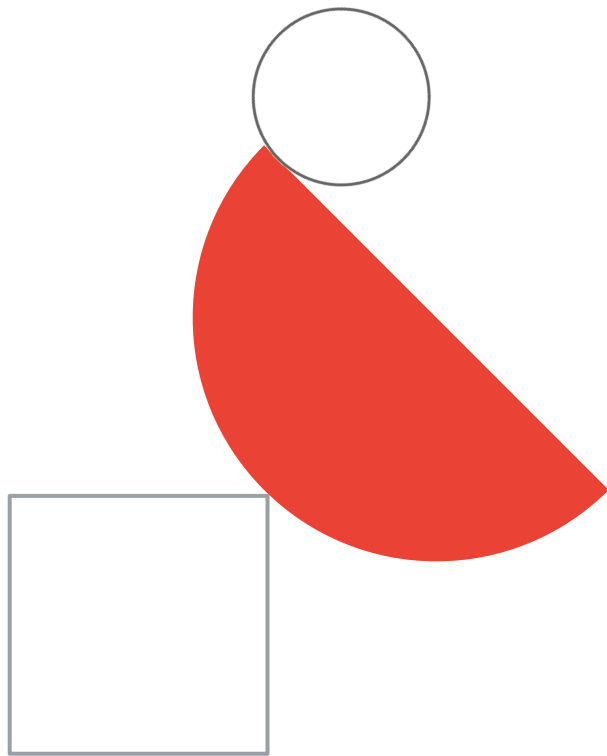
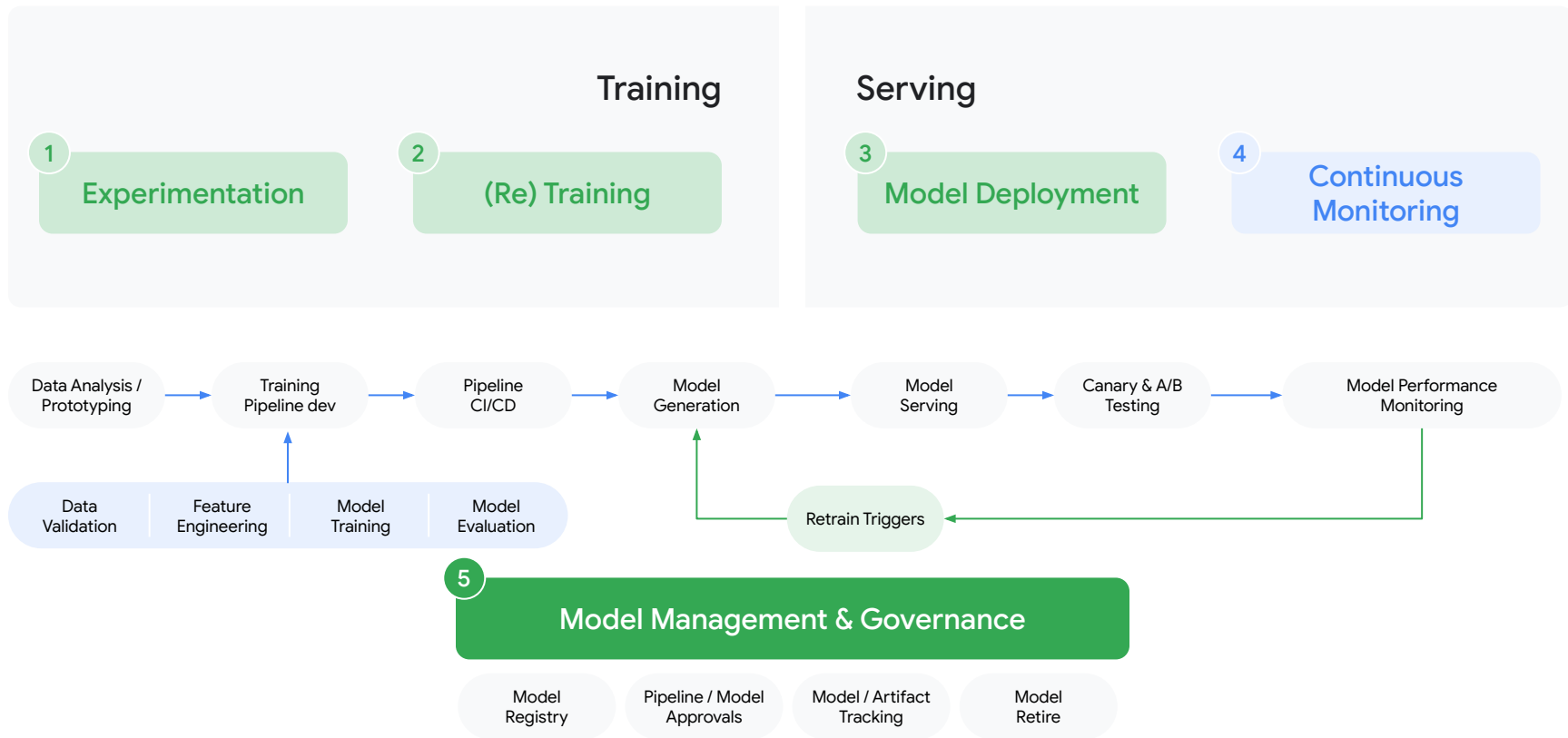


Model Monitoring with Vertex AI



Where does monitoring fit in?



Monitoring Challenges



While a model is static, the world around it changes

“The real world is dynamic and data changes fast. I need to know when live production data deviates from training data.”



Can't join predictions with actual outcomes

“I can not always collect the actual outcomes in a reasonable amount of time. I need other signals to monitor the model's performance.”



Bugs in model input generation

“Changes or bugs in upstream data or feature engineering pipelines can lead to incorrect model inputs. I need to catch these issues early.”



Diagnosing root cause is hard

“When model performance deviates, I need to quickly drill into what is causing it i.e. identify what changed, so I can fix it.”

Vertex AI

Applications

Vision and Video

Conversation

Language

Structured Data

Core

Notebooks

Data Labeling

Deep Learning Env

Experiments

Metadata

AutoML

Training

Explainable AI

Feature Store

Vizier (Optimization)

Prediction

Continuous Monitoring

Pipelines

AI Accelerators

Hybrid AI

Easy and proactive monitoring of model performance



Monitor and alert

Monitor signals for model's predictive performance, and alert when those signals deviate.



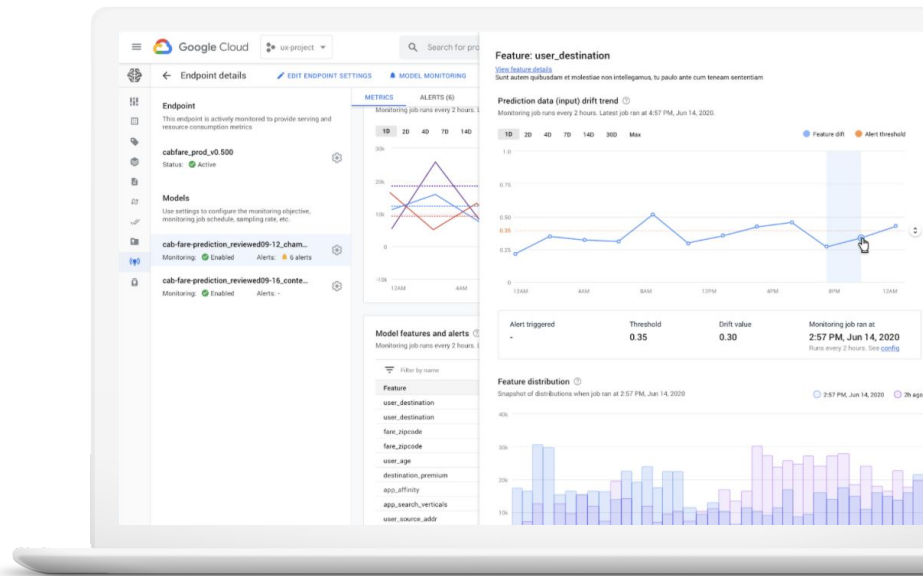
Diagnose

Help identify the cause for the deviation i.e. what changed, how and how much?



Update Model

Trigger model re-training pipeline or collect relevant training data to address performance degradation.



Calculate training-serving skew and prediction drift

Model Monitoring computes the statistical distribution of the latest feature values seen in production.

Calculate training-serving skew and prediction drift

Model Monitoring computes the statistical distribution of the latest feature values seen in production.

Baselines for skew and drift

Model Monitoring uses different baselines for skew detection and drift detection:

Calculate training-serving skew and prediction drift

Model Monitoring computes the statistical distribution of the latest feature values seen in production.

Baselines for skew and drift

Model Monitoring uses different baselines for skew detection and drift detection:

- For skew detection, the baseline is the statistical distribution of the feature's values in the training data.

Calculate training-serving skew and prediction drift

Model Monitoring computes the statistical distribution of the latest feature values seen in production.

Baselines for skew and drift

Model Monitoring uses different baselines for skew detection and drift detection:

- For skew detection, the baseline is the statistical distribution of the feature's values in the training data.
- For drift detection, the baseline is the statistical distribution of the feature's values seen in production in the recent past.

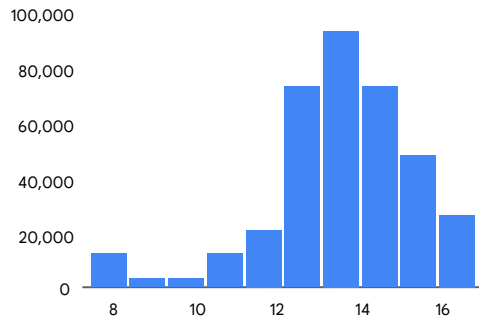
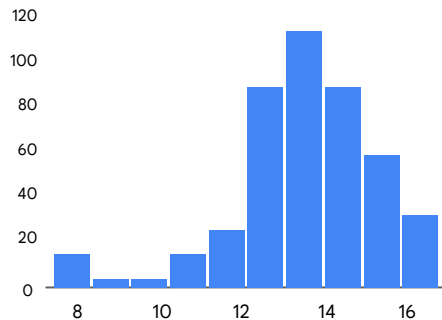
Example of feature monitoring

Feature distribution

Snapshot of distribution when job ran at Apr 29, 2021 10:00:00 AM

Latest prediction stats distribution

Hover over the chart to view stats



Training stats distribution

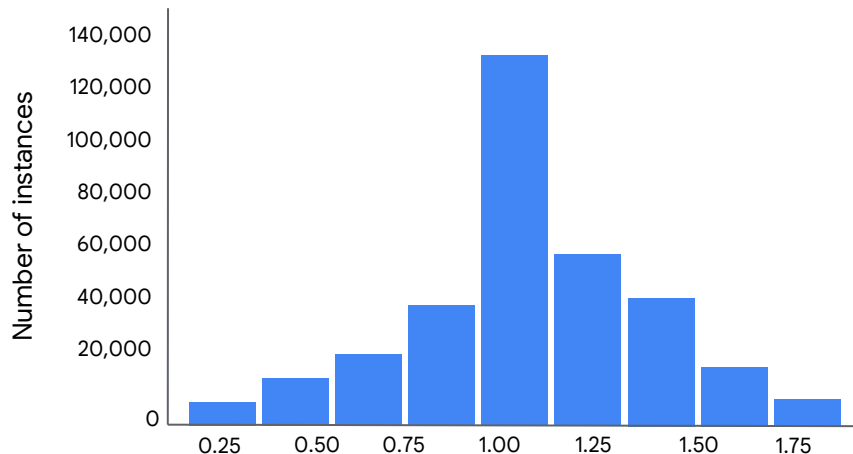
Monitoring jobs (up to last 50)

- Apr 29, 2021, 10:00:00 AM
- Apr 29, 2021, 9:00:00 AM
- Apr 29, 2021, 8:00:00 AM
- Apr 19, 2021, 9:00:00 PM
- Apr 19, 2021, 8:00:00 PM
- Apr 19, 2021, 7:00:00 PM
- Apr 19, 2021, 6:00:00 PM
- Apr 19, 2021, 5:00:00 PM
- Apr 19, 2021, 4:00:00 PM
- Apr 16, 2021, 10:00:00 PM
- Apr 16, 2021, 9:00:00 PM
- Apr 16, 2021, 8:00:00 PM
- Apr 16, 2021, 7:00:00 PM
- Apr 16, 2021, 6:00:00 PM
- Apr 16, 2021, 5:00:00 PM


Categorical and numerical features


For **categorical** features, the computed distribution is the number or percentage of instances of each possible value of the feature.


For **numerical** features, we divide the range of possible feature values into equal intervals, and compute the number or percentage of feature values that fall in each interval.





Create a job using the Cloud Console


 Vertex AI


 Features


 Labeling tasks


 Workbench


 Pipelines


 Training


 Experiments

 Models

 Endpoints

 Batch predictions

 Metadata

 Marketplace

Endpoints

+ CREATE ENDPOINT

REFRESH

Endpoints are machine learning models made available for online prediction requests. Endpoints are useful for timely predictions from many users (for example, in response to an application request). You can also request batch predictions if you don't need immediate results.

To create an endpoint, you need at least one machine learning model. [Learn more](#)

Region
us-central1 (Iowa)

Filter Enter a property name

<input type="checkbox"/>	Name	ID	Status	Models	Region	Monitoring	Most recent alerts	Last updated
<input type="checkbox"/>	my_usahousing_10.02.2021	563495311188688896	Active	1	us-central1	Disabled	—	Oct 2, 2021, 11:48:38 PM
<input type="checkbox"/>	credit_risk	1241146317619593216	Active	1	us-central1	Disabled	—	Aug 29, 2021, 12:18:02 AM
<input type="checkbox"/>	hello_endpoint	2976791392761675776	Active	1	us-central1	Disabled	—	Aug 6, 2021, 8:36:27 PM

Create a job using the Cloud Console

Vertex AI

- Features
- Labeling tasks
- Workbench
- Pipelines
- Training
- Experiments
- Models
- Endpoints**
- Batch predictions
- Metadata
- Marketplace

Endpoints [+ CREATE ENDPOINT](#) [REFRESH](#)

Endpoints are machine learning models made available for online prediction requests. Endpoints are useful for timely predictions from many users (for example, in response to an application request). You can also request batch predictions if you don't need immediate results.

To create an endpoint, you need at least one machine learning model. [Learn more](#)

Region: us-central1

Existing endpoint

Filter Enter a property name

<input type="checkbox"/>	Name	ID	Status	Models	Region	Monitoring	Most recent alerts	Last updated ↓
<input type="checkbox"/>	my_usahousing_10.02.2021	563495311188688896	✓ Active	1	us-central1	Disabled	—	Oct 2, 2021, 11:48:38 PM
<input type="checkbox"/>	credit_risk	1241146317619593216	✓ Active	1	us-central1	Disabled	—	Aug 29, 2021, 12:18:02 AM
<input type="checkbox"/>	hello_endpoint	2976791392761675776	✓ Active	1	us-central1	Disabled	—	Aug 6, 2021, 8:36:27 PM

Create a job using the Cloud Console

Vertex AI

- Features
- Labeling tasks
- Workbench
- Pipelines
- Training
- Experiments
- Models
- Endpoints**
- Batch predictions
- Metadata
- Marketplace

Endpoints [+ CREATE ENDPOINT](#) [REFRESH](#)

Endpoints are machine learning models made available for online prediction requests. Endpoints are useful for timely predictions from many users (for example, in response to an application request). You can also request batch predictions if you don't need immediate results.

To create an endpoint, you need at least one machine learning model. [Learn more](#)

Region: us-central1


Existing endpoint

Filter Enter a property name

<input type="checkbox"/>	Name	ID	Status	Models	Region	Monitoring	Most recent alerts	Last updated ↓
<input type="checkbox"/>	my_usahousing_10.02.2021	563495311188688896	✓ Active	1	us-central1	Disabled	—	Oct 2, 2021, 11:48:38 PM
<input type="checkbox"/>	credit_risk	1241146317619593216	✓ Active	1	us-central1	Disabled	—	Aug 29, 2021, 12:18:02 AM
<input type="checkbox"/>	hello_endpoint	2976791392761675776	✓ Active	1	us-central1	Disabled	—	Aug 6, 2021, 8:36:27 PM

Edit settings to enable monitoring

← credit_risk


 EDIT SETTINGS

<> SAMPLE REQUEST

Region


us-central1

Logs

[View Logs](#) 

Most recent monitoring job

—

<input type="checkbox"/>	Model	Status	Most recent alerts	Monitoring	Traffic split	Compute nodes	Type
<input type="checkbox"/>	credit_risk_202182831655	 Ready	—	Disabled	100%	Auto (1 minimum, 1 maximum)	Tabular

Toggle the switch to enable model monitoring

Edit endpoint

- 1 Define your endpoint
- 2 Model settings
- 3 Model monitoring

UPDATE

CANCEL



Settings in this step apply to **all models** deployed to the endpoint

Model monitoring

You can monitor the tabular and custom models deployed to this endpoint for changes in feature drift, training-serving skew and other objectives that help you understand how your model is performing to real world data.




Enable model monitoring for this endpoint

Toggle the switch to enable model monitoring

Edit endpoint

- 1 Define your endpoint
- 2 Model settings
- 3 Model monitoring

UPDATE CANCEL

 Settings in this step apply to **all models** deployed to the endpoint

Model monitoring

You can monitor the tabular and custom models deployed to this endpoint for changes in feature drift, training-serving skew and other objectives that help you understand how your model is performing to real world data.


☐ Enable model monitoring for this endpoint

Toggle the switch to enable model monitoring

Edit endpoint

- 1 Define your endpoint
- 2 Model settings
- 3 Model monitoring

UPDATE CANCEL

 Settings in this step apply to **all models** deployed to the endpoint

Model monitoring

You can monitor the tabular and custom models deployed to this endpoint for changes in feature drift, training-serving skew and other objectives that help you understand how your model is performing to real world data.

☐ Enable model monitoring for this endpoint

Monitoring job schedule

Model monitoring

You can monitor the tabular and custom models deployed to this endpoint for changes in feature drift, training-serving skew and other objectives that help you understand how your model is performing to real world data.

☒ Enable model monitoring for this endpoint

Monitoring job display name *

credit_risk_monitoring_gs



Define the display name of the monitoring job.

Monitoring job schedule

Monitoring window size *

24

Define the size of the time window to monitor when the monitoring job runs, in hours.

Monitoring window size defines the size of the time window to monitor when the monitoring job runs, in hours

Monitoring emails *

hello_world@xyz.com



Enter at least one valid email to receive email alerts.

Sampling rate

Sampling rate *

10



Define a percentage of the prediction input data that should be sampled when the monitoring job runs.

Input schemas

Optional

Monitoring job schedule

Model monitoring

You can monitor the tabular and custom models deployed to this endpoint for changes in feature drift, training-serving skew and other objectives that help you understand how your model is performing to real world data.

☒ Enable model monitoring for this endpoint

Monitoring job display name *

credit_risk_monitoring_gs



Define the display name of the monitoring job.

Monitoring job schedule

Monitoring window size *

24

Define the size of the time window to monitor when the monitoring job runs, in hours.

Monitoring emails *

hello_world@xyz.com



Enter at least one valid email to receive email alerts.

Monitoring email sends alert notifications to this email address

Sampling rate

Sampling rate *

10

%

Define a percentage of the prediction input data that should be sampled when the monitoring job runs.

Input schemas

Optional

Monitoring job schedule

- The time at which the monitoring job ran

Model monitoring

You can monitor the tabular and custom models deployed to this endpoint for changes in feature drift, training-serving skew and other objectives that help you understand how your model is performing to real world data.

☒ Enable model monitoring for this endpoint

Monitoring job display name *

credit_risk_monitoring_gs



Define the display name of the monitoring job.

Monitoring job schedule

Monitoring window size *

24

Define the size of the time window to monitor when the monitoring job runs, in hours.

Monitoring emails *

hello_world@xyz.com



Enter at least one valid email to receive email alerts.

Monitoring email sends alert notifications to this email address

Sampling rate

Sampling rate *

10

%

Define a percentage of the prediction input data that should be sampled when the monitoring job runs.

Input schemas

Optional

Monitoring job schedule

- The time at which the monitoring job ran
- The name of the feature that has skew or drift

Model monitoring

You can monitor the tabular and custom models deployed to this endpoint for changes in feature drift, training-serving skew and other objectives that help you understand how your model is performing to real world data.

☒ Enable model monitoring for this endpoint

Monitoring job display name *

credit_risk_monitoring_gs



Define the display name of the monitoring job.

Monitoring job schedule

Monitoring window size *

24

Define the size of the time window to monitor when the monitoring job runs, in hours.

Monitoring emails *

hello_world@xyz.com



Enter at least one valid email to receive email alerts.

Monitoring email sends alert notifications to this email address

Sampling rate

Sampling rate *

10

%

Define a percentage of the prediction input data that should be sampled when the monitoring job runs.

Input schemas

Optional

Monitoring job schedule

- The time at which the monitoring job ran
- The name of the feature that has skew or drift
- The alerting threshold as well as the recorded statistical distance measure

Model monitoring

You can monitor the tabular and custom models deployed to this endpoint for changes in feature drift, training-serving skew and other objectives that help you understand how your model is performing to real world data.

☒ Enable model monitoring for this endpoint

Monitoring job display name *

credit_risk_monitoring_gs



Define the display name of the monitoring job.

Monitoring job schedule

Monitoring window size *

24

Define the size of the time window to monitor when the monitoring job runs, in hours.

Monitoring emails *

hello_world@xyz.com



Enter at least one valid email to receive email alerts.

Monitoring email sends alert notifications to this email address

Sampling rate

Sampling rate *

10

%

Define a percentage of the prediction input data that should be sampled when the monitoring job runs.

Input schemas

Optional

Prediction request sampling rate

Model monitoring

You can monitor the tabular and custom models deployed to this endpoint for changes in feature drift, training-serving skew and other objectives that help you understand how your model is performing to real world data.

☒ Enable model monitoring for this endpoint

Monitoring job display name *

credit_risk_monitoring_gs



Define the display name of the monitoring job.

Monitoring job schedule

Monitoring window size *

24

Define the size of the time window to monitor when the monitoring job runs, in hours.

Monitoring emails *

hello_world@xyz.com



Enter at least one valid email to receive email alerts.

Sampling rate

Sampling rate *

10

Define a percentage of the prediction input data that should be sampled when the monitoring job runs.

Sampling rate defines a percentage of the prediction input data that should be sampled when the monitoring job runs

Input schemas

Optional

Monitoring objective

- ☒ **Training Prediction Skew Detection**
Skew is calculated between feature distributions of prediction input data and training data.
- ☐ **Prediction Drift Detection**
Uses prediction input data. Drift is calculated between feature distributions every time the monitoring job runs.

Training Prediction Skew Detection

Training data source

- ☒ Cloud Storage
- ☐ BigQuery table
- ☐ Vertex Dataset

 Training data location *

BROWSE

Supported file formats are: .csv, .tfrecord

Target Field

The name of the field from the training data the model is to predict, i.e. not in prediction input.

Target field *

Alert thresholds

Optional

Specify an alert threshold value for each feature that will be used to trigger alerts.

Monitoring objective

- ☒ **Training Prediction Skew Detection**
Skew is calculated between feature distributions of prediction input data and training data.
- ☐ **Prediction Drift Detection**
Uses prediction input data. Drift is calculated between feature distributions every time the monitoring job runs.

Training Prediction Skew Detection

Training data source

- ☒ Cloud Storage
- ☐ BigQuery table
- ☐ Vertex Dataset



Training data location *

BROWSE

Supported file formats are: .csv, .tfrecord

Target Field

The name of the field from the training data the model is to predict, i.e. not in prediction input.

Target field *

Alert thresholds

Optional

Specify an alert threshold value for each feature that will be used to trigger alerts.

Lab

`notebooks/model_monitoring/labs/model_monitoring_vertex.ipynb`

