DataRobot SmartBrief

Getting Smart on MLOps - September 2019

A Perfect Storm for ML in Production

AI-based applications promise to deliver new levels of competitiveness, intelligence, and automation for businesses. Despite increasing investments in data science, however, few companies can scale AI and ML initiatives to deliver transformational business value. The barriers are many, from finding and keeping data science talent, the endless variety of data science tools and production environments, and the lack of knowledge of ML within IT/Ops creating uncertainty and risk for new, mission-critical business applications.

Here are some of the discussion points you will hear about when talking with customers about their issues:

An Endless Variety of Tools and Environments - Each data scientist has their own preferred tools and languages. It is not unusual for members of a single data science team to use different notebooks, workbenches, and languages to create predictive models. This issue is only amplified when multiple data science teams try to deploy their work into production environments. These environments have their own complexities and variety from Kubernetes to Spark, Hadoop to a variation based on the companies preferred cloud provider.

Lack of Knowledge about Production Issues - On the IT/Ops side, the lack of knowledge of ML including the languages used and the behavior of ML applications means that IT is not able to properly manage the applications. Data scientists also lack knowledge on production issues and may put applications into production that do not meet the appropriate standards for production application performance, security, and reliability. This is amplified by a variety of frameworks and infrastructure options to create a nearly impossible mixture to manage for the business.

Lack of Production Monitoring – Traditional software monitoring systems were not designed for machine learning models. Typical software will show issues by consuming additional resources like memory or CPU or failing to respond to requests in a timely fashion. Machine learning models, however, can decay over time and provide increasingly inaccurate responses due to changing data conditions while still responding in a timely fashion. Traditional monitoring systems are not able to detect this data drift or other metrics that would indicate an issue with the model's predictions.

As a result of this inability to monitor model health, organizations that realize this issue may retrain models frequently to ensure that the models are current. For many, this may be a waste of resources. If the patterns in the underlying data that the model was trained on have not changed significantly, then the model will continue to perform well without retraining. Frequent retraining can become an issue as companies scale from tens to hundreds of models. If each model needs to be retrained each day, then this consumes a massive amount of computational and human resources which then prevents the teams from building and deploying new models into production.

"We just want our data scientists to do data science" - Data Scientists are a precious resource. Once machine learning-based models start moving into production situations, the only people who understand their languages and mechanics are the data scientists. Typically, the data scientists not only have to build the models but also lead the efforts to deploy, monitor, and troubleshoot models running in production environments. With only a few ML projects up and running, the data science team can find themselves spending the majority of their time on production issues rather than building new models. This is a retention issue as the data scientists are not doing what they like, and it is a scaling issue, as the number of new ML use cases delivered to the business grinds to a halt.

Lack of Governance – Regulated industries can have strict regulations about how models need to be managed when they are used in production applications. The financial services industry, for example, has numerous regulations for model risk management that require that models used to make decisions go through rigorous testing and validation before being put into production. Without these steps in place, even the best model developed by data science cannot be put into production.

Getting Smart on MLOps

MLOps is the combination of philosophy, practices, and tools that increase an organization's ability to deliver value from machine learning applications and services; providing a scalable and governed means to evolve machine learning applications rapidly in production environments. With the appropriate processes and technology in place, your teams can be more productive, deliver many more machine learning models, and manage risk and compliance.

Machine learning models deployed in production are long-lived entities that need care. MLOps provides the framework to deploy, monitor, manage and govern models running in production environments. Here are some key MLOps concepts you should know about.

- **Model Deployment** – In order to provide predictions a model must be deployed onto a production ready environment running on scalable infrastructure like Kubernetes. DataRobot provides a robust model deployment infrastructure that now uses Kubernetes for model deployment. In addition, models developed in DataRobot can be easily containerized and deployed to Kubernetes environments where they can now be monitored by the DataRobot MLOps product.

- **Monitor and Ensure Service Performance** - Anything that could impact the performance or reliability of the model must be monitored, and the right people must be informed if there are issues that need to be addressed. When looking at machine learning performance, this means that MLOps must ensure that all levels of the model infrastructure are monitored. This could be data drift between training and production data and model specific metrics that could highlight issues in model performance all the way to the infrastructure that the MLApps are running on which could become overloaded. Any of these could impact the applications SLAs with downstream business apps. DataRobot has robust monitoring for data drift, model metrics and service health.

- **Expect Continuous Improvement (Model Lifecycle Management)** – CI/CD practices are widely used in DevOps to support rapid innovation for traditional software applications. MLOps similarly need to be continuously updated due to retraining and model updates. In any case, the service must be updated without interrupting service for downstream business applications. DataRobot can seamlessly update models running in production without interrupting service.

- **Centralized Governance** – As predictive models help to make business decisions and become mission critical for business operations, oversite and security for these applications becomes vital. Production model governance focuses on the control of production models in the form of production access control, production model approval processes, audit trails for users and events and traceability of results back to a particular model. A centralized approach to MLOps is really the only way to enforce governance for models in production at scale. With multiple production ML systems, it becomes virtually impossible to ensure that access rights and approval policies are being followed uniformly.

Talking with Executives about MLOps

For most business leaders, how something works is secondary. The real key to starting an MLOps project is understanding the value that it can bring to an organization. Use this Q&A to help you have that value conversation with business leaders.

Q - How is MLOps going to help us to become an AI-driven enterprise?

A – MLOps will help you manage the production lifecycle of our AI models, which is actually a huge part that we need to get right. The majority of the life of a machine learning model is going to be spent in production, not in development. To scale your use of AI as a company, you need to be able not just to deploy a model one time, but you need to be able to see when a model needs to be updated, to

update the model without downtime, and to introduce new models that will improve the performance of business applications. All of these capabilities are part of MLOps.

Q – How does MLOps help us attract and retain data science talent?

A – You want your data scientists doing data science where they are building new models, not figuring out how to deploy their models or managing their models in production. With MLOps, you can automate significant portions of the deployment, model retraining, and management processes. You can also build a team in IT that will deploy and manage the models running in production. This will minimize the amount of time your data science team has to spend on production issues which will increase their satisfaction and make this a better place to work than companies that do not have this kind of process.

Q - Does MLOps lower our risk or help with regulatory compliance?

A – Yes, MLOps is critical to managing the risk for machine learning models in production. Financial services regulations, for example, require a complete history of how the machine learning model is being used in production applications including who is updating the model, what changes they are making and who approved the changes. Also, you will need to track what version of the model was used to make predictions at a given point in time in case there are any questions about bias in those predictions or other issues. MLOps allows you to track all of this information and provides an audit trail and connects with approval systems to make sure you are compliant.