Red Hat

Operationalize your AI models: 4 key elements

With the race to adopt artificial intelligence (AI) into a business, some organizations are under the misconception that simply deploying an AI model is the finish line. In reality, operationalizing an AI model through machine learning operations (MLOps) is a continuous process. The following 4 elements are key components of operationalizing your model after deployment, ideally through an AI development platform.

1 Monitoring

Effective monitoring helps make sure that an AI model continues to perform as expected in changing circumstances. It is crucial to identify and adapt to any deviations in data or performance early to maintain reliability and trust.

The following are commonly tracked metrics in operationalized AI models:

- **Accuracy and performance.** The portion of correct predictions the model makes.
- **Data quality.** This includes testing for data completeness, consistency, distribution, and more.
- **Resource utilization.** Monitoring for CPU/GPU and memory utilization by the model.
- **Latency and throughput.** The time it takes for the model to generate predictions.
- **Model drift.** The amount the model has drifted from baseline predictions, and how that compares to model accuracy.

After the metrics to be gathered are selected, an organization can then build a framework to effectively react to adverse values. The following components are a key part of creating that framework:

- **Determine metrics.** Identify the key metrics to be tracked.
- **Collection and storage.** Defined processes need to be put in place to capture data and store it.
- **Real-time alerts.** A system needs to be set up to alert relevant parties or trigger action, such as pipeline retraining, if monitored metrics go outside of norms.

2 Maintenance

Implementing proactive maintenance on AI models is a core part of operationalizing them. Maintenance involves updating models, fixing bugs, and adapting to changing environments.

The maintenance process includes setting up schedules for the following:

- **Perform model updates.** Set a timeline for regular tests of the AI models to perform retraining as needed.
- **Prioritize fixing bugs through a structured process.** Set a timeline to identify and correct errors in the model's data set.
- **Plan lifecycle management.** Create repeatable data science pipelines for model training and validation and integrate them with DevOps pipelines to deliver models across your enterprise.
- **Use automation.** Automated deployment processes allow for more consistent models to reduce the risk of errors and inconsistencies while also reducing time to market. It is important that you choose an automation platform that works well with your AI development platform.
- **Adopt strict version control.** When models are being constantly trained and retrained, the changes in data and models must be strictly controlled. Pulling from the wrong pool of data or using the wrong model can compromise the entire application.

3 Model retraining

Retraining is a process of refitting the model to proactively work to increase accuracy.

The MLOps process often involves the automatic training of a model on a schedule or in reaction to a trigger-driven event.

Retraining an AI model aims to make it consistently output the most accurate results tailored to your specific business needs, and works to reduce or reverse model drift.

This process involves data scientists conducting detailed analyses to establish metrics for performance and degradation, informing necessary adjustments and using tests such as A/B testing. The model monitoring systems doing real-time analysis and alerting can automatically initiate retraining.

To create a stable retraining process, choose a platform that allows for the creation of repeatable data science pipelines for retraining that can also be integrated with DevOps pipelines across your organization.

4 Governance

Governance is the process of establishing ethical guidelines and security practices, and adhering to regulatory requirements. Governance is not optional.

It is important to make sure that models fit in the ethical, security, and regulatory boundaries that you set for them when they are created, but also that repeated deployment and retraining of the models does not cause them to drift outside of those guardrails.

Core tenets of governance include:

Data governance, which focuses on the origin of the data being used, how it is collected, how accurate and up to date it is, and if the data exposes any personal or private information.

Process governance, which concerns itself with formalizing the MLOps process. It is often most important in industries such as finance with heavy regulatory burdens.

Key to the process of adhering to governance is:

- The use of regular and comprehensive model documentation and reporting with version control.
- Scheduled or automated auditing of AI and ML systems.
- A comprehensive system for documenting data.
- Strict management of AI and ML metadata.
- A process for validating AI models.

Setting up an MLOps process to operationalize your AI models requires an AI development platform designed around MLOps workflows. Red Hat® Consulting can help your organization navigate what fits your business needs.

Read more

Learn how Red Hat's engagement with the MLOps Foundation informs solution design and book a free discovery session.

Learn how

Discover how to develop and deploy AI/ML applications on Red Hat OpenShift® AI with our robust training course. Learn on site or virtually depending on your needs.