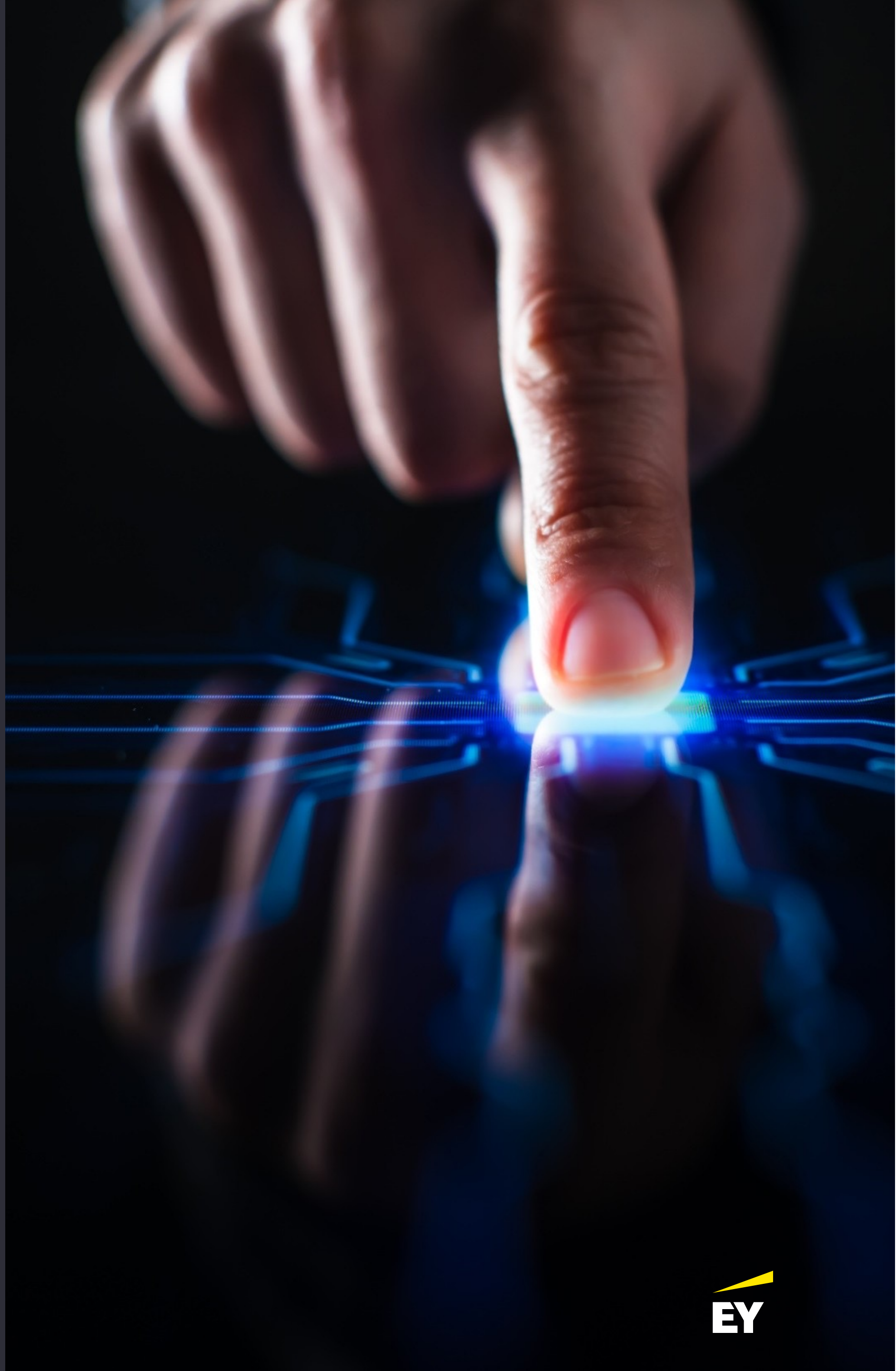
AI Models Ongoing Monitoring

Discussion Paper

Building a better
working world

AGENDA

Ongoing Monitoring Challenges and EY Response	3
► Overview of Ongoing Monitoring Needs and EY Response	4
► Need for Greater Systematic Diligence	5
► Complexity in Roles and Responsibilities	6
► Rapid Technical and Regulatory Evolvement	7
Ongoing Monitoring Framework	8
► Ongoing Monitoring Framework Overview	9
► Framework Components Deeper Dive	10
► Ongoing Monitoring Operating Model	17
► Ongoing Monitoring Roles and Responsibilities	18
Ongoing Monitoring Use Cases	19
► Ongoing Monitoring for AML Unsupervised Learning Models	20
► Ongoing Monitoring of Fairness for Credit Adjudication	22
► Ongoing Monitoring for AML Transaction Monitoring Supervised Learning Models	24
Ongoing Monitoring Technical Appendix	26
References	30



Ongoing Monitoring Challenges and EY teams Response

Overview of Ongoing Monitoring Needs and EY teams Response

Our understanding of ongoing monitoring needs:

- The current model risk management framework may not be well suited to AI models, e.g., risk coverage, transparency assessment, fairness assessment, model-specific risks and considerations, with fast-evolving regulatory guidelines
- Need for standardization of monitoring frameworks and alignment between IT, data, AI, business teams
- The traditional model life cycle is long and may be incompatible with rapidly evolving AI models
- Need for a centralized environment to perform validation and automate monitoring, documenting and reporting workflows

We provide solutions with synergized expertise and market experience:

Monitoring Framework

- ✓ Industry refined AI model standards aligned with regulations
- ✓ Runbooks and playbooks across a variety of use cases
- ✓ Subject Matter Experts in the field of Risk, Validation and Compliance

Facilitating Tools and Infrastructure

- ✓ All encompassing suite of tools built on trustworthy AI
- ✓ Platform to easily update models, monitoring triggers, view reports and statistics
- ✓ Support for regulatory compliant custom metrics

Automated Analytics & Reporting

- ✓ Automated monitoring and alerting process
- ✓ Customized report generation based on templates
- ✓ Deep dive into performance with visual analytics

Specific Challenges and Solutions for AI Model Ongoing Monitoring

1. Need for Greater Systematic Diligence

CHALLENGE: Need for Greater Systematic Diligence

DATA

- ▶ Data inputs volatility
- ▶ Unstructured data
- ▶ High dimensionality
- ▶ Covariate shift: raining data may not be representative of live data
- ▶ Real-time input monitoring for online learning features

MODEL

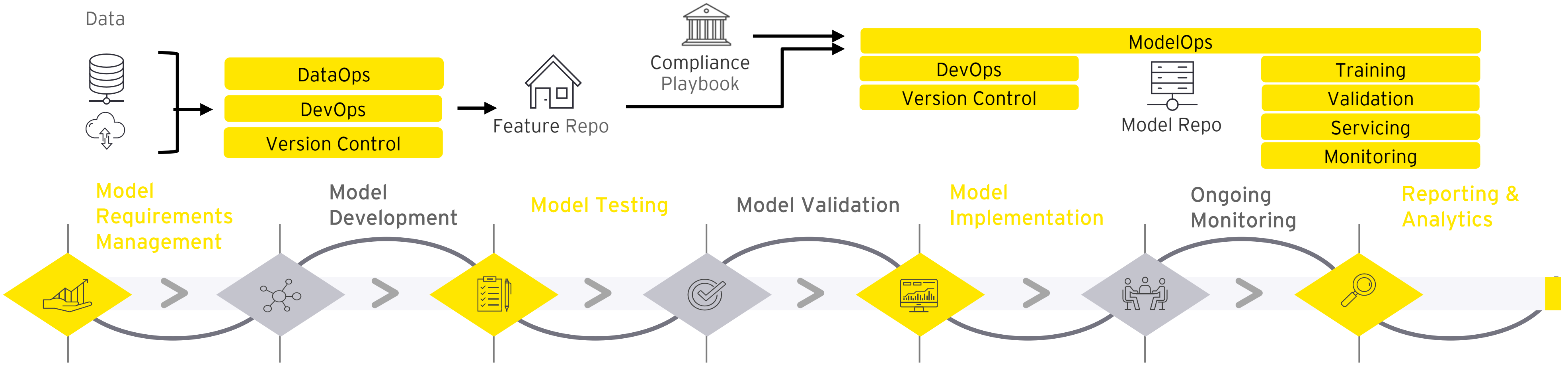
- ▶ Model complexity
- ▶ Greater reliance on model highlights greater need for performance monitoring
- ▶ Dynamic models: change in model itself through learning from data
- ▶ Additional considerations: explainability and fairness

CONFIGURATION INFRASTRUCTURE

- ▶ Slight configuration error causing radically different system behavior
- ▶ Lack of code standardization and management
- ▶ Live-data learning causing hyperparameters re-tuning
- ▶ Infrastructure processing power and data ingestion capacity

SOLUTION: Centralizing Xops with Playbooks

Centralization within a composable and agnostic platform makes it possible to rely on all metadata and artifacts produced across the institution to be able to automate and standardize the monitoring, control and risk assessment of the model



Data

DataOps

DevOps

Version Control

Feature Repo

Compliance Playbook

ModelOps

DevOps

Version Control

Model Repo

Training

Validation

Servicing

Monitoring

Model Requirements Management

Model Development

Model Testing

Model Validation

Model Implementation

Ongoing Monitoring

Reporting & Analytics

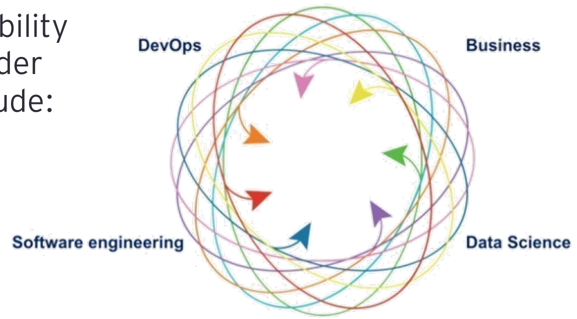
Specific Challenges and Solutions for AI Model Ongoing Monitoring

2. Complexity in Roles and Responsibilities

CHALLENGE: Demand for Multitude of Expertise and Control

AI models tend to have higher complexity, higher data consumption and dependency, lower explainability, and lower stability than traditional models. As a result, the model life cycle, including ongoing monitoring, requires the involvement of a wider range of experts that speak the same language and suitable controls in place. Examples of parties possibly involved include:

- Model Risk Management
- IT Operations
- Business
- Software Engineers
- Data Scientists



SOLUTION: Bringing Together Effective Multi-Skilled Teams

EY teams have rich experience in model risk management, regulatory compliance advisory, and business understanding.

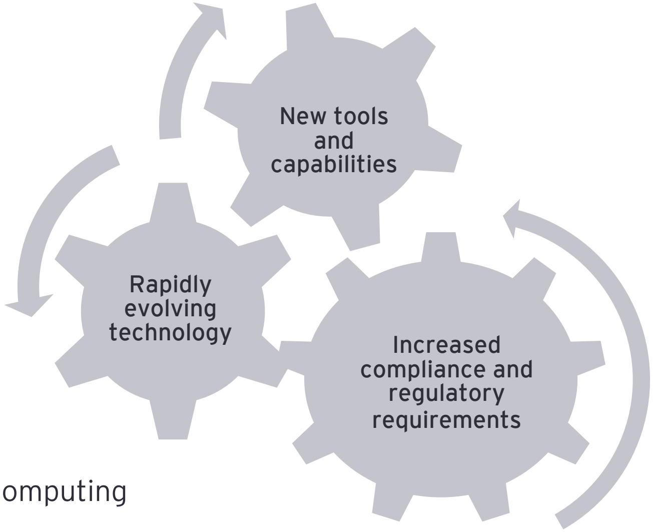
- EY teams have a seasoned core team with wide coverage of backgrounds and skillsets. Insights in peer practices through work collaborations are a significant value add.
- EY teams harvested IP on Trusted AI playbooks provides a detailed Standardized Model Lifecycle platform relying on pre-configured 3 levels of governance: standards, runbooks and processes.

Specific Challenges and Solutions for AI Model Ongoing Monitoring

3. Rapid Technical and Regulatory Evolvement

CHALLENGE: Field Evolving at a Higher Speed

- Traditional (SAS, Matlab, IBM etc.)
- AI-centric (IBM Watson, Datarobot, RapidMiner, H2O, etc.)
- New entrants (Algorithmia, ModelOp, Modzy, etc.)

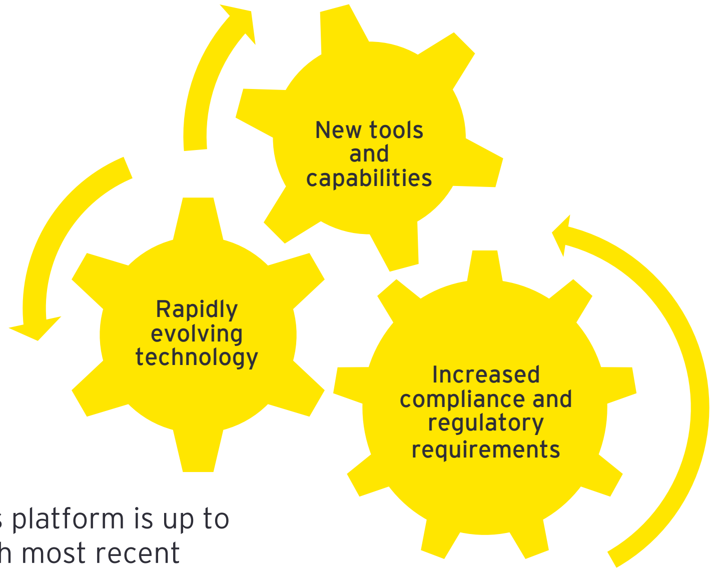

New tools and capabilities
Rapidly evolving technology
Increased compliance and regulatory requirements

- Cloud
- Unlimited computing capacity
- Containerisation
- Security
- DataOps, DevOps

- OSFI E-23
- ECB TRIM Guideline
- World Bank Credit Scoring Guidelines
- OECD AI Principles
- IOSCO Consultation report on AI

SOLUTION: EY teams Are Evolving with the Field

- EY teams academic collaborations and technology alliances facilitate prompt access to emerging AI tools and capabilities


New tools and capabilities
Rapidly evolving technology
Increased compliance and regulatory requirements

- EY XOps platform is up to date with most recent technologies available

- EY Trusted AI Standard details regulatory compliance guidelines for AI models
- EY teams have ongoing conversations with the regulators

Ongoing Monitoring Framework

Ongoing Monitoring Framework Overview

A Monitoring Metrics

- It defines three broad types of metrics: performance, stability and operations based on the use case
  - **Performance metrics:** Detection of Performance drift
  - **Stability metrics:** Prediction drift, data/feature drift, and concept drift
  - **Operations metrics:** Input/Output (IO), Memory, and Central Processing Unit (CPU) usage for predictions, latency when calling Machine Learning Application Programming Interface (ML API) endpoints

B Monitoring Frequency

- Models with self-learning and higher business impact should be monitored more frequently than static models
- More broadly, model risk tiering including materiality is critical in determining frequency

C Override Analysis

- Any override or overlay to the model should be documented and duly justified
- Substantial model overrides are signals that a model may require refinement

D Input Monitoring

- Data quality and validity checks
- Additional checks for unstructured data
- Feature Drift Detection: assess ranges, valid values, and distributions in both univariate and multivariate level; assess outlier volume/distribution in both univariate and multivariate level

E Output Monitoring

- Prediction Drift Detection: assess mean, median, min, max, and distributions for output, as well as stability of the predictions (distributions, volumes)

F Model Monitoring

- Performance Monitoring: directly measuring performance when the target is available; estimating performance when the actual target is compromised
- Model Stability: analyse stability of the model (concept drift detection, i.e., if the relationship between input and output changes)
  - Feature Importance Comparison
  - Model Benchmarking

G Trigger Review

- Quantitative Triggers – Test model accuracy and stability on the recent outcome and against a benchmark model (e.g., product sign up rate)
- Qualitative Triggers – Changes in business strategy / regulatory requirements (e.g., product characteristics, product offer conditions)
- Fallback triggers - Switch to benchmark/legacy model if there is a severe deterioration in model performance (e.g., fallback marketing)
- Processing and capacity triggers - Capture increased model usage or an increase of the data consumption (e.g., increased IO/Memory/CPU demands)

H Action Plan

- Diagnostics and deep dive for additional insights
- Remediation plan for the model retrain/recalibration/review

Key Ongoing Monitoring Considerations

With deployment of AI-powered models, EY teams ~~we~~ believe model monitoring should be refined with deeper insight and an automated triggering mechanism. A detailed monitoring plan should be in place to monitor any gradual or sudden changes that impact prediction quality and appropriateness of the model

Framework Components Deeper Dive (A, B, C)

Monitoring metrics , frequency and override analysis

A Monitoring Metrics

Proposed model monitoring metrics should be established based on the specific model requirements and business objectives.

- **Performance Metrics** to evaluate model performance in production
- **Stability Metrics:** (1) Prediction drift – Captures the distribution shift of the predicted outputs and/or dependent variable; (2) Data/Feature drift – Captures the distribution shift of each univariate feature and shift in multivariate space; (3) Concept drift - Captures the changes in relationships between input variables and the target variable.
- **Operations Metrics** to determine how the deployed model is performing from a usage point of view. E.g., IO/Memory/CPU usage when performing prediction, latency when calling AI API endpoints

B Monitoring Frequency

Monitoring **frequency** should be specified to monitor any gradual or sudden changes that impact prediction quality and the appropriateness of the model.

- For AI models, there is a need for a well-defined monitoring frequency **based on model risk tiering** (including factors such as materiality, complexity, regulations, ethical impact)
- In the **initial deployment** phase, **higher scrutiny** for AI models is needed.
- Models with **self-learning** and quantitative adjustment capabilities (e.g., reinforcement and active learning require additional scrutiny and controls) should be monitored more frequently than static models.

C Override Analysis

Any business overlays and overrides to the AI model should be documented and justified.

- **Business Overlays:** applied on top of AI model outputs in the implementation process. The business rationale should be included.
- **Model Overrides:** AI model outputs may be overridden by business users in practice. These occurrences should be documented and justified.

Framework Components Deeper Dive (D)

Input monitoring

D

- Input monitoring checks the **quality of input data** (data scheme changes, missing values, data formatting issues etc.) and detects possible **data/feature drifts** which could cause sudden or gradual decline of the model performance
- Monitoring the distribution of the model input can help identify the model performance decline **early**, even before the model performance falls out of the pre-defined acceptable range, i.e., **data/feature drift could happen before model performance becomes unacceptable**.
- In addition, data/feature drift monitoring can help **distinguish model underperformance from expected variations** and it can help with the **root cause analysis** when noticing a performance drop.
- Data/feature drift detection becomes especially **crucial** when the model performance cannot be measured directly (the ground truth is not available). Significant data drift likely indicates underlying model performance decaying.

Data Quality Check

- Data Scheme Check – captures mistakes/errors with data scheme changes
- Missing Data Comparison – captures abnormal missing value volumes
- Statistical Analysis – capture mistakes/errors with data formatting (e.g., unit mistake)

Data/Feature Drift Detection

- Univariate Feature Drift Detection
  - **Statistical Analysis**: min, max, median, 25% & 75% quartile, Q-Q Plots etc.
  - **Outlier Analysis** (volume, distribution): significant outlier increase could provide insight into model performance decay
  - **Distribution Comparison** – full population: capture distribution shift of each independent variables in the production data; the impact of the drift for each feature can be ranked based on the correlations with performance dropping in a given period of time
  - **Distribution Comparison** – target class population (classification): capturing the feature drift of the target population (i.e., population of interest)
- Multivariate Feature Drift Detection
  - **Inter-feature correlation analysis:** changes in correlations between features likely is an indication of multivariate drift
  - **Data Reconstruction with PCA**: Apply PCA data reconstruction to the reference data, using the average reconstruction error of the reference data to define the acceptable range for the reconstruction error; Apply the same PCA data reconstruction to the production data, calculating the average reconstruction error
  - **K-Means Clustering:** Train K-means clustering on reference data (i.e., training data) and create clusters predictions; assign the production data into these clusters; Percentage Analysis: compare how reference data and production data is split among clusters; Centroid-Distance Analysis: compare the average distances to centroid for reference and production data
  - **Multivariate outlier analysis:** capture the outlier volume/distribution shift in the multivariate dimension
- Data/Feature Drift Detection for Unstructured Data
  - Append to the old population a label of 0 and append to the new population a label of 1. Apply a supervised learning algorithm (where the label is the target) to the aggregate population

Framework Components Deeper Dive (E)

Output monitoring

E

- Output monitoring focuses on the stability of outputs (prediction) for the given production period, to detect the prediction drift. The prediction drift is an indicator of population structure changes, which could cause overall model performance decay.
- The Output Distribution Comparison analysis in both the full population or specific population, e.g. population around classification cut-off, can be used to detect prediction drift.

Output Distribution Comparison

- **Output Distribution Comparison – Full Population**
  - Comparing the distribution of recent output for the full population with that in the reference period (e.g. training data)
  - Shift in output distribution indicating possible underlying performance decline
- **Output Distribution Comparison – Population around cut-off** (classification models)
  - Comparing the distribution of recent output for population around cut-off (e.g., cut-off ± 10%) with that in the reference period (e.g. training data)
  - Shift in distribution around cut-off provides additional insights on model performance shift

Framework Components Deeper Dive (F)

Model monitoring

F

- **Performance monitoring** is the main pillar in traditional model monitoring. The aim of performance monitoring is to evaluate if the model performance is within a pre-defined acceptable performance range by comparing the model outputs **with the ground truths**.
- However, **the ground truth is often unavailable or compromised** (i.e., delayed or partially available) in practice. In such cases, estimated performance can provide additional insights on the possible model performance decline along with the input/output monitoring.
- In addition to performance monitoring, model stability tests, including feature importance comparison and model benchmarking etc., can be used to detect concept drift, i.e., changes in relationships between input variables and target variable.

Performance Monitoring

When the Ground Truth is Available:

- Classification Models
  - Accuracy, Precision and Recall, Specificity, F1-score, AUROC etc.
- Regression Models
  - Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-Squared, Adjusted R-Squared
- Clustering Models
  - Davies-Bouldin, Silhouette Width, Dunn, Calanski-Harabasz

When the Ground Truth is Not Available or Compromised:

- Classification Models: Confidence-based Performance Estimation (CBPE)
- Regression Model: Direct-Loss Estimation (DLE)

Model Stability

- Feature/Output Correlation Analysis
  - Perform the correlation analysis between feature and model output for training data and production data respectively
  - Comparing the correlation coefficients for training and production; a significant change in correlation indicating potential model performance drift
- Feature Importance Comparison:
  - Retrain the model using the production data
  - Compute the feature importance of the retrained model
  - Compare the feature importance of the retained model with that of the original model
  - The feature importance change is an indicator of concept drift
- Model Benchmarking:
  - Comparing the recent model output with the output of benchmark models
  - The deviation of benchmarking analysis from that in reference period (i.e. training data) can be an indicator of concept drift

Framework Components Deeper Dive (G)

Quantitative, qualitative, fallback and processing triggers deeper dive

G

- Set model-specific triggers to indicate thresholds supporting effective use of the model. Describe which mitigating actions should be taken in case a model performance incurs a trigger event

Quantitative Triggers

Against pre-set bounds or a benchmark model based on performance thresholds, stability and direction measurements

- Compare actual vs. thresholds and generate alert if moving average of % error is significant and break the pre-set bounds
- E.g., Model error $\leq \alpha1$, Performance as expected; $\alpha1 \leq$ Model error $\leq \alpha2$, Generate alert with a specific recommendation response.

Qualitative Triggers

Changes in business strategy / regulatory requirement

- **Low risk:** e.g., Minor to modest portfolio composition changes in terms of size and/or quality, which are unlikely to change the portfolio's balances or revenues
- **Medium risk:** e.g., Internal or external policy changes related to the portfolio, such as changes in regulatory requirement
- **High risk:** e.g., strategic changes leading to significant impact to portfolio development (e.g., model fairness regulation)

Fallback Triggers

In case of deterioration in performance (e.g., switch to benchmark / legacy model) is critical

- Establish metrics to trigger a **fallback to a baseline model** or legacy model if the AI model outputs violate business intuition
- Establish metrics to trigger a simple override of the model outputs in extreme scenarios
- Consider tracking the frequency of the end user's need to override the model outputs

Processing and Capacity Triggers

Increased usage and data consumption need to be captured by processing and capacity triggers to ensure infrastructure accommodation

- Establish metrics to capture increased **model usage**, or an increase of the **data throughput**
- Set model review trigger and define roles and responsibilities across AI model lifecycle participants to ensure the software and hardware accommodation capacity

Framework Components Deeper Dive (H)

Action Plan

Diagnostics

Upon identification of performance, output, and input deviation, it is helpful to investigate whether a business reason (e.g., strategy change), an operational error (e.g., system schema mapping), etc. is contributing to the issue. Sensitivity analysis and/or benchmarking are typically helpful tools. If that does not resolve the deviation, then a staggered action plan can be used.

Remediation Plan

Data Quality	Operation/IT
Connect with Data Team to investigate and resolve the outstanding issues with data quality	Connect with Technology team to investigate and resolve the outstanding issues with model operation in production

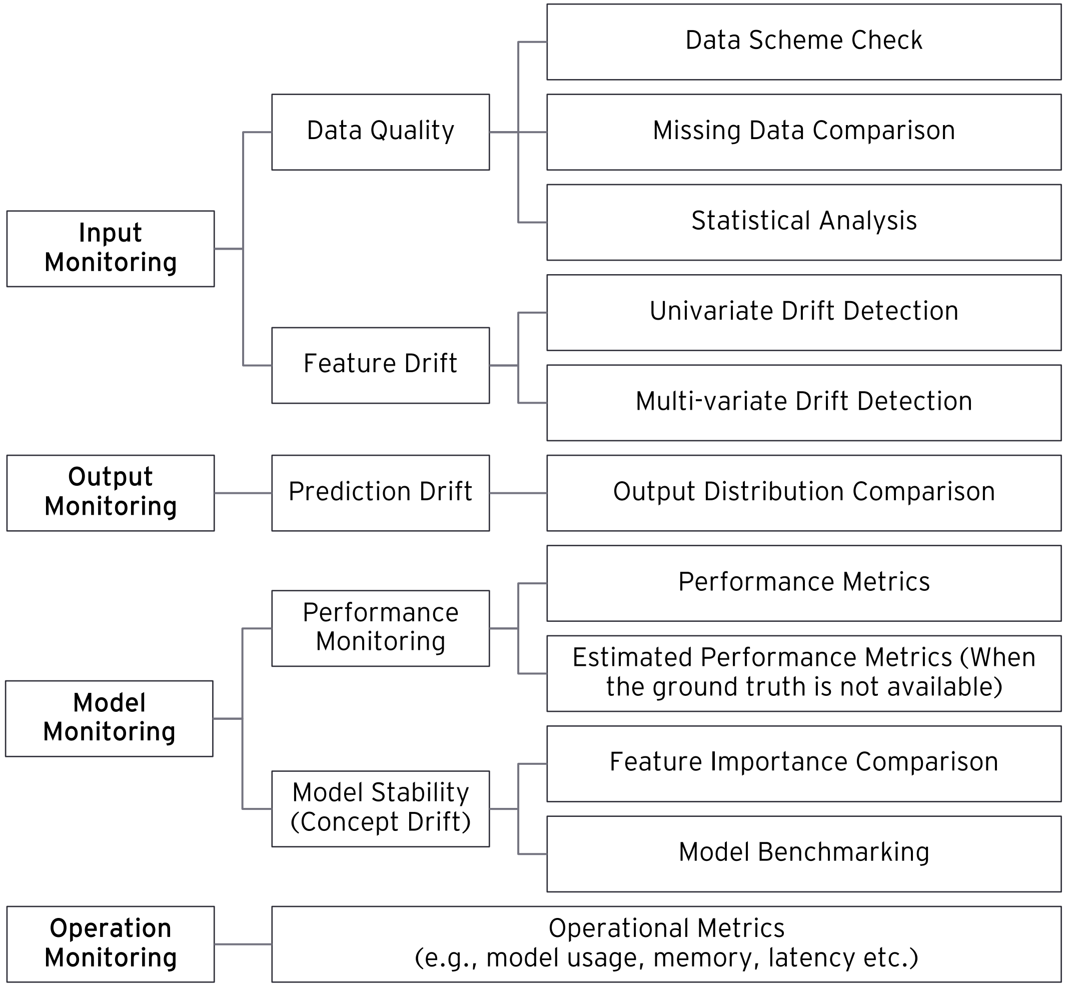
Model Updates/Changes

	1. Retraining and Recalibration	2. Model Architecture Review	3. Model Design Review
Description	▸ Re-run of the model based on the new data available (hyperparameters should be explicitly documented) ▸ Recalibration of hyperparameters and retraining	▸ Review of model architecture (e.g., decision tree vs XGBoost vs Neural Network)	▸ Review of fundamental model design, including changes to input features, target variable, machine learning paradigm, etc.
Occurrence	▸ Granularity defined by the following criteria: data update, model use frequency, performance window, etc.	▸ Triggered by an agile response if model recalibration is not satisfactory	▸ Well-defined frequency based on model materiality ▸ May be triggered if a model architecture review does not suffice
Approval	▸ Consider performance, swap-in swap-out, explainability of updates, etc. ▸ Ensure higher complexity models are supported by higher performance ▸ Ensure updates are communicated to Model Validation if applicable and approved		

Ongoing Monitoring Framework

Proposed ongoing monitoring process flow

Monitoring | Analyses



- Input Monitoring
  - Data Quality
    - Data Scheme Check
    - Missing Data Comparison
    - Statistical Analysis
  - Feature Drift
    - Univariate Drift Detection
    - Multi-variate Drift Detection
- Output Monitoring
  - Prediction Drift
    - Output Distribution Comparison
- Model Monitoring
  - Performance Monitoring
    - Performance Metrics
    - Estimated Performance Metrics (When the ground truth is not available)
  - Model Stability (Concept Drift)
    - Feature Importance Comparison
    - Model Benchmarking
- Operation Monitoring
  - Operational Metrics (e.g., model usage, memory, latency etc.)

Diagnostics

System Errors/Changes

e.g., Data base Infrastructure update, Changes in date formats, New 3rd party data provider or API, Change in default value

Natural Fluctuation

e.g., seasonal changes

Macroeconomic changes

e.g., Inflation, Interest rate changes, Unemployment rates

Adverse Events

e.g., disrupting events (pandemic, Wars, Energy crisis)

Business Drivers

e.g., product offerings change, portfolio acquisition, strategy change

Remediation Actions

Connect with **Data Team** to investigate and resolve the outstanding issues with data quality

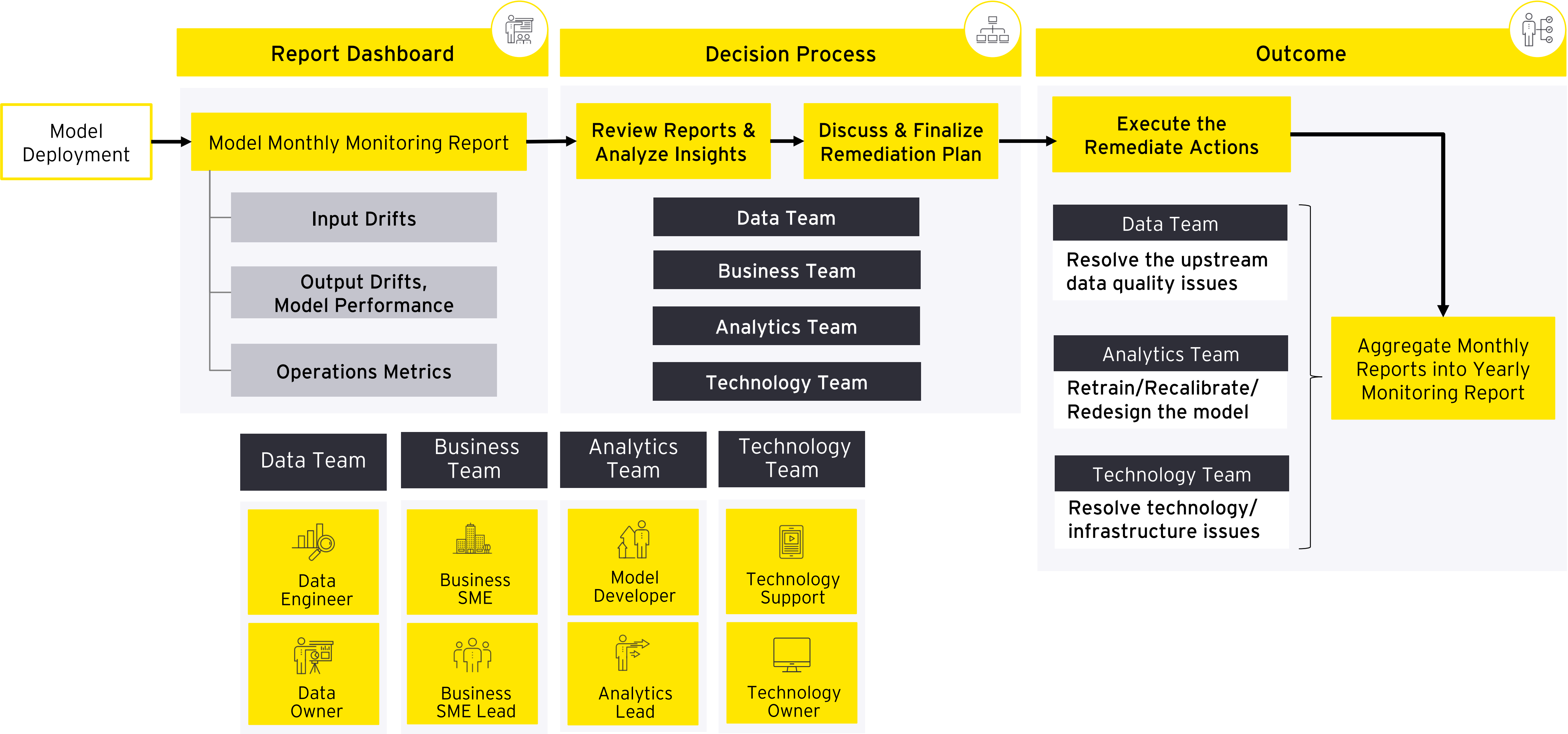
Model Updates/Changes

1. Retraining/Recalibration
2. Model Architecture Review
3. Model Redesign

Connect with **Technology Team** to investigate and resolve the outstanding issues with model operation

Ongoing Monitoring Operating Model

Proposed operating model for the application of the ongoing monitoring process flow



Report Dashboard

Decision Process

Outcome

Model Deployment

Model Monthly Monitoring Report

- Input Drifts
- Output Drifts, Model Performance
- Operations Metrics

Review Reports & Analyze Insights

Discuss & Finalize Remediation Plan

- Data Team
- Business Team
- Analytics Team
- Technology Team

Execute the Remediate Actions

Data Team: Resolve the upstream data quality issues

Analytics Team: Retrain/Recalibrate/ Redesign the model

Technology Team: Resolve technology/ infrastructure issues

Aggregate Monthly Reports into Yearly Monitoring Report

Data Team	Business Team	Analytics Team	Technology Team
Data Engineer	Business SME	Model Developer	Technology Support
Data Owner	Business SME Lead	Analytics Lead	Technology Owner

Ongoing Monitoring Roles and Responsibilities

Proposed RACI (Responsible, Accountable, Consulted, Informed) matrix for ongoing monitoring

Tasks associated with Model Monitoring		Roles associated with Model Monitoring					
Activity Group	Activities	Analytics Data Scientist	Analytics Lead	Business SME	Business Lead	Data Owner	Technology Owner
Monthly Monitoring	1. Execute model monitoring	R	A				
	2. Create Monthly Monitoring Report	R	A	I	I	I	I
	3. Review the reports and perform insight analysis	R	A	I	I	C	C
	4. Discuss monitoring results and finalize remediation plan	R	AR	R	AR	R	R
	5. Remediate Data Issues	C	C	I	I	A	
	6. Remediate Technology Issues	C	C	I	I		A
	7. Remediate Model Issues	R	A	C	C		I
Yearly Reporting	8. Create Yearly Monitoring Report by aggregating/consolidating the Monthly Monitoring Reports	R	A	C	C	C	C
	9. Submit Yearly Report to MV team if applicable	R	A	I	I		
Maintenance and Oversight	10. Ongoing Monitoring Framework Maintenance and Attestation	R	AR	C	C	C	C

Ongoing Monitoring Use Cases

Use Case 1: Ongoing Monitoring for AML Unsupervised Learning Models

Context, Problem Statement, Need, and Solution Approach

Context

As market conditions and customer behaviour vary over time, there can be frequent changes in regulatory and business requirements and data. As a result, AML models need to be consistently adjusted to keep relevant and mitigate risk.

Problem statement

- Changes in business/regulatory requirements increase the complexity of the ongoing monitoring process
- Changes in customer and transaction data distribution weaken the stability and performance of AML models
- The ongoing monitoring for AML unsupervised learning models is inefficient without a standardized process guideline

Need

Require a standardized process that helps to perform consistent and efficient ongoing monitoring for AML models

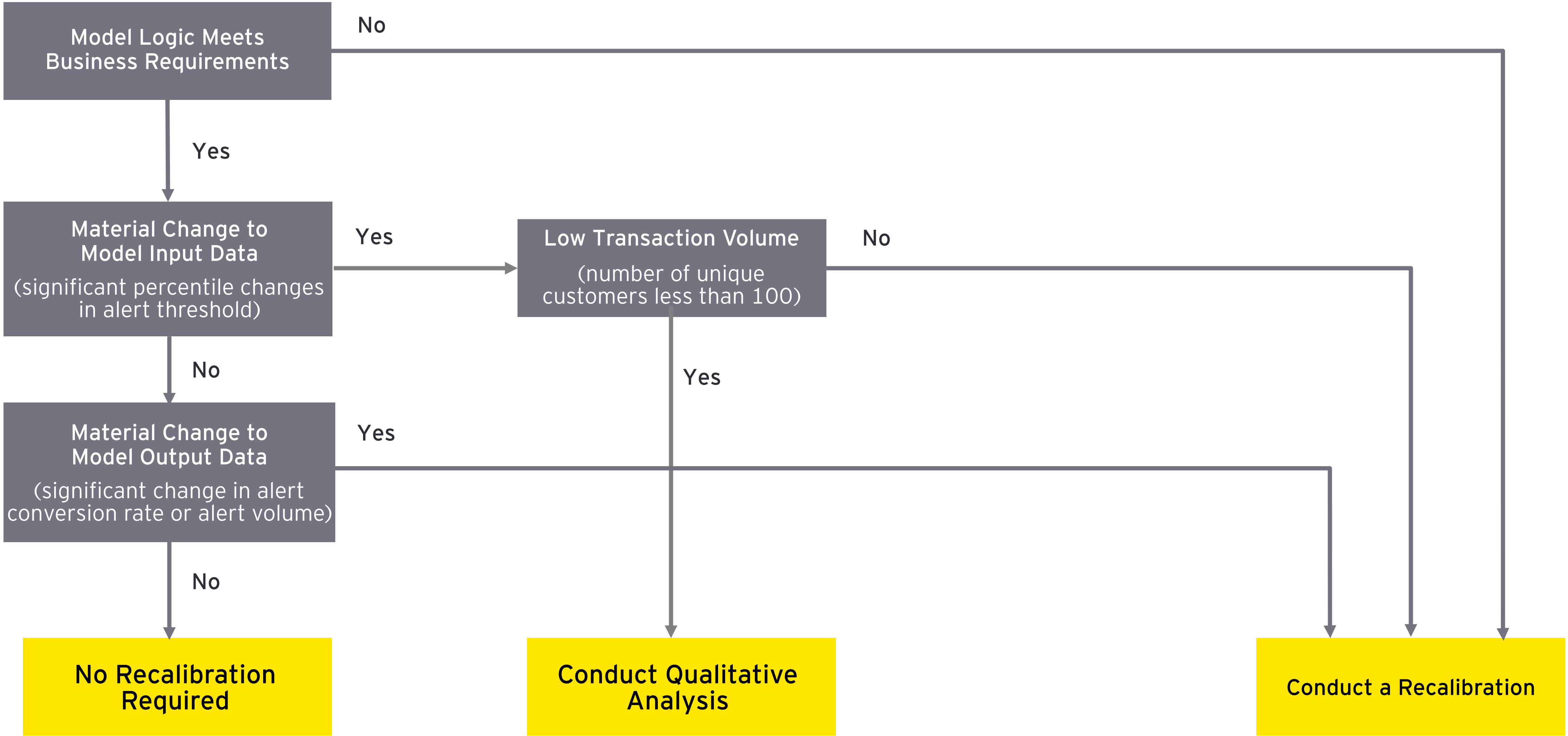
Solution Approach

Construct an ongoing monitoring decision tree for AML models to help standardize the ongoing monitoring process and increase its efficiency

Monitoring Metrics

- Input Data: Percentile change in alert generating thresholds
- Output Data:
  - Significant change in alert conversion rate
  - Significant change in alert generation volume

Use Case 1: Ongoing Monitoring for AML Unsupervised Learning Models


Model Logic Meets Business Requirements
No
Yes
Material Change to Model Input Data
(significant percentile changes in alert threshold)
Yes
Low Transaction Volume
(number of unique customers less than 100)
No
No
Yes
Material Change to Model Output Data
(significant change in alert conversion rate or alert volume)
Yes
No
No Recalibration Required
Conduct Qualitative Analysis
Conduct a Recalibration

Use Case 2: Ongoing Monitoring of Fairness for Credit Adjudication

Context, Problem Statement, Need, and Solution Approach

▶ Context

Fairness of a decision-making process encompasses two distinct notions: disparate treatment (decisions partly based on subjects' sensitive attributes) and disparate impact (decisions' outcome is disproportionately detrimental or beneficial to people with sensitive attributes). Fairness monitoring is especially important to banks as the bank's decision impact on the customer can be significant.

▶ Problem statement

Current MRM framework both lacks the thereof in the current validation practices as well as the need for it in ongoing post-prod monitoring.

This is true even for static models that are not retrained, which is eye-opening:

- Even a static model WILL have variance on fairness scores if the balance of runtime transactions (say hour by hour) sees dramatically more of the protected class than the privileged class.
- This can happen even if drift remains near zero, i.e., those records ARE part of the expected data distributions from training time. There is just a lot more of them at a specific point in time that are being biased against.

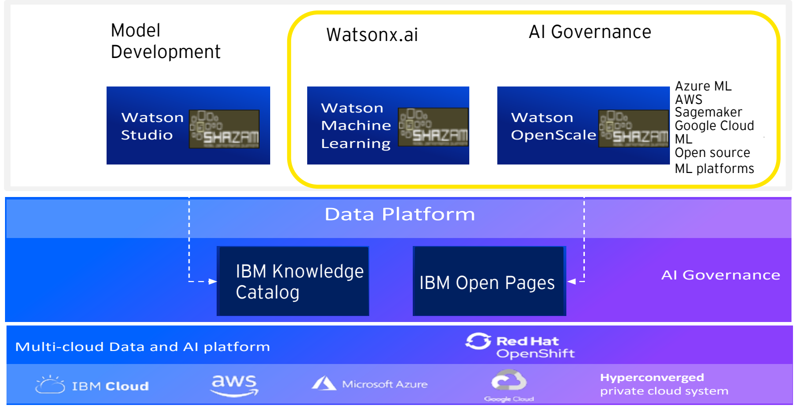
▶ Need

Requires to install new capabilities that equate this to "temporal fairness drift", which is different from data or accuracy drift.

▶ Solution

Develop a "dynamic validation" capability that couples regulatory-compliant playbooks on model validation and model change, and Trusted AI (standards, runbooks and workflows, pre-configuring the tests and the thresholds) with IBM's runtime monitoring engine

Use Case 2: Collaboration with IBM



Why

EY teams and IBM together identified opportunities to bring a high level of alignment between the 3 LoDs on model definition, risk tiering, performance measurement and ongoing monitoring, the third-party and open-source considerations.

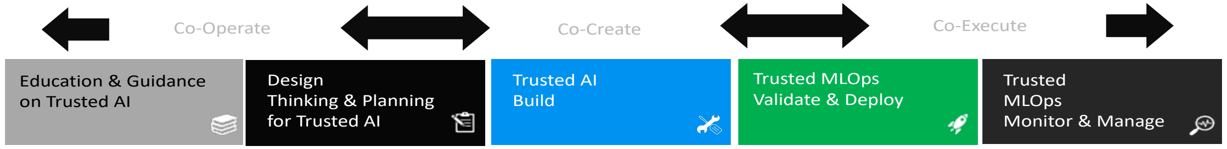
Value

IBM Trusted AI capabilities and deployment technology coupled with EY's harvested IP on Trusted AI playbooks provide an end-to-end Standardized Model Lifecycle platform relying on pre-configured 3 levels of governance: standards, runbooks and processes.

Objective

Develop specific solutions/client offerings to support Trusted MLOps Validate & Deploy and Trusted MLOps Monitor & Manage projects.

Execution


Co-Operate
Co-Create
Co-Execute
Education & Guidance on Trusted AI
Design Thinking & Planning for Trusted AI
Trusted AI Build
Trusted MLOps Validate & Deploy
Trusted MLOps Monitor & Manage

Proof-of-Concept		Description	Asset-based Approach	Benefits
EY XOps AI Model Development, Validation, Monitoring Playbooks integrated in **IBM Cloud Pak for Data**	Use Case 2b	IBM OpenScale features are enhanced by EY XOps AI Risk Standard to provide a cutting-edge ongoing model monitoring and **accelerated validation** solution for AI models	Importing EY Validation Runbook to extend the capability of IBM Cloud Pak for Data (CPD) with minimal execution risk.	▸ Time-to-market reduced by 80% ▸ Full validation of the deployed model on live data ▸ Reusability of pre-vetted deployed pipeline ▸ Model Change can be assessed "on-the-fly"
	Use Case 2a	IBM Cloud Pak for Data stores all model information in a repository for building personalized automated reports based on EY XOps AI model documentation engine	EY teams have developed an innovative IP to build regulatory-compliant automated documentation from IBM model repository for each client use case.	▸ 80% of documentation automated ▸ 2ndLoD Focus on analyses and remediation plan effectiveness ▸ Automated monitoring report

Use Case 3: Ongoing Monitoring for AML Transaction Monitoring Supervised Learning Models

Context, Problem Statement, Need, and Solution Approach

Context

The marketplace and customer behavior are constantly evolving. To keep up with these changes, banks are utilizing machine learning models for AML transaction monitoring. However, effective governance of these models requires ongoing monitoring of their inputs, outputs, and performance, as highlighted by continuously evolving regulatory guidelines.

Problem statement

- Changes in customer and transaction data distribution compromise the stability and performance of AML models
- Changes in business/regulatory requirements increase the complexity of the ongoing monitoring process

Need

Require an ongoing monitoring framework that helps to shed light on emerging customer portfolio changes, FIU alert volumes, conversion rates, and model performance

Solution Approach

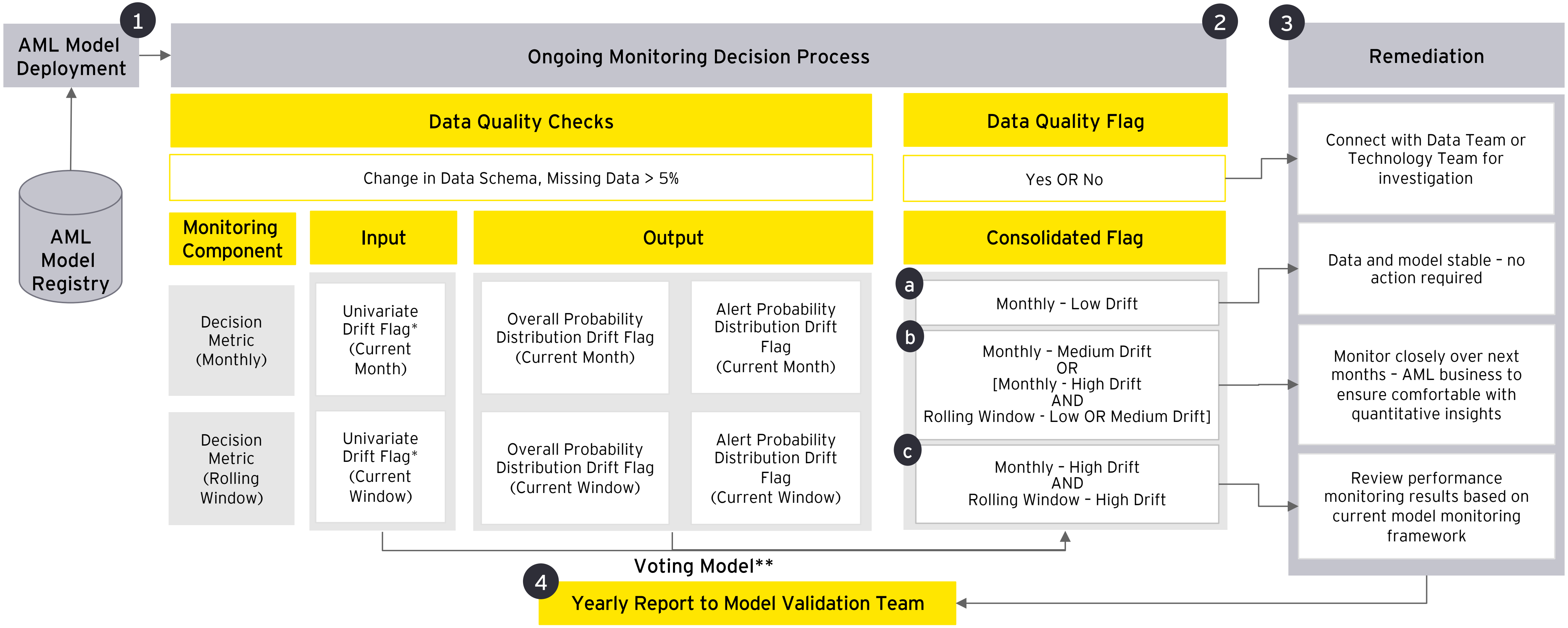
- Explore and test various techniques and metrics to identify the most informative and relevant analyses to gain insights on input drift, output drift, and performance
- Build an ongoing monitoring decision process to support consistent decisioning based on diagnostics

Monitoring Metrics

- Output Data:
  - Alert indicator – Alert Rate
  - Model performance indicator - Case Conversion Rate

Use Case 3: Ongoing Monitoring Decision Process

Mapping of input and output drift results to a remediation action



1 AML Model Deployment

AML Model Registry

2 Ongoing Monitoring Decision Process

Data Quality Checks

Change in Data Schema, Missing Data > 5%

Data Quality Flag

Yes OR No

Monitoring Component	Input	Output	
Decision Metric (Monthly)	Univariate Drift Flag* (Current Month)	Overall Probability Distribution Drift Flag (Current Month)	Alert Probability Distribution Drift Flag (Current Month)
Decision Metric (Rolling Window)	Univariate Drift Flag* (Current Window)	Overall Probability Distribution Drift Flag (Current Window)	Alert Probability Distribution Drift Flag (Current Window)

Consolidated Flag

a Monthly - Low Drift

b Monthly - Medium Drift
OR
[Monthly - High Drift
AND
Rolling Window - Low OR Medium Drift]

c Monthly - High Drift
AND
Rolling Window - High Drift

3 Remediation

Connect with Data Team or Technology Team for investigation

Data and model stable - no action required

Monitor closely over next months - AML business to ensure comfortable with quantitative insights

Review performance monitoring results based on current model monitoring framework

Voting Model**

4 Yearly Report to Model Validation Team

*The aggregate univariate drift flag is based on the weighted average PSI across features considered, where the weight corresponds to the feature coefficient in the logistic regression. Note the univariate drift was selected given its correlation with multi-variate drift and based on business input given its interpretability.

Drift flags rely on the PSI thresholds set out for low, medium, and high drifts. **These can be calibrated based on business risk appetite and updated on an ongoing basis through active learning.

***Note that model performance metrics are not included in the decision process given that they are lagging (~45 days for case conversion, ~3 months for STR conversion). Additionally, they may not be stable for new models immediately, till FIU investigations stabilize for new typologies

Ongoing Monitoring

Technical Appendix

Metrics for Distribution Comparison

Insights on relevant tests/metrics

Statistical Tests/Metrics	Variable Types	Notes
Kolmogorov-Smirnov (K-S) test	Numerical Variables	Equally Sensitive to mean and variance difference, but over-sensitive when the sample size is large, not recommended when sample size is larger than 1000.
Population Stability Index (PSI)	Numerical Variables; Categorical Variables	More sensitive to variance difference, especially sensitive to new values; binning strategy (for numerical features) could affect the results significantly
Wasserstein Distance (WD)	Numerical Variables	More Sensitive to mean difference; easy to interpret
Kullback-Leibler (K-L) divergence	Numerical Variables; Categorical Variables	More sensitive to variance difference; binning strategy (for numerical features) could affect the results significantly
Jensen-Shannon (J-S) distance	Numerical Variables; Categorical Variables	More sensitive to variance difference; binning strategy (for numerical features) could affect the results significantly
Chi-squared Test	Categorical Variables	Especially sensitive to the changes in low-frequency categories; not recommended for categorical features with many low-frequency categories or high cardinality features

Drift metrics need to be tailored to the application. Each metric has its own scale and sensitivity to different types of change, e.g., WD is more sensitive to the change in mean while PSI is more sensitive to the change in variance. Understanding scale, sensitivity to change, and the nuances of each metric are important when choosing the appropriate metrics for the designated use case.

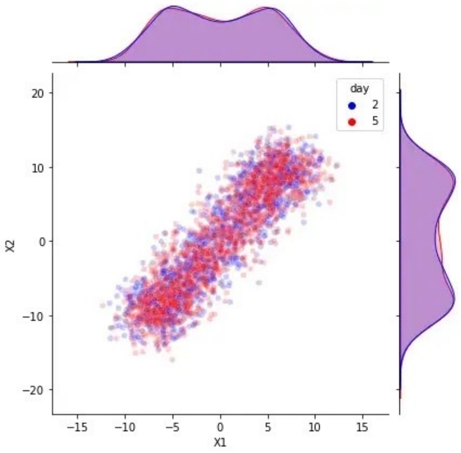
Univariate and Multivariate Drift Detection are complement

Trade-off: interpretability vs. compound effect

Univariate Drift Detection looks at each feature individually and check its distribution has changed compared to reference. While it is **simple to implement** and **easy to interpret**, the Univariate Drift Detection also has its shortcomings:

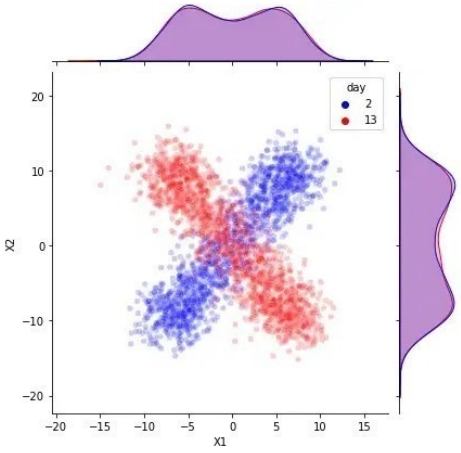
- It can be **redundant**. For example, if you have significantly correlated features, drift in all of them will be measured and counted multiple times in the overall metric.
- It **cannot** capture **multivariate drifts**. Drift can happen in such a way that while each feature by itself has the same distribution, the conditional distribution of 2 or more features together is drifting (see the graphs below)
- **Not** all drift metrics have the **same scale**. It can get complicated to average between different categorical and numerical features that use different drift metrics.

Multivariate Drift Detection can be used as a complementary analysis to address these shortcomings. It provides one aggregate metric reducing the risk of false alerts and detecting more subtle changes in the data structure that cannot be detected with univariate approaches.



Low Drift in both X1 and X2

Low Multivariate Drift



Low Drift in both X1 and X2

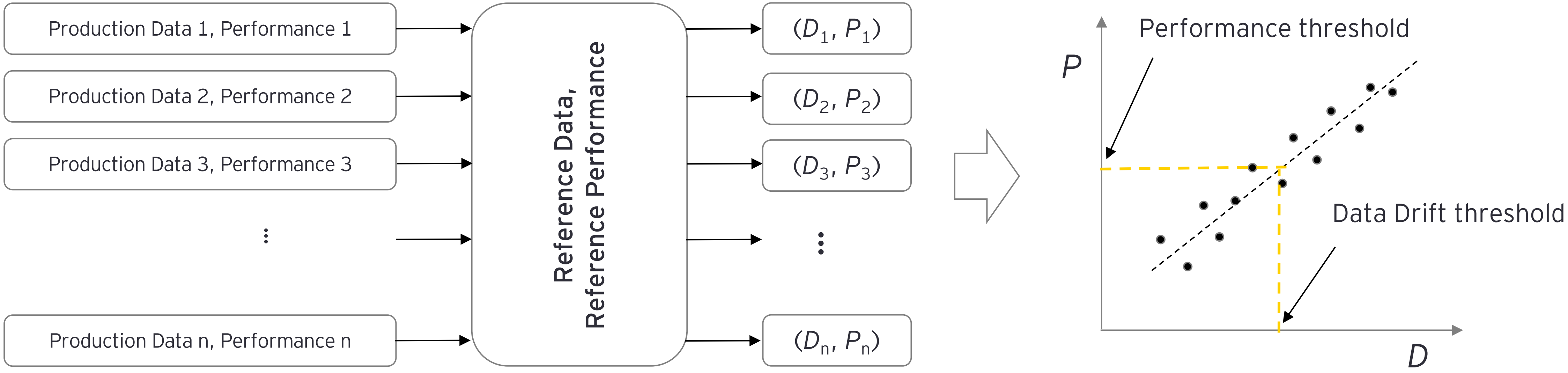
High Multivariate Drift

Bridging Data Drift with Model Performance

How Data Drift impacts Model Performance

While it is important to measure how significantly the data distribution has changed/drifted compared to reference using drift detection metrics, it is crucial to understand **the impact of such change/drift on the performance of the model.**

The relationship between the data drift and the model performance drift can be studied empirically as described below; calibration curves between performance drift and data drift metrics can be established. **Threshold(s)** on Data Drift Metrics can then be determined and tuned based on the **business expectations** on the model performance.



Production Data 1, Performance 1
Production Data 2, Performance 2
Production Data 3, Performance 3
⋮
Production Data n, Performance n
Reference Data, Reference Performance
(D_1, P_1)
(D_2, P_2)
(D_3, P_3)
⋮
(D_n, P_n)
Performance threshold
P
Data Drift threshold
D

$D_1, D_2, D_3, \ldots, D_n$ refers to the data drift metrics for each production period comparing to the reference;

$P_1, P_2, P_3, \ldots, P_n$ refers to the performance difference for each production period comparing to the reference

EY | Building a better working world

EY exists to build a better working world, helping to create long-term value for clients, people and society and build trust in the capital markets.

Enabled by data and technology, diverse EY teams in over 150 countries provide trust through assurance and help clients grow, transform and operate.

Working across assurance, consulting, law, strategy, tax and transactions, EY teams ask better questions to find new answers for the complex issues facing our world today.

EY refers to the global organization, and may refer to one or more, of the member firms of Ernst & Young Global Limited, each of which is a separate legal entity. Ernst & Young Global Limited, a UK company limited by guarantee, does not provide services to clients. Information about how EY collects and uses personal data and a description of the rights individuals have under data protection legislation are available via ey.com/privacy. EY member firms do not practice law where prohibited by local laws. For more information about our organization, please visit ey.com.

This publication contains information in summary form, current as of the date of publication, and is intended for general guidance only. It should not be regarded as comprehensive or a substitute for professional advice. Before taking any particular course of action, contact Ernst & Young or another professional advisor to discuss these matters in the context of your particular circumstances. We accept no responsibility for any loss or damage occasioned by your reliance on information contained in this publication.

ey.com/ca

Authors

Mario Schlener

Partner, Lead Financial Services Risk Management Practice and Enterprise Risk Strategy, EY Canada

EY Global FS Risk Technology, Alliance, Innovation Lead

mario.schlener@ca.ey.com

Jean-Francois Isabelle

Executive Director, Practice Lead, FinCrime Innovation, EY Canada

Jean-Francois.Isabelle@ca.ey.com

Yara Elias, Ph.D.

Senior Manager, AI Risk Lead, Financial Services Risk Management, EY Canada

yara.elias@ca.ey.com

Liang Hu, Ph.D.

Manager, Responsible AI and AI Risk , Financial Service Risk Management, EY Canada

liang.Hu@ca.ey.com

Vishaal Venkatesh

Senior, AI Risk, Financial Services Risk Management, EY Canada

vishaal.venkatesh@ca.ey.com

Pamina Lässing,

Senior Consultant, AI Risk, Financial Services Risk Management, EY Canada

Pamina.Laessing1@ca.ey.com