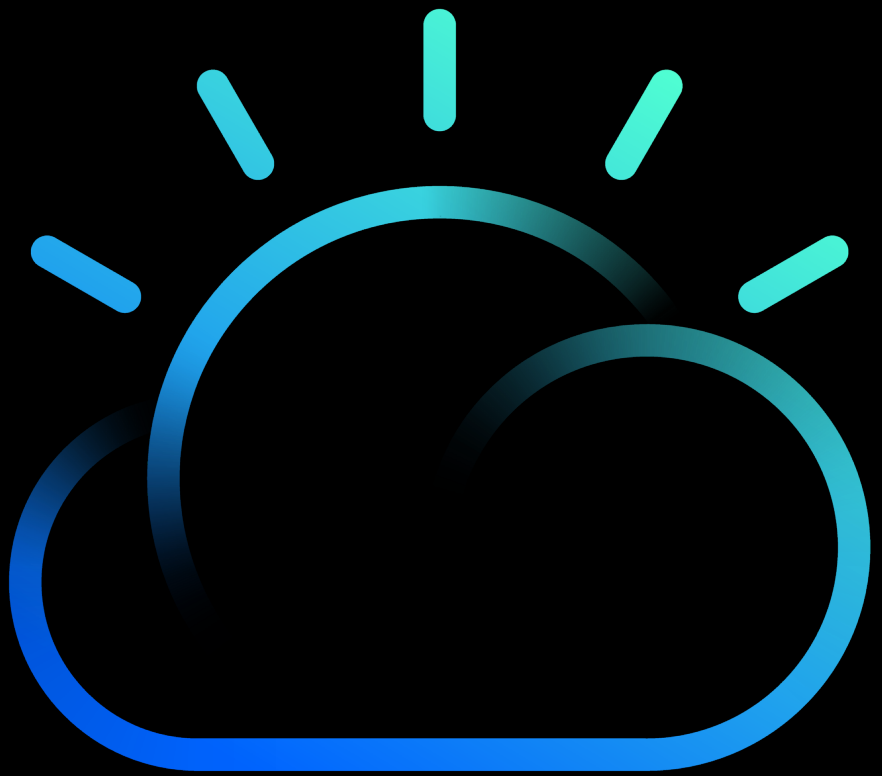
Manage and Monitor ML Models in Development, Test and Production

Naresh Olladapu
Developer Advocate
IBM – ISL(India Software Labs)
Email: nolladap@in.ibm.com



IBM **Cloud**

Overall Agenda

1. Introduction to AI Explainability and Bias Mitigation
2. IBM Watson OpenScale – Introduction and features
3. OpenScale Demo

Business stakeholders do not trust AI

60%

of companies see **regulatory constraints** as a barrier to implementing AI.

– *IBM IBV AI 2018*

63%

cite availability of **technical skills** as a challenge to implementation.

- IBM IBV AI 2018

Without expensive Data Science resources handholding multiple AI models in a production application:

1. No way to **validate** if AI models are **compliant with regulations** and will achieve expected business outcomes before deploying
2. Difficult to **track and measure** indicators of business success in production
3. Resource intensive and unreliable processes for **ongoing business monitoring and compliance**

Current Model Risk Management practices work for existing statistical or rule-based models

Processes rely on regular interaction between validator and model developer

- Easy to state and translate the regulatory guidelines into requirements
- Validators can reproduce the results relatively easily

Current systems focus more on the documentation and governance aspects of model validation

- Who developed the model?
- What data was used to test it?
- How did data and model evolve?
- Who approved it?
- Were approved techniques used?

Traditional statistical models are deterministic in nature or simpler to interpret and explain

- Business rules
- Descriptive statistics
- Standard financial formulae
- Excel spreadsheets
- Linear regressions
- Decision trees

Validation of AI models need to augment governance with proactive testing before and in production

Skills gap in order to test and validate ML/DL models

- Probabilistic nature of ML/DL models do not provide a straightforward result for interpretation
- Interpreting results and explaining to business managers and risk professionals require substantial time and effort from data scientists
- Model metrics do not convey business KPI impacts

Tests to be run on ML/DL models differ from straightforward application testing

- Model training data sets may not capture the range of data or combinations that could be encountered in runtime
- Any change in input data – also known as Drift – can cause the model to make inaccurate decisions, impacting business KPIs
- Training data may be cleaned to be free from bias but runtime data may induce biased behavior of model
- Difficult to run what-if scenarios on probabilistic models without additional data sets

Model Development Cycle

Challenges faced with ML/DL Models

Complexity and Assessment: Lack of knowledge of methods used by Model Developers / Vendors along with inconsistent documentation and increased volume of model change.

Tests to be run on ML/DL models differ from straightforward application testing

- **Drift**: Any change in input data – also known as Drift – can cause the model to make inaccurate decisions, impacting business KPIs
- **Bias**: Training data may be cleaned to be free from bias, but runtime data may induce biased behavior of model
- **Explainability**: Traditional statistical models are simpler to interpret and explain
- **Missing Validation/Test Data**: Model training data sets may not capture the range of data or combinations that could be encountered in runtime

Watson OpenScale

Validate and monitor AI models, deployed anywhere, to help comply with regulations, address internal safeguards, and mitigate business risk

Production monitoring for compliance and safeguards

Mitigate biased model behavior

Explain model decisions

Validate and control risk

Ensure that models are resilient to changing situations

Detect drift during runtime

Generate specific model retraining inputs

Align model performance with business outcomes

Actionable metrics and alerts

Open: Monitor and optimize models deployed on any model serve engine, beyond WML

Description:

Use Watson OpenScale to monitor models developed in a 3rd party IDE, open source framework and hosted in a 3rd party or private model serve engine

Value:

- Automate and Operationalize AI wherever your AI and applications reside
- Leverage flexibility and innovation of open source model building and training frameworks
- Retain existing data and model pipelines deployed to support applications in production
- Extend investments in skills and tools deployed as part of current data science and AI stacks

Model build/train frameworks

TensorFlow

scikit learn

Keras

PYTORCH

Caffe2

Spark MLlib

Model serving environments

kubernetes

aws

Azure ML

Multi-Cloud Deployment
Through IBM Cloud Pak for Data

- Watson OpenScale can be deployed wherever organizations need it
  - IBM Public Cloud
  - IBM Cloud Pak For Data
- Deployment through IBM Cloud Pak for Data provides a common installation and management experience wherever the offering is located

PUBLIC

Maximize on cloud agility and economics

PRIVATE

Cloud native, behind the firewall, client managed

Cloud Native Experience

Regardless of which cloud you choose, you get market leading Watson capabilities.

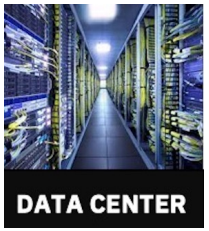
IBM SOFTLAYER

OPENSHIFT

amazon web services

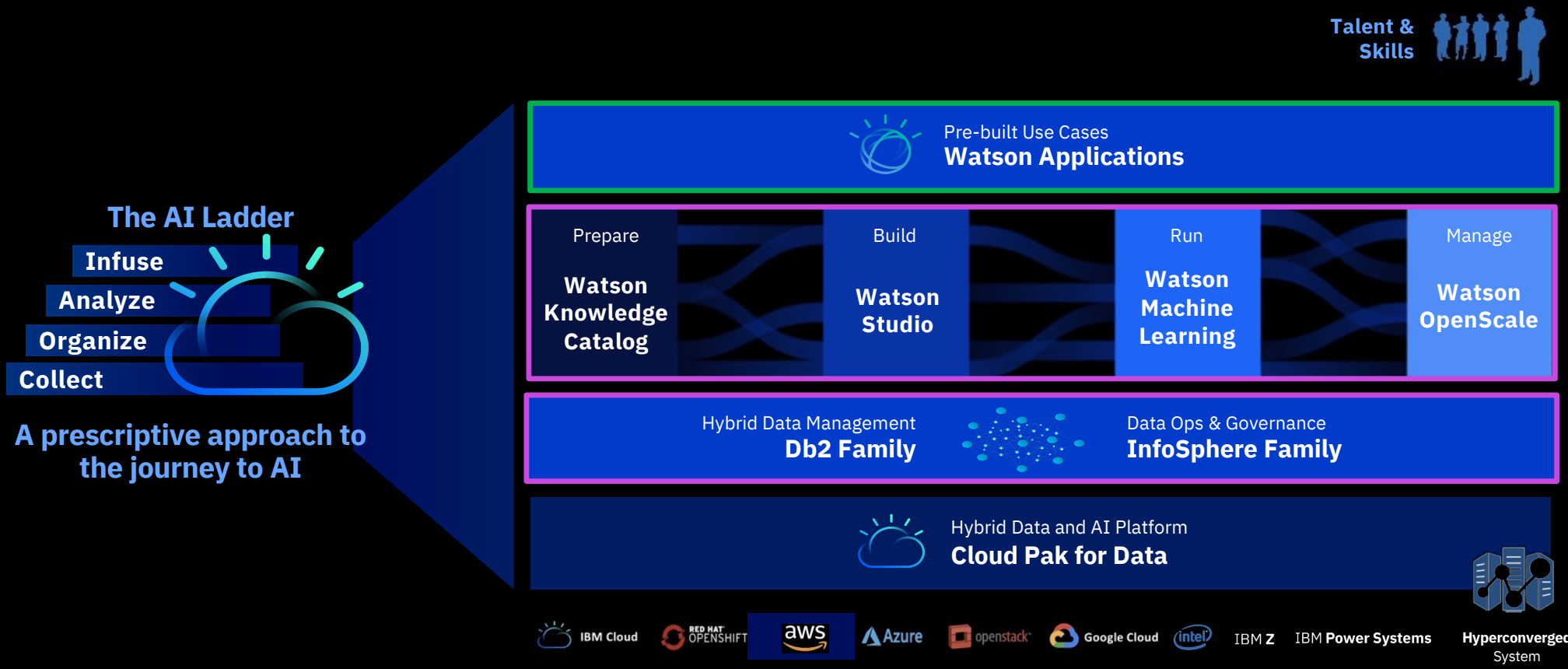
Microsoft Azure

openstack



The IBM Data and AI Portfolio

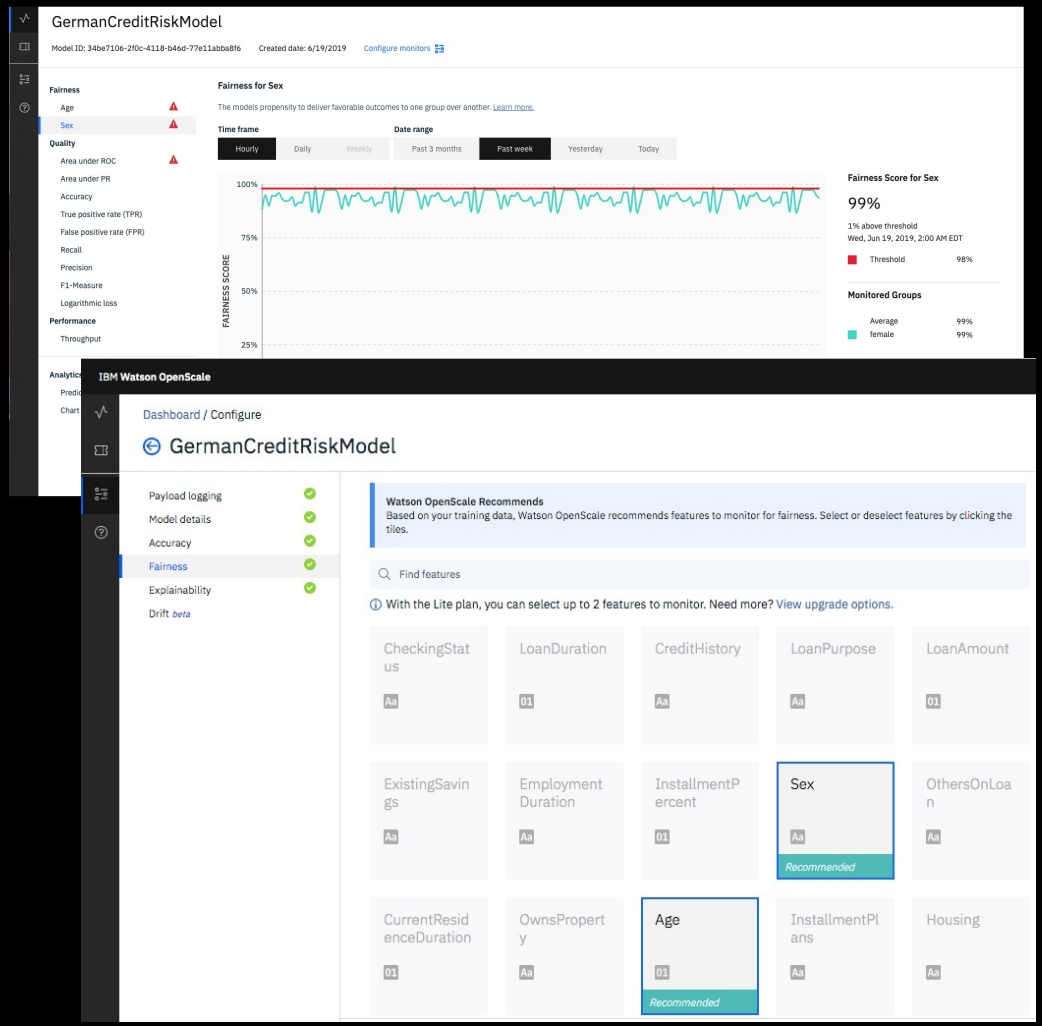
Everything you need for enterprise data and AI, on any cloud



Bias Detection

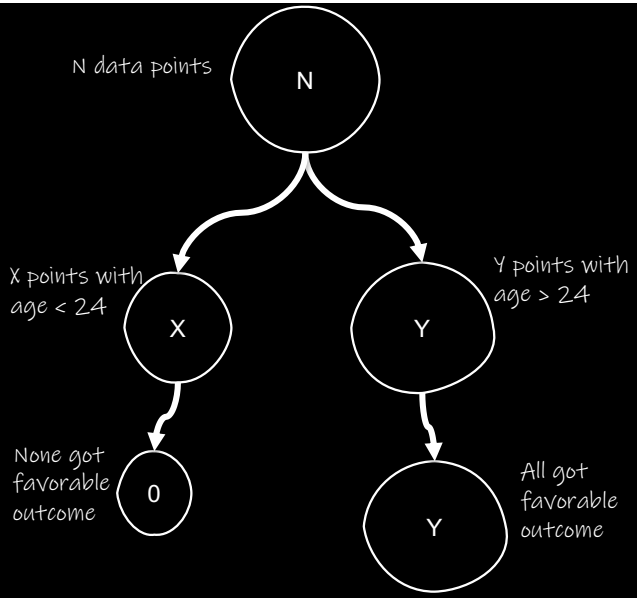
OpenScale enables enterprises to enforce fairness in their model's outcome by analyzing transactions in production and finding biased behavior by the model

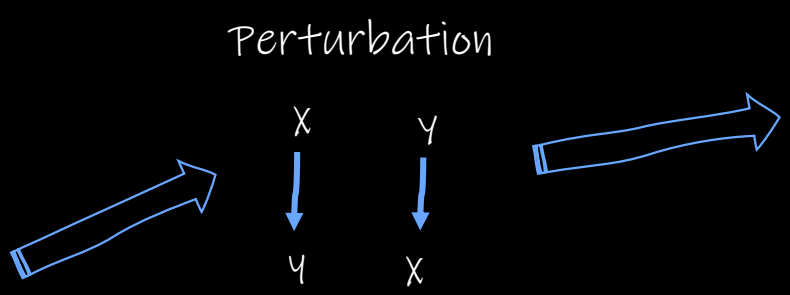
It pinpoints the source of bias and actively mitigates the biases found in production environment



Value:

- Automatically recommend common protected attributes to monitor during production
- Detect biases in runtime in order to catch impacts on business applications and compliance requirements without time consuming, manual data analysis
- Metrics and data to help data scientists further troubleshoot issues in data sets or models
- Mitigate biases in runtime in order to enforce regulatory or enterprise fairness guardrails in real time

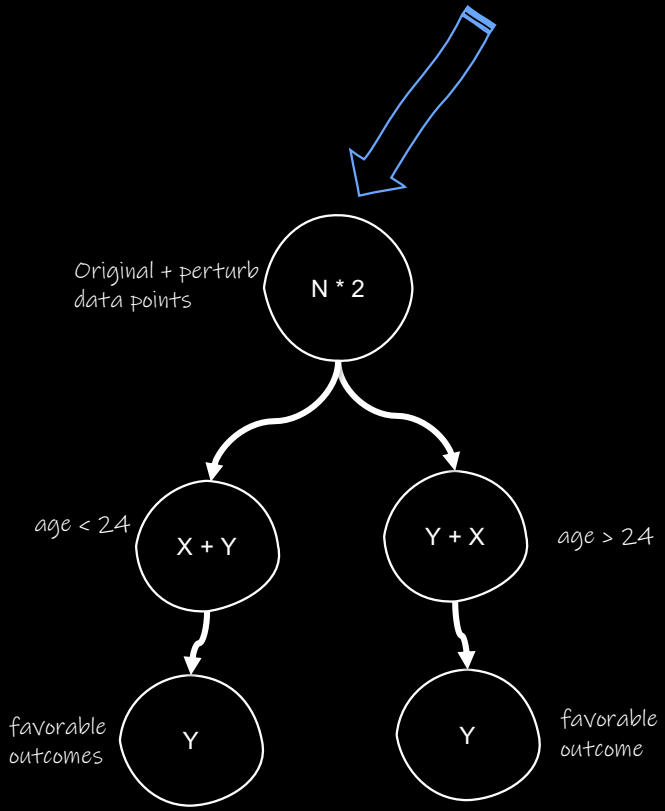




Perturbed records scoring

let's say

No change in prediction



disparate impact ratio = (0 / X) / (Y / Y) = 0%

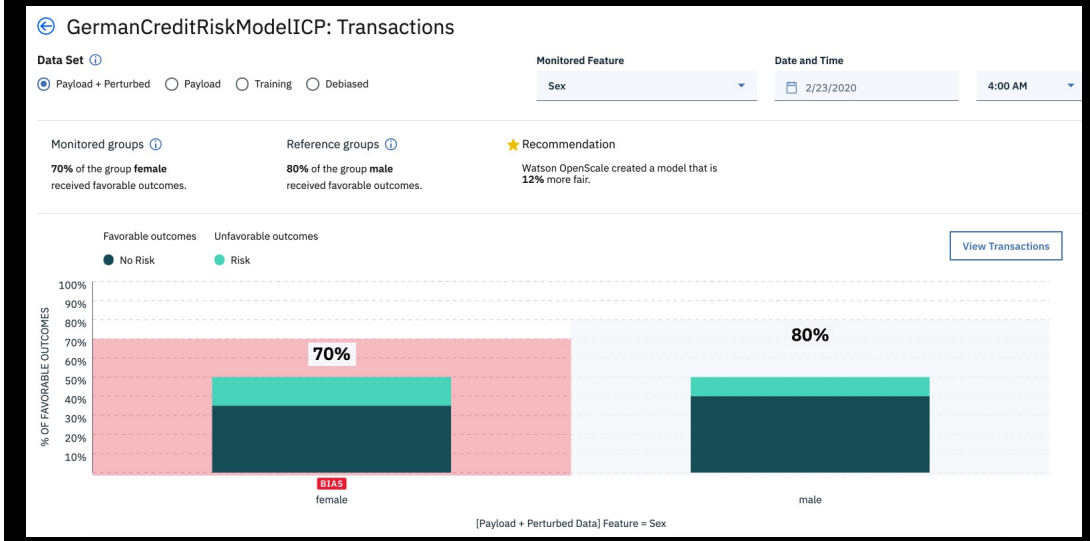
So, according to DI ratio, Model is highly biased.

But, is it really biased ??

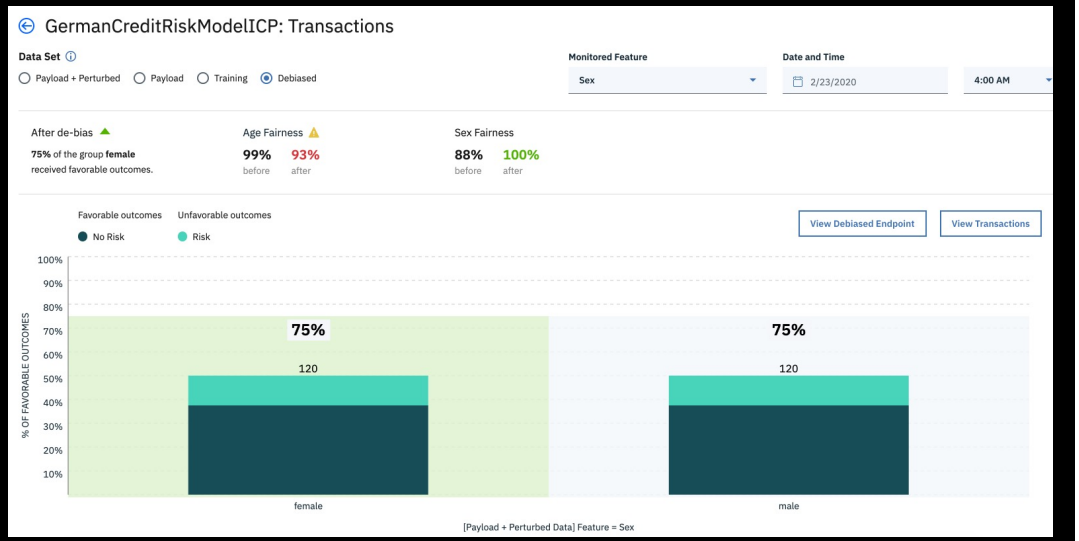
disparate impact ratio = (Y / (X + Y)) / (Y / (Y + X)) = 100%

You see.. model is not actually biased!

Bias Mitigation – Original Output



Bias Mitigation – De-biased output



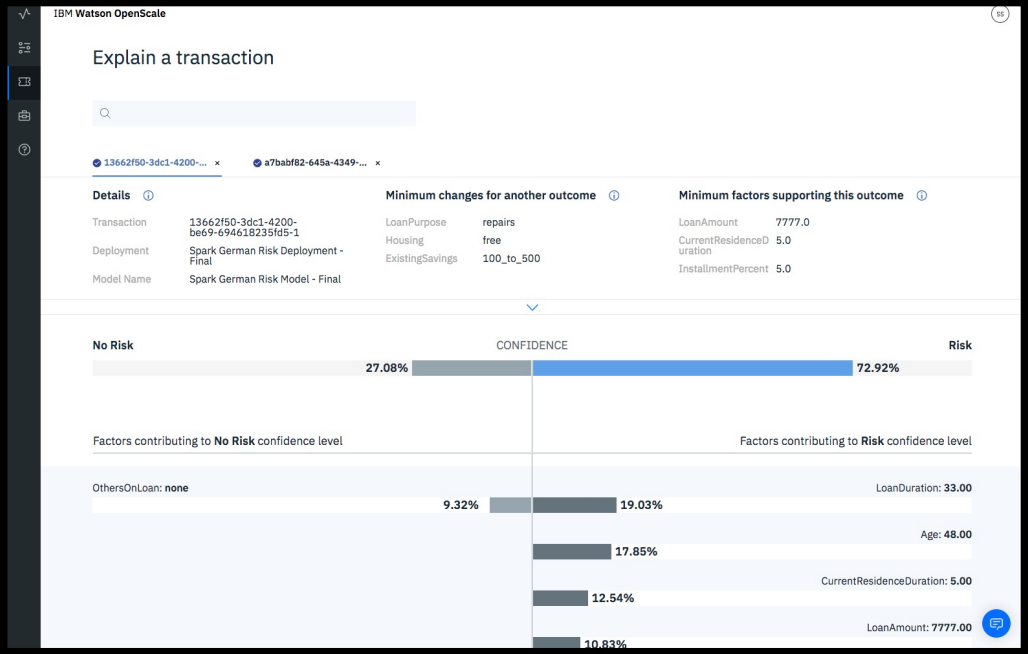
Explainability

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

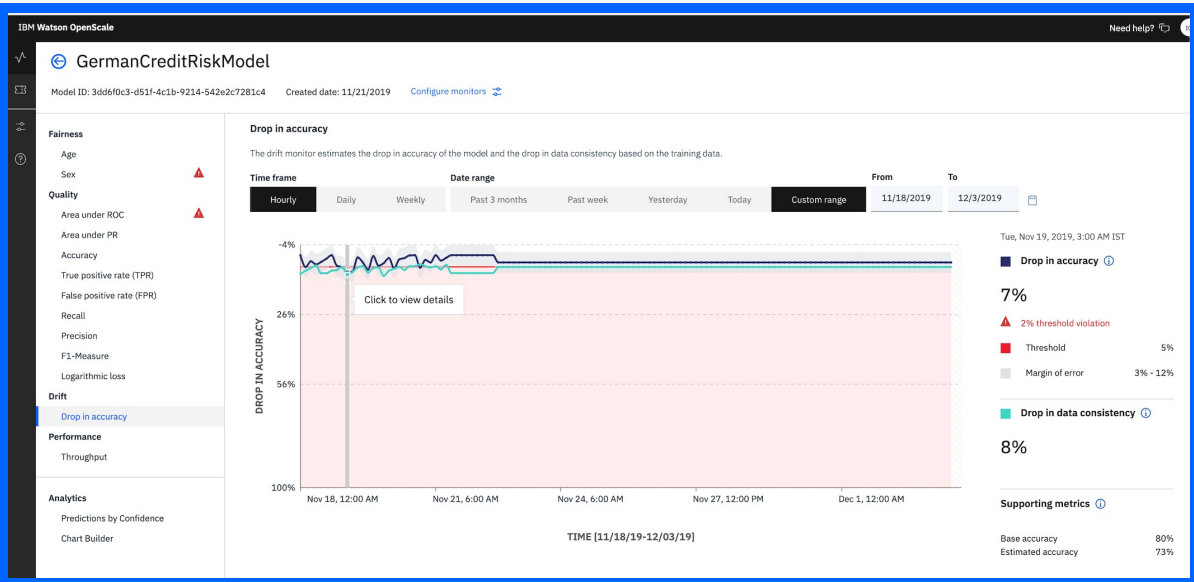
It provides a simple explanation that is user friendly and interactive

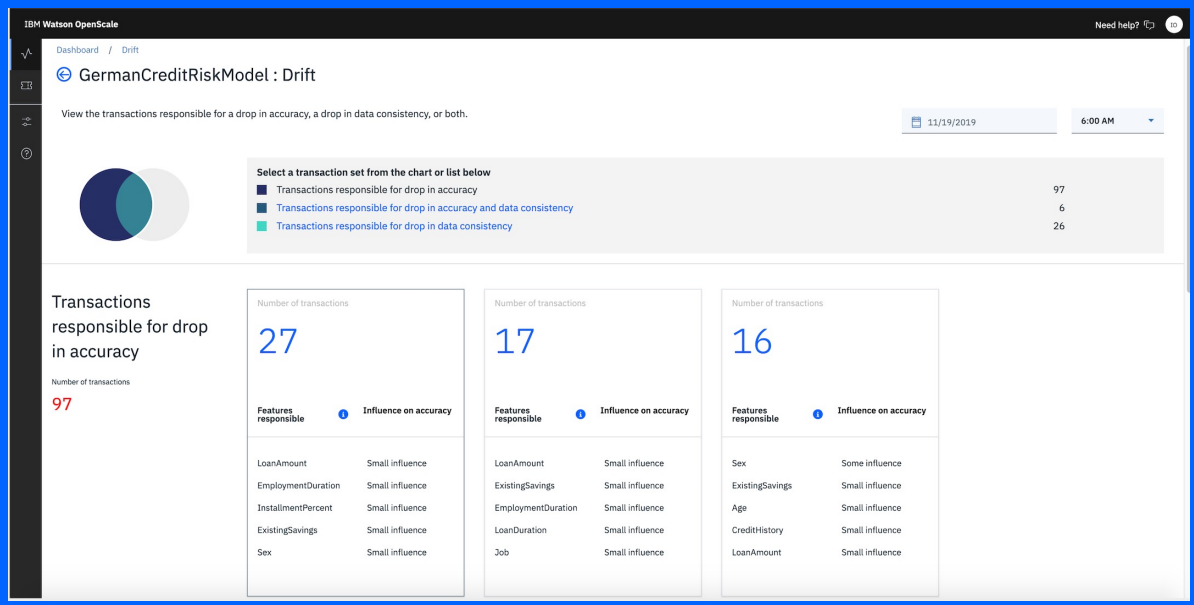
Value:

- 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



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





Emerging use case:

Model Risk Management in Financial Services

Problem:
Current risk management practices are not optimized for AI

- **Open source frameworks** not supported
- Additional **data science skills** required for validating AI models
- Processes rely on **manual interaction** between validator and model developer
- Current systems **focus more on the documentation and workflow** aspects of model validation, **no active testing**
- No focus on **active production monitoring**

Watson OpenScale will **automate active testing** of models for **validation and monitoring** and **synchronize results with governance platforms**

IBM Watson OpenScale

Need help? MB

Dashboard /

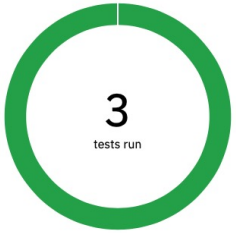
Credit Risk pre-prod deployment in B2 Evaluations

Actions

Model
Credit Risk pre-prod deployment in B2 Pre-production

Description
Credit Risk pre-prod deployment in B2

Model ID
21b4fe7b-5801-4266-9190-609ffe449646



Tests run
3

Tests passed
3

Tests failed
0

Evaluation date
Wed, Nov 27, 2019, 4:17 PM IST

Test data set
german_credit_data_biased_test.csv

Number of test records
200

Number of explanations
10

Fairness

Configure

95.10%

Green within threshold

100 records evaluated

Fairness by feature

Sex	95.10%
Age	95.20%

Quality

Configure

0.74

Green within threshold

100 records evaluated

Quality metrics

True positive rate (TPR)	0.58
Area under ROC	0.74
Precision	0.74
F1-Measure	0.65
Accuracy	0.81
Logarithmic loss	0.42
False positive rate (FPR)	0.09
Area under PR	0.65
Recall	0.58

Drift

Configure

1.30%

Green within threshold

100 records evaluated

Drift metrics

Drop in data consistency	7.75%
Drop in accuracy	1.30%
Predicted accuracy	78.20%
Base accuracy	79.50%

IBM Watson OpenScale + IBM OpenPages

Closed beta

Model Risk Management Solution for Financial Services

IBM OpenPages

Model Risk Governance

Store, manage and monitor a comprehensive model inventory

IBM Watson OpenScale

Active testing for model validation and continuous monitoring

OpenScale Demo


IBM