ML PLATFORMS

Model Factory Platforms

ML scientists spend months implementing ML workflows that encapsulate individual tasks like data collection, feature engineering, model training and tuning, model performance evaluation, and deployment to production. Data access is a challenge due to the plethora of data sources and feature engineering requires repetitive compute-intensive jobs to be run to generate features like aggregates and complex embeddings. Creating an ML workflow is tedious and time-consuming because a disparate set of tools are needed to perform different tasks in the workflow. Finally, deploying the workflow in production requires setting up periodic data extraction and model execution pipelines (batch/real-time), and integrating with Amazon's systems. Building the generic platforms for productionizing ML applications is critical, as it enables users to follow the highest standards with quality applications in production, ease of maintaining the applications, reduced operational overhead, re-use of the components reducing the resource usage, and most importantly saving users time. Challenges include keeping up with the faster growth and dynamically changing ML technologies, building scalable, low-latency, high performant platforms which can help business grow, and creating solutions to enable users to adopt the new platforms with ease by integrating with Amazon's internal systems. Model Factory is a suite of platforms built internally to solve these problems of ML Scientists and Model deployment and productionization issues. Also, to cater to the community of people who are not familiar with ML, we have built a prediction service which automates the sciences and engineering behind building the ML models and productionize it without any engineering effort. This is our attempt at an Auto-ML solution.

Alki Service

There are many ML use cases inside Amazon that require Computer Vision related models, like visual similarity based product recommendation, product image quality validation and catalog attribute extraction from image. These models need to evaluate hundreds of millions of product images in a timely manner, either as real-time service or as a daily/weekly batch update, and cost significant engineering effort to build a solution. Alki Service was created to streamline the production of these models. Scientists can easily onboard a new Computer Vision model to the service, and partner teams can easily call the service, pass in information and get back results, without dealing with the complexity of ML and big data processing. We are leveraging AWS and EMR to build the solution and exploring GPU acceleration to further improve the scalability.