

Managing the end-to-end machine learning lifecycle with MLflow

BOSS'20 workshop (<https://boss-workshop.github.io/boss-2020>)

Abstract

Building and deploying a machine-learning model can be difficult to accomplish. Enabling data scientists to reproduce ML pipelines is equally challenging. Moreover, doing so can impact data science teams' productivity, leading to a significant waste of time and resources.

Getting models up to speed in the first place is significant enough that it can be easy to overlook long-term management. What does this involve in practice? In essence, we have to compare the results of different versions of ML models to track what's running where, and to redeploy and rollback updated models as needed. Each of these requires its own specific tools, and it's these changes that make the ML lifecycle so challenging compared to traditional software development lifecycle (SDLC) management.

Paulo Gutierrez and Denny Lee will walk you through MLflow, an open source platform to manage the ML lifecycle, including experimentation, reproducibility, deployment, and a central model registry.

Outline

- MLflow quick start
- Track machine learning training runs
- Log, load and deploy MLflow models
- Run MLflow Projects on Databricks
- Manage the Lifecycle of MLflow Models in MLflow Model Registry

Technology

- A laptop with 8–16 GB of memory with Chrome or Firefox browsers installed
- A pre registration for the Databricks Community Edition

List of presenters

Paulo Gutierrez

Databricks

Paulo Gutierrez is a Solutions Architect at Databricks. Before joining Databricks, he led data teams that built a petabyte-scale analytics platform for Japan's largest e-commerce site. Prior to that, he worked on network management systems and other broadcast network infrastructure solutions. Paulo contributes to open source projects through code, documentation, feature requests and discussions. Knows a little too much about coffee, cameras and mechanical keyboards. Born and raised in the Philippines, living in Japan.

Denny Lee

Databricks

Denny Lee is a Staff Developer Advocate at Databricks. He is a hands-on distributed systems and data sciences engineer with extensive experience developing internet-scale infrastructure, data platforms, and predictive analytics systems for both on-premise and cloud environments. He also has a Masters of Biomedical Informatics from Oregon Health and Sciences University and has architected and implemented powerful data solutions for enterprise Healthcare customers.

Website

<https://databricks.com/>

<https://mlflow.org/>