BADM 4830 / BAIM 4200 Advanced Business Analytics

Module 6 - MLOps – In Practice

Ivan Portilla
Jesus.Portilla@Colorado.edu
Portilla@gmail.com
github.com/jiportilla/giveback

Objectives of This Module

Upon completion of this module, you will understand:

1. DevOps Lifecycle
2. Natural Language Understanding
3. Model Bias Considerations
4. Ethics in AI
5. Lab: Social Media use case

Fall 2021
BADM 4830 / BAIM 4200 Advanced Business Analytics

- This course will give students the language, knowledge, and actionable methods to work alongside technical and non-technical members of your team to create AI solutions.

- Students will explore what it means to design artificial intelligence systems as a team, guided by a clear intent and a focus on people. This course will give you the framework and tools you need to recognize responsible AI design, align your team, and work with data sources to start building AI solutions.

- Students will learn the tools, technology, and practices that enable cross-functional AI teams to efficiently deploy, monitor, retrain, and govern models in production systems.

Re-cap

1. Model Evaluation
2. Model Bias Considerations – Trust & Transparency
3. Ethics in AI

Model Bias

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.

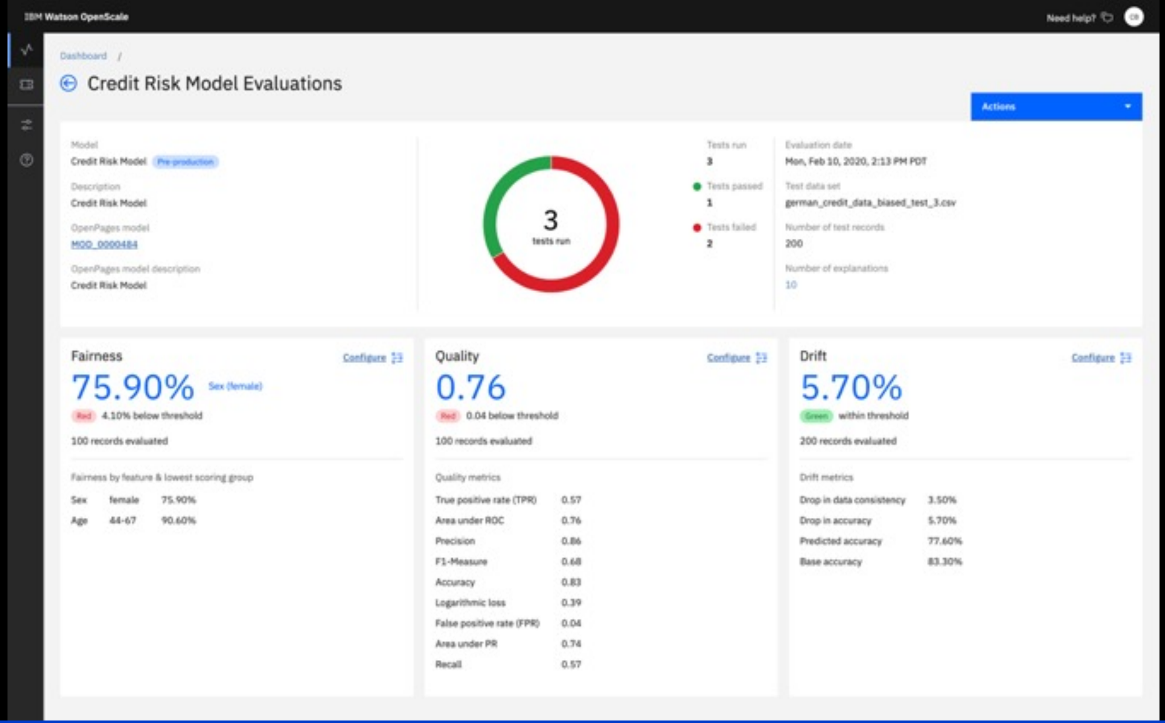
- IBM IBV AI 2018

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**

Manage AI Model Risk

Financial institutions need to manage risk associated with the usage of AI models



Credit Risk Model Evaluations

Model
Credit Risk Model Pre-production

Description
Credit Risk Model

OpenPages model
MOD_0000484

OpenPages model description
Credit Risk Model

3 tests run

Tests run
3

● Tests passed
1

● Tests failed
2

Evaluation date
Mon, Feb 10, 2020, 2:13 PM PDT

Test data set
german_credit_data_biased_test_3.csv

Number of test records
200

Number of explanations
10

Fairness

75.90% Sex (female)

Red 4.10% below threshold

100 records evaluated

Fairness by feature & lowest scoring group

Sex	female	75.90%
Age	44-67	90.60%

Quality

0.76

Red 0.04 below threshold

100 records evaluated

Quality metrics

True positive rate (TPR)	0.57
Area under ROC	0.76
Precision	0.86
F1-Measure	0.68
Accuracy	0.83
Logarithmic loss	0.39
False positive rate (FPR)	0.04
Area under PR	0.74
Recall	0.57

Drift

5.70%

Green within threshold

200 records evaluated

Drift metrics

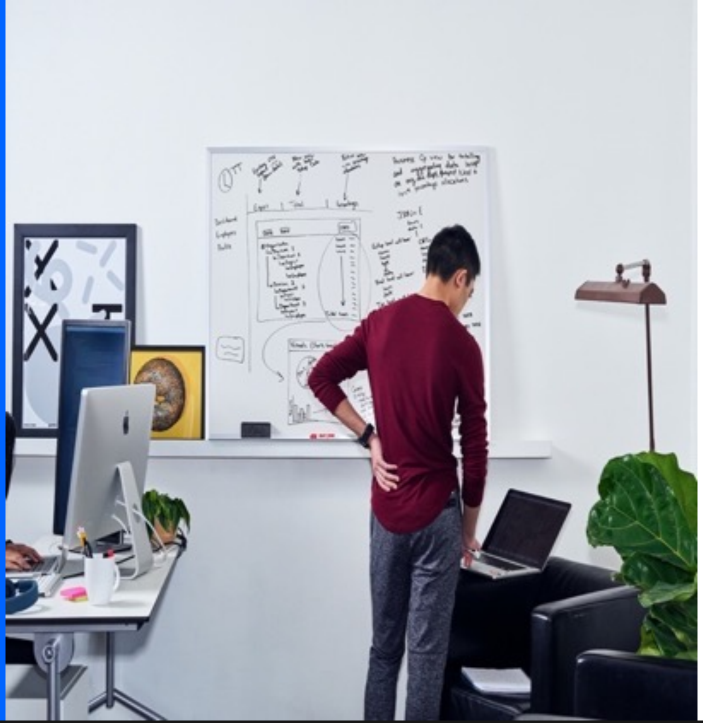
Drop in data consistency	3.50%
Drop in accuracy	5.70%
Predicted accuracy	77.60%
Base accuracy	83.30%

Problem:

Current risk management practices are not optimized for AI:

- **Traditional statistical models** are deterministic in nature or simpler to interpret and explain
- Current systems **focus more on the documentation and governance** aspects of model validation, **no active testing**
- No focus on **active production monitoring**
- Processes rely on **manual interaction** between validator and model developer

Model performance impact is not just technical: Business KPIs are impacted, too.



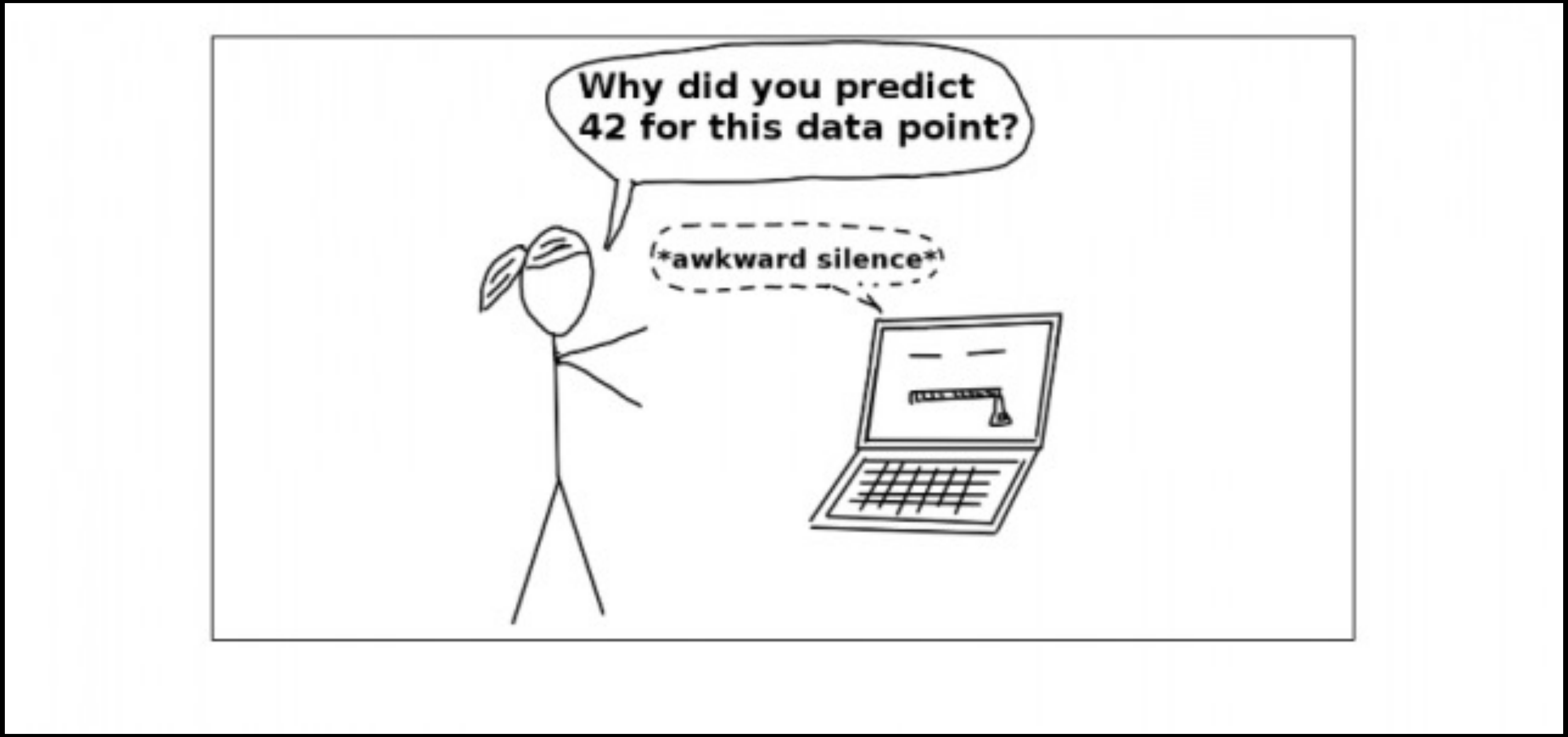
“The model’s output no longer features my products in customer emails. You’re hurting my team’s performance”

Business Stakeholder

“This is an amazing model. I was able to create a model that sends emails with individualized content optimized to increase clicks”

Data Scientist

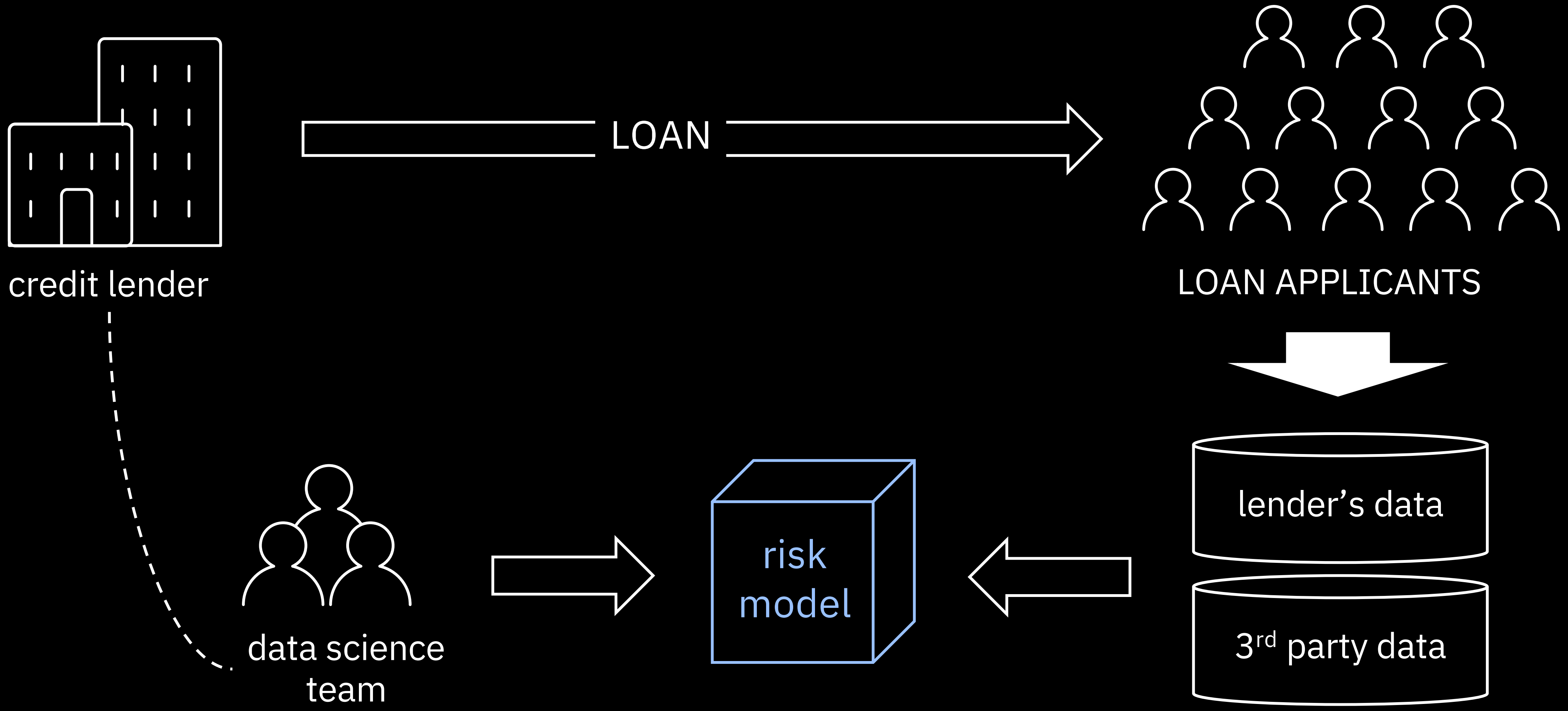
How can I trust this risk model?



https://app.mural.co/invitation/mural/badm48308688/1617567059330?sender=uc418b717494c964835994700&key=1d5a6232-e939-4722-96a7-b1853de6d62d

Credit Risk Model

low-cost financial services with transparent credit risk modeling



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

Watson OpenScale

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

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

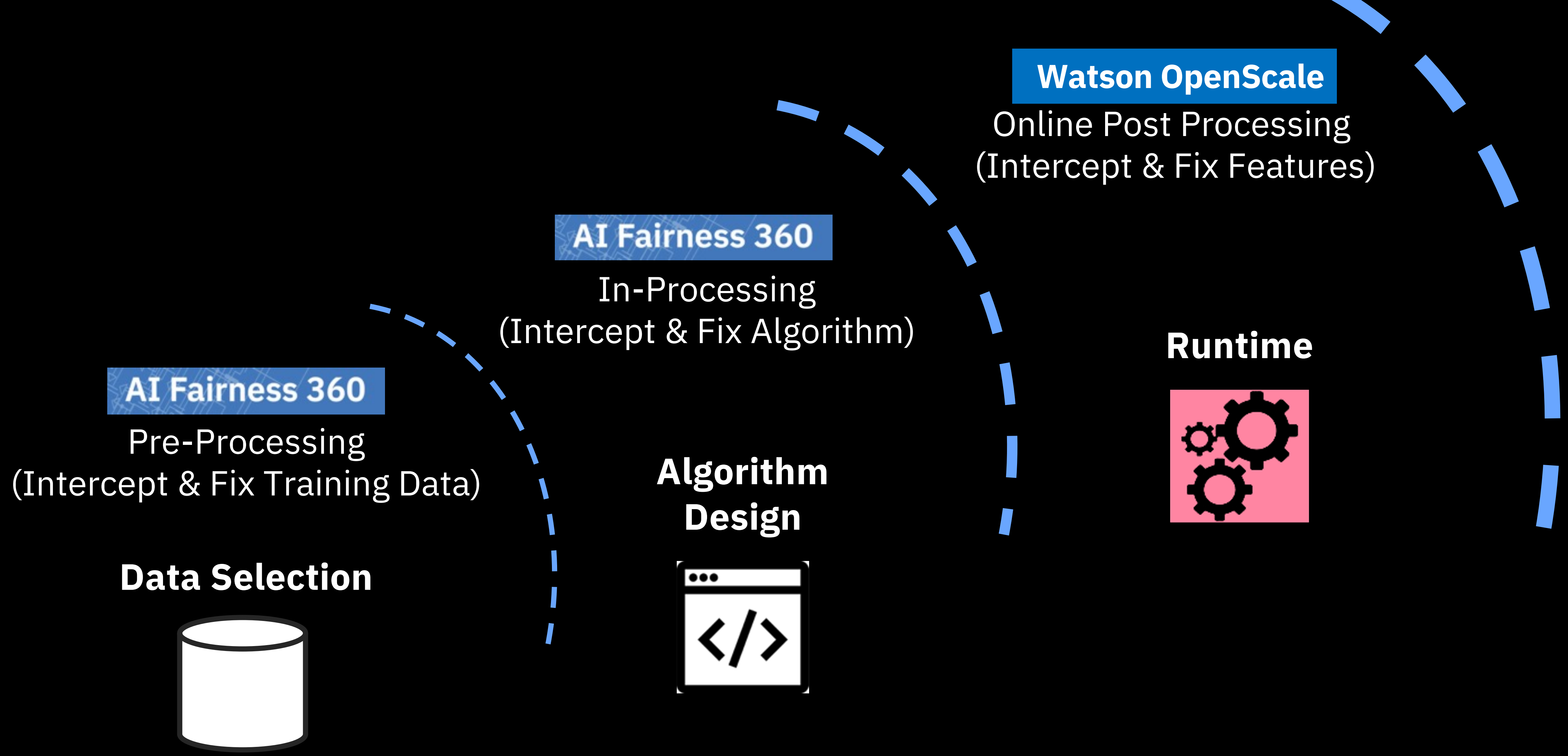
Generate specific model retraining inputs

Align model performance with business outcomes

Correlate model metrics and business KPIs

Actionable metrics and alerts

Using Watson OpenScale & Toolkits Together



Watson OpenScale

Online Post Processing
(Intercept & Fix Features)

AI Fairness 360

In-Processing
(Intercept & Fix Algorithm)

Runtime

AI Fairness 360

Pre-Processing
(Intercept & Fix Training Data)

Algorithm Design

Data Selection

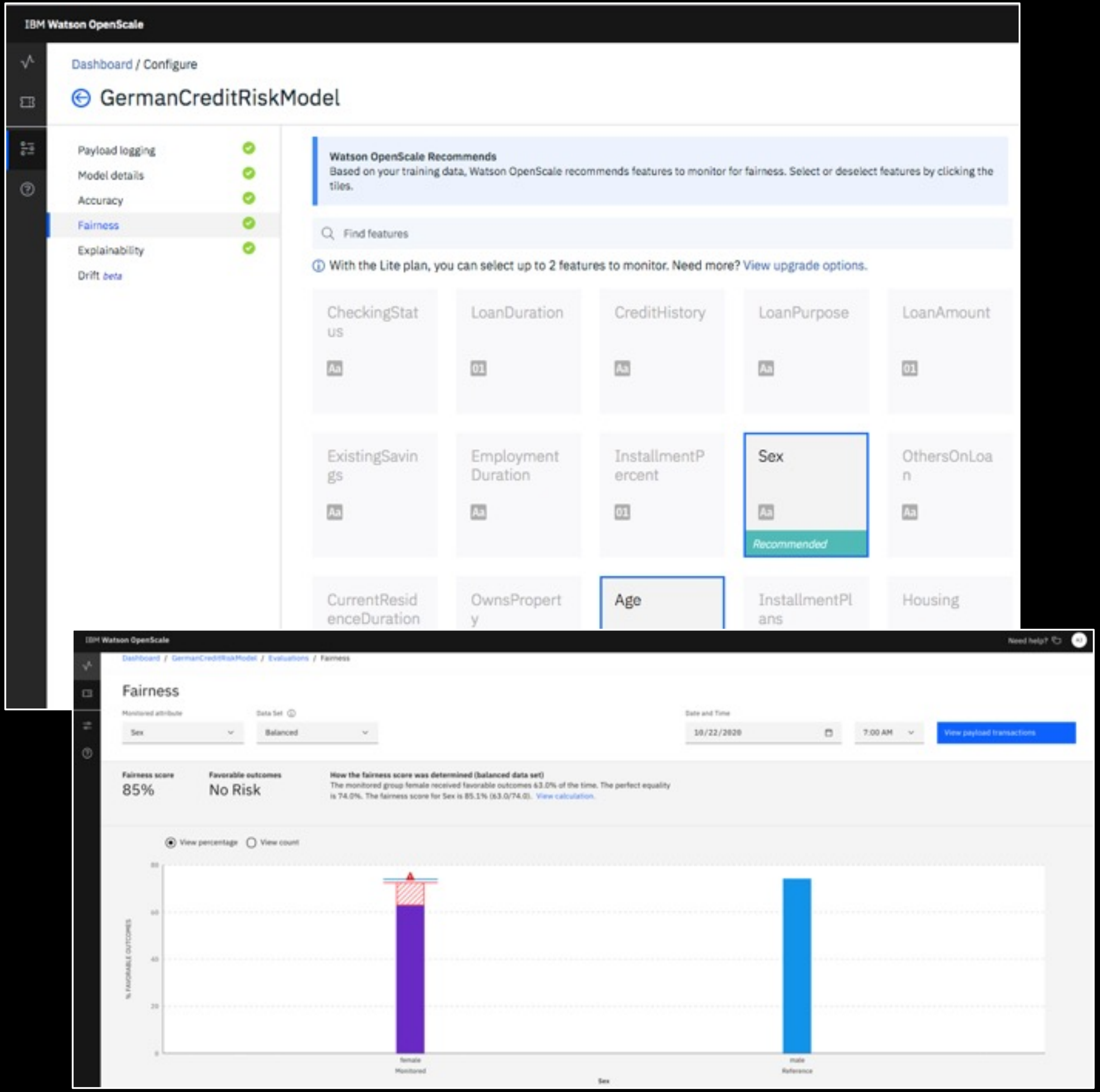
Bias Detection

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



IBM Watson OpenScale

Dashboard / Configure

GermanCreditRiskModel

Watson OpenScale Recommends
Based on your training data, Watson OpenScale recommends features to monitor for fairness. Select or deselect features by clicking the tiles.

With the Lite plan, you can select up to 2 features to monitor. Need more? View upgrade options.

IBM Watson OpenScale

Dashboard / GermanCreditRiskModel / Evaluations / Fairness

Fairness

Monitored attribute: Sex
Data Set: Balanced
Date and Time: 10/22/2020 7:00 AM

Fairness score: 85%
Favorable outcomes: No Risk

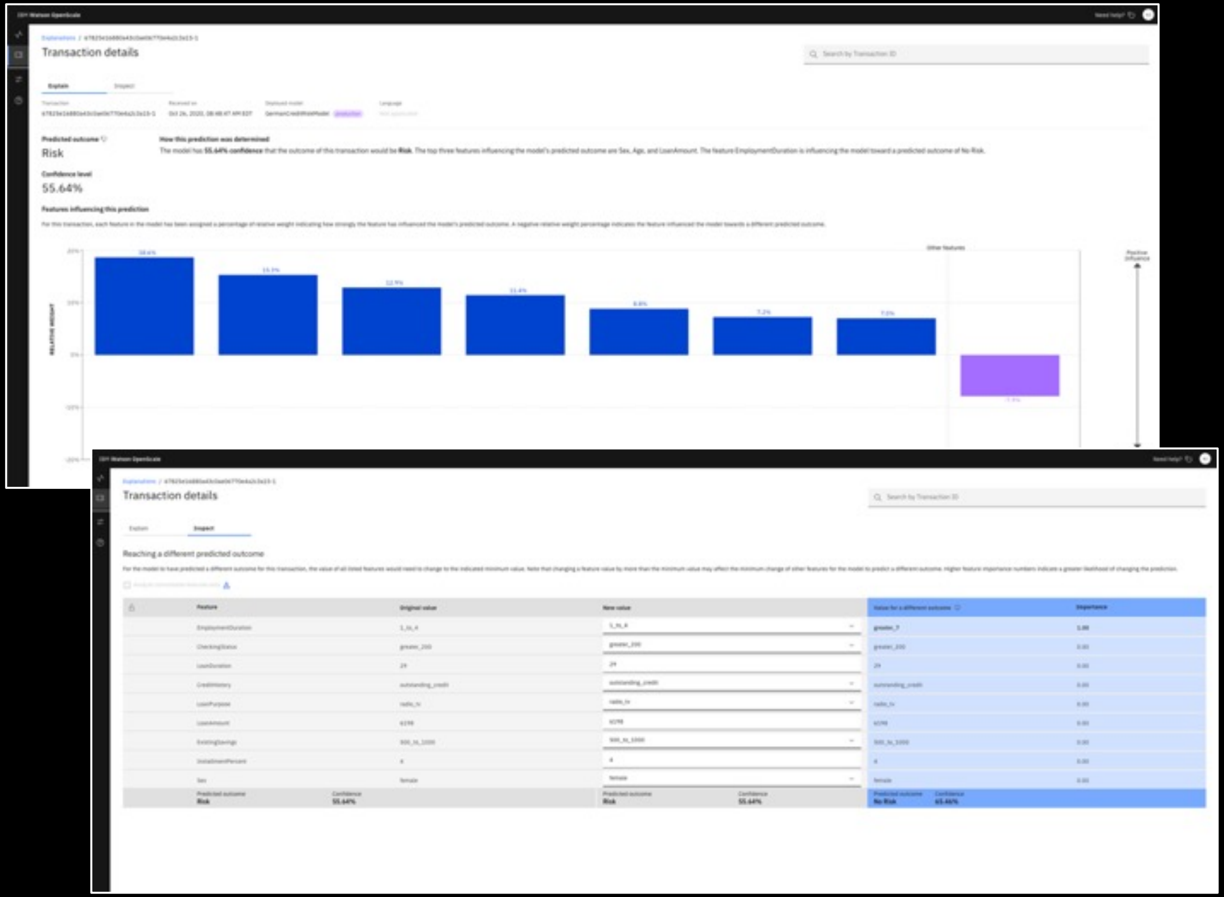
How the fairness score was determined (balanced data set)
The monitored group female received favorable outcomes 63.0% of the time. The perfect equality is 74.0%. The fairness score for Sex is 85.1% (63.0/74.0). View calculation.

● View percentage ○ View count

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



Transaction details

Predicted outcome
Risk

Confidence level
55.64%

How this prediction was determined
The model has **55.64% confidence** that the outcome of this transaction would be **Risk**. The top three features influencing the model's predicted outcome are Sex, Age, and LoanAmount. The feature EmploymentDuration is influencing the model toward a predicted outcome of No Risk.

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.

Transaction details

Reaching a different predicted outcome
For the model to have predicted a different outcome for this transaction, the value of all listed features would need to change to the indicated minimum value. Note that changing a feature value by more than the minimum value may affect the minimum change of other features for the model to predict a different outcome. Higher feature importance numbers indicate a greater likelihood of changing the prediction.

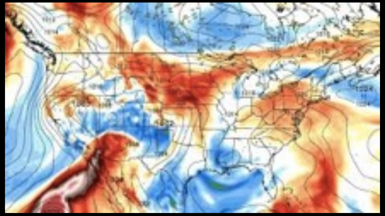
Feature	Original value	New value	Value for a different outcome	Importance
EmploymentDuration	1_to_4	1_to_4	greater_7	1.00
CheckingStatus	greater_200	greater_200	greater_200	0.00
LoanDuration	29	29	29	0.00
CreditHistory	outstanding_credit	outstanding_credit	outstanding_credit	0.00
LoanPurpose	radio_tv	radio_tv	radio_tv	0.00
LoanAmount	6198	6198	6198	0.00
ExistingSavings	500_to_1000	500_to_1000	500_to_1000	0.00
InstallmentPercent	4	4	4	0.00
Sex	female	female	female	0.00
Predicted outcome Risk	Confidence 55.64%	Predicted outcome Risk Confidence 55.64%	Predicted outcome No Risk Confidence 65.46%	

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

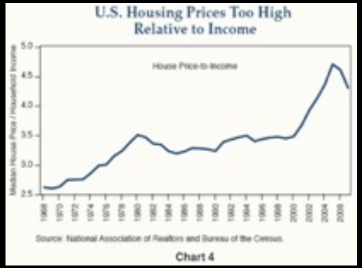
Business environments are dynamic leading to “drift” in data and cause inaccuracies in model prediction

Weather data changes in short term can affect long term climate models

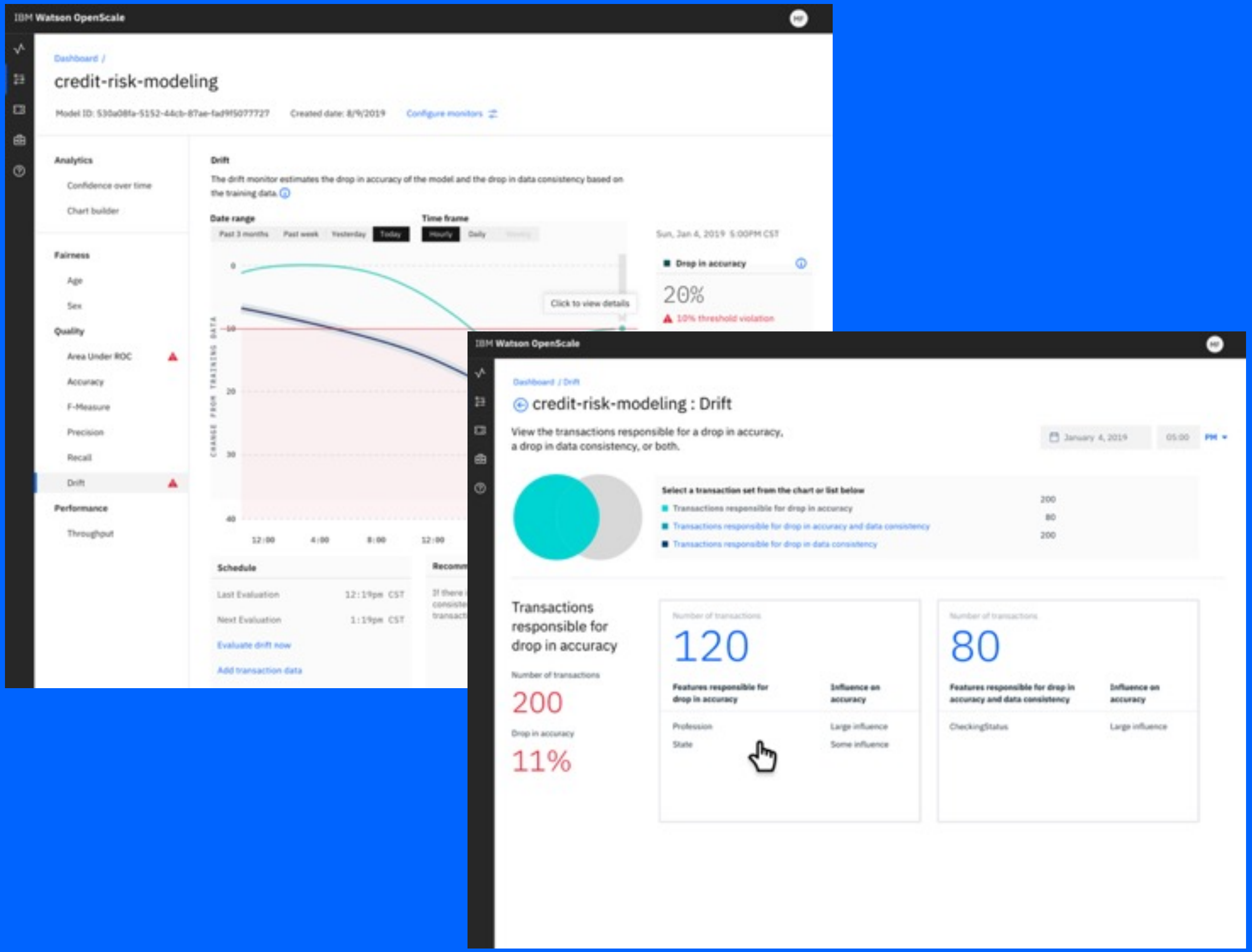


Online shopping behavior

Rise in income levels in specific geos can throw off global models



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



credit-risk-modeling

Model ID: 530a08fa-5152-44cb-87ae-fad9f5077727 Created date: 8/9/2019

Drift

The drift monitor estimates the drop in accuracy of the model and the drop in data consistency based on the training data.

Sun, Jan 4, 2019 5:00PM CST

Drop in accuracy

20%

10% threshold violation

Schedule

Last Evaluation	12:19pm CST
Next Evaluation	1:19pm CST

credit-risk-modeling : Drift

View the transactions responsible for a drop in accuracy, a drop in data consistency, or both.

January 4, 2019 05:00

Select a transaction set from the chart or list below

Transactions responsible for drop in accuracy	200
Transactions responsible for drop in accuracy and data consistency	80
Transactions responsible for drop in data consistency	200

Transactions responsible for drop in accuracy

Number of transactions

200

Drop in accuracy

11%

Number of transactions

120

Features responsible for drop in accuracy	Influence on accuracy
Profession	Large influence
State	Some influence

Number of transactions

80

Features responsible for drop in accuracy and data consistency	Influence on accuracy
CheckingStatus	Large influence

A Customs Department in the Middle East & Africa

Customs Department reduces manual effort spent in import risk processes

Industry: Government
Geography: Middle East & Africa

Business Challenge

Client’s current rule-based import risk identification system is not accurate enough to identify scenarios where importing an item is a real risk. Most times, the system identifies cases as risk, which later on are found to be not a real risk situation. Custom Officers spend their time more in non-real risk cases than the real ones

Solution

The solution is to use AI models to predict the import risk. In addition to that, the Customs Officers are supported with the ability to know the reason (explainability) for the model predicting risk vs. no risk.
Officers can make quick decisions on appropriate mitigation steps.

Outcome

- Reduction of the number of cases with legal imports identified as risk can help save the Custom Officers time and effort.
- Officers can take quick and informed decisions on risks and corresponding mitigation steps using explainability. It saves manual inspection effort down the line.
- The AI-based risk import model can be productized and used by the Customs department of other countries for more significant benefits.

A Major North American Retailer

Helping a major client to proactively mitigate bias in their hiring process

Business Challenge

Difficult to determine if ML model is biased against groups when the model is a black box to the user. It is hard to explain decisions/results from an ML model in human understandable form

Use Case

Leverage Watson OpenScale to accelerate identification of any bias in hiring and “explain” decisions made by AI models

Expected Outcome

- Applying AI to quickly and consistently detect and counteract bias in the candidate assessment activities and support hiring decisions, by using AI automation. IBM’s Watson OpenScale on Cloud Pak for Data monitors the health and actively fairness in our client’s hiring systems
- This enables transparency in the decision making, both for uniformity of decisions and to “explain" any specific decision. The monitoring and explanations allows our client to inspect, understand, and correct decisions, maintaining its commitments as a fair employer, and fairness for each of its candidates and new employees

Industry: Retail and Consumer Products
Geography: North America

KPMG: Stewarding responsible AI with Watson OpenScale

KPMG identified four trust imperatives hindering AI adoption:

- **Integrity:** How do we ensure data quality throughout its lifecycle?
- **Fairness:** How do we reject prejudice or bias towards groups, sets of individuals, or data attributes?
- **Explainability:** How can we in business terms explain the decisioning of how AI came to its conclusion?
- **Resiliency:** How can we shield AI and AI infused services against cyber threats or adversarial attacks?

KPMG uses Watson OpenScale to help clients:

- Better understand unconscious bias
- Be more proactive to regulatory change
- Provide the trust they need to drive AI adoption

> "Watson OpenScale is one of the only technologies in the marketplace that gives transparency in business terms to our clients by articulating how and why an AI model came to its determination as well as what data attributes were used."
> — Kelly Combs, KPMG

AI picks the most exciting moments at the US Open without bias

Business challenge:

- The US Open consists of 254 tennis matches with tens of thousands of points, many occurring in parallel
- It is impossible for editors – and for fans – to see most of the best points in the tournament
- Manually editing highlights is time-consuming and resource-intensive

Solution:

IBM built an AI system that automatically clips and creates highlight videos and assigns a fair excitement score to each, using Watson OpenScale to debias the scores based on factors such as court, player rank, player age, crowd size, and so on

Benefits:

- All candidate highlight videos are available and scored for fairness within two minutes of the end of each match
- Debiasing for crowd size – making large and small crowds more comparable – made the model 52 percent more fair
- Controlling for age bias increase the overall fairness of highlight scores from 42 to 91 percent

AI picks the most exciting moments at the US Open without bias

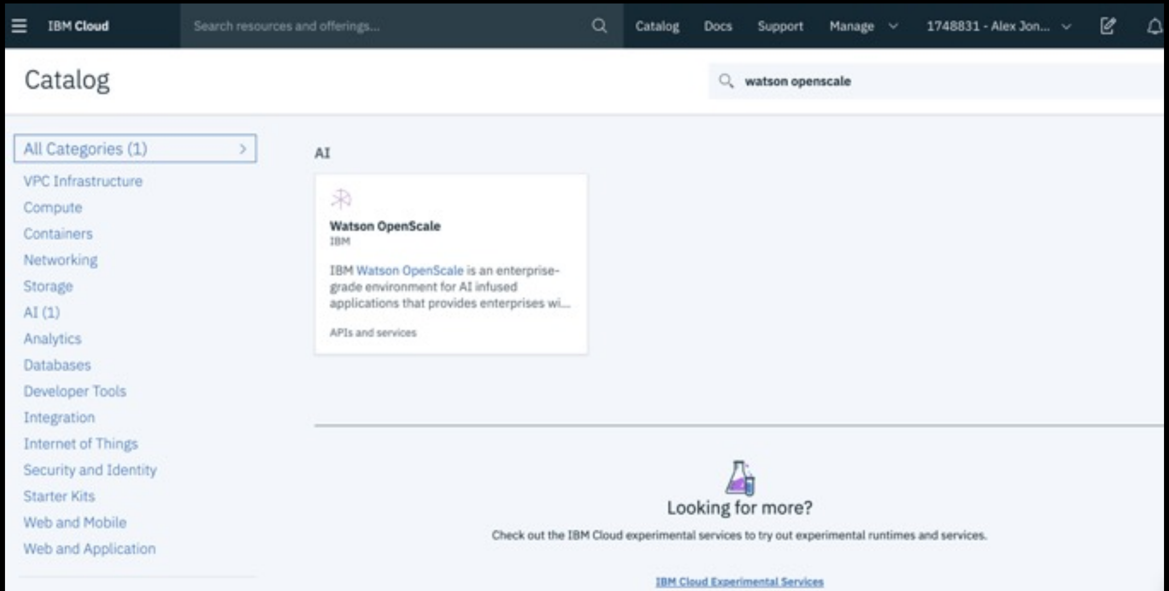
https://www.youtube.com/watch?v=Bw9eIfike_s

https://www.ibm.com/blogs/journey-to-ai/2021/03/turning-the-future-into-a-sure-win/

How does AI identify match highlights?

Try it out for free

- Free, lite version of Watson OpenScale is available on IBM cloud
- Full functionality, limited to evaluation scale workloads
- Check out our samples & try with your own models
- https://cloud.ibm.com/catalog/services/watson-openscale
- Fairness and Explainability 360 toolkits
- https://aif360.mybluemix.net
- https://aix360.mybluemix.net



Watson OpenScale
IBM

IBM Watson OpenScale is an enterprise-grade environment for AI infused applications that provides enterprises wi...

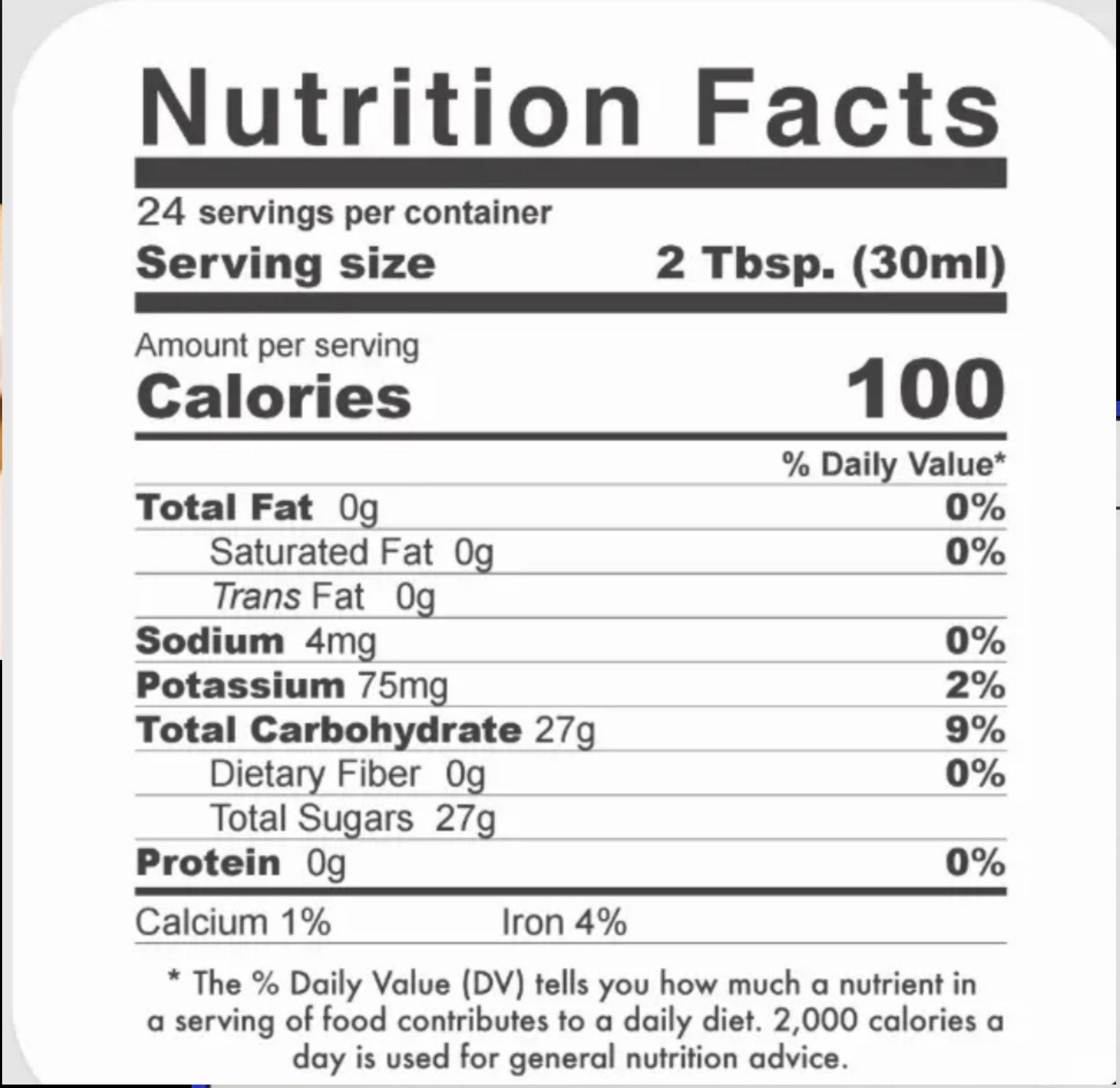
APIs and services

Looking for more?

Check out the IBM Cloud experimental services to try out experimental runtimes and services.

IBM Cloud Experimental Services

How do I trust this ML Model



Nutrition Facts

24 servings per container

Serving size **2 Tbsp. (30ml)**

Amount per serving

Calories **100**

	% Daily Value*
Total Fat 0g	**0%**
Saturated Fat 0g	**0%**
Trans Fat 0g	
Sodium 4mg	**0%**
Potassium 75mg	**2%**
Total Carbohydrate 27g	**9%**
Dietary Fiber 0g	**0%**
Total Sugars 27g	
Protein 0g	**0%**

Calcium 1% Iron 4%

* The % Daily Value (DV) tells you how much a nutrient in a serving of food contributes to a daily diet. 2,000 calories a day is used for general nutrition advice.

Lab - Factsheets

https://www.youtube.com/watch?v=-4d3kEVsu-s

https://aifs360.mybluemix.net/examples/max_text_sentiment_classifier

https://developer.ibm.com/exchanges/models/all/max-text-sentiment-classifier

Ivan Portilla

ivanp@us.ibm.com

@iportilla

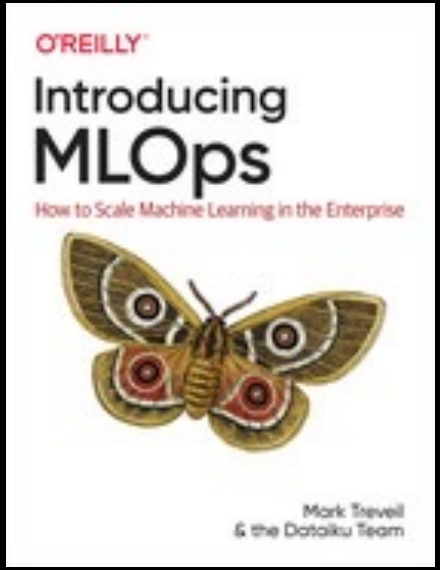
Design Thinking in Software and AI Projects

Proving Ideas Through Rapid Prototyping

—

Robert Stackowiak
Tracey Kelly

Apress®



O'REILLY®

Introducing MLOps

How to Scale Machine Learning in the Enterprise

Mark Treveil
& the Dataiku Team

https://medium.com/inside-machine-learning/ai-ops-managing-the-end-to-end-lifecycle-of-ai-3606a59591b0