


Monitoring AI models in IBM z/OS

Galina Ford
Machine Learning for IBM z/OS Technical Lead
and Architect
gorelik@us.ibm.com



AI on IBM Z:
make more
valuable
outcomes
possible
for every
industry

Financial Services



FRAUD
DETECTION



ANTI-MONEY
LAUNDERING



RISK SCORING




CREDIT
DECISIONING

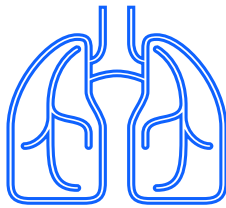
Insurance



CLAIMS FRAUD
DETECTION



INTELLIGENT
UNDERWRITING

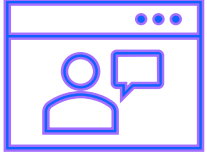


TEXT/IMAGE
PROCESSING




PRODUCT
RECOMMENDATIONS

Government




GEOSPATIAL IMAGE
ANALYSIS



FRAUD & SENTIMENT
ANALYSIS



AUDIT &
COMPLIANCE



CHAT SERVICE

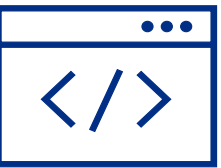
Others



RETAIL
INVENTORY/DEMAND
FORECASTING



SYSTEM ADMIN
ASSISTANT

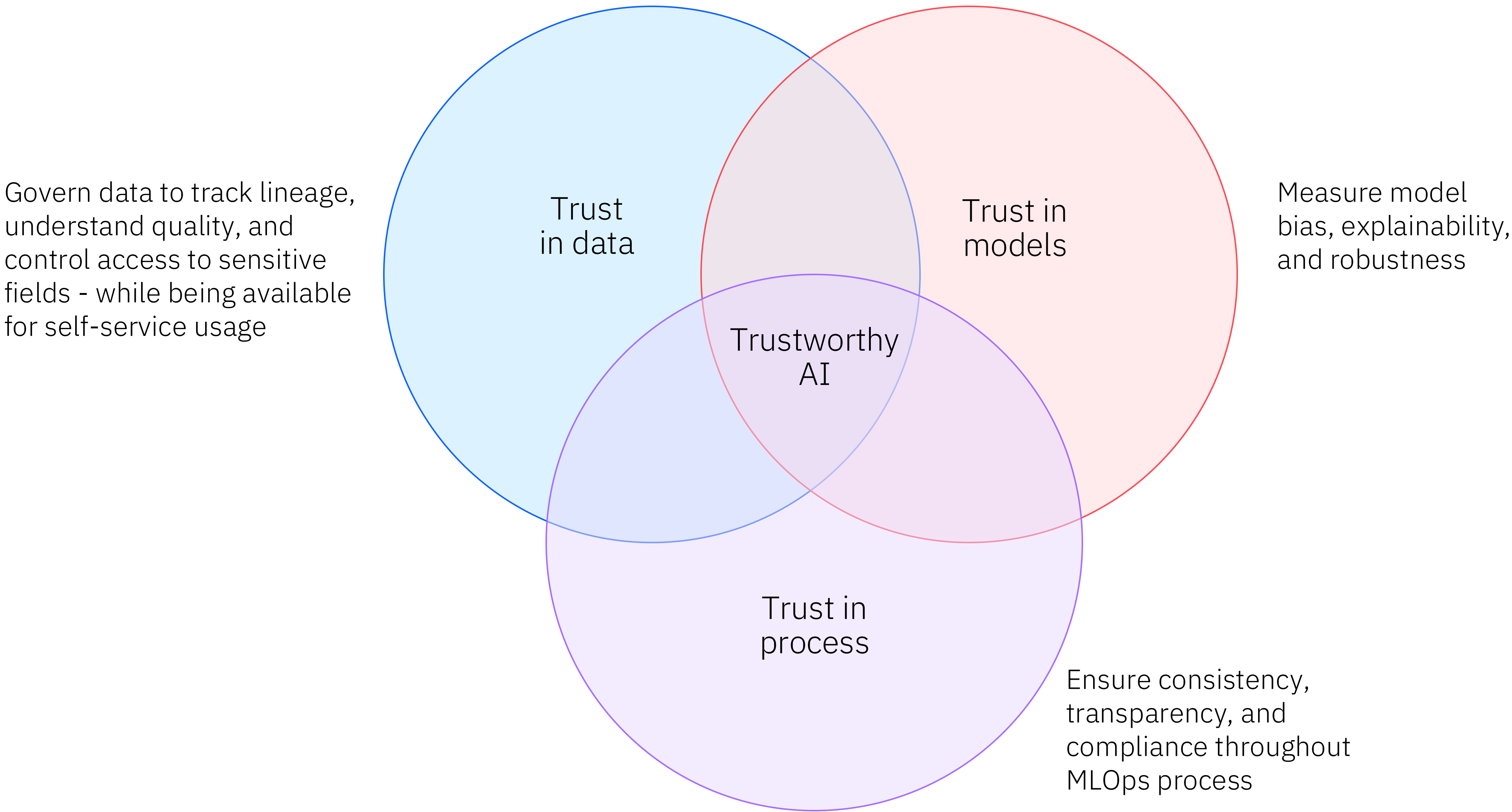


CODE
ASSISTANT



TRANSPORTATIO
N LOGISTICS

Trustworthy AI

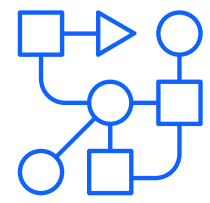


Machine Learning for IBM z/OS

Accelerated inferencing at scale for Transactional AI workloads on z/OS

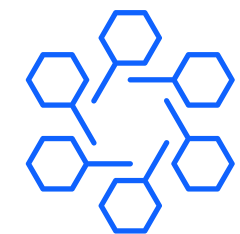
The full-featured machine learning platform on z/OS designed to infuse AI into mission critical IBM Z enterprise applications by leveraging AI models trained anywhere or on IBM Z and deploying them on z/OS including a multi-model architecture, co-located with enterprise applications, transaction data, and business logic for high throughput and extremely low latency to modernize applications and drive business insights at scale

Benefits



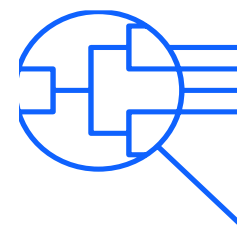
End-end platform

GUI and script driven features for end-end AI model lifecycle management



Scalable AI

Leverage the latest accelerator to score transactions natively in CICS, IMS and batch applications at scale



Easy import and deploy

Leverage predictive models and Encoder LLMs trained anywhere or with watsonx.ai and deploy with MLz

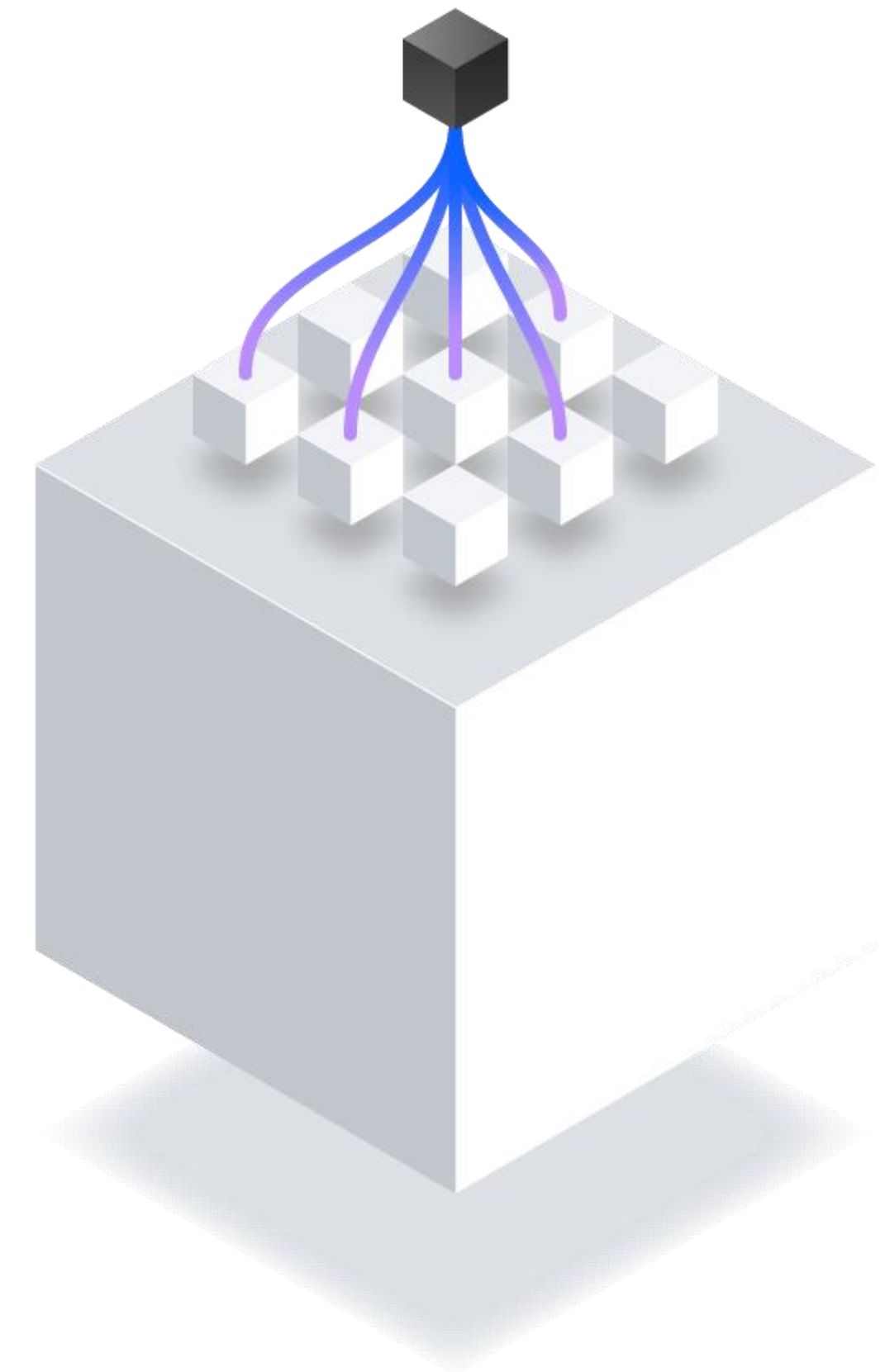


Trustworthy features

Leverage built in explainability and drift detection features or enterprise AI governance with watsonx.governance

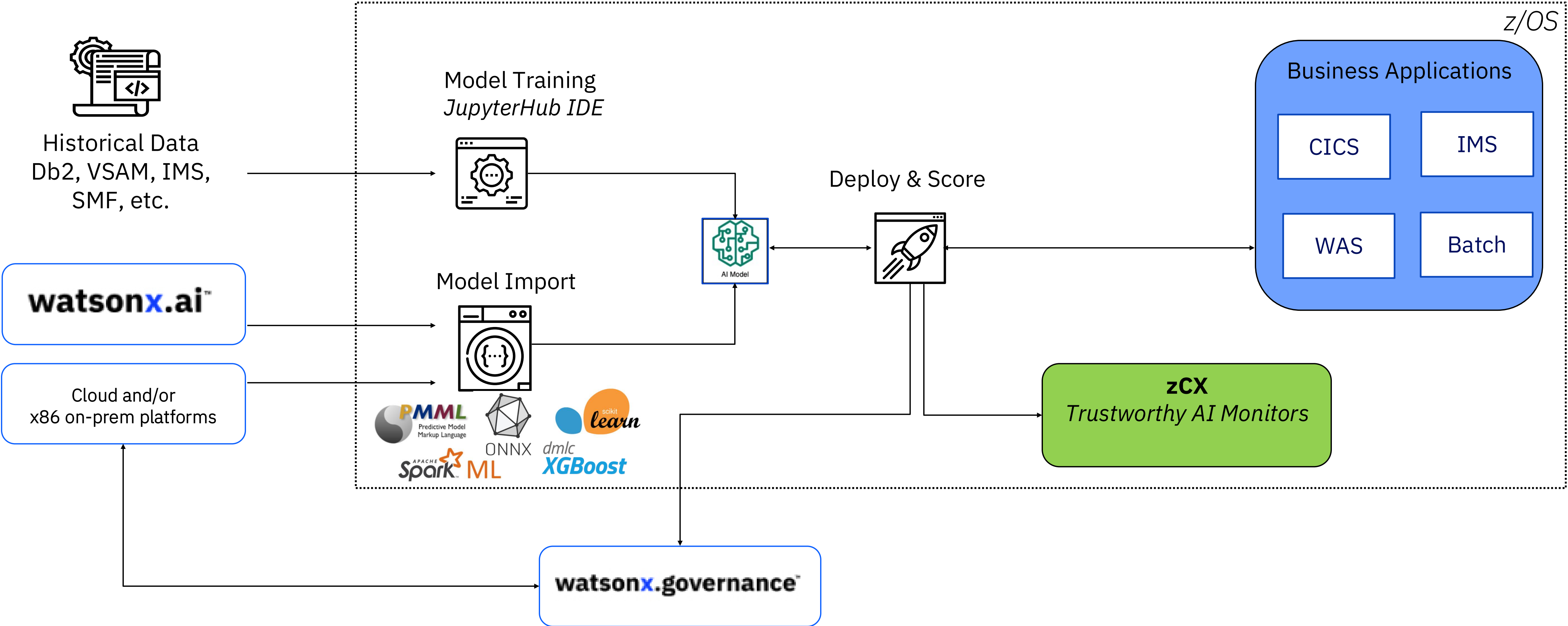
Transactional AI

Infuse AI into every transaction



Train anywhere – Deploy on IBM Z

Machine Learning for IBM z/OS



Key Features of Machine Learning for IBM z/OS v3.2 – Enterprise Edition

1

GUI Configuration

Web-based Configuration Tool for single instance and HA configuration

2

Model training tool

Integrated Jupyter server for model training on Z
Leverage IBM Z Spark 3.5 and Python AI Toolkit for training and scoring

3

Scoring Engines

Online scoring for various types of machine learning models and deep learning and encoder LLMs
Leverage on-chip AI accelerator for scoring acceleration

4

Integrated Scoring

In-transaction scoring through native CICS and WOLA interface for CICS, IMS and BATCH COBOL applications

5

Model management

Web-based UI and REST APIs for end-to-end model lifecycle management

6

Trustworthy AI

Deploy your models with the utmost trust leveraging the integrated trustworthy AI capabilities

Trustworthy AI Capabilities

*Supports multiple scoring interfaces:
REST, CICS, and WOLA (e.g., from IMS
transactions and batch jobs)*

*Supports multiple model types:
SparkML, Python based scikit-learn
and XGBoost, PMML, ONNX models*

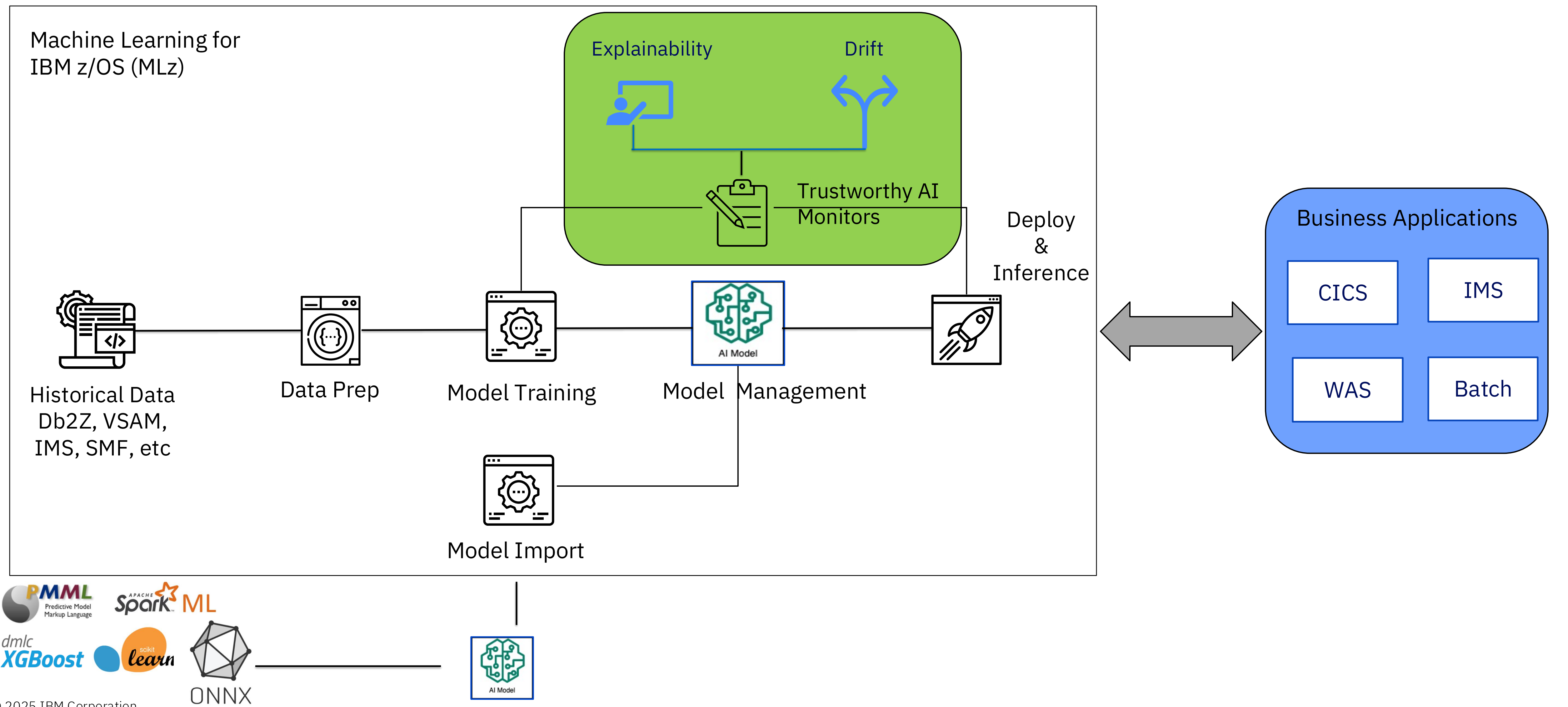
Explainability

*Enables continuous monitoring
of z/OS transactions that are
scored by AI models and
provides explanations on the
model's output at the individual
transaction level*

Drift Detection

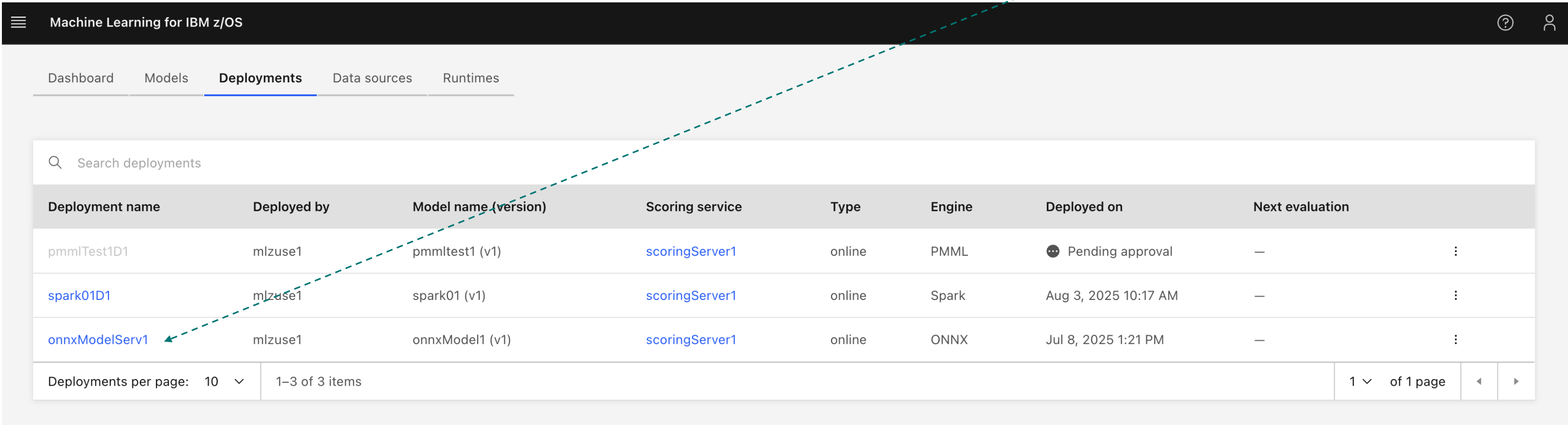
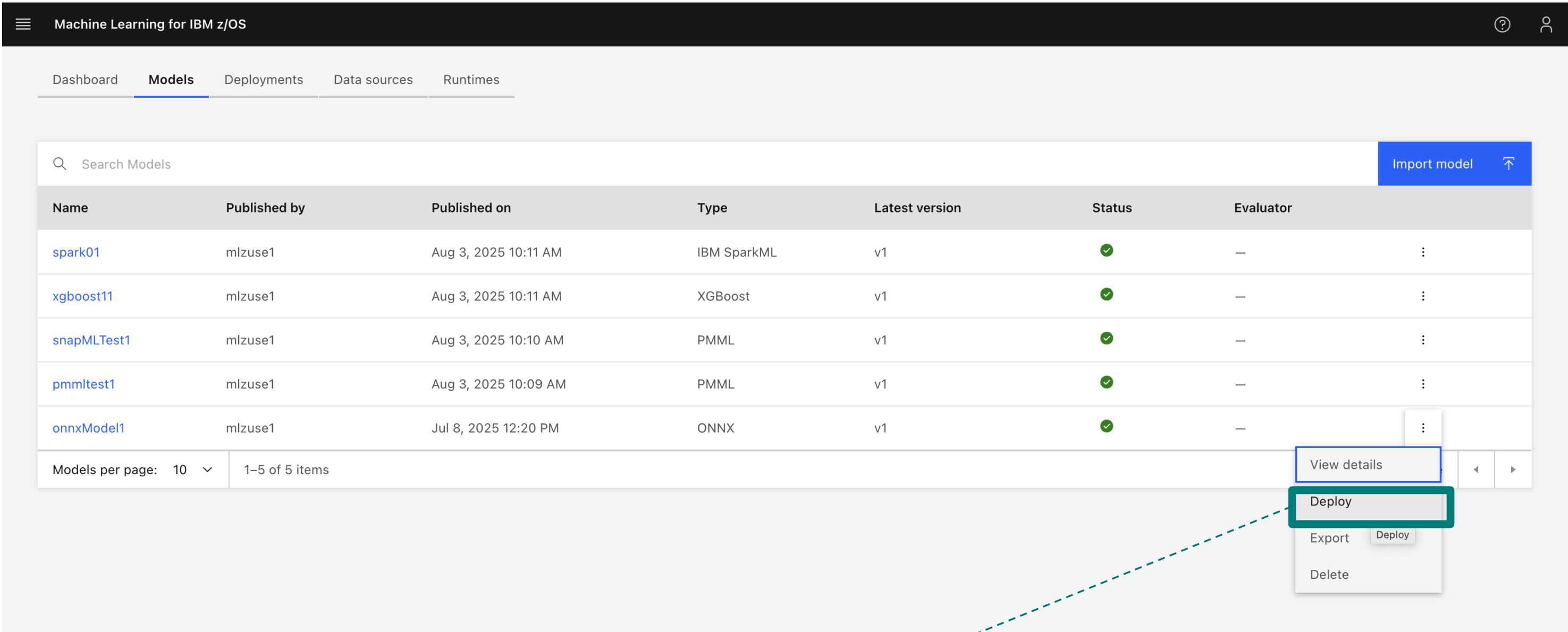
*Enables monitoring and
evaluation of both output drift
and feature drift, providing
data scientists with insights
from different perspectives*

Trustworthy AI – in Machine Learning for IBM z/OS



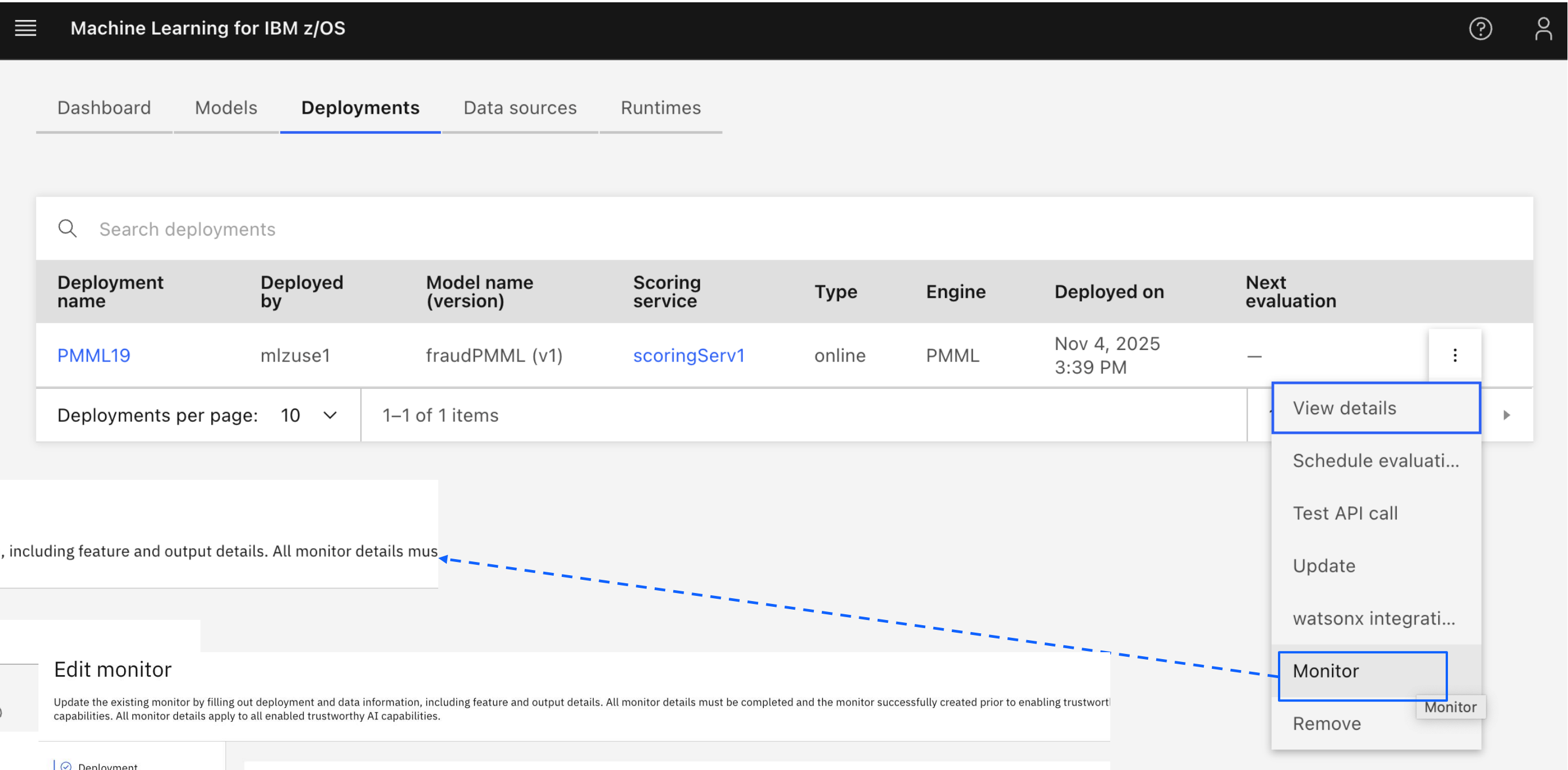
Model deployment and management

- Intuitive web user interface for AI model life cycle management
- REST APIs available for automation
- Dual control available for deployment related activities



Model Monitoring

- Requires a [z/OS Container Extensions instance](#)
- [Intuitive web user interface](#) for AI model life cycle management
- [REST APIs](#) available for automation



Create monitor

Create a new monitor by filling out deployment and data information, including feature and output details. All monitor details must be completed and the monitor successfully created prior to enabling trustworthy AI capabilities. All monitor details apply to all enabled trustworthy AI capabilities.

Deployment

Features

Monitor name ⓘ

Enter monitor name

Deployment scoring endpoint ⓘ

https://127.0.0.1:12345

Algorithm type

Binary classification

Data collection ⓘ

☐ Store samples

Upload training data

Trustworthy AI capabilities require access to training data. Max file size is 100 MB. Supported file types are .csv and .json.

Drag and drop files here or click to select files

training.csv ×

Edit monitor

Update the existing monitor by filling out deployment and data information, including feature and output details. All monitor details must be completed and the monitor successfully created prior to enabling trustworthy AI capabilities. All monitor details apply to all enabled trustworthy AI capabilities.

Deployment

Features

Feature columns

Shown are the features and label derived from the deployed model. Select the Categorical checkbox for any features that should be identified as categorical data. The Type column is automatically set based on your model schema, as well as the Label column, which represents the ground truth data your model is trying to predict. You cannot change the input for Type and Label from this table.

Feature	Type	Categorical	Label
fraud	integer	<input type="checkbox"/>	<input checked="" type="checkbox"/>
transaction_amount	real	<input type="checkbox"/>	<input type="checkbox"/>
merchant_category	real	<input type="checkbox"/>	<input type="checkbox"/>
customer_age	real	<input type="checkbox"/>	<input type="checkbox"/>
transaction_location	real	<input type="checkbox"/>	<input type="checkbox"/>
customer_type	real	<input type="checkbox"/>	<input type="checkbox"/>
transaction_day_of_year	real	<input type="checkbox"/>	<input type="checkbox"/>
transaction_day_of_month	real	<input type="checkbox"/>	<input type="checkbox"/>
transaction_day_of_week	real	<input type="checkbox"/>	<input type="checkbox"/>
transaction_hour_of_day	real	<input type="checkbox"/>	<input type="checkbox"/>

Monitor settings

monitorFraud1

Monitor details

✓

Evaluations ⓘ

Explainability

✓

Drift

✓

Model output

Select the prediction column (the column containing the predicted value) from the training data.

Feature	Type
probability(0)	number
probability(1)	number

Explainability allows you to explain specific transactions and understand what influenced the model to come to a decision for that transaction. Enabling explainability for this monitor requires selecting explainer type(s).

☒ Enable explainability for this monitor

Explainer type ⓘ

2 × LIME,SHAP

☒ LIME

☒ SHAP

Monitoring: real time or add-hoc?

- Collected transactions
Drift evaluation will run at configured time intervals if the minimum number of transactions have occurred
- Explainability will support selection from stored transactions

- Add-hoc input
Drift evaluation and Explainability require data to be imported

Monitor settings
monitorFraud1

Monitor details

Evaluations

Explainability

Drift

Drift allows you to monitor for degradation in model performance or accuracy. Enabling drift evaluations for this monitor requires setting the schedule for evaluations, choosing the important features, and defining thresholds for alerts.

☒ Enable drift evaluations for this monitor

Evaluation criteria
Define the criteria for an evaluation to run. If samples are being stored, this includes the time interval for how often to check if the minimum sample size has been met.

Time interval 1 hours Minimum sample size 10

Important features
Select the important features to score for drift.

Feature	Type
<input checked="" type="checkbox"/> customer_age	number
<input checked="" type="checkbox"/> customer_type	number

Items per page: 10 1-10 of 13 items 1 of 2 pages

Thresholds
Define alert thresholds for each type of drift. If the drift score goes past the threshold, you will be alerted. Thresholds must be between 0-1; a higher drift score indicates more drift.

Output drift threshold 0.5 Feature drift threshold 0.5

Edit monitor

Update the existing monitor by filling out deployment and data information, including feature and output details. All monitor details must be completed and the monitor successfully created prior to enabling trustworthy AI capabilities. All monitor details apply to all enabled trustworthy AI capabilities.

Deployment

Features

Monitor name monitorFraud1

Deployment name PMML19

Algorithm type Binary classification

Data collection ☒ Store samples

Upload training data
Trustworthy AI capabilities require access to the data used to train the deployed model. Max file size is 300MB. Supported file types are .csv and .json.

Drag and drop files here or click to upload

training.csv

Storing samples (inferences) allows the system to present a database of transactions to explain and continually monitor for drift. If unchecked, you will not be able to schedule automatic evaluations and will need to import the necessary data to explain or monitor for drift at the time of evaluation.

Method ☒ Select from database ☐ Input via record

Search and filter for transactions
Search and filter for transactions available to this monitor. Narrow your search as much as possible to reduce load time and number of transactions returned.

Select search parameters

All Search transactions

Date Range Start date 02/23/2025 End date 02/23/2025

Dynamic Filter Feature x1 Operator Choose an operator Base value

Add dynamic filter

Reset search & filters Apply search & filters

Sample transactions
To find a specific transaction beyond this list, use the provided search and filters. You may select up to five transactions to explain.

Transaction ID	Date & time
<input type="checkbox"/> HZZU00IN_0	2025-02-22 10:51:03
<input type="checkbox"/> 9AyfWBX_0	2025-02-22 10:50:32

Items per page: 10 1-2 of 2 items 1 of 1 page

Cancel Create explanations

Explainability

LIME

Model-agnostic
local, interpretable
explanations -
explains single
predictions of a model

SHAP

Model-specific
global explanations -
the effect of a feature
on the target variable

demoMonitor Explained transactions

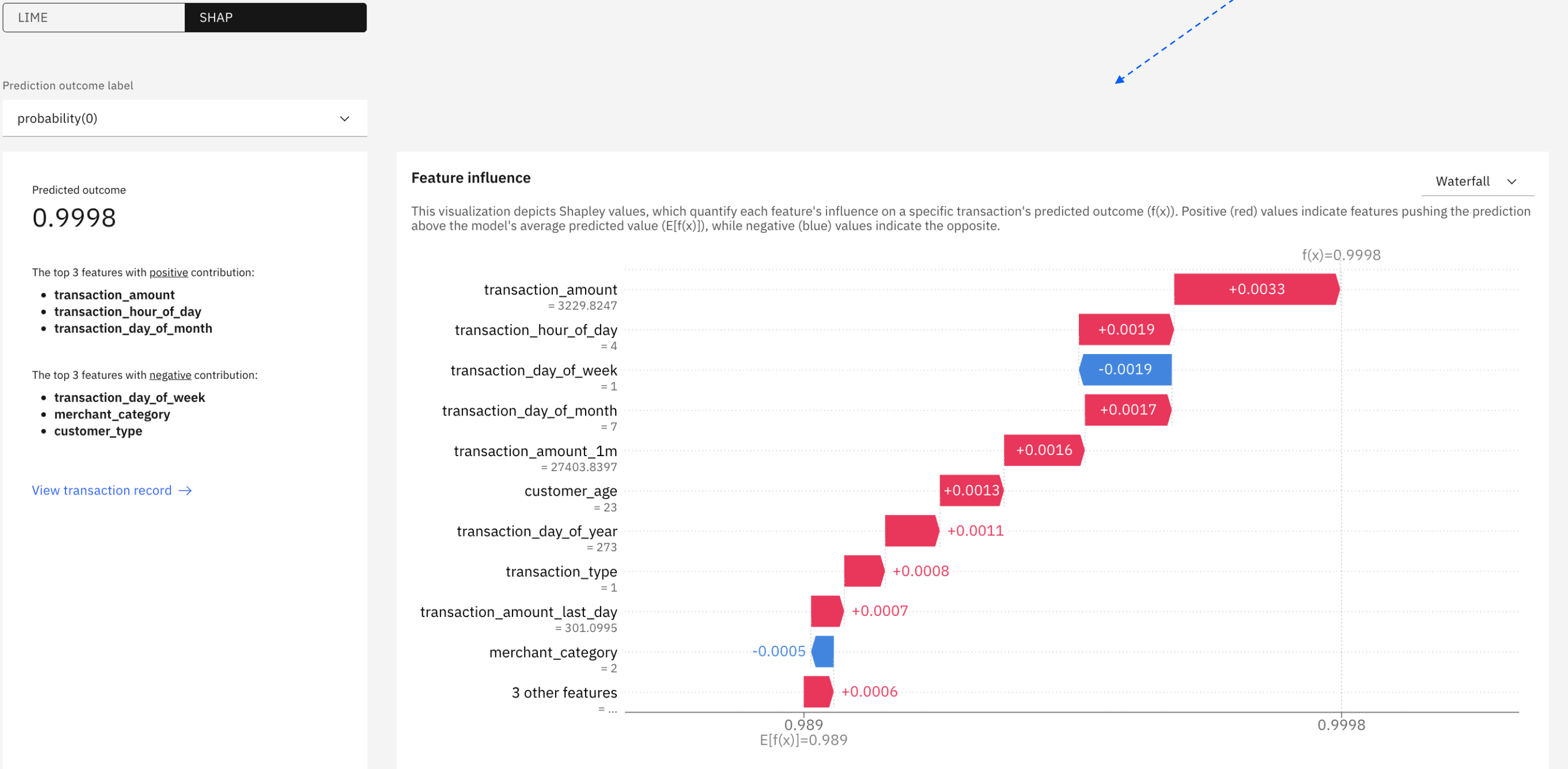
View the status and results for explained transactions, and add new transactions to be explained. You may close the page while explanations are loading and return once they are available for viewing.

Search explained transactions

Add transactions

Transaction ID	Transaction date & time	Explanation date & time	↓	Status		
HZZU00IN_0	2025-02-23 09:00:16	2025-02-23 09:00:16		<div></div>	View results	<div></div>
9AyfWrBX_0	2025-02-23 09:00:16	2025-02-23 09:00:16		<div></div>	View results	<div></div>
9AyfWrBX_0	2025-02-22 10:51:57	2025-02-22 10:51:57		<div></div>	View results	<div></div>
HZZU00IN_0	2025-02-22 10:51:57	2025-02-22 10:51:57		<div></div>	View results	<div></div>
uhwE4pQl	2025-02-22 10:39:52	2025-02-22 10:39:52		<div></div>	View results	<div></div>

1 of 1 page

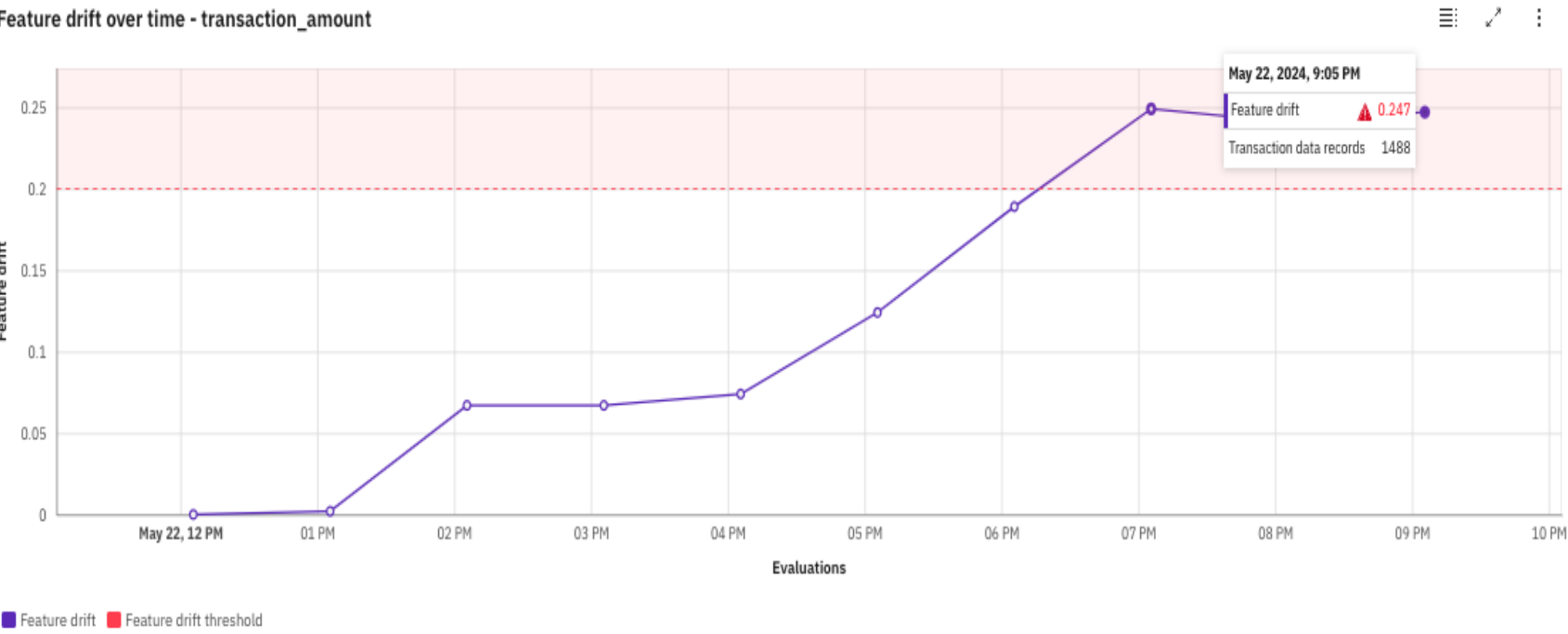
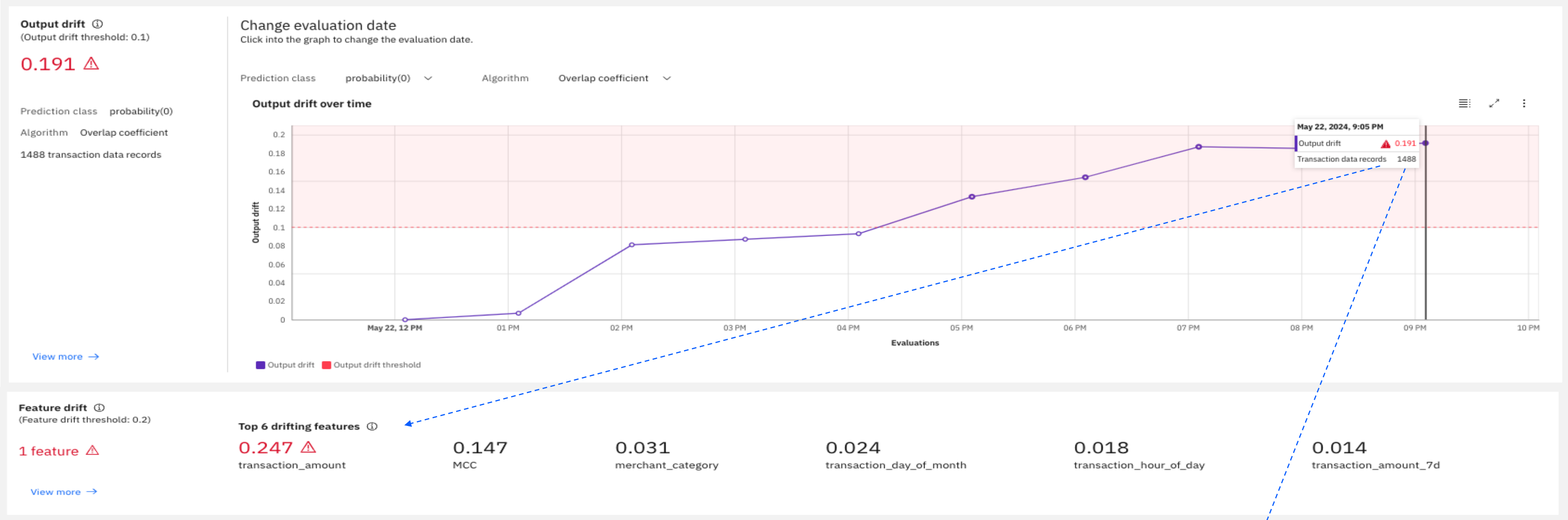


Drift

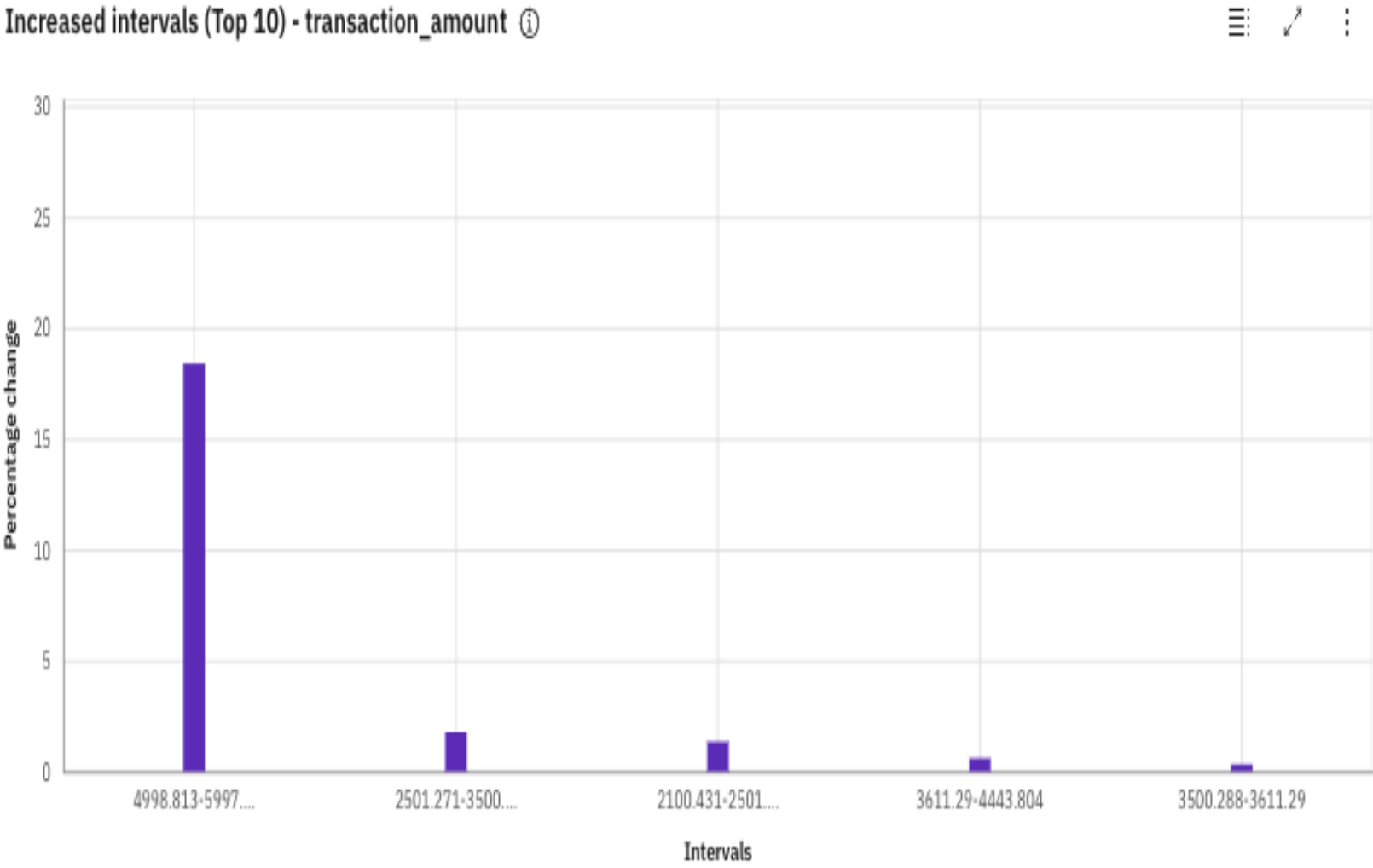
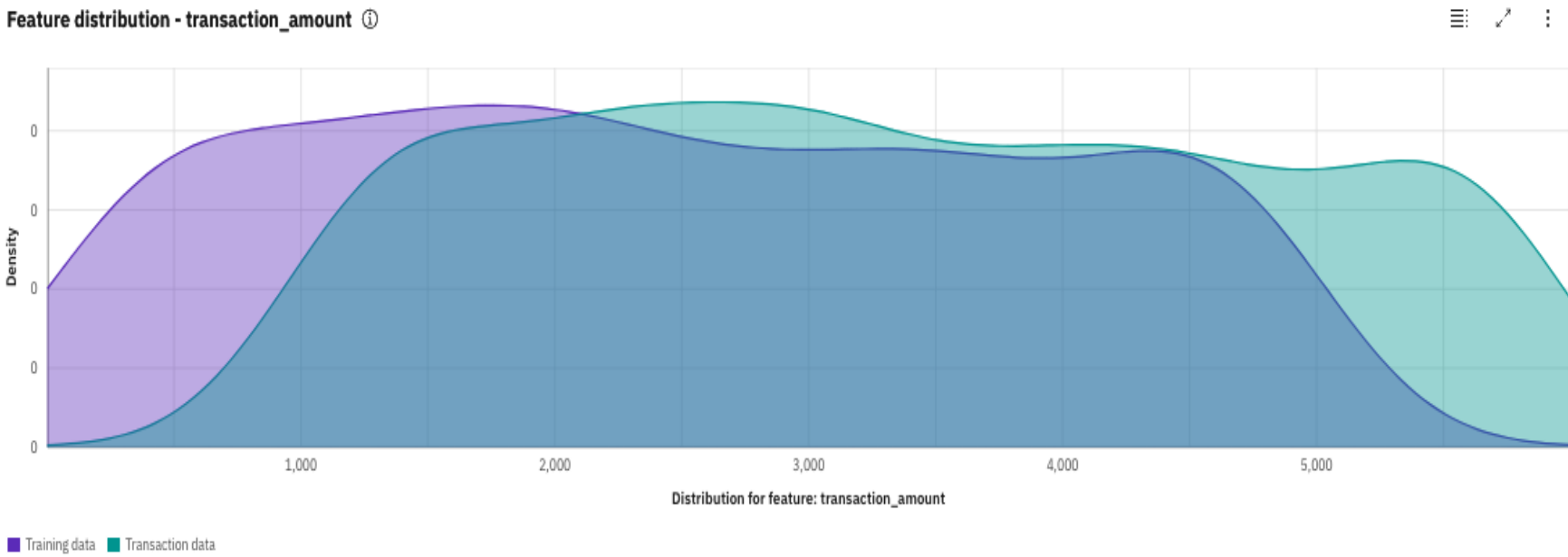
Output Drift
Measured by *Overlap coefficient* and *Total variation distance*

Feature Drift
Categorical features are measured by *Jensen-Shannon* and *Total variation distance*

Non-categorical features are measured by *Overlap coefficient* and *Total variation distance*



View the density distribution plots for the selected feature and the corresponding intervals that changed most significantly compared to the training data.



Thank you

© 2025 International Business Machines Corporation

IBM, the IBM logo, IBM Z, DataStage and Db2 are trademarks of IBM Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on <https://www.ibm.com/legal/copyright-trademark>.

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.

Client examples are presented as illustrations of how those clients have used IBM products and the results they may have achieved. Actual performance, cost, savings or other results in other operating environments may vary.

Not all offerings are available in every country in which IBM operates.

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

All names and references for organizations and other business institutions used in this deliverable’s scenarios are fictional. Any match with real organizations or institutions is coincidental. All names and associated information for people in this deliverable’s scenarios are fictional. Any match with a real person is coincidental. Videos may show information about or belonging to third parties. Such videos do not suggest endorsement of third parties or their products or services.