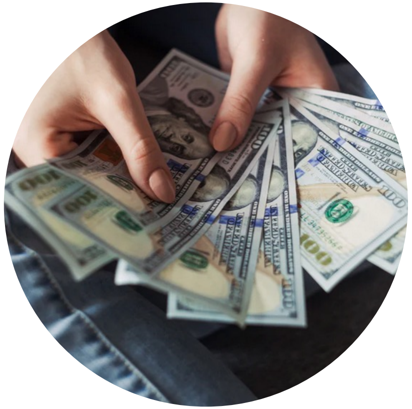
Trustworthy AI in action

July 26, 2022

John Thomas
VP & DE, IBM Expert Labs
@johnjaithomas

Ashley Casovan
Executive Director
Responsible AI Institute
@ResponsibleAI

With AI increasingly powering critical workflows…



credit

employment

customer management

fraud

*…**trust** is essential*

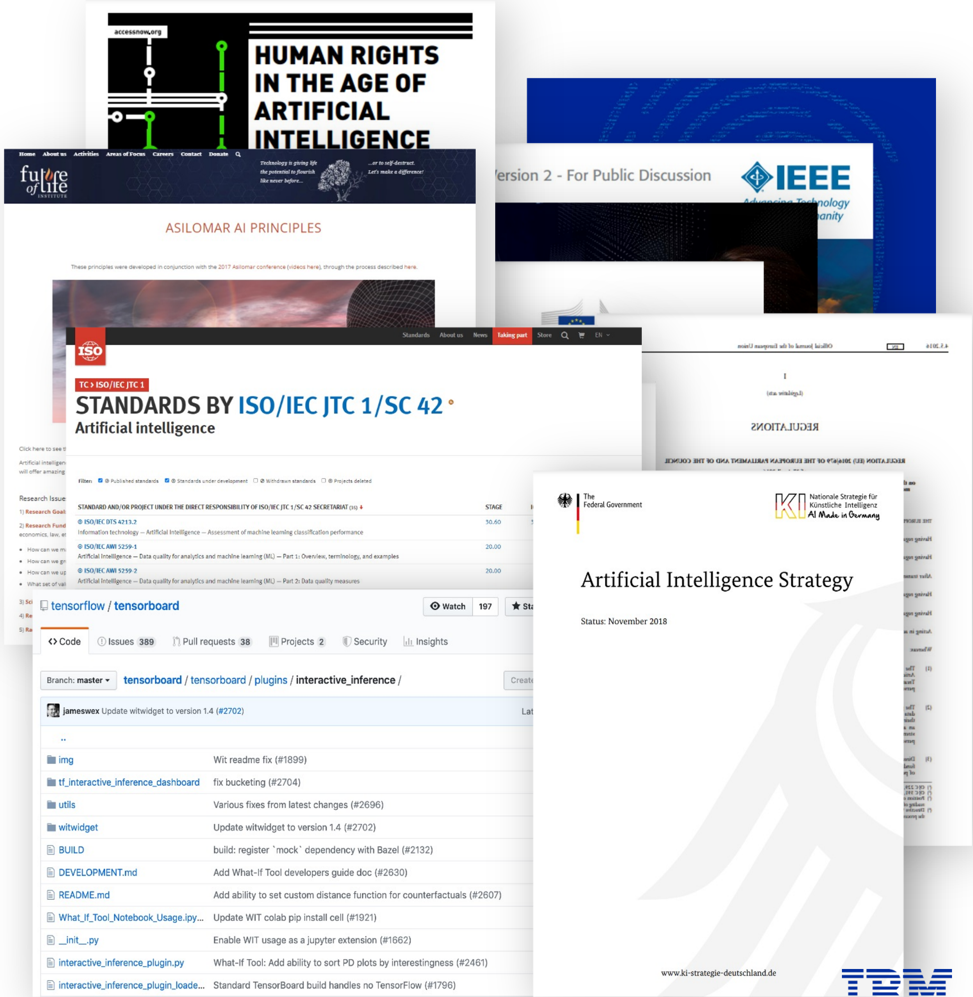
Simplifying how to implement responsible AI

> “Only about a quarter (28%) of citizens are willing to trust AI systems in general. Two out of five citizens are unwilling to share their information or data with an AI system and a third are unwilling to trust the output of AI systems.”

- University of Queensland and KPMG, 2021

> “Fewer than 20% of executives strongly agree that their organizations’ practices and actions on AI ethics match (or exceed) their stated principles and values.”

- IBM and Oxford Economics, 2021



accessnow.org

HUMAN RIGHTS IN THE AGE OF ARTIFICIAL INTELLIGENCE

Version 2 - For Public Discussion

IEEE

Home About us Activities Areas of Focus Careers Contact Donate

Technology is giving life the potential to flourish like never before… …or to self-destruct. Let's make a difference!

future of life INSTITUTE

ASILOMAR AI PRINCIPLES

These principles were developed in conjunction with the 2017 Asilomar conference (videos here), through the process described here.

Standards About us News Taking part Store EN

ISO

TC › ISO/IEC JTC 1

STANDARDS BY ISO/IEC JTC 1/SC 42

Artificial intelligence

Filter: ☑ Published standards ☑ Standards under development ☐ Withdrawn standards ☐ Projects deleted

STANDARD AND/OR PROJECT UNDER THE DIRECT RESPONSIBILITY OF ISO/IEC JTC 1/SC 42 SECRETARIAT (35)	STAGE
ISO/IEC DTS 4213.2 Information technology — Artificial Intelligence — Assessment of machine learning classification performance	30.60
ISO/IEC AWI 5259-1 Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 1: Overview, terminology, and examples	20.00
ISO/IEC AWI 5259-2 Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 2: Data quality measures	20.00

tensorflow / tensorboard

Watch 197

Code Issues 389 Pull requests 38 Projects 2 Security Insights

Branch: master ▾ tensorboard / tensorboard / plugins / interactive_inference /

jameswex Update witwidget to version 1.4 (#2702)

..	
img	Wit readme fix (#1899)
tf_interactive_inference_dashboard	fix bucketing (#2704)
utils	Various fixes from latest changes (#2696)
witwidget	Update witwidget to version 1.4 (#2702)
BUILD	build: register `mock` dependency with Bazel (#2132)
DEVELOPMENT.md	Add What-If Tool developers guide doc (#2630)
README.md	Add ability to set custom distance function for counterfactuals (#2607)
What_If_Tool_Notebook_Usage.ipy...	Update WIT colab pip install cell (#1921)
__init__.py	Enable WIT usage as a jupyter extension (#1662)
interactive_inference_plugin.py	What-If Tool: Add ability to sort PD plots by interestingness (#2461)
interactive_inference_plugin_loade...	Standard TensorBoard build handles no TensorFlow (#1796)

The Federal Government

Nationale Strategie für Künstliche Intelligenz
AI Made in Germany

Artificial Intelligence Strategy

Status: November 2018

www.ki-strategie-deutschland.de

Organizations must consider Regulatory Compliance

Sarbanes-Oxley Act

USA

2021 – National AI advisory committee

2022—Algorithmic Accountability Act of 2022

2022 – American Data Privacy and Protection Act

Canada

2017- National AI Strategy

2020—Directive on Automated Decision Making

2022—Artificial Intelligence and Data Act

European Union

2018— Coordinated Plan on AI

2021 – Draft AI Act

United Kingdom

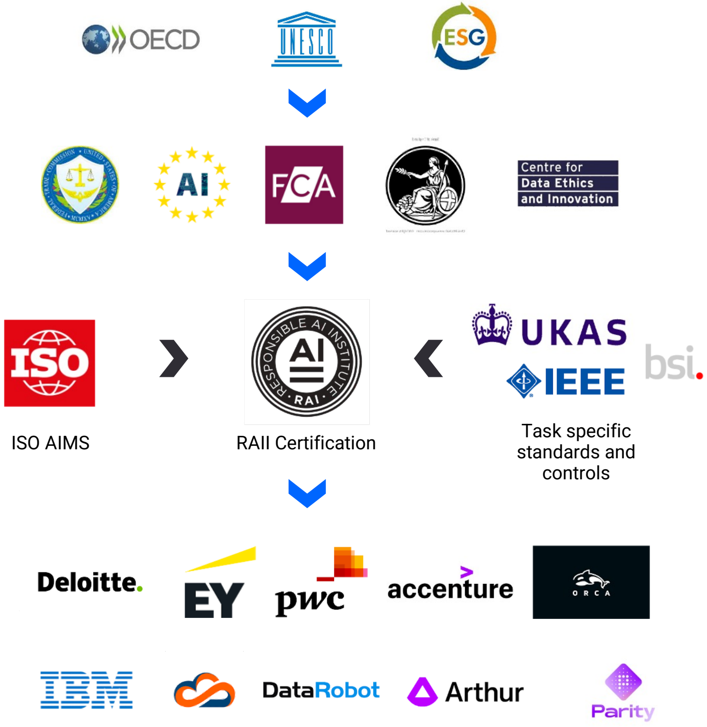
2021— CDEI AI Assurance Guide

Standards Bodies

National standards and accreditation bodies are working on AI specific standards and frameworks to help implement responsible AI (eg. ISO, NIST, SCC, CEN-CENELEC, BSI, UKAS, ANSI, CETA, IEEE, etc.)

AI Regulatory and Standards Landscape

AI Principles	● International NGOs ● Corporate values ● ESG objectives
AI Regulations	● National ● State, Regional, and Local ● Corporate policies
Standards, Certifications, and Industry Best Practices	● Accreditations ● Management Standards ● Use case & function specific certifications ● Task specific standards & controls ● Industry best practices
Evaluations	● Point-in-time audits ○ Manual ○ Semi-automated ● Ongoing monitoring ○ Statistical evaluations ○ Data quality evaluations ○ Automated policy evaluations ○ Document tool chain



We need a multidisciplinary, multidimensional approach to trustworthy AI

From principles to actions

what should be done

principles, values, norms, laws, regulations

how to instrument it

techniques, algorithms, software, best practices

how to operationalize it

mechanisms, systems, and processes to keep AI trustworthy

What does it take to Trust a decision made by an AI?

We started from these HUMAN-CENTRIC questions

Is it easy to understand?

EXPLAINABILITY

Easy to understand outcomes/decisions

Why did the AI arrive at an outcome? When would it have been different?

Is it fair?

FAIRNESS

Impartial and addressing bias

Are privileged groups at a systematic advantage compared to other groups?

Did anyone tamper with it?

ROBUSTNESS

Handle exceptional conditions effectively

Can we evaluate and defend against a variety of threats?

Is it accountable?

TRANSPARENCY

Open to inspecting facts and details

Can we increase understanding of why and how AI was created?

Does it safeguard data?

PRIVACY

High integrity data & business compliance

How do we ensure owners retain control of data and insights?

AI governance enables trustworthy AI

Foundational for strategy and execution of AI solutions

Strategy

Who?
Business, AI ethics board, data/AI leaders, people responsible for internal policy and regulations (CPO etc.)

Planning

Who?
Business, AI ethics board, data/AI leaders, people responsible for internal policy and regulations (CPO etc.), ecosystem

Development and deployment

Who?
Dev teams, IT leaders, CDO, people responsible for software and data scientists

Operate

Who?
IT leaders, MLOps teams

Monitor + portfolio management

Business outcomes
Who? Business leaders + MLOps teams

Governance of models
Who? Business leaders + MLOps teams

AI strategy
Humans in the loop from the beginning

AI governance tools and process
Ensuring governance of AI models through technology and data

Decide and drive the AI strategy for the organization. Establish AI policies for the organization. (This may include AI principles, regulations, laws, etc.) => AI Ethics Board

Encode the policies into business rules, guidelines and transparent reporting mechanism. Determine appropriate guardrails and parameters.

Enable data collection and transparent reporting to make needed information available to all stakeholders.

RAII Implementation Framework Dimensions and Sub-Dimensions

1 Systems Operations

1.1 System Scope and Function
1.2 Human-in-the-Loop
1.3 Model is Fit for Purpose
1.4 Data Relevance and Representativeness
1.5 Data Quality

2 Explainability & Interpretability

2.1 Communication About the Outcome
2.2 Notification
2.3 Recourse
2.4 Understanding the AI System's Decisions or Functions

3 Accountabilty

3.1 Organizational Governance
3.2 Team Governance

4 Consumer Protection

4.1 Transparency to the User and Data Subject
4.2 Harms to Individuals
4.3 Protections

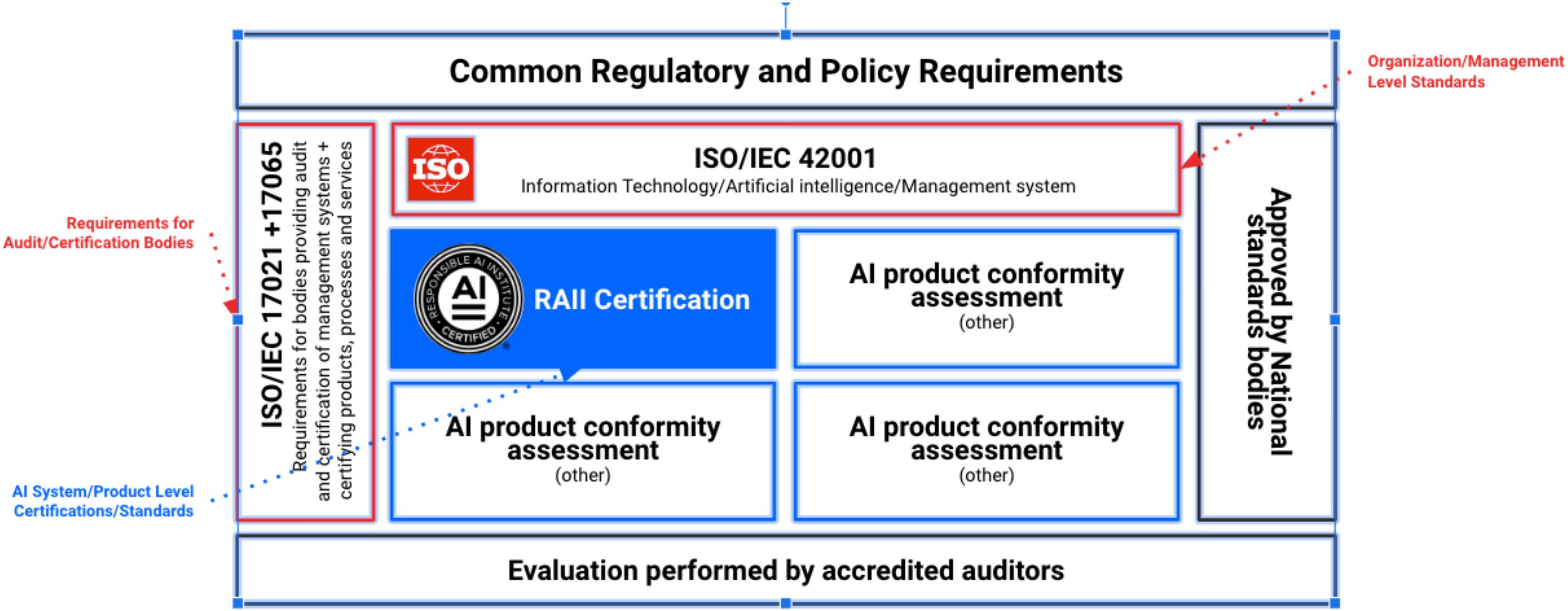
5 Bias & Fairness

5.1 Bias Impacts
5.2 Bias Training
5.3 Bias Testing

6 Robustness

6.1 Data Drift
6.2 System Acceptance Test is Performed
6.3 Contingency Planning

Complying with AI regulations



Common Regulatory and Policy Requirements

Organization/Management Level Standards

ISO/IEC 42001

Information Technology/Artificial intelligence/Management system

Requirements for Audit/Certification Bodies

ISO/IEC 17021 +17065

Requirements for bodies providing audit and certification of management systems + certifying products, processes and services

RAII Certification

AI product conformity assessment (other)

AI product conformity assessment (other)

AI product conformity assessment (other)

Approved by National standards bodies

AI System/Product Level Certifications/Standards

Evaluation performed by accredited auditors

Sample Control Development

OECD AI Guiding Principle on Fairness (Global Level)

AI systems should be designed in a way that respects the rule of law, human rights, democratic values and diversity, and they should include appropriate safeguards – for example, enabling human intervention where necessary – to ensure a fair and just society.

Enterprise AI Policy (Industry Level)

2. Ensure Fairness in AI Systems

2.1 Identify unwanted bias. 2.2. Test for unwanted bias. 2.3 Perform bias training. 2.4 Ensure recourse is implemented

Client Bias Guideline (Company Level)

2.1 Complete a harms mapping to understand the unintended consequences and identify mitigation measures

2.2 Perform appropriate bias test for all protected classes, ensure that it meets the acceptable threshold.

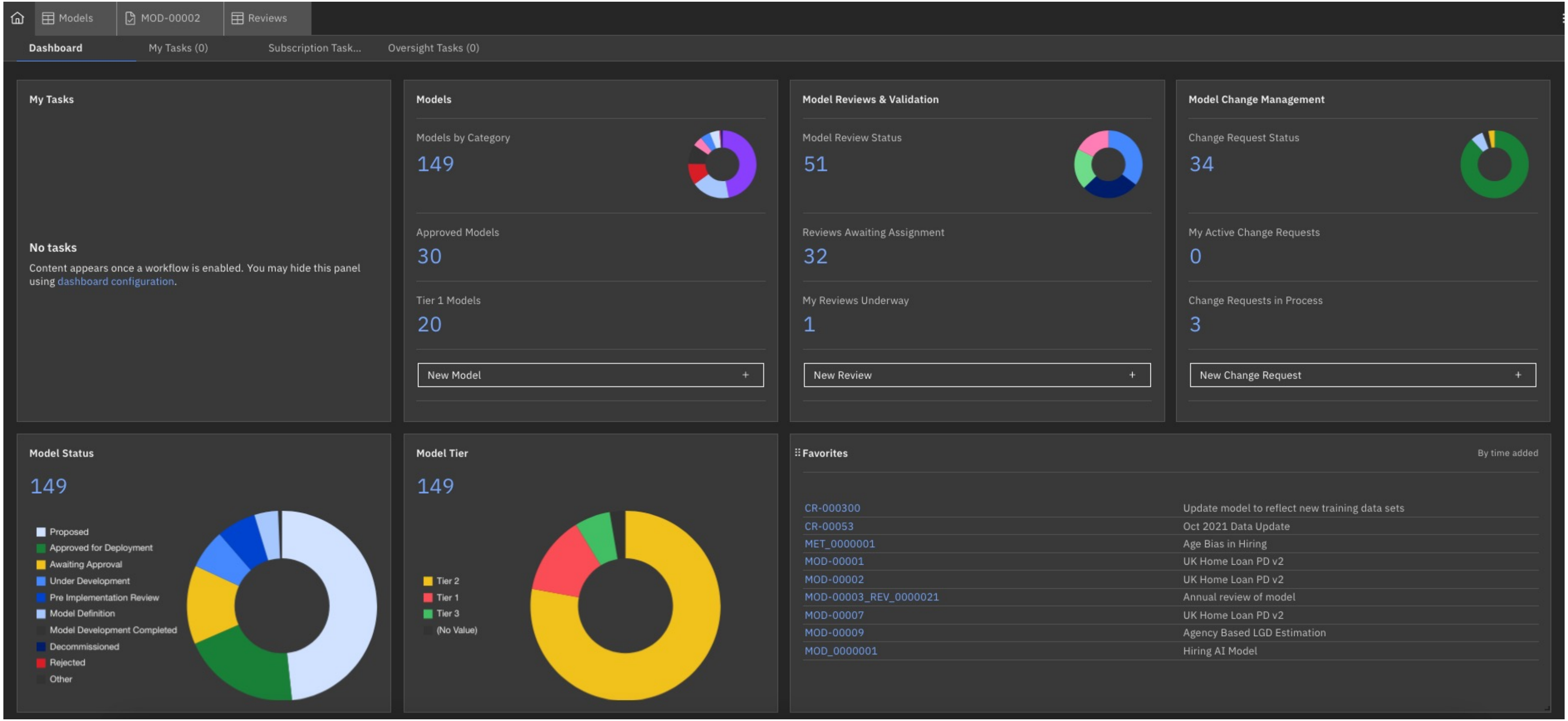
2.3 Require bias training for all employees involved in the production and deployment of an AI system

2.4 All AI systems must have a recourse plan in the case of the system not working as intended

Bias Control (AI System/UC Level)

2.2 Perform demographic parity and equalized odds test for protected classes. Acceptable result is 0.85.

Enterprise Governance View of AI/ML models



My Tasks

No tasks

Content appears once a workflow is enabled. You may hide this panel using dashboard configuration.

Models

Models by Category
149

Approved Models
30

Tier 1 Models
20

Model Reviews & Validation

Model Review Status
51

Reviews Awaiting Assignment
32

My Reviews Underway
1

Model Change Management

Change Request Status
34

My Active Change Requests
0

Change Requests in Process
3

Model Status

149

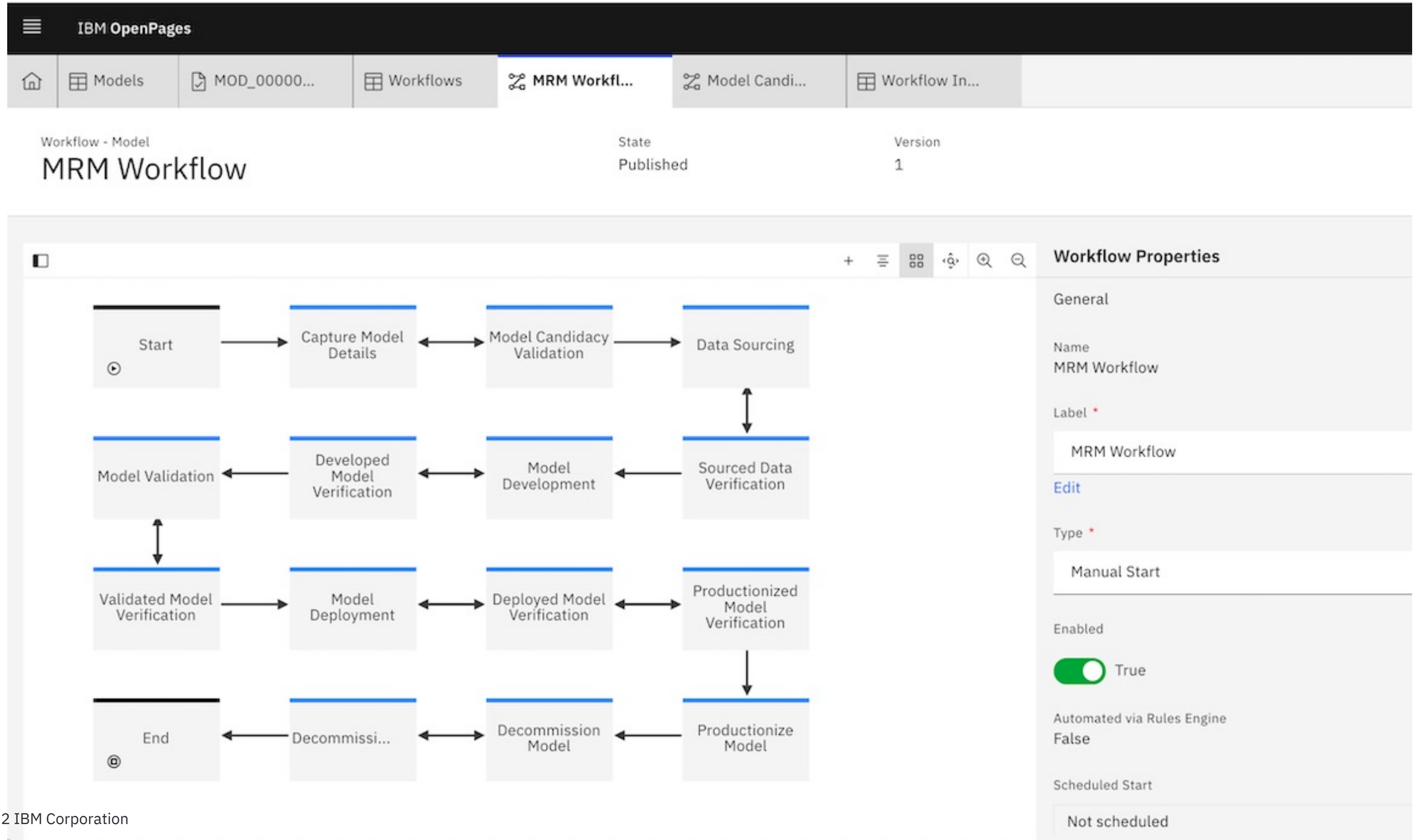
Model Tier

149

Favorites

CR-000300	Update model to reflect new training data sets
CR-00053	Oct 2021 Data Update
MET_0000001	Age Bias in Hiring
MOD-00001	UK Home Loan PD v2
MOD-00002	UK Home Loan PD v2
MOD-00003_REV_0000021	Annual review of model
MOD-00007	UK Home Loan PD v2
MOD-00009	Agency Based LGD Estimation
MOD_0000001	Hiring AI Model

Enterprise workflows provide consistency with flexibility



Workflow - Model
MRM Workflow

State
Published

Version
1

Workflow Properties

General

Name
MRM Workflow

Label *
MRM Workflow

Type *
Manual Start

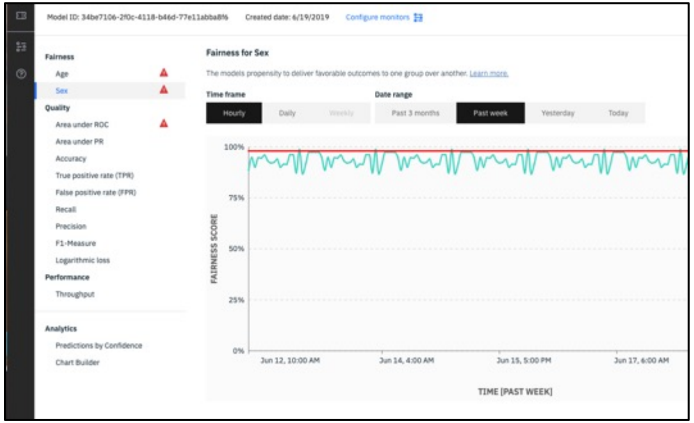
Enabled
True

Automated via Rules Engine
False

Scheduled Start
Not scheduled

Example of a runtime guardrail

Bias Monitoring and Mitigation with Watson OpenScale



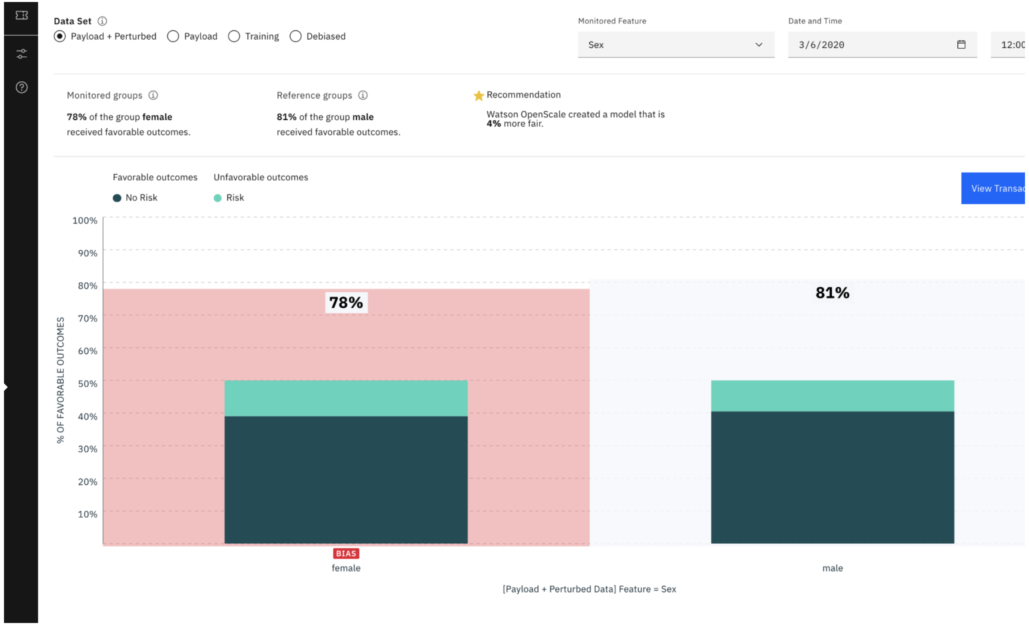
Model ID: 34be7106-2f0c-4118-b46d-77e11abba8%   Created date: 6/19/2019   Configure monitors

Fairness for Sex

The models propensity to deliver favorable outcomes to one group over another. Learn more.

Time frame: Hourly

Date range: Past week



Data Set ● Payload + Perturbed ○ Payload ○ Training ○ Debiased

Monitored Feature: Sex

Date and Time: 3/6/2020 12:0

Monitored groups

78% of the group **female** received favorable outcomes.

Reference groups

81% of the group **male** received favorable outcomes.

Recommendation

Watson OpenScale created a model that is **4%** more fair.

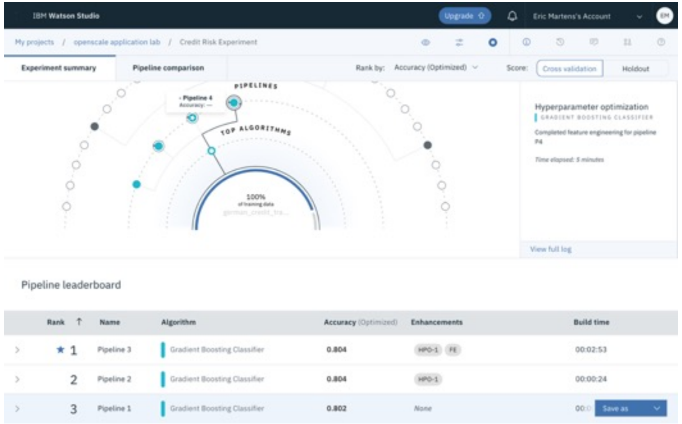
Setup ongoing monitoring of deployed model
Define monitored and reference groups

Calculate Disparate impact Value

78% of the monitored group (female) have a favorable output
81% of the reference group (male) get a favorable output

Disparate impact Value: 96%
Mitigation based on policy

Guardrails across the AI lifecycle

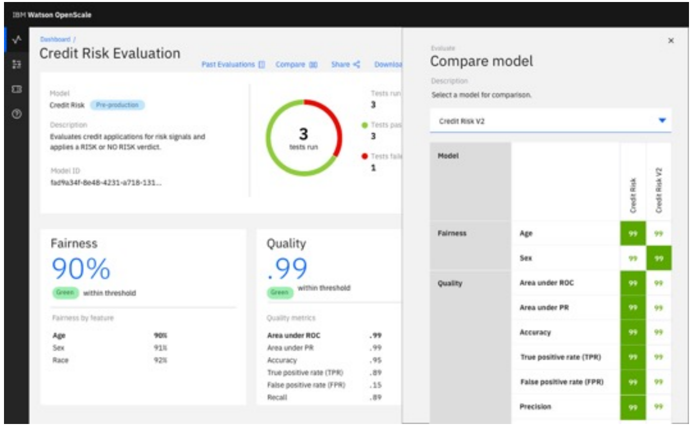


Hyperparameter optimization
GRADIENT BOOSTING CLASSIFIER
Completed feature engineering for pipeline P4
Time elapsed: 5 minutes

Pipeline leaderboard

Rank	Name	Algorithm	Accuracy (Optimized)	Enhancements	Build time
★ 1	Pipeline 3	Gradient Boosting Classifier	0.804	HPO-1 FE	00:02:53
2	Pipeline 2	Gradient Boosting Classifier	0.804	HPO-1	00:00:24
3	Pipeline 1	Gradient Boosting Classifier	0.802	None	00:

Challenger models



Credit Risk Evaluation

Model
Credit Risk

Description
Evaluates credit applications for risk signals and applies a RISK or NO RISK verdict.

Model ID
fad9a34f-8e48-4231-a718-131...

Tests run 3
Tests pas 3
Tests fail 1

Fairness
90%
within threshold

Fairness by feature

Age	90%
Sex	91%
Race	92%

Quality
.99
within threshold

Quality metrics

Area under ROC	.99
Area under PR	.99
Accuracy	.95
True positive rate (TPR)	.89
False positive rate (FPR)	.15
Recall	.89

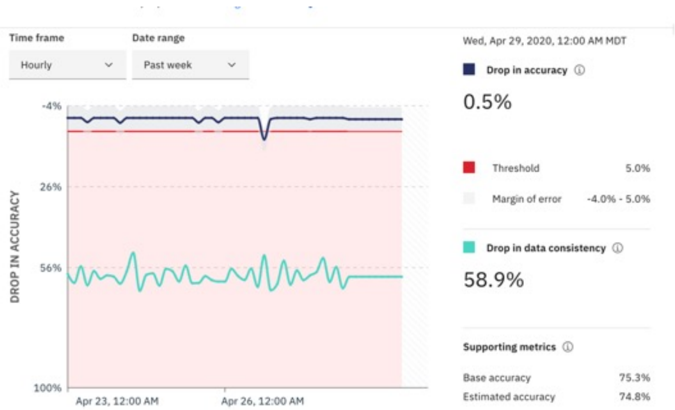
Compare model

Description
Select a model for comparison.

Credit Risk V2

Model		Credit Risk	Credit Risk V2
Fairness	Age	99	99
	Sex	99	99
Quality	Area under ROC	99	99
	Area under PR	99	99
	Accuracy	99	99
	True positive rate (TPR)	99	99
	False positive rate (FPR)	99	99
	Precision	99	99

Validation, Model Risk Management



Time frame: Hourly
Date range: Past week

Wed, Apr 29, 2020, 12:00 AM MDT

Drop in accuracy
0.5%

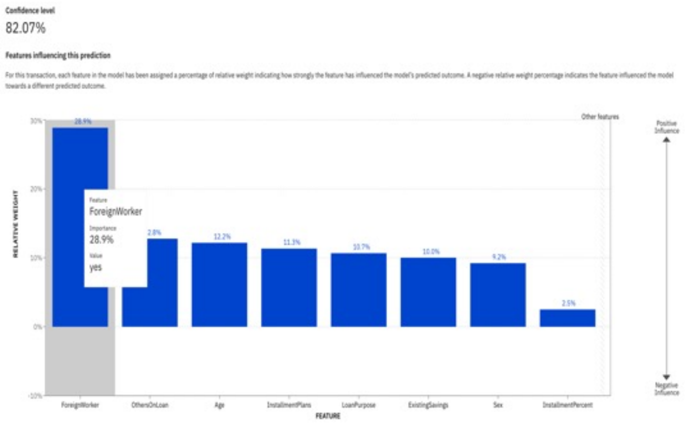
Threshold	5.0%
Margin of error	-4.0% - 5.0%

Drop in data consistency
58.9%

Supporting metrics

Base accuracy	75.3%
Estimated accuracy	74.8%

Drift in data consistency, Drift in accuracy



Confidence level
82.07%

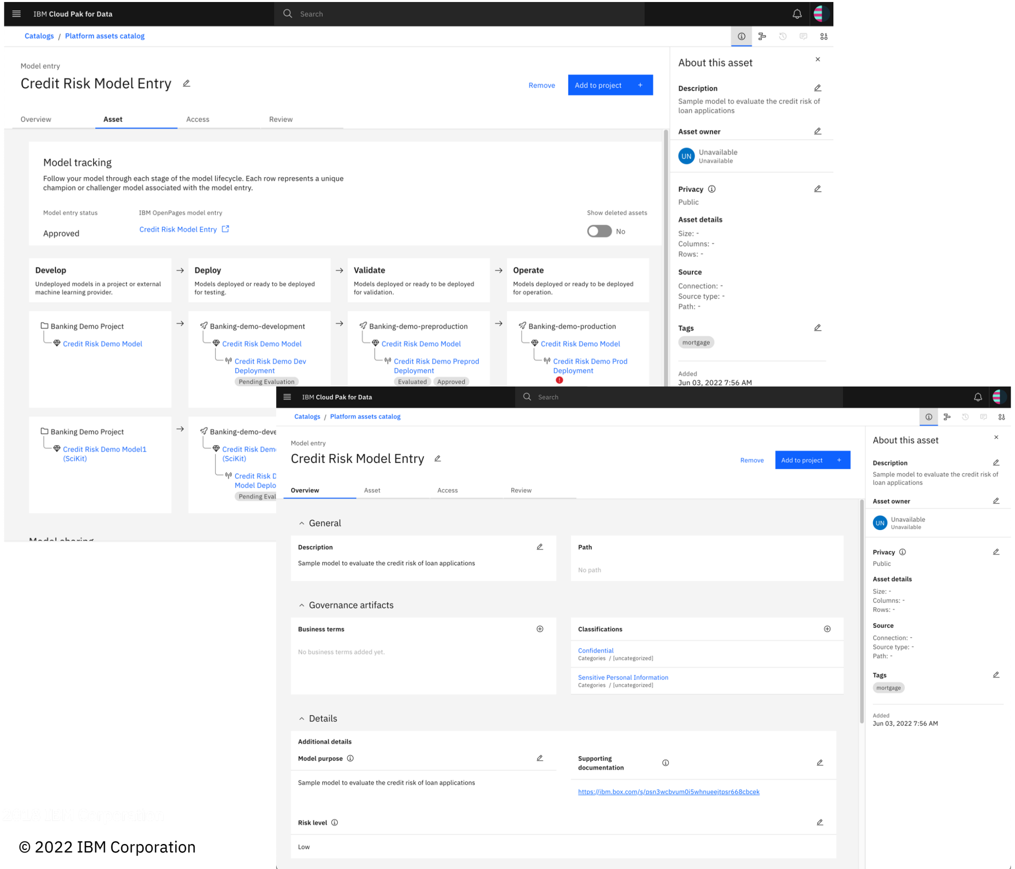
Features influencing this prediction

For this transaction, each feature in the model has been assigned a percentage of relative weight indicating how strongly the feature has influenced the model's predicted outcome. A negative relative weight percentage indicates the feature influenced the model towards a different predicted outcome.

Feature: ForeignWorker
Importance: 28.9%
Value: yes

Local and contrastive explanations

Seamlessly gather facts across the AI lifecycle



IBM Cloud Pak for Data

Search

Catalogs / Platform assets catalog

Model entry

Credit Risk Model Entry

Remove

Add to project +

Overview | **Asset** | Access | Review

Model tracking

Follow your model through each stage of the model lifecycle. Each row represents a unique champion or challenger model associated with the model entry.

Model entry status	IBM OpenPages model entry	Show deleted assets
Approved	Credit Risk Model Entry	No

Develop	Deploy	Validate	Operate
Undeployed models in a project or external machine learning provider.	Models deployed or ready to be deployed for testing.	Models deployed or ready to be deployed for validation.	Models deployed or ready to be deployed for operation.
Banking Demo Project Credit Risk Demo Model	Banking-demo-development Credit Risk Demo Model Credit Risk Demo Dev Deployment Pending Evaluation	Banking-demo-preproduction Credit Risk Demo Model Credit Risk Demo Preprod Deployment Evaluated Approved	Banking-demo-production Credit Risk Demo Model Credit Risk Demo Prod Deployment
Banking Demo Project Credit Risk Demo Model1 (SciKit)	Banking-demo-deve Credit Risk Dem (SciKit) Credit Risk D Model Deplo Pending Eval		

Model sharing

About this asset

Description

Sample model to evaluate the credit risk of loan applications

Asset owner

UN Unavailable
Unavailable

Privacy

Public

Asset details

Size: -
Columns: -
Rows: -

Source

Connection: -
Source type: -
Path: -

Tags

mortgage

Added
Jun 03, 2022 7:56 AM

IBM Cloud Pak for Data

Search

Catalogs / Platform assets catalog

Model entry

Credit Risk Model Entry

Remove

Add to project +

Overview | Asset | Access | Review

General

Description

Sample model to evaluate the credit risk of loan applications

Path

No path

Governance artifacts

Business terms

No business terms added yet.

Classifications

Confidential
Categories / [uncategorized]

Sensitive Personal Information
Categories / [uncategorized]

Details

Additional details

Model purpose

Sample model to evaluate the credit risk of loan applications

Supporting documentation

https://ibm.box.com/s/psn3wcbvum0i5whnueejtpsr668cbcek

Risk level

Low

About this asset

Description

Sample model to evaluate the credit risk of loan applications

Asset owner

UN Unavailable
Unavailable

Privacy

Public

Asset details

Size: -
Columns: -
Rows: -

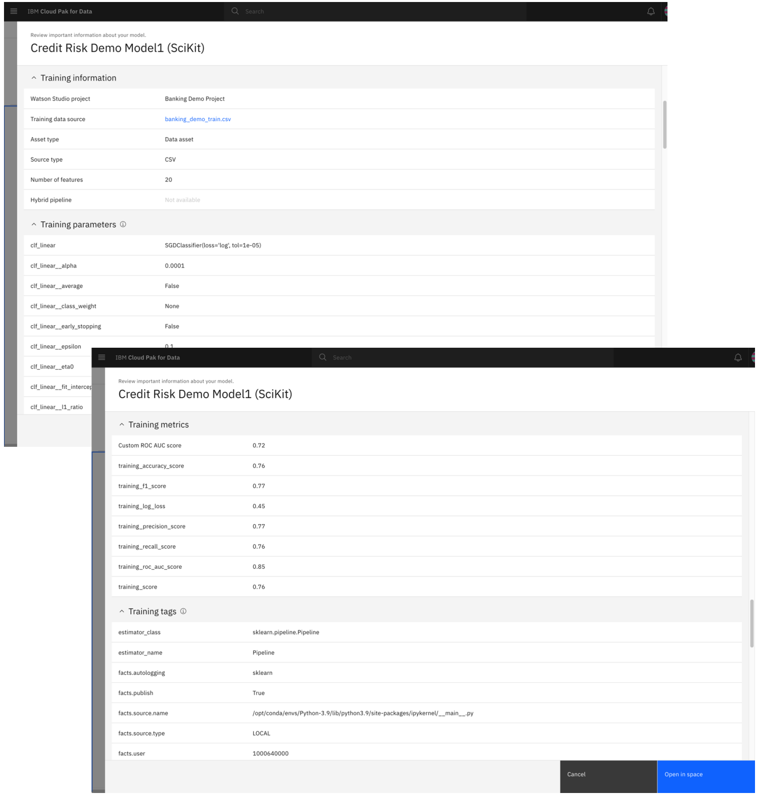
Source

Connection: -
Source type: -
Path: -

Tags

mortgage

Added
Jun 03, 2022 7:56 AM



IBM Cloud Pak for Data

Search

Review important information about your model.

Credit Risk Demo Model1 (SciKit)

Training information

Watson Studio project	Banking Demo Project
Training data source	banking_demo_train.csv
Asset type	Data asset
Source type	CSV
Number of features	20
Hybrid pipeline	Not available

Training parameters

clf_linear	SGDClassifier(loss='log', tol=1e-05)
clf_linear__alpha	0.0001
clf_linear__average	False
clf_linear__class_weight	None
clf_linear__early_stopping	False
clf_linear__epsilon	0.1
clf_linear__eta0	
clf_linear__fit_intercep	
clf_linear__l1_ratio	

IBM Cloud Pak for Data

Search

Review important information about your model.

Credit Risk Demo Model1 (SciKit)

Training metrics

Custom ROC AUC score	0.72
training_accuracy_score	0.76
training_f1_score	0.77
training_log_loss	0.45
training_precision_score	0.77
training_recall_score	0.76
training_roc_auc_score	0.85
training_score	0.76

Training tags

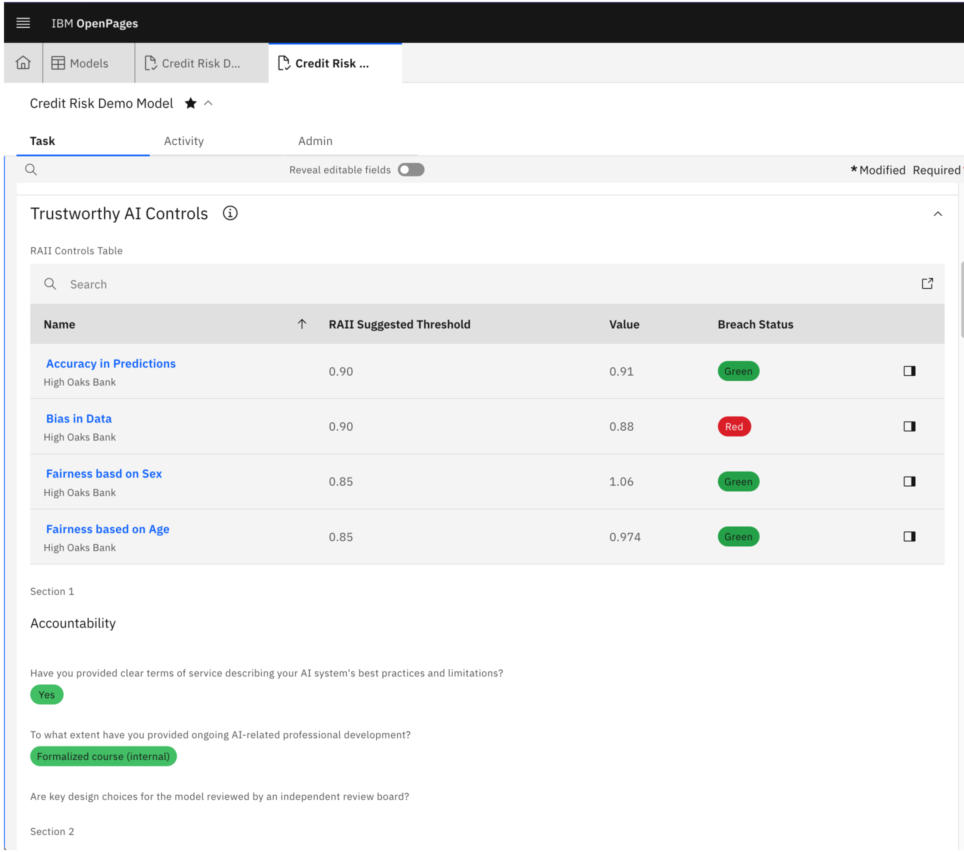
estimator_class	sklearn.pipeline.Pipeline
estimator_name	Pipeline
facts.autologging	sklearn
facts.publish	True
facts.source.name	/opt/conda/envs/Python-3.9/lib/python3.9/site-packages/ipykernel/__main__.py
facts.source.type	LOCAL
facts.user	1000640000

Cancel

Open in space

How to test and evaluate AI standards

- Were the following key design choices for the model reviewed by an independent review board? (y/n)
- Who has been informed about the AI system’s potential or perceived risks? (select all which apply)
- What was the result of the demographic parity test for the data used to train the model?



Trustworthy AI Controls

RAII Controls Table

Name	RAII Suggested Threshold	Value	Breach Status
Accuracy in Predictions High Oaks Bank	0.90	0.91	Green
Bias in Data High Oaks Bank	0.90	0.88	Red
Fairness basd on Sex High Oaks Bank	0.85	1.06	Green
Fairness based on Age High Oaks Bank	0.85	0.974	Green

Section 1

Accountability

Have you provided clear terms of service describing your AI system's best practices and limitations?
Yes

To what extent have you provided ongoing AI-related professional development?
Formalized course (internal)

Are key design choices for the model reviewed by an independent review board?

Section 2

IBM Cloud Pak for Data enables a governed, automated AI lifecycle

Model Development

Watson Studio (Build)

MLOps
AutoAI

Watson Studio (Deploy)

3rd party ML engines

Azure ML
AWS Sagemaker
Google Cloud ML
Open-source ML platforms

Model train, serve
ModelOps

AI Fact collection, Model Inventory

AI Factsheets

Catalog, policies, enforcement

Watson Knowledge Catalog

Risk and Compliance Workflow

IBM Open Pages

Validation, Testing and Monitoring

Watson Studio (Trust)

AI Governance Solution

Multi-cloud Data and AI platform

Red Hat OpenShift

IBM **Cloud**
aws
Microsoft Azure
Google Cloud
Hyperconverged private cloud system

Trustworthy AI requires a multidisciplinary approach

From principles to actions

what should be done

principles, values, norms, laws, regulations

how to instrument it

techniques, algorithms, software, best practices

how to operationalize it

mechanisms, systems, and processes to keep AI trustworthy

How IBM can help you get started with trustworthy AI

Co-Operate

Co-Create

Co-Execute

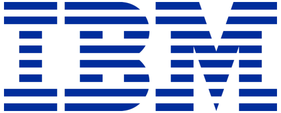
Education & Guidance on Trustworthy AI	Design Thinking & Planning for Trustworthy AI	Trusted AI Build	Trusted MLOps Validate & Deploy	Trusted MLOps Monitor & Manage
1 Week	2 Weeks	6 Weeks	6 Weeks	6 Weeks
✓ Build customer's baseline knowledge ✓ Guide customer through hands-on work	✓ Identify, scope and prioritize use cases ✓ Detailed task analysis for trustworthy AI	✓ Explore, cleanse, prepare datasets for use case ✓ Build & test trustworthy AI/ML model	✓ Implement Best Practices for model validation ✓ Setup deploy pipeline integrated with CI/CD	✓ Implement monitors for Bias/Fairness, Drift, etc. ✓ Model Explainability On Demand
❖ Outcome: Ability to move to Plan-Build stages	❖ Outcome: Action Plan for Build/Deploy/Manage	❖ Outcome: Working AI/ML model	❖ Outcome: Validation and Deployment pipeline	❖ Outcome: Monitoring and Management setup

Contact IBM at www.ibm.com/products/expertlabs/trustworthy-ai

RAII: How to get involved in the community

For more updates, connect with
the Responsible AI Institute at:

- linkedin.com/company/responsible-ai-institute
- twitter.com/ResponsibleAI
- www.responsible.ai


IBM