Use Gartner's MLOps Framework to Operationalize Machine Learning Projects

Published 25 August 2022 - ID G00774981 - 23 min read

By Analyst(s): Farhan Choudhary, Shubhangi Vashisth, Erick Brethenoux

Initiatives: Artificial Intelligence; Evolve Technology and Process Capabilities to Support D&A

> Organizations struggle to integrate ML solutions with existing production applications, where operationalization mostly comes as an afterthought to model development. Data and analytics leaders can greatly reduce the risk of such failures with three stages that create a framework for MLOps.

Additional Perspectives

- Summary Translation: Use Gartner's MLOps Framework to Operationalize Machine Learning Projects
  (23 September2022)

Overview

Key Findings

- As per the 2021 Gartner AI in Organizations Survey, over 70% of organizations define artificial intelligence (AI) success measures before implementation. The survey also indicates that organizations struggle with measuring the business value of AI projects and lack the understanding of AI benefits and uses.

- Most machine learning (ML) and data science (DS) projects fail because operationalization is not a primary consideration of data scientists. Failure to promote DevOps practices leads to significant rework when promoting to production.

- ML development does not stop once a model is live. The many modes of failure mean it is critical to ensure maintainable and scalable model governance practices.

Recommendations

To achieve long-term DS and machine learning project success, data and analytics leaders responsible for AI strategy should:

- Establish a systematic machine learning operationalization (MLOps) process through Gartner's MLOps framework.
- Review and revalidate ML model operational performance by ensuring that deployed models meet the goals of integrity (technical and economic), transparency and sustainability.
- Minimize the technical debt and complex maintenance procedures by adopting complete DevOps practices at a person and process level, not just technical.

Strategic Planning Assumptions

By 2024, use of synthetic data and transfer learning will halve the volume of real data needed for machine learning.

By 2025, 80% of the largest global organizations will have participated at least once in federated ML to create more accurate, secure and environmentally sustainable models.

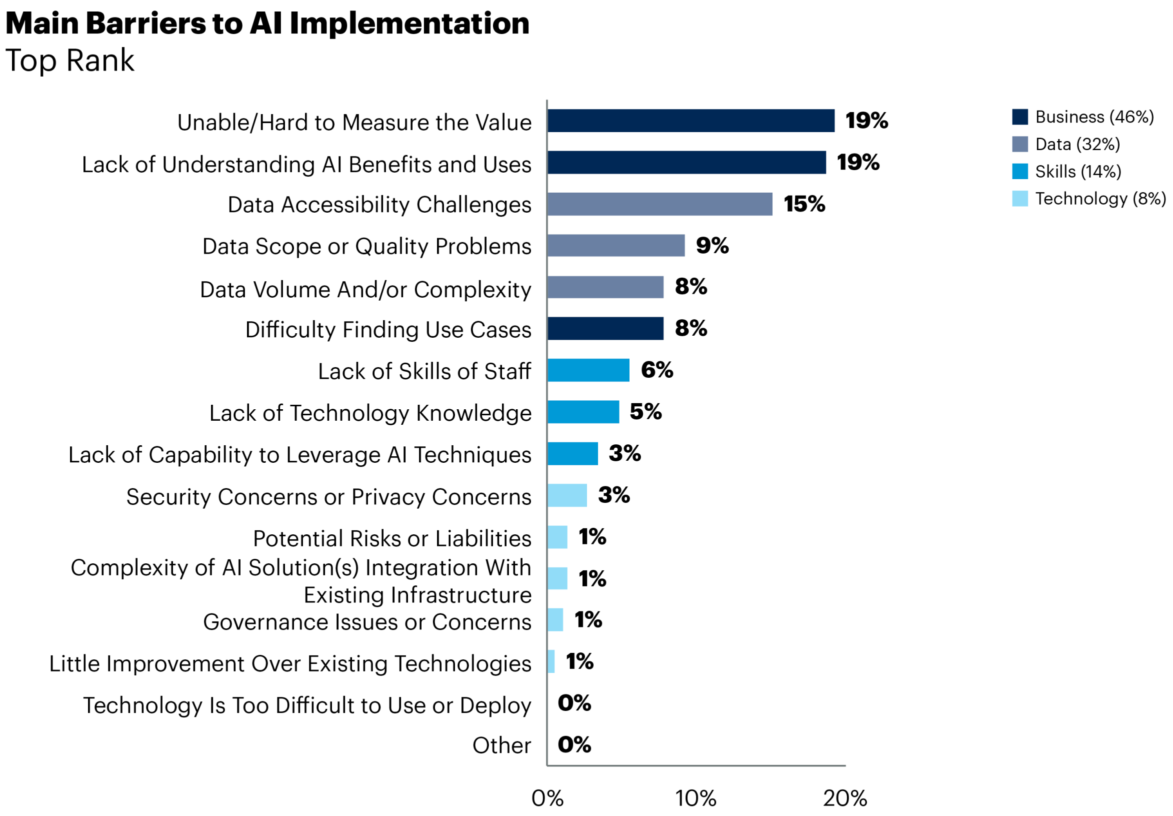
By 2024, 70% of organizations relying solely on ML for AI initiatives will spend more money per model than those leveraging composite AI techniques.

Introduction

This research is the first part of a two-part series on MLOps. For a comprehensive overview of organizational best practices accompanying MLOps, please refer to the companion piece, Use 3 MLOps Organizational Practices to Successfully Deliver Machine Learning Results.

According to the 2021 Gartner AI in Organizations Survey, the main barriers preventing implementation of AI are difficulty measuring value and lack of understanding of AI benefits and uses (see Figure 1).

Figure 1: Main Barriers to AI Implementation



n = 698; excluding 'not sure'

Q. What are or will be the top 3 barriers to the implementation of AI techniques within your organization?

Source: 2021 Gartner AI in Organizations Survey

Note: Percenatges have been rounded off

774981_C

In the last decade, organizations have often learned that publishing a model is not enough. That "publish moment" is followed by many steps that make the analytical model operationalization cycle as important as the analytical model development cycle.

The focus of data science teams has traditionally been on developing analytical assets (see Note 1), while dealing with the operationalization of these assets has been an afterthought.

Many reasons underlie that afterthought:

- A lack of a formal operationalization methodology

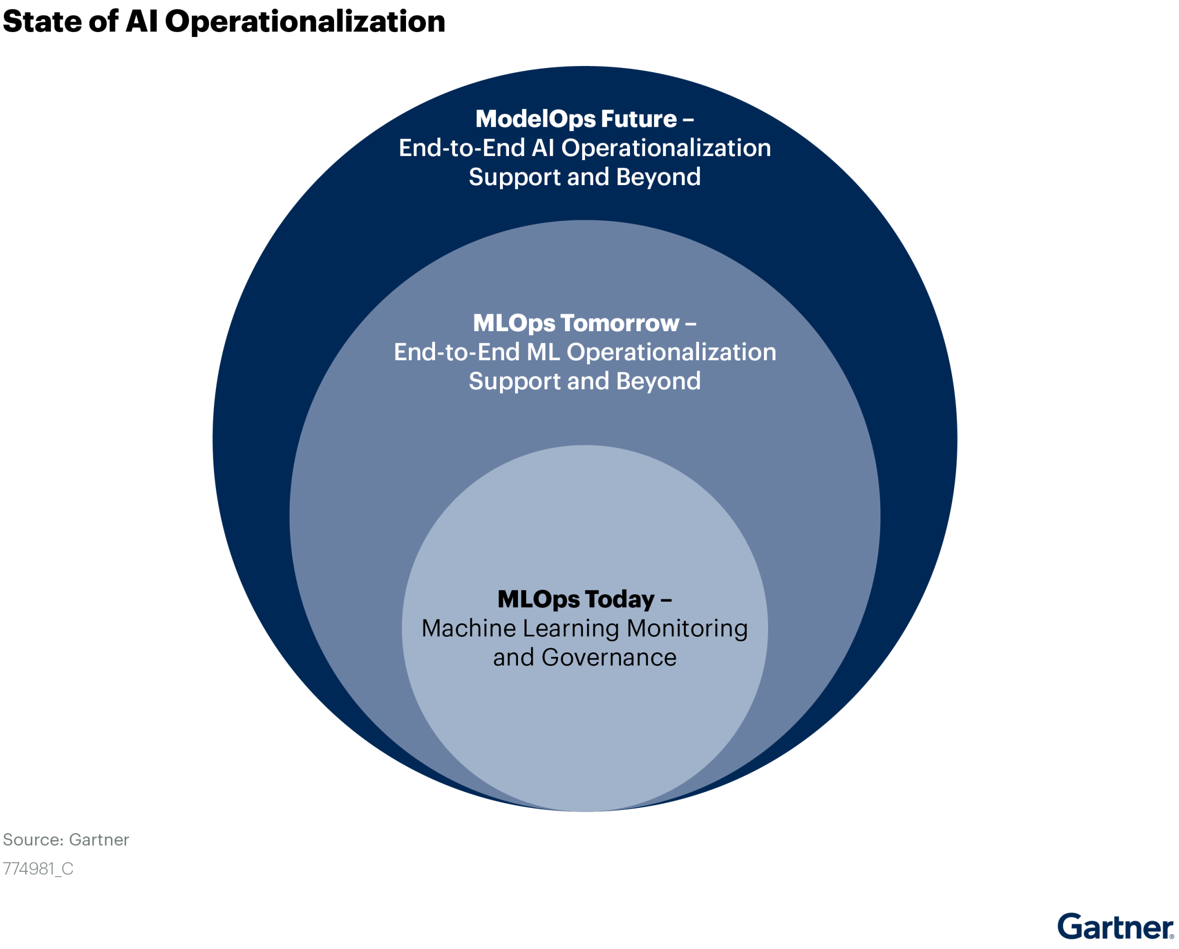
- A lack of data science teams' understanding, interest and collaboration in the engineering discipline of production and the context of the application that the ML model is being put into
- A deficit of formal communication channels among the data science team, the IT operational team and the line-of-business stakeholders
- A lack of joint testing and validations with different stakeholders, including business, software engineering, DevOps and I&O
- An unwillingness to deal with the operational aspect of advanced analytics, along with a misunderstanding of the practical business impact of operationalization

Machine learning operationalization is a critical step in aligning analytics investments with strategic business objectives — the "last mile" toward business value. Establishing best practices beyond the analytical asset development process starts with understanding the organization's business priorities and then systematically applying a model life cycle management (MLM) discipline. Parts of the MLOps framework can be achieved via an MLOps platform, creating a custom-made MLOps framework within the organization, extending DevOps functionality to support the MLOps function or a combination of these approaches. In extension to MLOps, ModelOps (see Note 2) refers to the operationalization of all AI models and includes MLOps that deals with the operationalization of ML models. This research specifically refers to the end-to-end life cycle of ML models, which we define as MLOps below.

MLOps (machine learning operationalization): MLOps enables the operationalization of the end-to-end machine learning model life cycle that supports the continuous delivery and continuous integration of ML models in a production environment. Core capabilities include feature curation, feature management (store), model governance (in some cases through ModelOps), model release, activation, monitoring, performance tracking, management, logging, reuse and maintenance.

However, from an AI perspective (i.e., dealing with the wide variety of AI models beyond ML), consider a ModelOps approach (see Figure 2).

Figure 2. State of AI Operationalization


State of AI Operationalization
ModelOps Future – End-to-End AI Operationalization Support and Beyond
MLOps Tomorrow – End-to-End ML Operationalization Support and Beyond
MLOps Today – Machine Learning Monitoring and Governance
Gartner

Source: Gartner
774981_C

The current analytical open-source movement is producing a wide range of analytical assets that will eventually have to be consumed — but current open-source deployment techniques provide “dissemination means,” not “managed production means.” From that perspective, techniques such as containerization or serverless deployment are efficient and effective in pushing (deploying) models into production, but they do not relieve the organization from adopting a formal operationalization process.

Analysis

Establish a Systematic Operationalization Process

The model development cycle shown in Figure 2 is a variation on the cross-industry standard process for data mining (CRISP-DM) methodology that has been used for two decades. That methodology, simple and powerful, imposes a development discipline that promotes the integrity and robustness of the resulting analytical models.

D&A leaders must have a strong motivation and a use case for commissioning the model life cycle management. They must identify both technical and business dependencies, as well as create a baseline governance framework that can be improved upon iteratively during the development stages. D&A leaders must be mindful that model management and governance is not an exercise that needs to be performed at the end of the MLM but a continuous exercise that ensures the reliability of ML workflows.

There are 3 phases you must consider:

1. Model development and validation
2. Model release or experimentation stage for preproduction checks
3. Model activation stage to operate models in real business conditions and
    1. Review and revalidate ML model operational performance
    2. Review and revalidate ML model business performance

These are discussed and illustrated below.

1. Validation Phase of Model Development Cycle Process

As per the 2021 AI in Organization Survey, organizations face significant challenges with data scope, quality and accessibility challenges that stall AI projects. Since data is foundational to ML pipelines, data can also be boosted for better performance of ML models. Acquiring quality datasets, synthetic data (see Innovation Insight for Synthetic Data), leveraging data labeling and annotation, using external data or multistructured data for a higher degree of contextual awareness assists in better ML performance (see Three Steps to Boost Data for AI and How to Improve the Performance of AI Projects). This leads to better intended outcomes that are balanced, reliable and bias-controlled.

Furthermore, some mature AI organizations can also consider logical feature stores to address the need for feature reusability, reproducibility and reliability in ML portfolios. Code used to create features for ML can be stored, versioned and accompanied by additional metadata. The majority of feature store capabilities include a repository of these features from which users can pull training and testing datasets, accelerating ML model development. Some feature stores can also orchestrate feature transformations and monitor data serving models in production (see Feature Stores for Machine Learning (Part 1): The Promise of Feature Stores and Feature Stores for Machine Learning (Part 2): Current State and Future Directions).

For successful MLOps, it is important to introduce validation checkpoints throughout the development and operationalization cycles of a project workflow. Collaboration among all team members (see A CTO's Guide to Top Artificial Intelligence Engineering Practices) and the ability to reproduce results with the model are key aspects to ensure successful validation. The model testing and validation step in the development cycle (see Figure 3) includes:

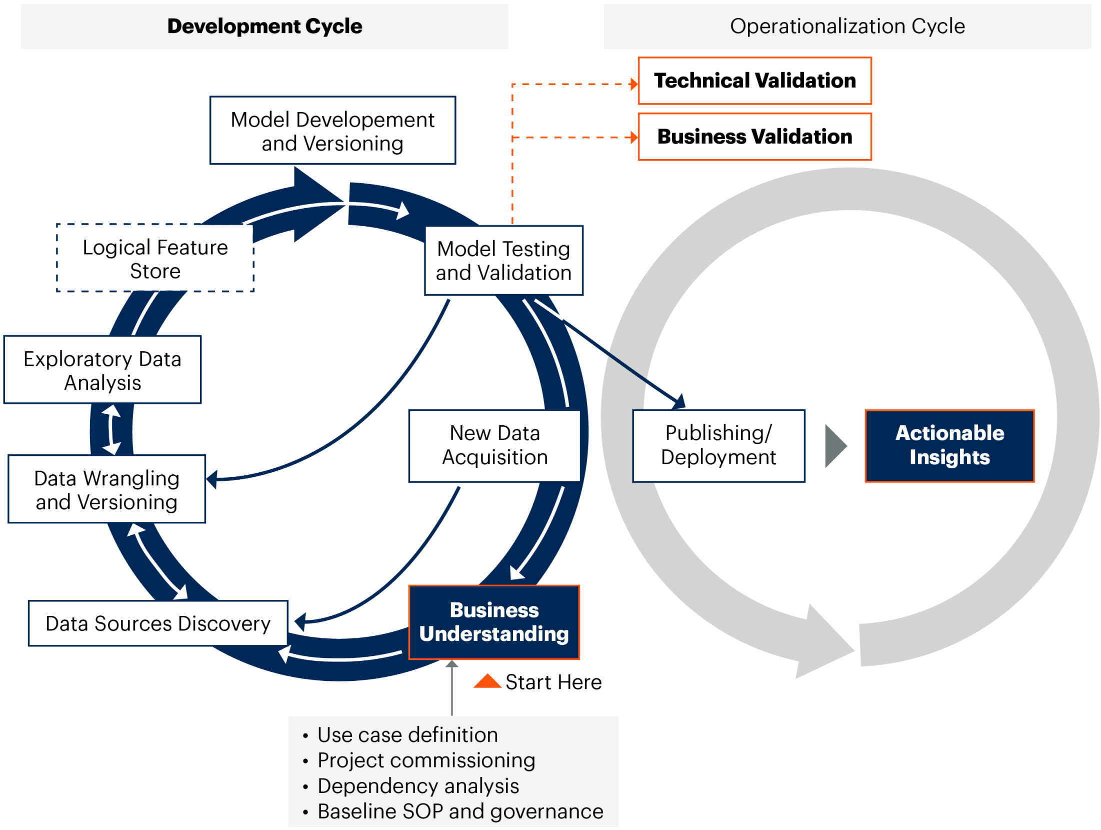
- **Business Validation**. The model is developed on the basis of specific desired business outcomes and needs to operate within specified thresholds while continuously monitoring for unintended outcomes. Business validation will ensure that the model actually delivers the value determined in the business understanding step (see 4 Machine Learning Best Practices to Achieve Project Success).
- **Technical Validation**. The model is technically sound while tested against the dataset, and is set apart at the beginning of the model development process. Model engineering should also satisfy all interpretability requirements.

Model testing and validation is a necessary step before introducing the model to a preproduction environment. There are several mechanisms and guardrails that need to be introduced on the basis of model type, model usage, domain and industry-specific metrics as well. For instance, testing of a NLP model would differ from a computer vision model, where an NLP model would be tested for minimum functionality, invariance or directional expectation whereas computer vision models are highly domain- and use-case-specific. NLP model testing is very use-case-specific as the models are different from each other, as is the training and validation data. For translation purposes, models are typically different from text analytics models or speech-to-text models.

Another area where NLP models are often quite different from other ModelOps is in the languages themselves. Testing often needs to be performed in each supported language. On the other hand, a computer vision model would have to pass scenario tests, ensure high predictability in noisy images or have high accuracy in testing environments. Hardware and AI on the edge poses additional challenges with respect to integration and interoperability in machine vision use cases. On the other hand, emotion detection models are culturally and demographically sensitive and can't be generalized.

Figure 3: Validation Stage of MLM

Validation Stage of MLM



Source: Gartner

Note: Logical feature stores is an optional step, typically important for organizations with a large ML footprint

774981_C

Gartner sees a new role coming up, that of a model validator — an individual who did not participate in the development of the model. This exercise ensures model validation with a reduced bias by virtue of not being included in the build process.

2. Model Operationalization Cycle Process

Once a model has been published out of the development cycle, its introduction into the operationalization cycle takes place in two main phases: the release phase and the activation phase, depicted in Figures 4 and 5.

Depending on their project size and complexity, not all models developed by organizations undergo those two phases. It is also possible that some organizations (especially while implementing real-time model operationalizations) might integrate the release phase as part of their development cycle, but a large majority of analytically mature organizations productionize models through those two phases.

It is important to note at this point that models in production are often part of an ensemble of models working together to provide insights within a particular process, or embedded in one or more applications (internal or customer/client facing). These models have a direct impact on a multitude of business KPIs such as compliance adherence, operational metrics, revenue generation and so on; hence, operationalization is key to driving ML projects and initiatives.

Release Phase: Testing Models in Business Conditions

The goal of the release phase is to set the published model "free" inside a guarded perimeter to verify that the assumptions made while developing the model still hold true in the real world of the actual business process. There are also vendor solutions that allow organizations to create best model configurations. This is also referred to as the "experimentation" stage. We have identified seven steps in this process (see Figure 3):

- **Model risk review.** Models must be retested to ensure that they don't expose the organization to business (financial, reputational) risks and have no open backdoors from which the model can be reverse-engineered. A thorough security review of the model is a must-have to ensure safer operations of systems. This includes built-in protection mechanisms against, and ongoing monitoring for, data poisoning checks, adversarial attack resistance, query attacks to determine model logic and model manipulation attacks (through perturbations, ill-intended inputs or malicious intents by benign actors).
- **Model release.** The model is promoted to the release phase, ready to be undertaken by the operationalization team and labeled as a candidate model (that is, development-vetted, but not yet fully production-ready).

- **Endpoint identification.** This step refers to validation of the decision points where the model will be delivering its insight. Those analytics endpoints could be within an existing application, within a business process, as part of an ensemble of models and/or as an input to another decision-modeling mechanism (like a business rule) or on the edge devices.

- **Parameter testing.** Target business processes might be subject to technical constraints where the velocity, shape, volume and quality of the input data might not exactly align with the data in sandboxes used to develop the models. This step aims at testing that alignment.

- **Integration testing.** Provided that the expected data matches the development assumptions, integration assumptions (that is, REST APIs, microservices call, code integration) also have to be tested to ensure proper performance.

- **Instantiation validation.** As models in production are often part of model ensembles, even slight variations in those elemental models (such as connected models instantiated across multiple states or regions in the same country) can produce radically different results.

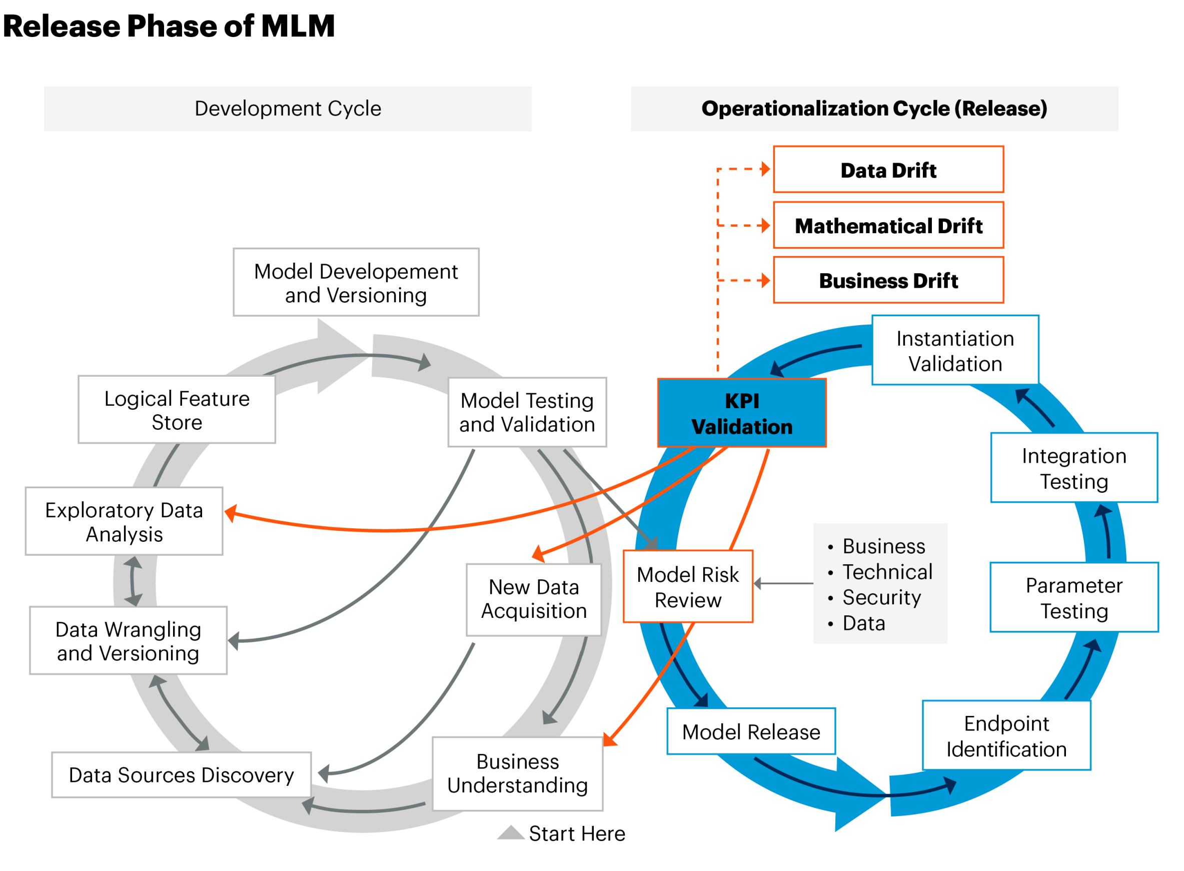
- **KPI validation.** Model performance should not only be measured against technical parameters (such as precision) but also against the agreed-on KPIs. Validate models for:

    - **Business drift.** This includes the deviations in business KPIs (measurable business outcomes) set forth as early as the business understanding step in the development process. The model could be drifting due to new market conditions that do not necessarily affect the data directly.

    - **Mathematical drift.** Deviations in technical parameters (such as precision) result in degraded model performance. The model drift could be due to a radical change in the pattern being monitored.

    - **Data drift.** Any shift between the data that has been used to build the model and the actual data in production could render the model inefficient or incompetent.

Figure 4: Release Phase of MLM



Source: Gartner
774981_C

Depending on the outcome of the KPI validation step, four paths are possible:

- The model fails due to data drift. In this case, the process can proceed back to the new data acquisition step in the development process.
- The model fails due to business drift. In this case, it might be wise to reevaluate the original business assumptions through the business understanding step back at the beginning of the development cycle.
- The model fails due to mathematical drift. In this case, proceed back to the analysis step in the development cycle.
- The model delivers as promised and is ready to move to the activation phase of the operationalization cycle.

3. Activation Phase: Operating Models in Real Business Conditions

Now that the model has been tested in the real world and is performing as intended, it is ready for deployment. The goal is to activate that model within existing business processes across the organization at the endpoints identified (and validated) in the release phase of the operationalization process. We have identified seven steps in this process (see Figure 4):

- **Management and governance.** Once a model is ready for activation, it should be cataloged, documented and versioned (including the original set of variables used to develop it). That model should also be submitted to the governance rules adopted by the data science team and the production committee (and, in particular, application product managers). ML models must continuously integrate, deploy, test and verify (CI-CD-CT-CV) while ensuring reproducibility, reusability and reliability.

- **Model activation.** This step is the hand-off of the operationalization-process-validated models to the production team as activated models. At this point, the models are production-ready, fully documented and compliant with the governance rules.

- **Model deployment.** Depending on how those models will be executed (on-premises, in the cloud or both,), measures need to be taken to guarantee the smooth processing of the transactions leveraging the model (or the ensemble the activated model is part of). There are multiple deployment strategies that can be undertaken depending on ML model or application, but they're broadly classified into:
  - *Static Deployment*
    - Basic deployment, in which the ML model is completely replaced by the older model.
    - Recreation, in which the existing version of the model will be scaled down before scaling up the newer model version.
    - Blue-green, in which the blue environment has the current model deployed and is live to serve current requests. The green environment acts as a staging area that contains the new model where it is tested against live data for expected performance and functionality. Once the testing is complete, the data traffic is shifted to the green environment and the blue is kept as a rollback or decommissioned for next deployment.
    - Canary deployment, in which the new model rollout happens incrementally until full deployment is achieved.
    - A/B Testing, in which different versions (with different features) of the model are deployed in production to compare performance of models against one another with an aim to increase the conversion rate, and the best gets deployed.
    - Champion-challenger: Multiple models run in parallel against live data, constantly competing against each other on model performance; the champion model always stays deployed.
  - *Dynamic deployment*
    - Multiarmed bandits are an advanced version of A/B testing (see Note 4). For more information, see Google's note on Application deployment and testing strategies.

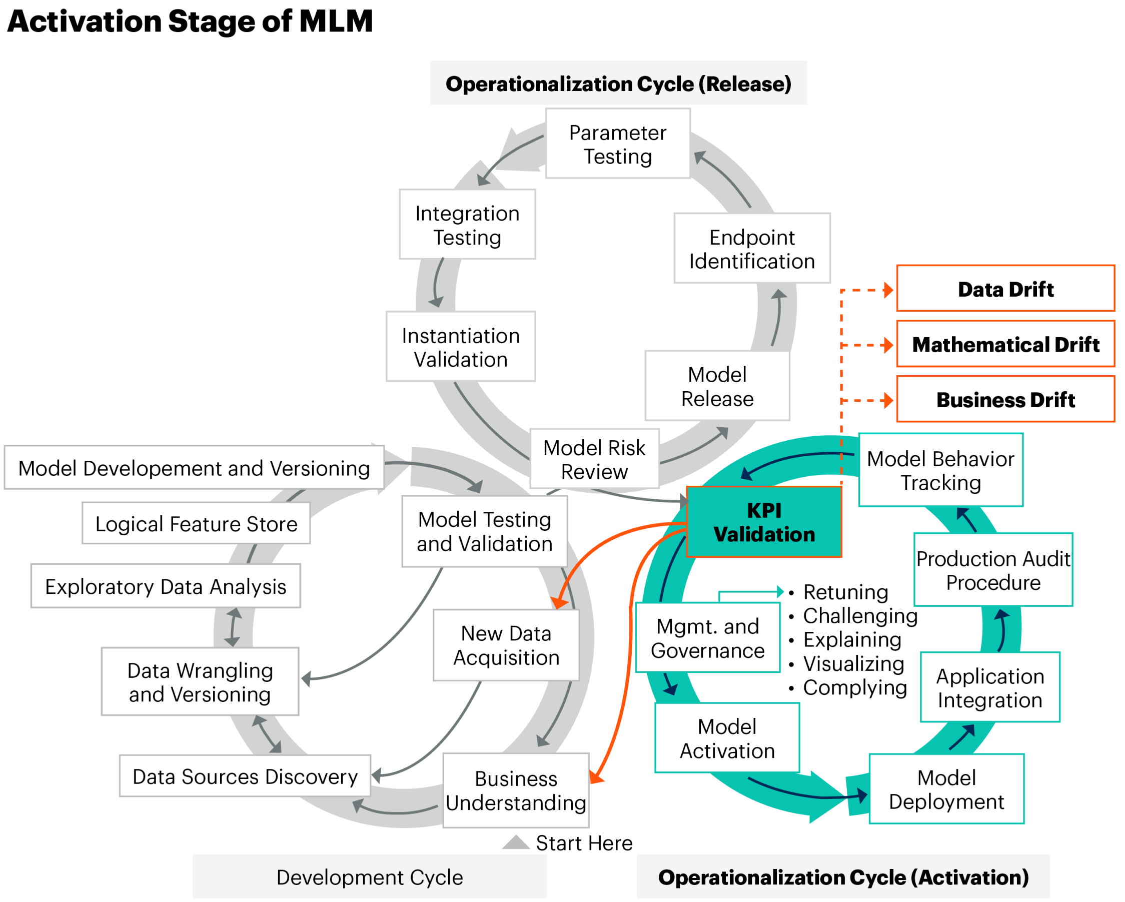
- **Application integration.** Insights are delivered through decision endpoints, the large majority of which are part of existing applications. Extra coding is sometimes necessary to properly embed a model, or its input, within an application, a business process or another set of insights to enhance a business outcome. This is where the model will finally deliver its business value.

- **Production audit procedures.** Model telemetry, along with various performance metrics in terms of accuracy, response time, input data variations and infrastructure performance instrumentation, including possible implementation of software agents, have to be implemented to gather the necessary data to monitor models in production. Various instruments (see Note 3), such A/B testing methods, multivariate testing, multiarmed bandits (MABs; see Note 4) and shadow models can be implemented to appreciate the performance of models.

- **Model behavior tracking.** Alerts and monitoring methods have to be established to track the performance of models in production. Performance thresholds and notification mechanisms are implemented in this step to systematically flag any divergence or suspicious behavior.

- **KPI validation.** Extended from the release phase and fed by the two previous steps, the KPI validation step consistently measures the business contribution of the models (or ensemble models) in production. The idea is to obtain, as precisely as possible, the business value that can be attributed to the model. To determine that value, data from the production audit procedures step should prove invaluable.

Figure 5: Activation Stage of MLM



Source: Gartner
774981_C

3a. Review and Revalidate ML Model Operational Performance

From this point, the models remain in the operationalization cycle as long as they are performant, meet attribution thresholds and deliver business value derived from the models in production. As long as models are performing, they keep their place in the production cycle. Otherwise, management rules are in place in the management and governance step to reevaluate those models in the light of changing circumstances captured in the three previous operationalization steps. Those models are then sent back into the development cycle process. Successful organizations have adopted various technologies to provide the foundation to manage operational decision services in a reliable, repeatable, scalable and secure way. That governance (see Note 5) foundation should provide:

- **A catalog.** A centralized way to store and secure ML assets to make it easier for analysts to collaborate and to allow them to reuse or exchange models or other assets as needed. It could be a secured community or a collaboration space as well as a centralized *model inventory* or repository with appropriate role-based controls and guardrails on it.

- **Governance.** Protocols to ensure adherence to all internal and external standards, procedures and regulations, not just for compliance reasons, but as an increasing amount of data gets aggregated — for example, to address potential privacy issues.

- **Capabilities.** Automated versioning, fine-grained traceable model scoring and change management capabilities (including champion/challenger features) to closely test, monitor and audit analytical asset life cycles from a technical as well as a business performance perspective (through KPIs).

- **Coherence.** Simple protocols established to:

  - Provide functional bridges between the development and operationalization cycles.

  - Enhance cooperation and consultations between development and operationalization teams.

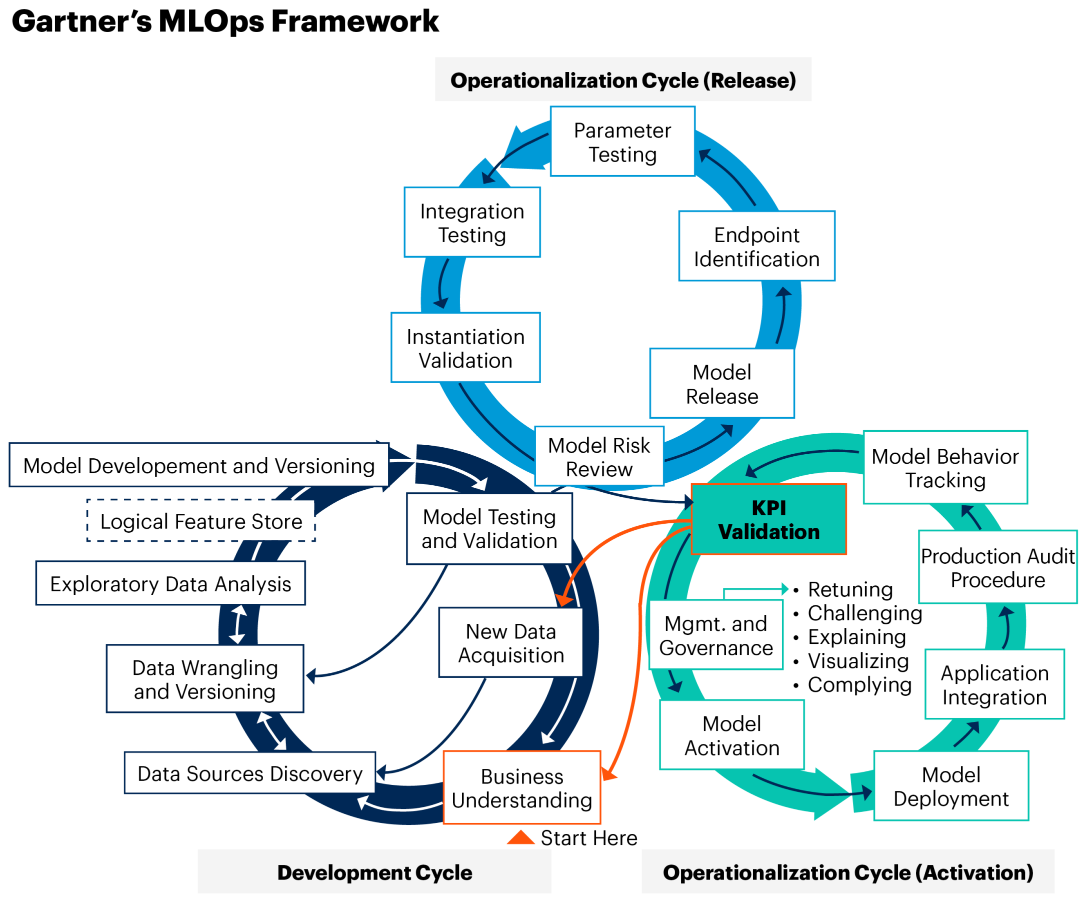
  - Provide efficient liaison services between data science and lines of business.

  - Improve the transparency of deployed analytical assets.

  - Control the exchanges between the production and the development teams (such as when to retune models, expose challenger models and levels of model explanations).

The combined processes are illustrated in Figure 6.

Figure 6. Gartner's MLOps Framework



Source: Gartner

Note: Logical feature stores is an optional step, typically important for organizations with a large ML footprint

774981_C

Machine learning systems carry a significant technical debt that has to be addressed upfront. To confront the problem early, data and analytics leaders should:

- Start as early as the development process to address the KPIs and detect operationalization risks. Most successful organizations balance risk and ROI.
- Establish a well-instrumented and disciplined model operationalization process to scrupulously track the behavior and performance of ML models deployed across business processes.
- Heed the management and governance step in the operationalization process as its rules and procedures will eventually define the success or failure of your deployed analytical assets.

3b. Review and Revalidate ML Model Business Value

The KPI validation step of the operationalization process needs to be constantly revalidated while the model is in production. Beyond the organization's changes of business strategy and tactics, market conditions can shift and customer behaviors or competitive pressure can evolve, impacting the business performance of models. While all of those factors can contribute to models' degradation, others to watch for include:

- Concept drift (where seasonality, decision impacts, local data particularities, dynamic data transitions and so on can induce a change in class distribution)
- Bias reinforcements (where decisions are increasingly subjective through models' self-fulfilling recommendations)
- Feedback loops (where models influence their own behavior as they evolve over time)

KPIs should also be subjected to a higher level of scrutiny that allows them to track not only the business process context of the active model, but across the business processes that are impacted by the decisions influenced by models (or collection of models). This will avoid the local optimum problem in which a few decisions are optimized for local decisions, eventually leading to global issues. By auditing the development, refinement and usage of models and other assets throughout their life cycle, the formalization of the operationalization cycle provides a bigger-picture advantage: more consistent organizational decisions. For a list of sample vendors that offer niche capabilities in MLOps, refer to Note 6.

Recommendations:

- Secure the commitment of the business stakeholders at the business understanding step for the continuous validation of the business KPIs while the model is in production.
- Track the models' (or model ensembles') degradation signals by constantly monitoring the business indicators influenced by productized models.
- Continuously measure and harmonize the overall business impact and collective contribution, within or across processes, of integrated model collections.

Evidence

2021 Gartner AI in Organizations Survey: This survey was conducted to understand the keys to successful AI implementations and the barriers to the operationalization of AI. The research was conducted online from October through December 2021 among 699 respondents from organizations in the U.S., Germany and the U.K. Quotas were established for company size and for industries to ensure a good representation across the sample. Organizations were required to have developed AI or intended to deploy AI within the next three years. Respondents were required to be part of the organization's corporate leadership or report into corporate leadership roles, and have a high level of involvement with at least one AI initiative. Respondents were also required to have one of the following roles when related to AI in their organizations. determine AI business objectives, measure the value derived from AI initiatives or manage AI initiatives development and implementation. The survey was developed collaboratively by a team of Gartner analysts and Gartner's Research Data, Analytics and Tools team. Disclaimer: Results of this survey do not represent global findings or the market as a whole, but reflect the sentiments of the respondents and companies surveyed.

Notes

Note 1: Analytical assets include machine learning (predictive) and mathematical models, and precalculated propensities (such as predictive scores).

Note 2: ModelOps (AI Model Operationalization): ModelOps is focused primarily on the governance and life cycle management of a wide range of operationalized AI and decision models (including machine learning, knowledge graphs, rules, optimization, linguistic and agent-based models). Core capabilities include CI/CD integration, model development environments, champion-challenger testing, model versioning, model store and rollback. ModelOps also enables the governance (see Note 5) and procedures for retuning, reusing, retraining or rebuilding AI models, aimed at providing an uninterrupted flow between the development, operationalization and full maintenance of AI models. Adopting a ModelOps strategy should facilitate the performance, scalability and reliability of AI models.

Note 3: Model Evaluation Techniques

An often-preferred online model evaluation and selection technique is to use a multiarmed bandit algorithm. A/B testing has one major drawback. The number of test results in each group, A and B, needed to find the value of the A/B test is high. This means that a significant part of the users routed to the suboptimal model would experience suboptimal behavior of the product for a long time. Ideally, user exposure to a suboptimal model should be as few times as possible.

At the same time, users should be exposed to each of the two models several times to get reliable estimates of both models' performance. This is known as the exploration-exploitation dilemma. The performance of the models should be explored enough to be able to reliably choose the best one. However, at the same time, the performance of the best model should be exploited as much as possible to reduce the negative effect of exposing the users to a suboptimal model.

Note 4: Multiarmed Bandits

Multiarmed bandits (MABs) are a way to compare one or more versions of the model and select the best-performing one in the production environment. MABs have an interesting property. After an initial exploration period during which the MAB algorithm gathers enough evidence to evaluate the performance of each model (arm), eventually, the best-performing arm is played all the time. This means that, after the convergence of the MAB algorithm, all users are routed to the version of the software running the best model.

This property of the MAB algorithm allows users to deploy the new model while keeping the old one and waiting for the MAB algorithm to converge. This gives users the information on whether the new model performs better than the old one. At the same time, it lets the MAB algorithm replace the old model with the new one once it is certain that the new model performs better.

Note 5: Gartner's Definition of Governance

- Setting decision rights and accountability, as well as establishing policies that are aligned to business objectives (preservation and growth of shareholder value)
- Balancing investments in accordance with policies and in support of business objectives (coherent strategy realization)
- Establishing measures to monitor adherence to decisions and policies (compliance and assurance)
- Ensuring that processes, behaviors, and procedures are in accordance with policies and within tolerances to support decisions (risk management)

Note 6: Operationalization at Scale

Sample vendors that support operationalization at scale include Algorithmia, DataRobot (ParallelM), Datatron, Iguazio, ModelOp, Saagie and Seldon.

Document Revision History

Use Gartner’s 3-Stage MLOps Framework to Successfully Operationalize Machine Learning Projects - 2 July 2020

How to Operationalize Machine Learning and Data Science Projects - 3 July 2018

Recommended by the Authors

Some documents may not be available as part of your current Gartner subscription.

Understanding MLOps to Operationalize Machine Learning Projects

Case Study: Realizing the Promise of Analytics and BI Platforms (Dow)

Case Study: How to Apply Ethical Principles to AI Models (Danish Business Authority)

Case Study: Make AI Models Credible, Not Explainable (Unity Health Toronto)