CLOUDERA FORRESTER

MLOPS IN ENTERPRISE PRODUCTION MACHINE LEARNING

Q&A with Kjell Carlsson, PhD of Forrester Research, Inc

Q&A With Kjell Carlsson, PhD of Forrester Research, Inc.

Kjell Carlsson is the Principal Analyst at Forrester serving application development and delivery professionals, covering data science, artificial intelligence, and advanced analytics. Following up on our recent webinar with Kjell, we asked him to answer our questions about production machine learning (ML) and ML model operations — the current state of MLOps, trends driving use cases, and its challenges and promises for optimizing the data and ML lifecycle.

What is MLOps? How is it different from ModelOps, Model DevOps, etc.?

At its core, MLOps (or Machine Learning Ops) is about capabilities to deploy, manage, monitor, retrain, and govern machine learning models at scale, i.e. operationalizing ML models. It's about taking your models from wherever they are developed, deploying them in the systems and environments where they perform best, monitoring them to ensure that they are performing correctly, and retraining them when they start to degrade. And all this in a transparent, reproducible and secure (i.e. governed) fashion. Above all, MLOps is about scale – enabling the organization to move beyond operationalizing a handful of models in an ad hoc fashion, to an ongoing capability for operationalizing hundreds or thousands of models and beyond.

MLOps actually extends to cover the entire lifecycle of a machine learning model including the data management and development phases as well, because steps and decisions made in these phases have strong implications for how the model can be operationalized, how it will perform and whether it is reproducible at a later date. However, since, for most organizations, the need is greatest when it comes to capabilities after a model has been developed, MLOps today focuses on this part of the model lifecycle.

What about ModelOps (or Model DevOps as some folks have called it)?

ModelOps includes MLOps, but is about the operationalization of all models, not just machine learning ones. It is, thus, also about deploying and managing decisioning models based on business rules, optimization models, etc.. In practice, however, the term ModelOps is interchangeable with MLOps because operationalizing ML models is the largest gap for most organizations as processes and tools for operationalizing other types of models are more mature.

Why do organizations need MLOps?

While developing machine learning models can be valuable on its own for discovering new business insights, most of the business value comes when the models are put into production and are making accurate, ongoing predictions and decisions (e.g. about a customer's lifetime value, and whether to engage them with a promotion or a sales call). MLOps capabilities help an organization get their ML models into production faster, increasing time to value, and dramatically increasing their ability to iterate through the ML lifecycle. Business stakeholders rarely know exactly what they will need until it has been developed, and by helping iterate faster, MLOps also drives innovation and value in terms of delivering solutions that serve the business better. Further they help ensure the performance of those solutions by allowing the organization to monitor these models, and it makes your data scientists, engineers and developers more productive by cutting down on the amount of manual effort required to maintain and manage these models. In addition to reducing risk, for enterprises in regulated industries, MLOps can help them comply with regulatory requirements by making it easier to explain and reproduce their models.

What are the key capabilities (both organizational and platform) required for productionalizing machine learning at scale? (Not just last mile, but full ML lifecycle)

There are a host of capabilities that are necessary for productionalizing ML at scale, but the most important ones fall into three categories:

- **Deployment & serving:** orchestrating the deployment of the model including the provisioning of infrastructure, staging the model, managing dependencies, orchestrating the multiple steps that occur when a model is called, and serving the model in a robust, scalable, and high availability fashion. As part of this, organizations increasingly need to support the deployment of models developed using multiple different ML frameworks. It is especially critical to support open source ML programming languages and frameworks like Python and R both to futureproof the organization's investments in machine learning, as well as tap the growing ranks of data science talent who are trained on these tools. However, an organization's MLOps capabilities should support proprietary ML tools as well to take advantage of all best of breed solutions. Further, organizations need to develop the ability to deploy models across an ever-growing range of platforms and hybrid cloud environments to take advantage of the cost, scalability, latency and security tradeoffs of each.
- **Monitoring**: beyond logging calls to, the data consumed by and predictions generated by models, organizations need MLOps capabilities that evaluate the performance of models in terms of operational KPIs (e.g. latency, failed calls to the model, pipeline errors), accuracy KPIs (e.g. data drift, concept drift, declines in accuracy, precision and recall), and ultimately business KPIs (e.g. revenue or cost impact of correct and incorrect predictions). With the ability to create, schedule, monitor, and notify based on these KPIs, an organization can monitor the performance of their models and rapidly take the right actions, such as retraining the models, when the data, systems, or user behavior requires it.
- **Collaboration & governance**: as the aphorism goes machine learning "is a team sport", and organizations need the ability to ensure that everyone is playing the right role and working together effectively. That means having feature catalogs, model catalogs and versioning, lifecycle lineage, and security capabilities — so that folks can access what each other have done, replicate it, validate it, and build on it. It also means having tools -- such as visual interfaces and explainability features -- that bridge the gap between different roles and skillsets.

What benefits can organizations expect from investing in an MLOps platform?

Organizations can expect more business impact from their machine learning projects because they will be able to accelerate and scale their ability to operationalize their machine learning models. Their machine learning projects will better meet the needs of the business because they can undergo more iterative cycles of improvement, and because their models will stay more accurate because they now get retrained when they degrade. With greater transparency and control comes lower business risk from faulty or biased predictions. They will also be able to detect, and adjust to changes in the market faster as they will be reflected in the distributions of the data their models are consuming, the predictions their models are making, as well as the resulting outcomes. In addition, they can expect more productive (and hence valuable) data scientists -- as well as engineers and developers involved in ML projects -- as they will not need to spend as much manual effort deploying and monitoring models, and they can expect them to be happier too – as they will spend more of their time utilizing their skills and see more results from their efforts.

What happens if my organization doesn't develop our MLOps capabilities?

Without MLOps, organizations will continue to struggle to get their models into production, they will continue to be less accurate in production, they will fail to realize when their models are becoming less accurate, and they will struggle to re-train their models. In other words, organizations will suffer more ML projects stuck in PoC purgatory, wasted effort by data scientists, longer time to value on ML projects, lower quality predictions, and a ticking time bomb in terms of models that are in production that are making worse-and-worse predictions and decisions. Though it's not possible to separate correlation from causation, in one of my recent surveys fast growing firms were 3x more likely to be able to get their models into production and to monitor and retrain them on an ongoing basis. If you don't develop your MLOps capabilities, it is a good bet that your business competitiveness will decline.

Learn more about MLOps in this on demand webinar, Enabling Production MLOps at Scale – Hands on with Cloudera Machine Learning

About Cloudera

At Cloudera, we believe that data can make what is impossible today, possible tomorrow. We empower people to transform complex data into clear and actionable insights. Cloudera delivers an enterprise data cloud for any data, anywhere, from the Edge to AI. Powered by the relentless innovation of the open source community, Cloudera advances digital transformation for the world's largest enterprises.