MLOps 101: The Foundation for Your AI Strategy

DataRobot

Table of Contents

What Is MLOps?

The definition of MLOps (machine learning operations) includes the culmination of people, processes, practices and underpinning technologies that automate the deployment, monitoring, and management of machine learning (ML) models into production in a scalable and fully governed way to finally provide measurable business value from machine learning. Laying an MLOps foundation allows data, development, and production teams to work collaboratively and leverage automation to deploy, monitor, and govern machine learning services and initiatives within an organization.

Depending on an organization's maturity level, their MLOps infrastructure can be represented by something as simple as a set of vetted and maintained processes. For example:

- How models are created, trained and approved
- Where models are stored
- How models are deployed
- How models are evaluated and monitored in production
- How models are either corrected or removed from the production environment to minimize risks
- How these processes repeat and intertwine to enable a cyclical machine learning operations process

A mature MLOps foundation can be represented by an automated system that streamlines all of the model's lifecycle steps, from its training and deployment, through its production life, to its retirement and storage for compliance and risk management reasons, offering full transparency into the process. Needless to say, the above needs to function with complete and seamless integration to all current Ops and Cloud services and processes.

Follow this link to learn more.

Why Do Organizations Need an MLOps Infrastructure?

Many organizations are dipping their toes into machine learning and artificial intelligence (AI). Some are already reaping some of the rewards of artificial intelligence through increased productivity and revenue. However, for most organizations embarking on this transformational journey, the results are yet to be seen and for those who are already underway, scaling their results appears as completely uncharted waters.

According to a survey by NewVantage Partners, only 15% of leading enterprises have deployed AI capabilities into production at any scale. Most of these leading organizations have significant AI investments, but their path to tangible business benefits is challenging, to say the least. There are a number of reasons for this that we find to be reoccuring practically everywhere.

To find out who needs MLOps, visit this page.

There's a skill, motivation, and incentive gap between teams developing machine learning models (data scientists) and operators of those models (DevOps, software developers, IT, etc.). There are a plethora of issues here, which vary by organization and business unit. Here are a few examples:

- Lack of available data science talent means that when organizations find someone with the right experience, they allow these individuals to operate in an environment that's most suitable for them, which leads to the next problem.
- Models are typically created using DS-friendly languages and platforms, who are typically suboptimal or more so, unfamiliar to Ops teams and their services, who were designed for regular software languages and platforms.
- Ops teams are geared towards optimizing run time environments upon their cloud, resource managers, role based services, etc. Data Science teams are not only unaware of any considerations these dependencies require, but are typically oblivious to them altogether and hence models they create do not take these into consideration at all.
- Lack of a proper native governance structure as pertains to Machine Learning models, with system, lifecycle, user logs, stifles troubleshooting, as well as legal and regulatory reporting.
- Organizations that don't properly monitor their models end up introducing what could possibly become immense risk to their organizations due to production models that don't reflect the ever-changing patterns in data, user/consumer behavior, and a host of other issues that may affect the accuracy of the model and that will not be altered upon when they happen.
- Many operations professionals are not aware of the unique characteristics and sensitivities of Machine Learning, and as their role is focused on managing mission critical production environments, are very wary and concerned with the implications that deploying Machine Learning into production might have on their work. Lacking relevant ML management tools and processes only compounds the risks associated with ML adoption. These justifiable concerns serve as trivial causes for vast delays in progress and deployment. Data scientists are not familiar with the principles of production-grade code, so they spend their time babysitting their models when something eventually goes wrong.

MLOps allows organizations to alleviate these and many other issues by providing a technological backbone for managing the machine learning lifecycle through automation and scalability. It provides for seamless collaboration between the data teams responsible for generating models with the teams traditionally managing the services running upon production environments, thus streamlining the path to strategic goals that organizations want to achieve with AI.

The Pillars of MLOps

Any MLOps solution capable and agile enough to handle complex situations has to be proficient in the following four critical areas to safely deliver machine learning applications in production at scale:

- Model deployment
- Model monitoring
- Model lifecycle management
- Production model governance

Without these pillars, no MLOps system can maintain a level of sophistication required for maximum impact. The absence of these four critical areas would lead to more manual processes, spaces for human error, or even procedural blindspots, where critical issues might go unnoticed for days and months, amplifying corporate and operational risks. Let's take a closer look at each one of these pillars.

What Is Model Deployment?

The Challenge

A machine learning model's lineage can be traced to numerous machine learning platforms and various programming languages, the creators of which are typically agnostic of actual production environments and their mission critical considerations. When organizations try to implement them in an environment suited for normal software applications, the manual integration creates friction, to the point where the model can't be deployed without jeopardizing the stability of the production environment. As a result, the organization loses the option to scale this activity, as well as revenue and forfeits cost savings.

The Solution

MLOps simplifies model deployment by streamlining the processes between modeling and production deployments. It should not matter which platform or language the model was built on. An enterprise-grade MLOps system should allow organizations to plug in their models and generate consistent API access for application teams on the other end, regardless of deployment environments and choice of cloud services and providers.

What Is Model Monitoring?

The Challenge

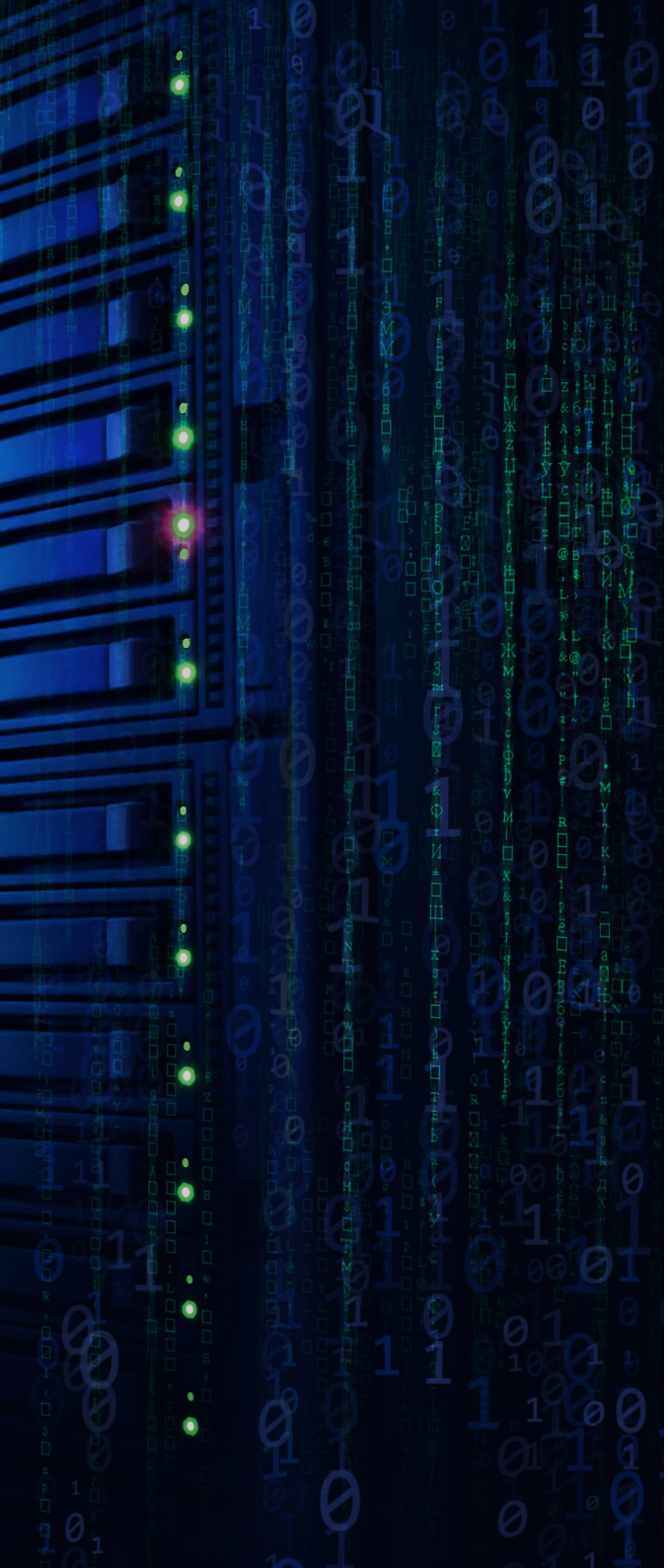
There are many reasons for machine learning model degradation or other performance related issues over time. For example, you could be making live predictions on a dataset with customer data, but the behavioral patterns of that customer may have changed, due to economic crisis, market volatility, natural disaster, or just simply the weather. The performance of these models won't be useful and may even be harmful to your business. Models trained on older data that doesn't represent the current reality may be not only inaccurate but irrelevant. The root problem is that you wouldn't know or be able to tell when this happens. Without dedicated production monitoring explicitly designed for machine learning, you could expose your business to risks cascading from not even knowing about completely irrelevant predictions.

The Solution

MLOps allows both production and AI teams to monitor models in ways specific to machine learning. A robust monitoring infrastructure should be able to proactively monitor data drift, feature importance, and model accuracy issues. Advanced capabilities may include features built to increase trust towards models in production even further. For example, the principle of humility in AI dictates that models should be able to inform not only when predictions are possibly going bad, but also when they're not confident in the quality of their predictions.

Learn Some of the Best Practices in Production Model Monitoring

Listen to the Podcast



What Is Model Lifecycle Management?

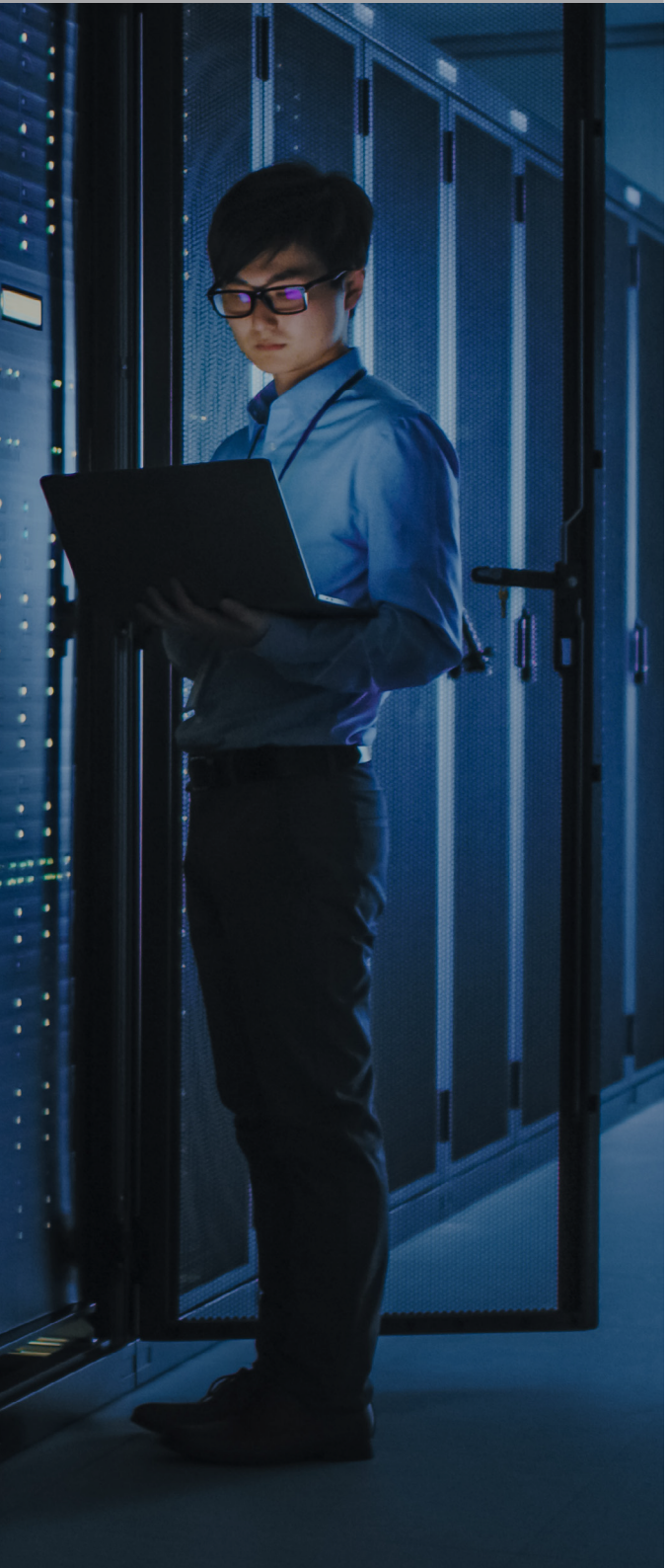
The Challenge

As organizations build out robust machine learning initiatives, the number of models actually in production may grow exponentially, making managing these models and their lifecycle a cumbersome task. A great deal of time is spent on manual attempts to bring order to this process which may include everything that has to do with ensuring the phases any model has to go through are streamlined, approved via a flexible workflow, and are as automated as possible. Some examples are:

- Champion/ challenger model gating: It is now becoming a best practice to introduce any new model (aka 'challenger') by first running it in production and measuring its performance vis-a-vis its predecessor (aka 'champion') for a defined timeframe, in order to determine whether the new model is worthy of becoming the new 'champion' by outperforming it. While this process ensures increasing levels of quality and continuous improvement of predictions and model stability, the important thing is for this process to become completely automated.
- Troubleshooting and triage: Simply monitoring models is not sufficient without the ability to triage, troubleshoot, and rectify suspicious or poorly performing models.
- Model approval: A formal process designed to minimize risks associated with deploying the model, ensuring that all relevant business or technical stakeholders have signed off on the model.
- Model updates in production: The ability to swap models without disturbing the production workflow is key to business continuity.

The Solution

MLOps allows for a production model lifecycle management system that automates processes, such as champion/challenger gating, troubleshooting and triage, hot-swap model approvals, and offers a secure workflow to ensure the efficient management of your models' lifecycle as you scale.



What Is Production Model Governance?

The Challenge

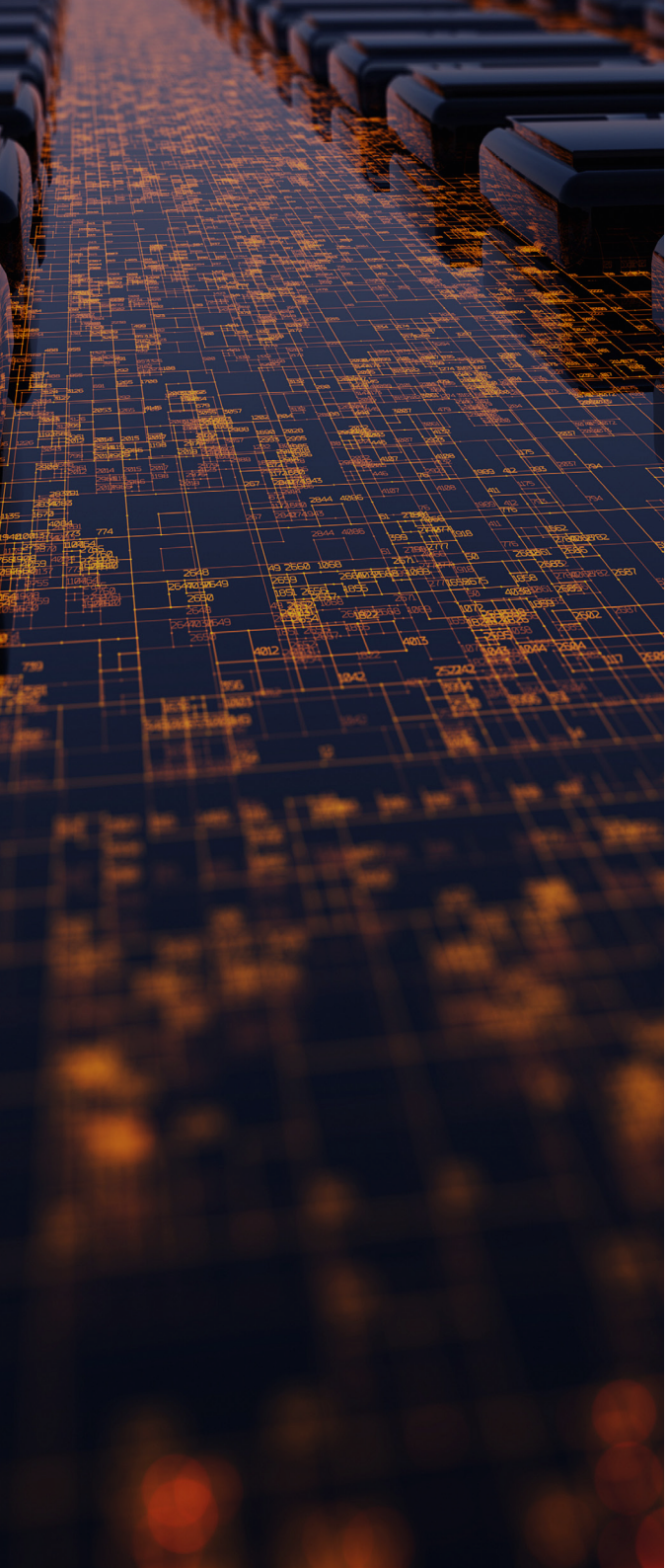
Organizations putting machine learning models into production are dealing with regulatory, compliance, and corporate risk minefields, especially after the introduction of regulations like CCPA, EU/UK GDPR, and others. This issue becomes especially critical for organizations operating on a global scale, where the maze of rules and laws becomes almost impossible to navigate. In these situations, organizations need to maintain complete model lineage tracking (approvals, model interactions, versions deployed, updates, etc.), something that is practically impossible to perform manually.

The Solution

MLOps offers an enterprise-grade production model governance solution, which can deliver:

- Model version control
- Automated documentation
- Complete and searchable lineage tracking and audit trails for all production models

This allows companies to minimize corporate and legal risks, maintain a transparent production model management pipeline, minimize and even eliminate model bias, and deliver a host of other benefits.



MLOps for Data Scientists

Data scientists and data science teams working on machine learning and AI initiatives can immensely benefit from MLOps. On one hand, the solution will automate many parts of their day to day life, and on the other, it will help Data Scientists effectively collaborate with their Ops counterparts, offloading much of the burden of day to day model management.

MLOps allows data scientists to focus on what's important — discovering new use cases to tackle, working on feature discovery, and building more in-depth business expertise. Data scientists should not be wasting time maintaining models or reviewing their performance manually. No more manual testing, monitoring, and validation.

Discover the benefits of MLOps for Data Science Leaders and Practicing Data Scientists.

MLOps for Business Leaders and Executives

With MLOps capabilities in place, organizations can start focusing on things that really matter, scaling AI capabilities throughout the organization, while simultaneously tracking KPIs that matter to each team and department.

Business leaders are now looking for assurance that their predictions are fast, accurate, and above all can be unbiased and trusted. Ensuring that decisions are made on the predictions they receive is imperative, since the ability to allow the company to conduct deeper and more meaningful analysis means better business results. However, there are still many obstacles on the path to AI with ROI. To achieve savings, organization and competitive benefits, and measurable ROI from AI projects, companies need to have MLOps capabilities in place to operationalize AI and machine learning at scale.

Discover all of the benefits of MLOps for Line of Business Leaders and Executives.

MLOps for Software Developers

In too many cases, trying to get a machine learning model to work in an application or service turns into an endless ETL nightmare. A robust MLOps system simplifies this process for developers by supplying a straightforward deployment and versioning system, backed by a clear and easy-to-manage API.

Developer-focused MLOps features in such systems include, but are not limited to:

- Clear and simple API (REST)
- Developer support for machine learning operations (documentation, examples, etc.)
- Versioning and lineage for all production models
- Portable Docker images and more

To learn more about MLOps for Software Developers, visit this page.

MLOps for IT Operations Teams

MLOps was built precisely to allow IT teams to take full control of production models, ensuring that any model can be easily deployed in a matter of clicks, no matter the origin or programming language. It was designed to look and feel like familiar IT solutions, and seamlessly integrated with leading IT management solutions. Eliminating the friction and disconnect between IT and AI teams.

MLOps can grant IT teams the opportunity to take ownership of AI in production by offering:

- Alerting and connection to service, IT management systems and ticketing systems
- Repo integration
- Change approval workflows
- Seamless model update rollouts
- Robust access control management (LDAP, RBAC, etc.) and more

To learn more about MLOps for IT Operations Teams, visit this page.

MLOps for DevOps and Data Engineers

MLOps allows DevOps and data engineering teams by adding the way they manage the actual machine learning models' layer, handling everything from testing and validation to updates and performance metrics, in a single place. This allows your business to generate more value from AI by being able to seamlessly scale internal deployment and monitoring capabilities and monitor service health over time to meet latency, throughput, and reliability SLAs.

MLOps for DevOps teams and Data Engineers should include a variety of capabilities:

- No-code prediction GUI
- Anomaly, performance, and bias warnings
- Accessible and optimized API
- Swappable models with automated gating of choice, guaranteeing smooth transitioning and 100% uptime

MLOps for Risk and Compliance Teams

An MLOps infrastructure allows risk and compliance teams to streamline their internal processes and improve the quality of oversight for complex machine learning projects. A mature MLOps solution should support a variety of features, like customizable governance workflow policies, approval processes and alerting, to raise awareness of these issues immediately as well as ensure all lineage is tracked and can be found whenever needed.

This is also where the concept of humility in AI comes into play again. A system that can self-diagnose problems with predictions and notify relevant risk management stakeholders allows for tighter enterprise controls over machine learning projects in production. Additional key MLOps elements that risk and compliance teams should be on the lookout for are built-in prediction explanation capabilities, predictions-over-time analysis, and audit logs.

To learn more about MLOps for Risk and Compliance Teams, visit this page.

MLOps for AI and Machine Learning Centers of Excellence

Building a coordinated, strategically aligned, and scalable AI and machine learning-driven operation is difficult. From dealing with organizational silos to going against the technological core of the company and "the way things are always done," this is a monumental task. MLOps allows AI and Ops teams to embed cutting edge predictive models in an efficient and value-driven way.

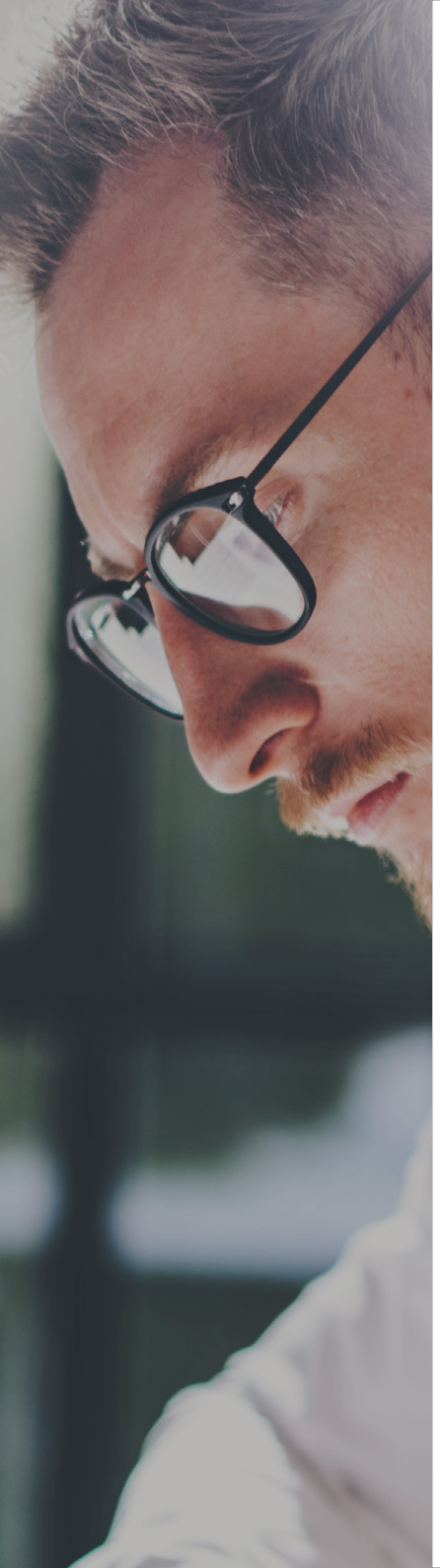
- Work and experiment with different types of models created on any platform and in any language inside a single MLOps solution.
- Use and build upon the foundation you already have, regardless of run-time environments, on premise or multi-cloud scenarios.
- Deploy reliable, trustworthy, and unbiased models.

To learn more about MLOps for AI and Machine Learning Centers of Excellence, visit this page.

DataRobot MLOps

DataRobot MLOps allows organizations to deploy, manage, monitor, and govern their machine learning models from a single place, empowering the different stakeholders to seamlessly collaborate around the common goal of scaling and managing trusted ML models in production. As an origin-agnostic and destination-agnostic platform, MLOps can work with models no matter what environments or languages they were developed in, or where they are going to be deployed. The platform includes proprietary model health monitoring and accommodates for changing conditions by continuously learning through automated challenger models that are designed to test the effectiveness of the existing models in production, as well as automatically generating new challenger models as needed.

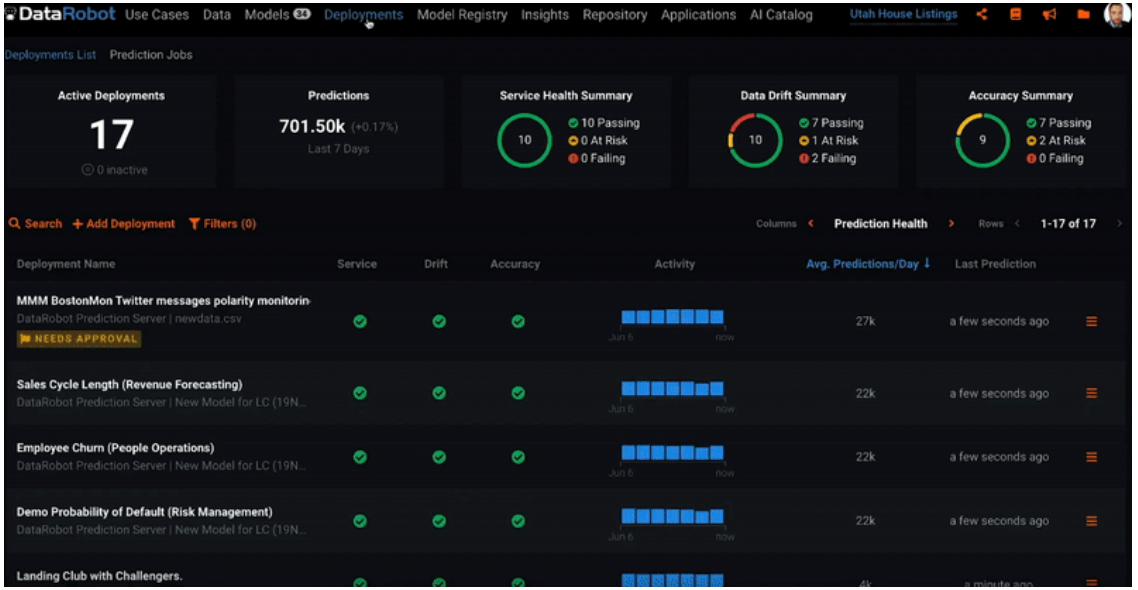
DataRobot ensures all centralized production machine learning processes work under a robust governance framework across your organization, leveraging and sharing the load of production management with additional teams you already have in place.



Build and Run Your Models Anywhere

Our MLOps capabilities enable the deployment of practically any model to virtually any production environment. This can be any cloud, on-prem, or hybrid environment. On top of it all, DataRobot's monitoring capabilities ensure that the production models you have already deployed are transmitting live performance updates to a single and centralized machine learning operations system.

- **MLOps Monitoring Agents** – MLOps uniquely applies the agent concept familiar in the DevOps world to run your models in your preferred infrastructure while monitoring all models centrally within a single pane of glass.
- **Central and Scalable ML Monitoring System** – Proprietary monitoring built from the ground up specifically for such scenarios, provide instant visibility into the performance of hundreds of models, deployed anywhere and built by anyone across your organization.
- **Flexible Model Deployment** – Easily deploy models in any open-source language or library and expose a production-quality REST API with support for real-time or batch predictions.



Automated Model Health Monitoring and Lifecycle Management

- **Built-in Data Science Expertise** – Fully productized mechanism designed to surface drift in data, accuracy and bias.
- **Continuous Model Competitions** – Automatically gate new models via champion/challenger processes.
- **Production Diagnostics** – Monitor service health over time to meet your latency, throughput, and reliability SLAs.

Embedded Governance, Humility, and Fairness

- **Humble and Trusted AI** – Ensure the ethical use of your AI by comparing model predictions and accuracy across different ethics guidelines that are configurable to your industry.
- **Model Approval Workflows** – Maintain thorough reviews of model updates with less tedious manual work, using customizable and governed review cycles and approval workflows.
- **ML Audit Trail and Logging** – For regulatory compliance purposes, MLOps tracks and preserves a full lineage of prediction activity and any model updates so that you always know what model was created, used, when it was updated, and by whom.

MLOps Education

DataRobot University allows anyone to learn more the basic concepts around MLOps, as well as take a deeper dive into DataRobot MLOps and how it allows enterprises to scale their AI and machine learning efforts. You can choose from a selection of both free and paid MLOps courses now.

Visit this page to start mastering MLOps by accessing our courses, learning sessions, and resources.

DataRobot

DataRobot helps enterprises embrace artificial intelligence (AI). Invented by DataRobot, automated machine learning enables organizations to build predictive models that unlock value in data, making machine learning accessible to business analysts and allowing data scientists to accomplish more faster. With DataRobot, organizations become AI-driven and are enabled to automate processes, optimize outcomes, and extract deeper insights.

Sign up for a free trial today to find out how DataRobot can help your organization at **datarobot.com/trial**