Monitoring AI models in IBM z/OS

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

IBM

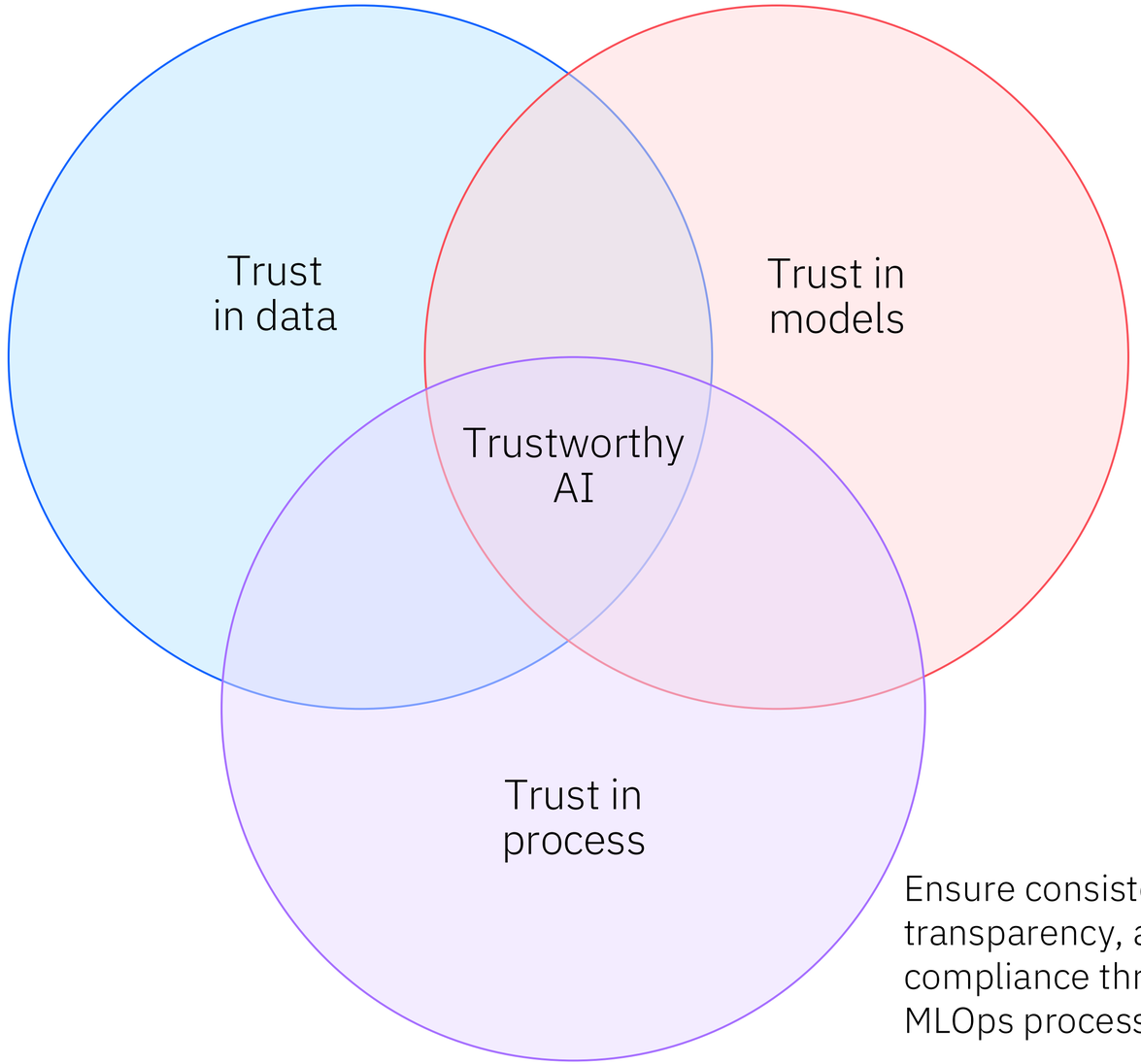
AI on IBM Z: make more valuable outcomes possible for every industry

Financial Services	Insurance	Government	Others
FRAUD DETECTION	CLAIMS FRAUD DETECTION	GEOSPATIAL IMAGE ANALYSIS	RETAIL INVENTORY/DEMAND FORECASTING
ANTI-MONEY LAUNDERING	INTELLIGENT UNDERWRITING	FRAUD & SENTIMENT ANALYSIS	SYSTEM ADMIN ASSISTANT
RISK SCORING	TEXT/IMAGE PROCESSING	AUDIT & COMPLIANCE	CODE ASSISTANT
CREDIT DECISIONING	PRODUCT RECOMMENDATIONS	CHAT SERVICE	TRANSPORTATIO N LOGISTICS

Transactional AI

Trustworthy AI

Govern data to track lineage, understand quality, and control access to sensitive fields - while being available for self-service usage



Measure model bias, explainability, and robustness

Ensure consistency, transparency, and compliance throughout MLOps process

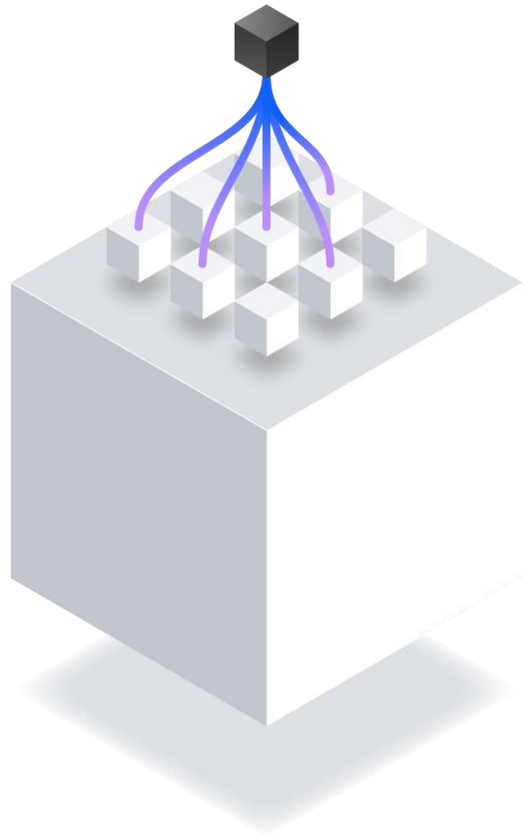
Machine Learning for IBM z/OS

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

Transactional AI

Infuse AI into every transaction



Benefits

End-end platform

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

Easy import and deploy

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

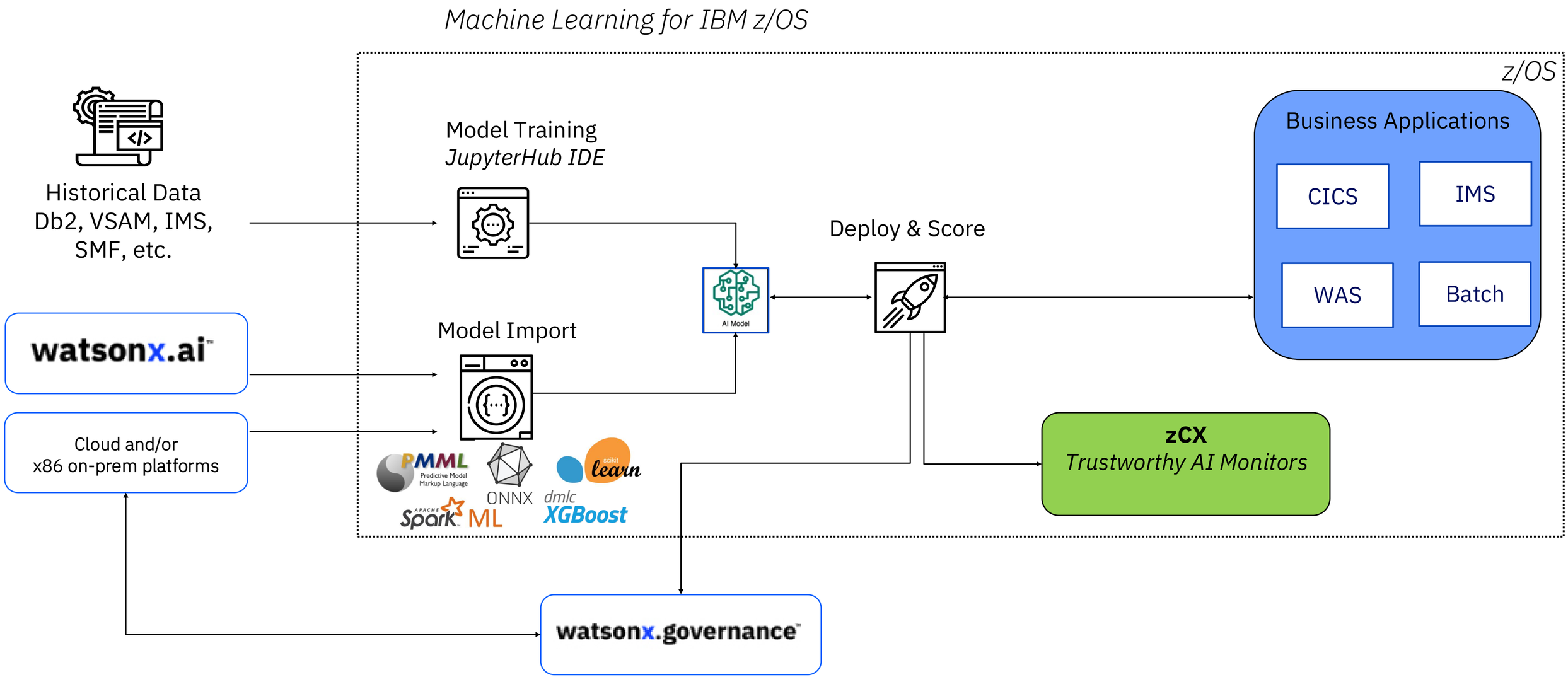
Scalable AI

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

Trustworthy features

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

Train anywhere – Deploy on IBM Z



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

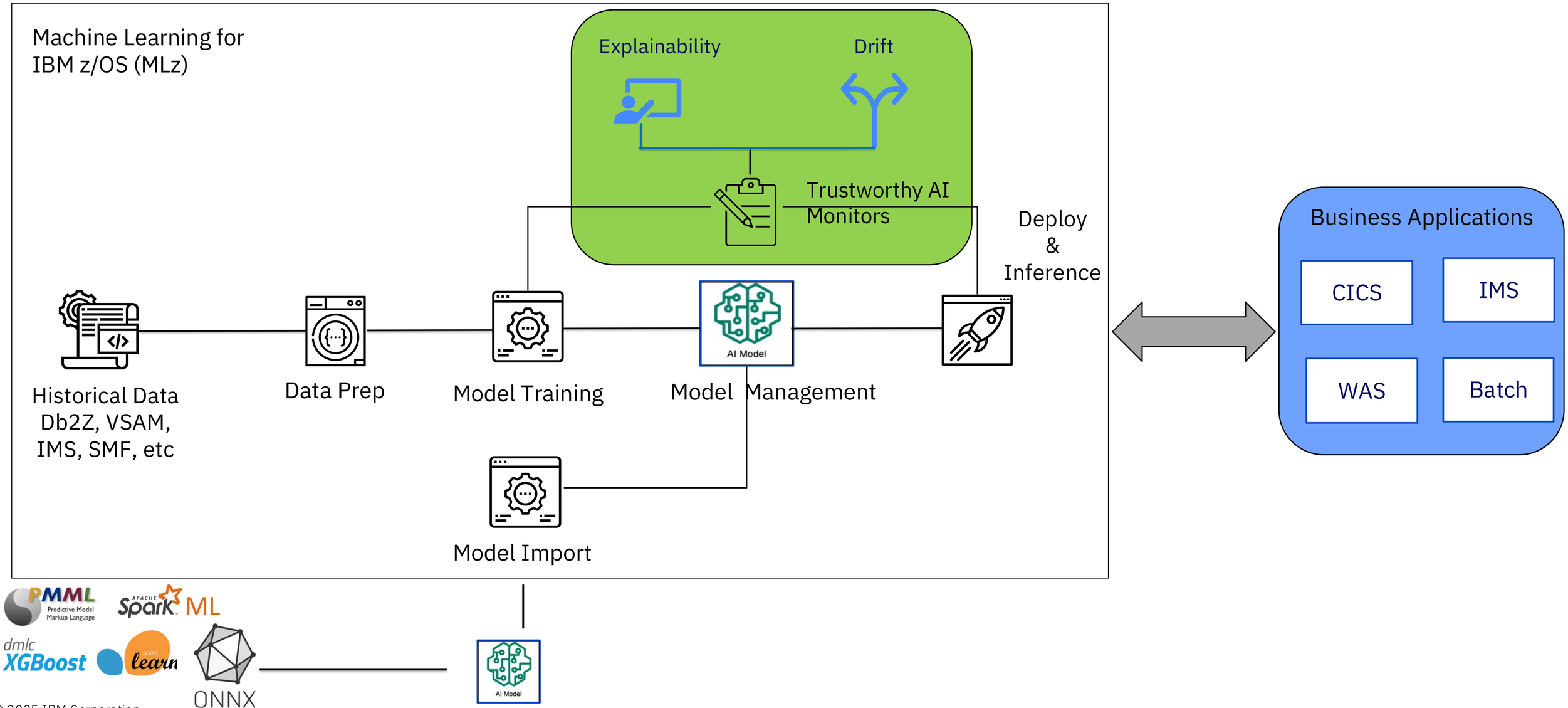
Explainability

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

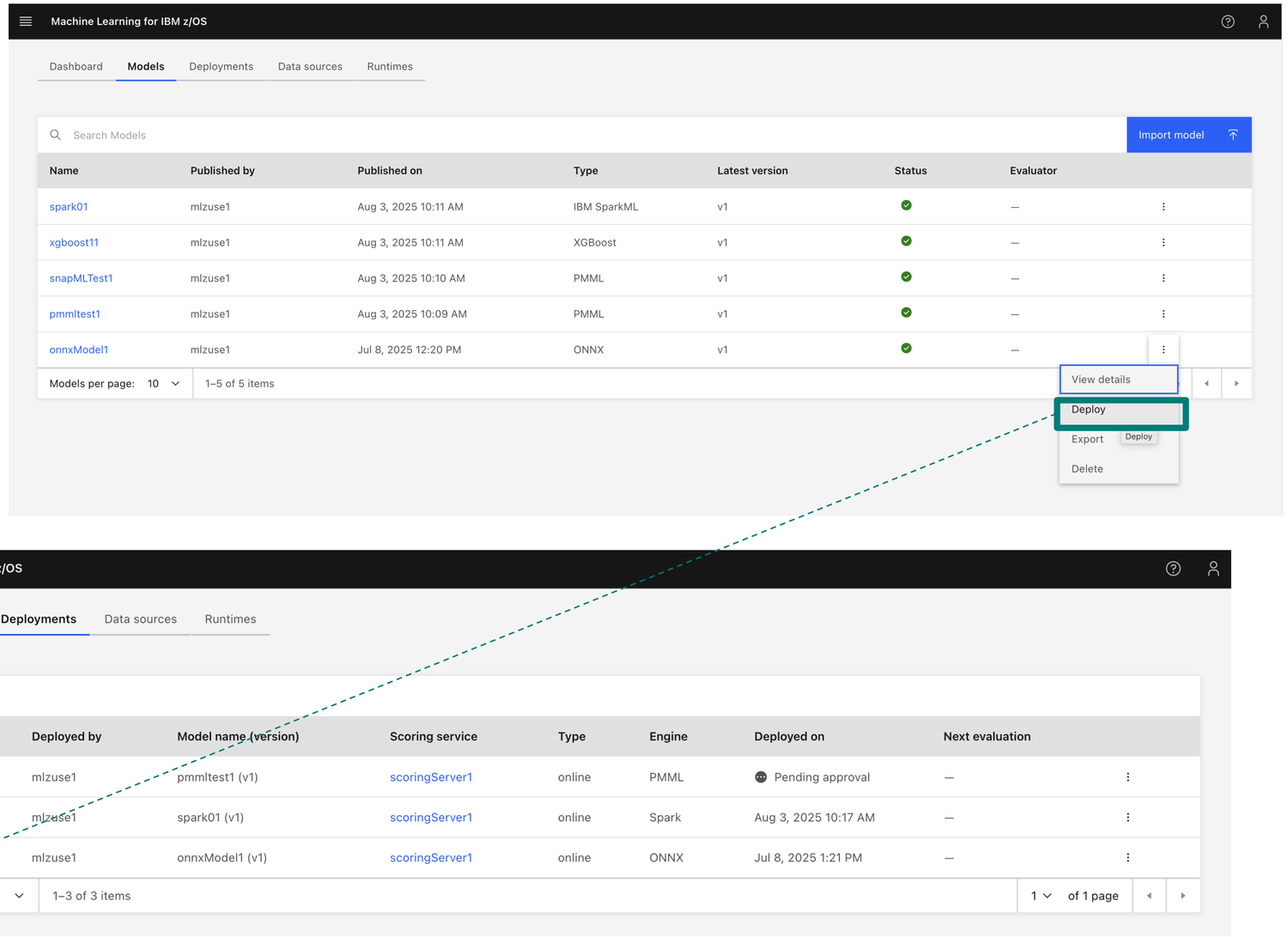
Enables monitoring and evaluation of both output drift and feature drift, providing data scientists with insights from different perspectives

Trustworthy AI – in Machine Learning for IBM z/OS



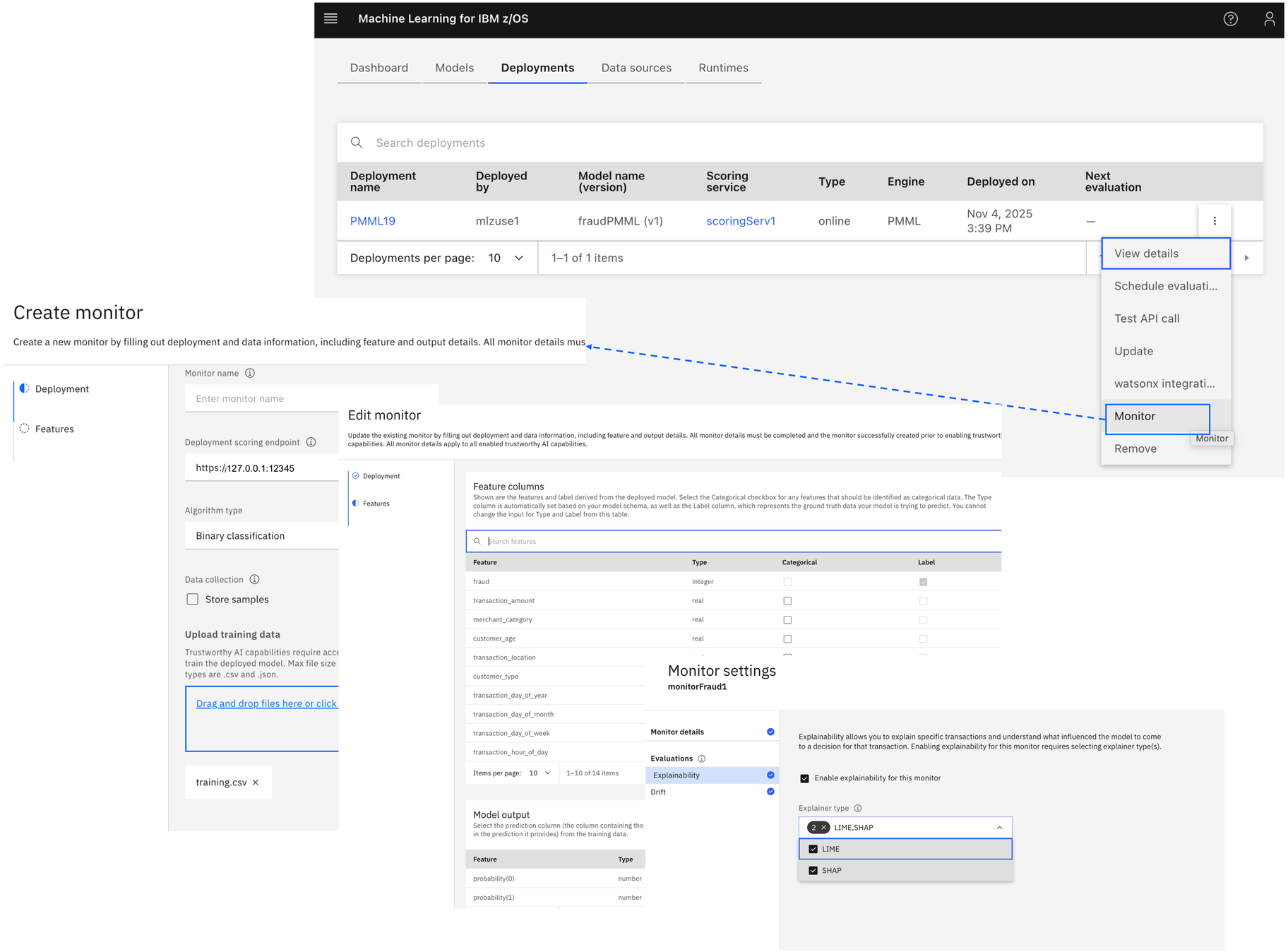
Model deployment and management

- 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

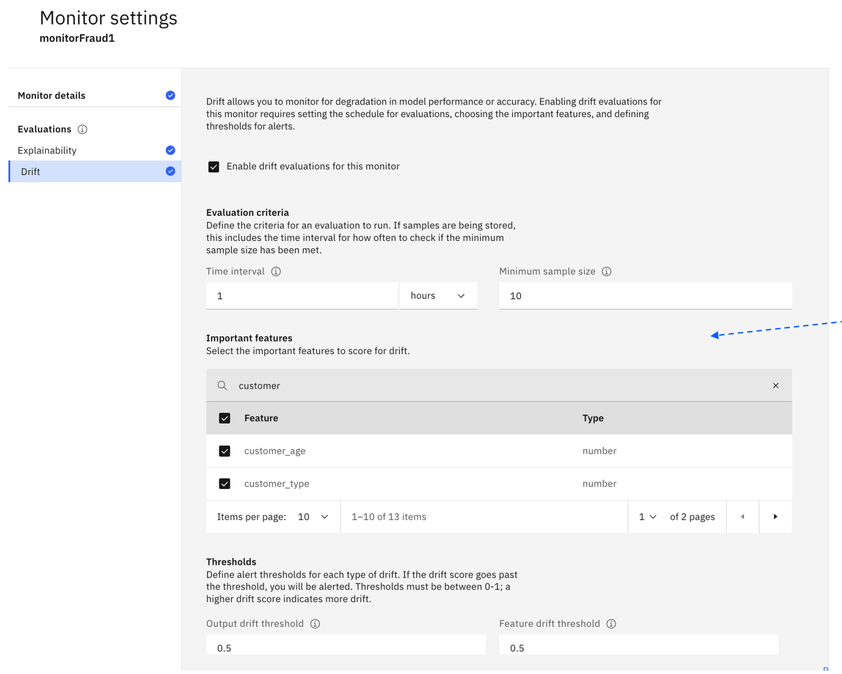


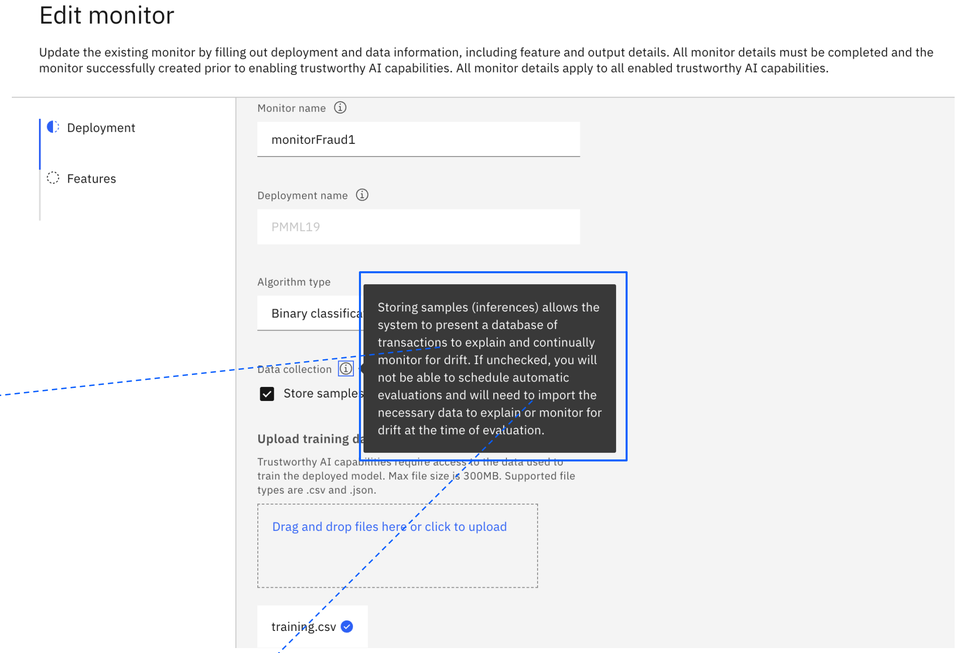
Monitoring: real time or add-hoc?

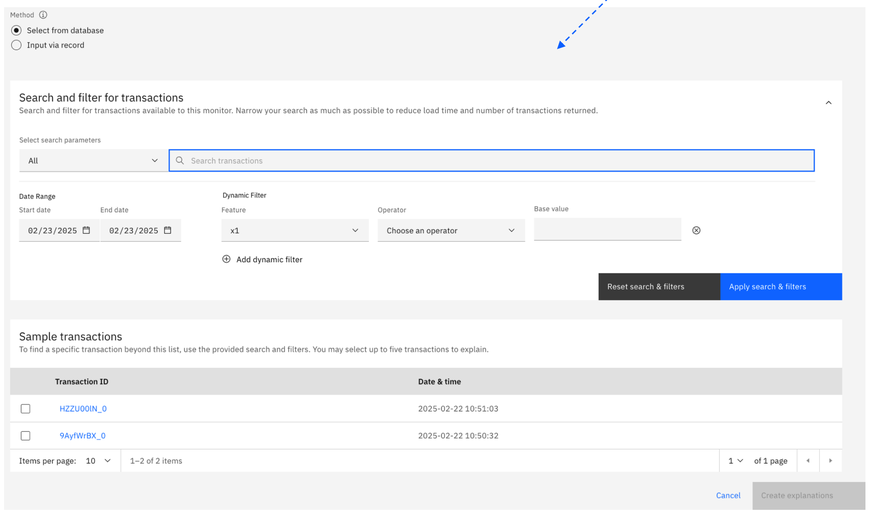
- Collected transactions
  Drift evaluation will run at configured time intervals if the minimum number of transactions have occurred

  Explainability will support selection from stored transactions

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







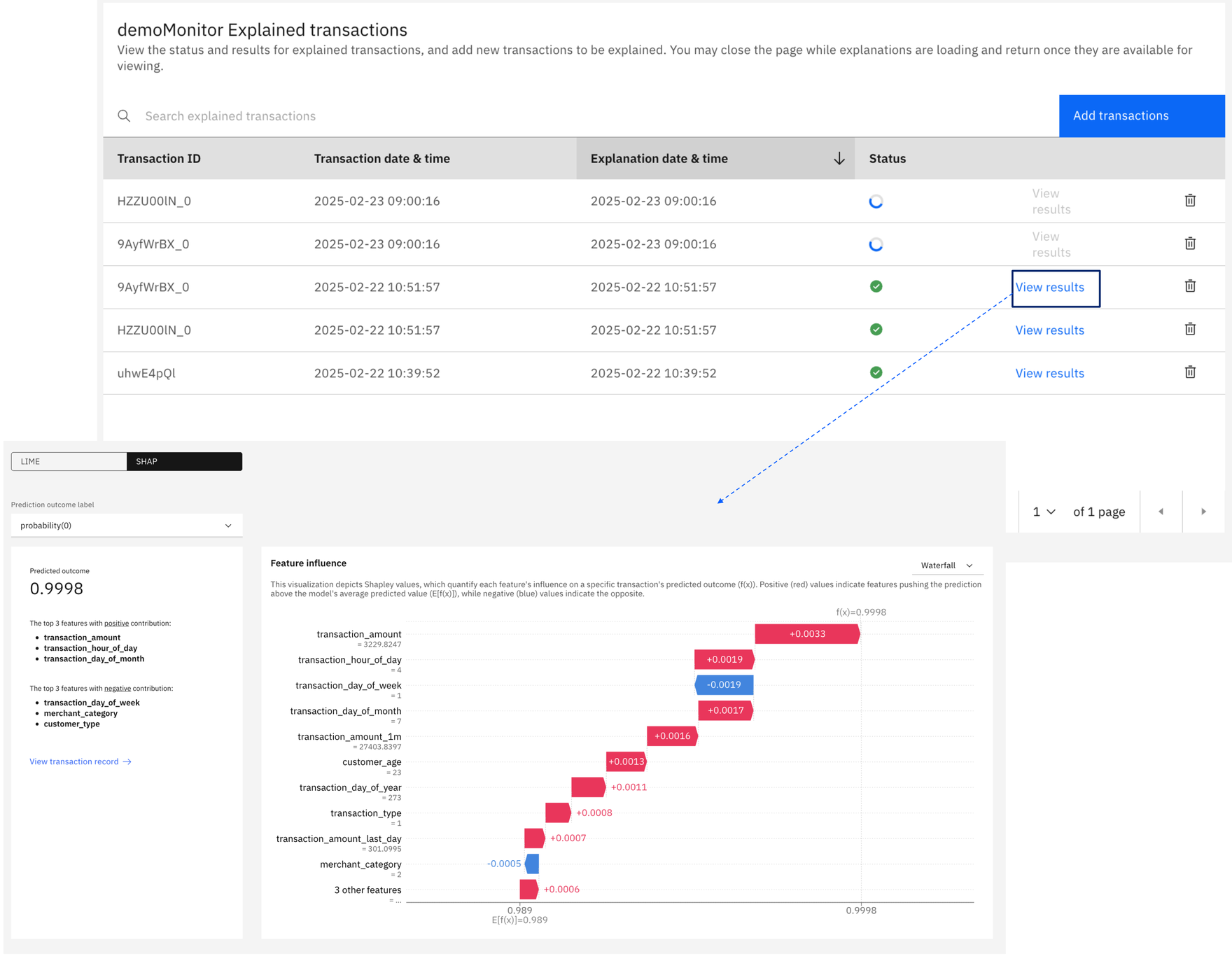
Explainability

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



Drift

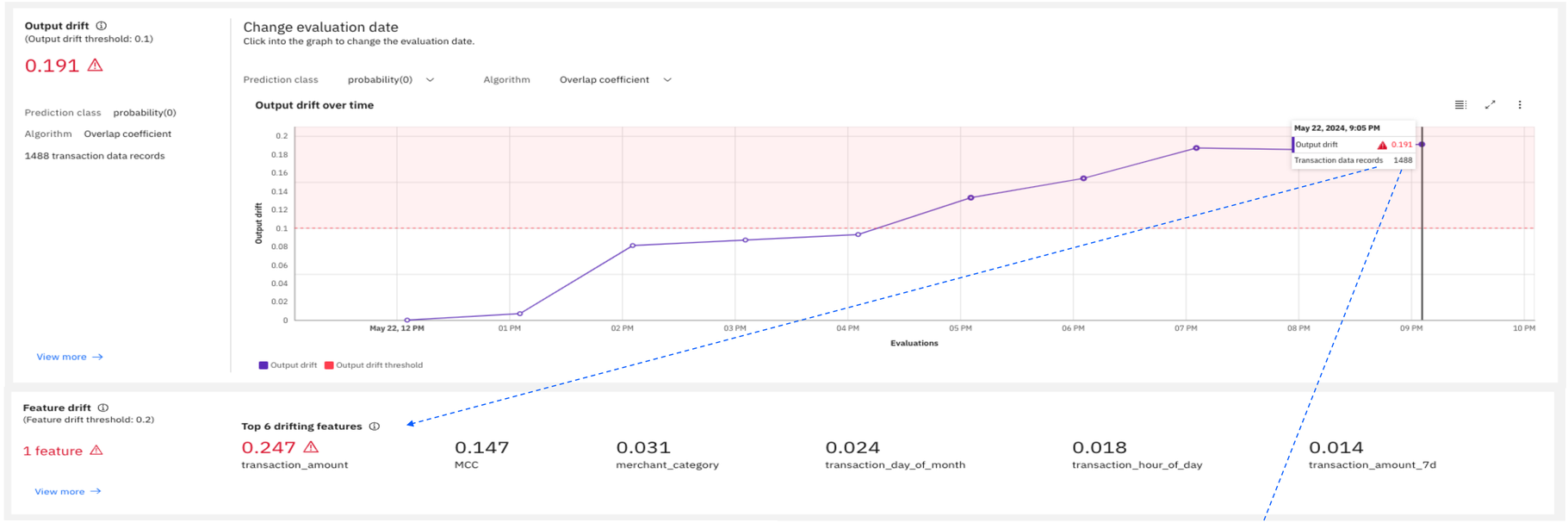
Output Drift

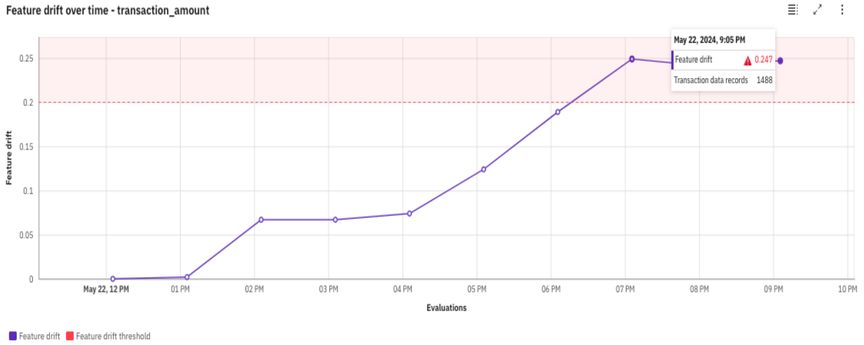
Measured by *Overlap coefficient* and *Total variation distance*

Feature Drift

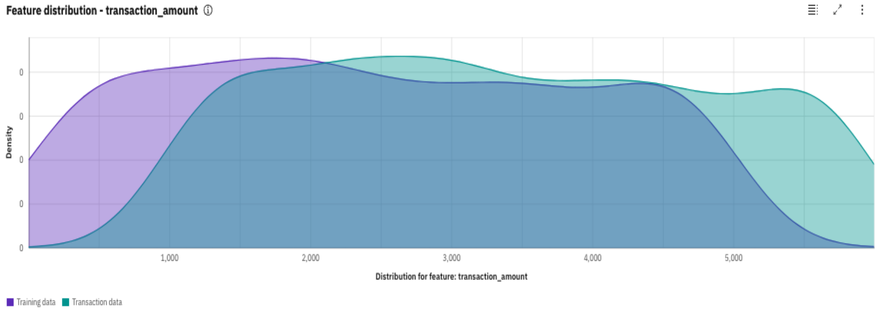
Categorical features are measured by *Jensen-Shannon* and *Total variation distance*

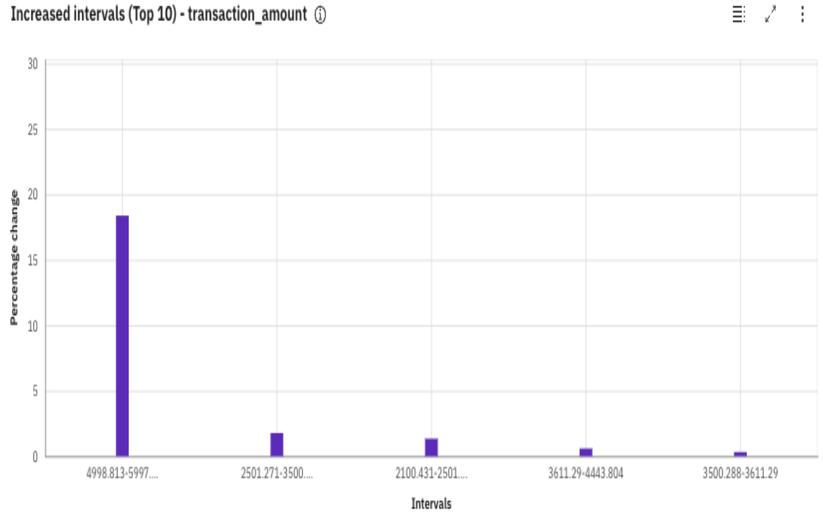
Non-categorical features are measured by *Overlap coefficient* and *Total variation distance*





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 IBM Corporation