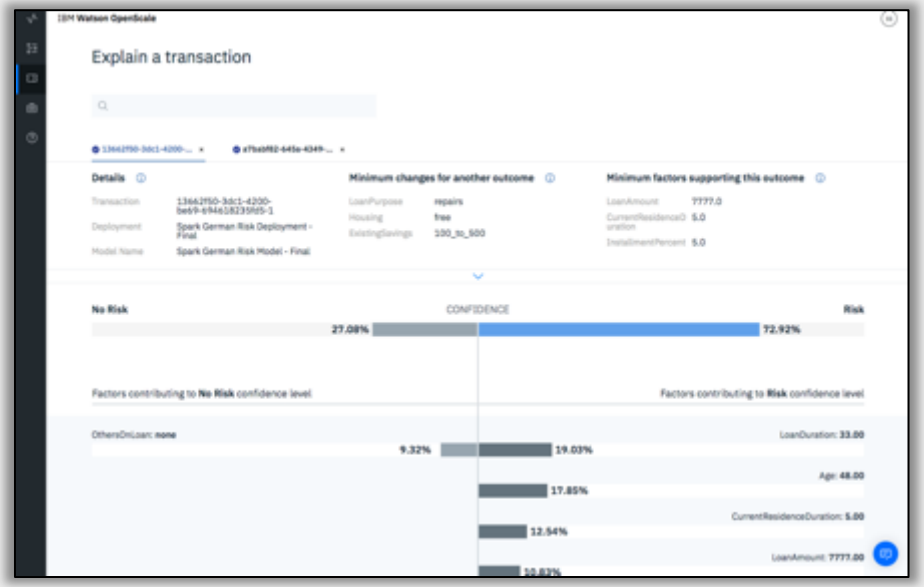
- Predictions are 4% more fair in this example
- 81% of the protected group (female) and of the reference group (male) get a favorable output
- Disparate impact Value: 100%



Explainability

OpenScale records every individual transaction and drills down into its working to explain how the model makes decisions

It provides a simple explanation that is user friendly and interactive



Value:

- Explain individual transaction level decisions made by the model in run time, including details about most important attributes and their values in order to assist in compliance and customer care situations
- Analyze individual transactions in a what-if manner in order to understand how model behavior will change in different business situations

LIME and Contrastive Explanations

Lime Output:

– Set of features which played a positive role or negative role in the prediction.

– Also identifies the feature weights which helps to identify the most important or least important features

Contrastive Explanation:

– Explains the behavior of the model in the vicinity of the data point whose explanation is being generated.

– Assumption:

- The most common value is the least interesting from an explanation point of view
- E.g., If median salary is between $70-90K, then someone who has a salary of $80K it is not very “interesting” or to say it differently it is “normal”.
- However, if someone has salary of $200K, it is very “interesting”

Drift Detection in OpenScale

Drift Monitor in OpenScale measures two types of drifts:

- **Drop in accuracy**: It estimates the drop in accuracy of the model at runtime. The model accuracy could drop if there is an increase in transactions similar to those which the model was unable to evaluate correctly in the training data.
- **Drop in data consistency**: It estimates the drop in consistency of the data at runtime as compared to the characteristics of the data at training time.

A drop in model accuracy and data consistency may lead to a negative impact on the business outcomes associated with the model.

OpenScale measures the drift without requiring labelled data. Accuracy computation using labelled data can be expensive and might not be comprehensive

OpenScale does Drift detection on the entire payload data

OpenScale will automatically detect drifted transactions and pinpoint datapoints that contribute to drift

Topics

- What is Machine Learning?
- Model Fairness
  - AIF360 Toolkit
- Model Explainability
  - AIX360 Toolkit
- Model Robustness
  - Adverarial Robustness Toolbox
- Watson OpenScale
- **Lab Overview**

Lab Overview

- Lab-1: Setup Environment
- Lab-2: AIF360 Toolkit
- Lab-3: AIX360 Toolkit
- Lab-4: Adversarial Robustness Toolbox
- Lab-5: Watson OpenScale

Lab Overview - www.github.com/bleonardb3/TR_POT_01-28-2021

When deploying machine learning models, other factors besides the accuracy of the model needs to be considered. Is the model biased? Can the decisions made by the model be explained?. Is the model robust to adversarial attacks? IBM has developed 3 toolkits to help address these questions.

This session will provide a brief overview of the toolkits and 1 lab on each toolkit. In addition, a lab on Watson Openscale which monitors accuracy, bias, and model drift will also be included. The attendees will use Watson Studio to complete the labs.

Lab-1 - This lab will walk through the steps to create a Watson Studio project. A Watson Studio project is a way to organize your data and analytical assets for an analytics project.

Lab-2 - This lab will feature IBM's AI Fairness 360 (AIF360), a comprehensive open-source toolkit of metrics to check for unwanted bias in datasets and machine learning models, and state-of-the-art algorithms to mitigate such bias.

Lab-3 - This lab will feature IBM's AI Explainability 360 (AIX360), a comprehensive open source toolkit of state-of-the-art algorithms that support the interpretability and explainability of machine learning models.

Lab-4 - This lab will feature IBM's Adversarial Robustness Toolbox (ART). ART is a library dedicated to adversarial machine learning. Its purpose is to allow rapid crafting and analysis of attacks and defense methods for machine learning models. ART provides an implementation for many state-of-the-art methods for attacking and defending classifiers.

Lab-5 - This demo will feature IBM Watson OpenScale. IBM Watson OpenScale is an open platform that helps remove barriers to enterprise-scale AI by supporting bias mitigation, accuracy, and explainability of outcomes among other features.

Lab Overview - www.github.com/bleonardb3/TR_POT_01-28-2021

Lab-1: Setup Environment

Introduction:

This lab will set up the Watson Studio environment for subsequent labs and introduce you to the Project features of Watson Studio. Watson Studio is an integrated platform of tools, services, data, and meta-data to help companies and agencies accelerate their shift to be data driven organizations. The platform enables data professionals such as data scientists, data engineers, business analysts, and application developers collaboratively work with data to build, train, deploy machine learning and deep learning models at scale to infuse AI into business to drive innovation. Watson Studio is designed to support the development and deployment of data and analytics assets for the enterprise.

Objectives:

Upon completing the lab, you will have:

1. Created a project
2. Created an object storage instance and associate it with the project
3. Initiate the IBM Watson OpenScale Auto setup.

Step 1. Please click on the link below to download the instructions to your machine.

Instructions.

Summary

- ✓ **What is Machine Learning?**
- ✓ Model Fairness - AIF360 Toolkit
- ✓ Model Explainability - AIX360 Toolkit
- ✓ Model Robustness - Adversarial Robustness Toolbox
- ✓ Watson OpenScale
- ✓ Labs

A3 Center Link:
https://www.ibm.com/industries/federal/analytics

Notices and disclaimers

U.S. Government Users Restricted Rights — use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM.

This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates.

Information in these presentations (including information relating to products that have not yet been announced by IBM) has been reviewed for accuracy as of the date of initial publication and could include unintentional technical or typographical errors. IBM shall have no responsibility to update this information. **This document is distributed "as is" without any warranty, either express or implied. In no event, shall IBM be liable for any damage arising from the use of this information, including but not limited to, loss of data, business interruption, loss of profit or loss of opportunity.** IBM products and services are warranted per the terms and conditions of the agreements under which they are provided. The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

IBM products are manufactured from new parts or new and used parts. In some cases, a product may not be new and may have been previously installed. Regardless, our warranty terms apply."

Any statements regarding IBM's future direction, intent or product plans are subject to change or withdrawal without notice.

Performance data contained herein was generally obtained in a controlled, isolated environments. Customer examples are presented as illustrations of how those customers have used IBM products and the results they may have achieved. Actual performance, cost, savings or other results in other operating environments may vary.

References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business.

Workshops, sessions and associated materials may have been prepared by independent session speakers, and do not necessarily reflect the views of IBM. All materials and discussions are provided for informational purposes only, and are neither intended to, nor shall constitute legal or other guidance or advice to any individual participant or their specific situation.

Notices and disclaimers continued

It is the customer's responsibility to insure its own compliance with legal requirements and to obtain advice of competent legal counsel as to the identification and interpretation of any relevant laws and regulatory requirements that may affect the customer's business and any actions the customer may need to take to comply with such laws. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the customer follows any law.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products about this publication and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products. IBM does not warrant the quality of any third-party products, or the ability of any such third-party products to interoperate with IBM's products. **IBM expressly disclaims all warranties, expressed or implied, including but not limited to, the implied warranties of merchantability and fitness for a purpose.**

The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents, copyrights, trademarks or other intellectual property right.

Thank you

Bernard Beekman
Executive IT Architect
beekmanb@us.ibm.com

Michael Cronk
IT Architect
Michael.cronk@ibm.com