Monitor a model for quality and fairness

Overview

AI is used in everyday life to support human decision-making.

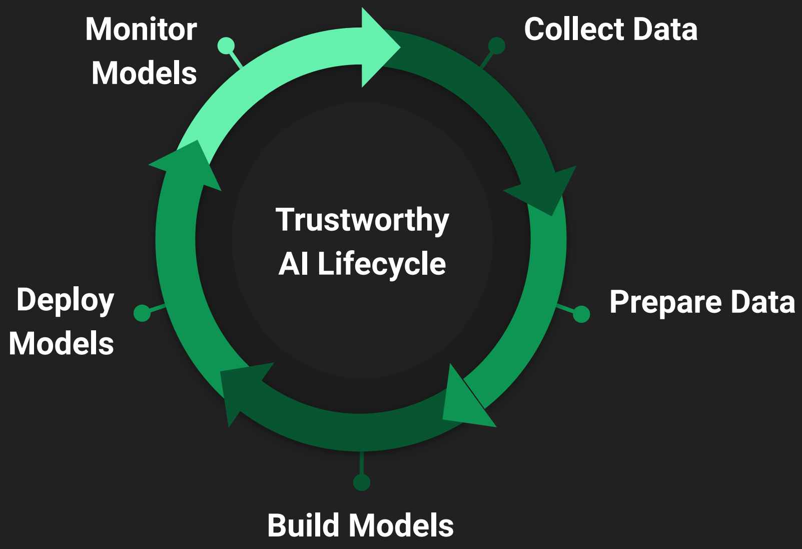
AI is computer algorithm. For many people, it's a block box.

How do we operationalize AI with confidence?

What is Fairness?

How to measure bias?

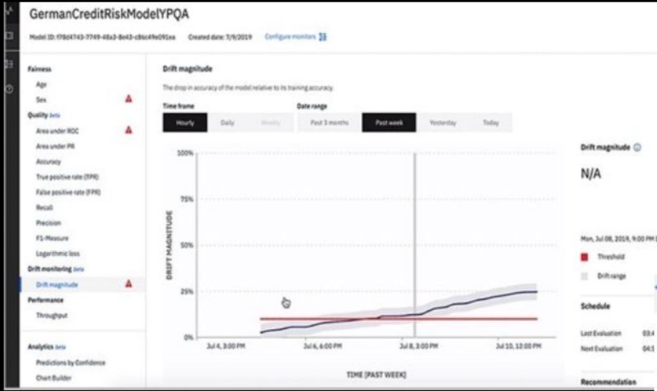
How to enable responsible use of AI?



IBM Watson OpenScale

Implement responsible, explainable AI

Mitigate drift, bias, and model risk



GermanCreditRiskModelYPQA

Model ID: f78d4743-7749-48a3-8e43-c86c49e091ea Created date: 7/9/2019 Configure monitors

Drift magnitude

The drop in accuracy of the model relative to its training accuracy

Time frame: Hourly | Daily | Weekly

Date range: Past 3 months | Past week | Yesterday | Today

Drift magnitude

N/A

Mon, Jul 08, 2019, 9:00 PM IST

🟥 Threshold	N/A
▨ Drift range	

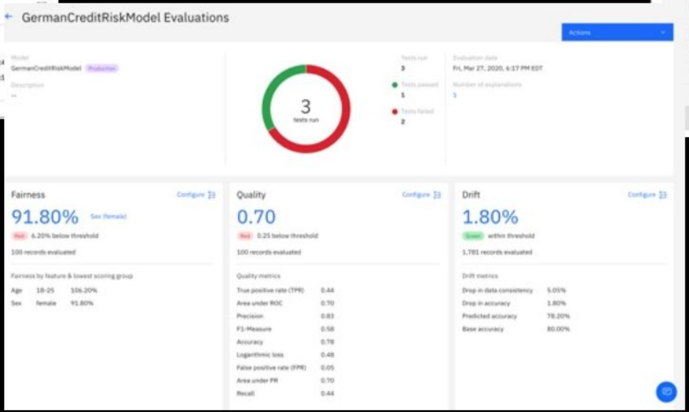
Schedule

Last Evaluation 03:4

Next Evaluation 04:1

Recommendation

Drift: monitor model drift by hourly, daily or weekly



GermanCreditRiskModel Evaluations

Model: GermanCreditRiskModel Production

Description: --

Tests run: 3

Tests passed: 1

Tests failed: 2

Evaluation date: Fri, Mar 27, 2020, 6:17 PM EDT

Number of explanations: 1

Fairness Configure

91.80% Sex (female)

Red 6.20% below threshold

100 records evaluated

Fairness by feature & lowest scoring group

Age	18-25	106.20%
Sex	female	91.80%

Quality Configure

0.70

Red 0.25 below threshold

100 records evaluated

Quality metrics

True positive rate (TPR)	0.44
Area under ROC	0.70
Precision	0.83
F1-Measure	0.58
Accuracy	0.78
Logarithmic loss	0.48
False positive rate (FPR)	0.05
Area under PR	0.70
Recall	0.44

Drift Configure

1.80%

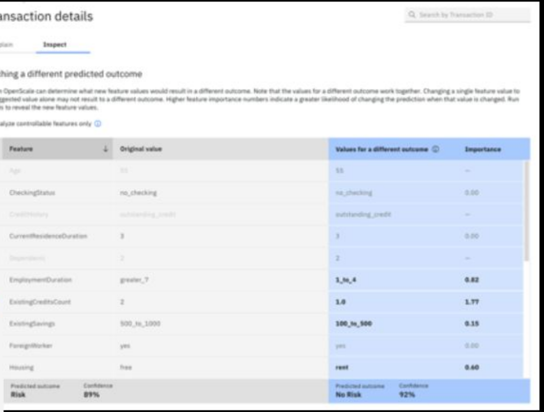
Green within threshold

1,781 records evaluated

Drift metrics

Drop in data consistency	5.05%
Drop in accuracy	1.80%
Predicted accuracy	78.20%
Base accuracy	80.00%

Model risk evaluation: fairness, quality and drift metrics to share model insights



Transaction details

Reaching a different predicted outcome

Watson OpenScale can determine what new feature values would result in a different outcome. Note that the values for a different outcome work together. Changing a single feature value to the suggested value alone may not result to a different outcome. Higher feature importance numbers indicate a greater likelihood of changing the prediction when that value is changed. Run analysis to reveal the new feature values.

☑ Analyze controllable features only

	Feature	Original value	Values for a different outcome	Importance
🔒	Age	55	55	--
	CheckingStatus	no_checking	no_checking	0.00
🔒	CreditHistory	outstanding_credit	outstanding_credit	--
	CurrentResidenceDuration	3	3	0.00
🔒	Dependents	2	2	--
	EmploymentDuration	greater_7	**1_to_4**	**0.82**
	ExistingCreditsCount	2	**1.0**	**1.77**
	ExistingSavings	500_to_1000	**100_to_500**	**0.15**
	ForeignWorker	yes	yes	0.00
	Housing	free	**rent**	**0.60**
	Predicted outcome **Risk**	Confidence **89%**	Predicted outcome **No Risk**	Confidence **92%**

Explain transactions: Determine what features reach different outcomes

Demo

1. Tutorial scenario: https://github.com/IBM/ai-data-workshop/tree/main/monitor-model-with-openscale
2. Notebook
3. Evaluation
4. Insight Dashboard
    a. Fairness monitor
    b. Quality monitor
    c. Explain transaction

AI Fairness 360

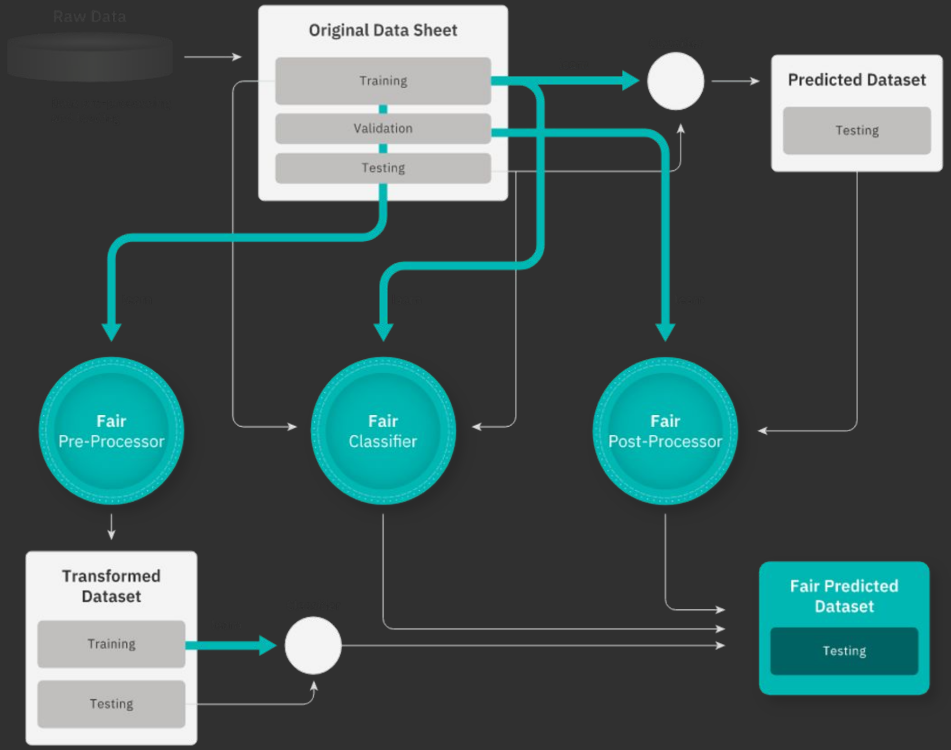
AI Fairness 360 (Open Source)

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 Demo

1. Import Notebook
2. Run notebook