SageMaker

Curated FAQs

Amazon SageMaker Model Hosting FAQ

Q: What is a model server?

A: SageMaker endpoints are HTTP REST endpoints that use a containerized web server, which includes a model server. These containers are responsible for loading up and serving requests for a machine learning model. They implement a web server that responds to /invocations and /ping on port 8080.

Common model servers include TensorFlow Serving, TorchServe and Multi Model Server. SageMaker framework containers have these model servers built in.

Q: What is Bring Your Own Container with Amazon SageMaker?

A: Everything in SageMaker Inference is containerized. SageMaker provides managed containers for popular frameworks such as TensorFlow, SKlearn, and HuggingFace.with the SageMaker Model Registry.

Q: Do I need to train my models on SageMaker to host them on SageMaker endpoints?

A: No, SageMaker offers the capacity to bring your own trained framework model that you've trained outside of SageMaker and deploy it on any of the SageMaker hosting options.

Q: How should I structure my model if I want to deploy on SageMaker but not train on SageMaker?

A: SageMaker requires your model artifacts to be compressed in a .tar.gz file, or a *tarball*. SageMaker automatically extracts this .tar.gz file into the /opt/ml/model/ directory in your container.

Amazon SageMaker MLOps FAQs

Q. **Do I Need to Create a Project to Use SageMaker Pipelines?**

No. SageMaker pipelines are standalone entities just like training jobs, processing jobs, and other SageMaker jobs. You can create, update, and run pipelines directly within a notebook by using the SageMaker Python SDK without using a SageMaker project.

Projects provide an additional layer to help you organize your code and adopt operational best practices that you need for a production-quality system.

Q. Can I use SageMaker Experiments with SageMaker Pipelines?

Yes. SageMaker Pipelines is natively integrated with SageMaker Experiments..

Q. SageMaker Project templates have a model deploy repository that uses CloudFormation (CFN) to create an endpoint. Are there ways to deploy the model without using CloudFormation?

You can customize the deploy repository in the project template to deploy the model from the model registry any way you like. The template uses CloudFormation to create a real-time endpoint, as an example. You can update the deployment to use the SageMaker SDK, boto3, or any other API that can create endpoints instead of CFN.

Q. What's the recommended way to manage dependencies for different SageMaker Pipelines steps?

You can use a SageMaker Projects template to implement image-building CI/CD. With this template, you can automate the CI/CD of images that are built and pushed to Amazon ECR. Changes in the container files in your project's source control repositories initiate the ML pipeline and deploy the latest version for your container.

Q. What's the best way to reproduce my model in SageMaker?

SageMaker's Lineage Tracking service works in the backend to track all the metadata associated with your model training and deployment workflows. This includes your training jobs, datasets used, pipelines, endpoints, and the actual models. You can query the lineage service at any point to find the exact artifacts used to train a model. Using those artifacts, you can recreate the same ML workflow to reproduce the model as long as you have access to the exact dataset that was used. A trial component tracks the training job. This trial component has all the parameters used as part of the training job. If you don't need to rerun the entire workflow, you can reproduce the training job to derive the same model.

Amazon SageMaker Model Registry FAQ

Q. How should I organize my models into Model Groups and model packages in the SageMaker Model Registry?

A model package is the actual model that is registered into the Model Registry as a versioned entity. Please note there are two ways you can use model packages in SageMaker

a. One is withSageMaker Marketplace — these model packages are not versioned.
b. The other is with the SageMaker Model Registry, in which the model package *must* be versioned.

The Model Registry receives every new model that you retrain, gives it a version, and assigns it to a Model Group inside the Model Registry.

Q. How does the SageMaker Model Registry differ from Amazon Elastic Container Registry (Amazon ECR)?

The SageMaker Model Registry is a metadata store for your machine learning models. Amazon Elastic Container Registry is a repository that stores all of your containers. Within the Model Registry, models are versioned and registered as model packages within Model Groups. Each model package contains an Amazon S3 URI to the model files associated with the trained model and an Amazon ECR URI that points to the container used while serving the model.

Q. How do I tag model packages in the SageMaker Model Registry?

Model packages in the SageMaker Model Registry do not support tags—these are versioned model packages. Instead, you can add key value pairs using CustomerMetadataProperties. Model package groups in the model registry support tagging.

Evaluate, explain, and detect bias in models

Q: What happens in the background when Sagemaker Model monitor is enabled?

Amazon SageMaker Model Monitor automates model monitoring alleviating the need to monitor the models manually or building any additional tooling. In order to automate the process, Model Monitor provides you with the ability to create a set of baseline statistics and constraints using the data with which your model was trained, then set up a schedule to monitor the predictions made on your endpoint. Model Monitor uses rules to detect drift in your models and alerts you when it happens. The following steps describe what happens when you enable model monitoring:

a. **Enable model monitoring**: For a real-time endpoint, you have to enable the endpoint to capture data from incoming requests to a deployed ML model and the resulting model predictions. For a batch transform job, enable data capture of the batch transform inputs and outputs.
b. **Baseline processing job**: You then create a baseline from the dataset that was used to train the model. The baseline computes metrics and suggests constraints for the metrics. For example, the recall score for the model shouldn't regress and drop below 0.571, or the precision score shouldn't fall below 1.0. Real-time or batch predictions from your model are compared to the constraints and are reported as violations if they are outside the constrained values.
c. **Monitoring job**: Then, you create a monitoring schedule specifying what data to collect, how often to collect it, how to analyze it, and which reports to produce.
d. **Merge job**: This is only applicable if you are leveraging Amazon SageMaker Ground Truth. Model Monitor compares the predictions your model makes with Ground Truth labels to measure the quality of the model. For this to work, you periodically label data captured by your endpoint or batch transform job and upload it to Amazon S3.

   After you create and upload the Ground Truth labels, include the location of the labels as a parameter when you create the monitoring job.

When you use Model Monitor to monitor a batch transform job instead of a real-time endpoint, instead of receiving requests to an endpoint and tracking the predictions, Model Monitor monitors inference inputs and outputs. In a Model Monitor schedule, the customer provides the count and type of instances that are to be used in the processing job. These resources remain reserved until the schedule is deleted irrespective of the status of current execution.

Q: What is Data Capture, why is it required, and how can I enable it?

In order to log the inputs to the model endpoint and the inference outputs from the deployed model to Amazon S3, you can enable a feature called Data Capture.

Q: Does enabling Data Capture impact the performance of a real-time endpoint ?

Data Capture happens asynchronously without impacting production traffic. After you have enabled the data capture, then the request and response payload, along with some additional meta data, is saved in the Amazon S3 location that you specified in the DataCaptureConfig. Note that there can be a delay in the propagation of the captured data to Amazon S3.

You can also view the captured data by listing the data capture files stored in Amazon S3.

Q: Why is Ground Truth needed for model monitoring?

Ground Truth labels are required by the following features of Model Monitor:

- **Model quality monitoring** compares the predictions your model makes with Ground Truth labels to measure the quality of the model.
- **Model bias monitoring** monitors predictions for bias. One way bias can be introduced in deployed ML models is when the data used in training differs from the data used to generate predictions. This is especially pronounced if the data used for training changes over time (such as fluctuating mortgage rates), and the model prediction is not as accurate unless the model is retrained with updated data. For example, a model for predicting home prices can be biased if the mortgage rates used to train the model differ from the most current real-world mortgage rate.

Q: For customers leveraging Ground Truth for labeling, what are the steps I can take to monitor the quality of the model?

Model quality monitoring compares the predictions your model makes with Ground Truth labels to measure the quality of the model. For this to work, you periodically label data captured by your endpoint or batch transform job and upload it to Amazon S3. Besides captures, model bias monitoring execution also requires Ground Truth data. In real use cases, Ground Truth data should be regularly collected and uploaded to the designated Amazon S3 location. To match Ground Truth labels with captured prediction data, there must be a unique identifier for each record in the dataset.

Model Card - FAQs

Q. Can I customize a model card?

Amazon SageMaker Model Cards have a defined structure to them that cannot be modified. While you cannot change the structure of the model card, there is some flexibility introduced through custom properties in the **Additional information** section of the model card.

Q. Can I edit a model card once it is created?

Model cards have versions associated with them. A given model version is immutable across all attributes other than the model card status.

Q. Can I create model cards for models that were not trained using SageMaker?

A: Yes. You can create model cards for models not trained in SageMaker, but no information is automatically populated in the card. You must supply all the information in the card.

Q. Can I export or share model cards?

A: Yes. You can export each version of a model card to a PDF, downloaded, and share it.

Q. Do I need to register my model in the Model Registry to use model cards?

A: No. You can use model cards independently of the Model Registry.

Model Dashboard FAQ

Refer to the following FAQ topics for answers to commonly asked questions about Amazon SageMaker Model Dashboard.

Q. What is Model Dashboard?

Amazon SageMaker Model Dashboard is a centralized repository of all models created in your account. The models are generally the outputs of SageMaker training jobs, but you can also import models trained elsewhere and host them on SageMaker. Model Dashboard provides a single interface for IT administrators, model risk managers, and business leaders to track all deployed models and aggregates data from multiple AWS services to provide indicators about how your models are performing. You can view details about model endpoints, batch transform jobs, and monitoring jobs for additional insights into model performance. The dashboard's visual display helps you quickly identify which models have missing or inactive monitors so you can ensure all models are periodically checked for data drift, model drift, bias drift, and feature attribution drift. Lastly, the dashboard's ready access to model details helps you dive deep so you can access logs, infrastructure-related information, and resources to help you debug monitoring failures.

Q. What are the prerequisites to use Model Dashboard?

You should have one or more models created in SageMaker, either trained on SageMaker or externally trained. While this is not a mandatory prerequisite, you gain the most value from the dashboard if you set up model monitoring jobs via Amazon SageMaker Model Monitor for models deployed to endpoints.

Q. Who should use Model Dashboard?

Model risk managers, ML practitioners, data scientists and business leaders can get a comprehensive overview of models using the Model Dashboard. The dashboard aggregates and displays data from Amazon SageMaker Model Cards, Endpoints and Model Monitor services to

display valuable information such as model metadata from the model card and model registry, endpoints where the models are deployed, and insights from model monitoring.

Q. How do I use Model Dashboard?

Model Dashboard is available out of the box with Amazon SageMaker and does not require any prior configuration. However, if you have set up model monitoring jobs using SageMaker Model Monitor and Clarify, you use Amazon CloudWatch to configure alerts that raise a flag in the dashboard when model performance deviates from an acceptable range. You can create and add new model cards to the dashboard, and view all the monitoring results associated with endpoints. Model Dashboard currently does not support cross-account models.

Q. What is Amazon SageMaker Model Monitor?

With Amazon SageMaker Model Monitor, you can select the data you want to monitor and analyze without writing any code. SageMaker Model Monitor lets you select data, such as prediction output, from a menu of options and captures metadata such as timestamp, model name, and endpoint so you can analyze model predictions. You can specify the sampling rate of data capture as a percentage of overall traffic in the case of high volume real-time predictions. This data is stored in your own Amazon S3 bucket. You can also encrypt this data, configure fine-grained security, define data retention policies, and implement access control mechanisms for secure access.

Q. What types of model monitors does SageMaker support?

- SageMaker Model Monitor provides the following types of model monitors:
- *Data Quality*: Monitor drift in data quality.
- *Model Quality*: Monitor drift in model quality metrics, such as accuracy.
- *Bias Drift for Models in Production*: Monitor bias in your model's predictions by comparing the distribution of training and live data.
- *Feature Attribution Drift for Models in Production*: Monitor drift in feature attribution by comparing the relative rankings of features in training and live data.

Q. What inference methods does SageMaker Model Monitor support?

Model Monitor currently supports endpoints that host a single model for real-time inference and does not support monitoring of multi-model endpoints.

Q. How does Model Monitor work?

Amazon SageMaker Model Monitor automatically monitors machine learning models in production, using rules to detect drift in your model. Model Monitor notifies you when quality issues arise through alerts. To learn more, see How Amazon SageMaker Model Monitor works.

Q. When and how do you bring your own container (BYOC) for Model Monitor?

Model Monitor computes model metrics and statistics on tabular data only. For use cases other than tabular datasets, such as images or text, you can bring your own containers (BYOC) to monitor your data and models. For example, you can use BYOC to monitor an image classification model that takes images as input and outputs a label.

Q. Are there any performance concerns using DataCapture?

When turned on, data capture occurs asynchronously on the SageMaker endpoints. To prevent impact to inference requests, DataCapture stops capturing requests at high levels of disk usage. It is recommended you keep your disk utilization below 75% to ensure DataCapture continues capturing requests.