databricks

Model Deployment Fundamentals

LECTURE

Model Deployment Strategies

This lecture introduces machine learning model deployment, its place in the full ML lifecycle, and the different deployment modes used in practice

Machine Learning **Model Deployment**

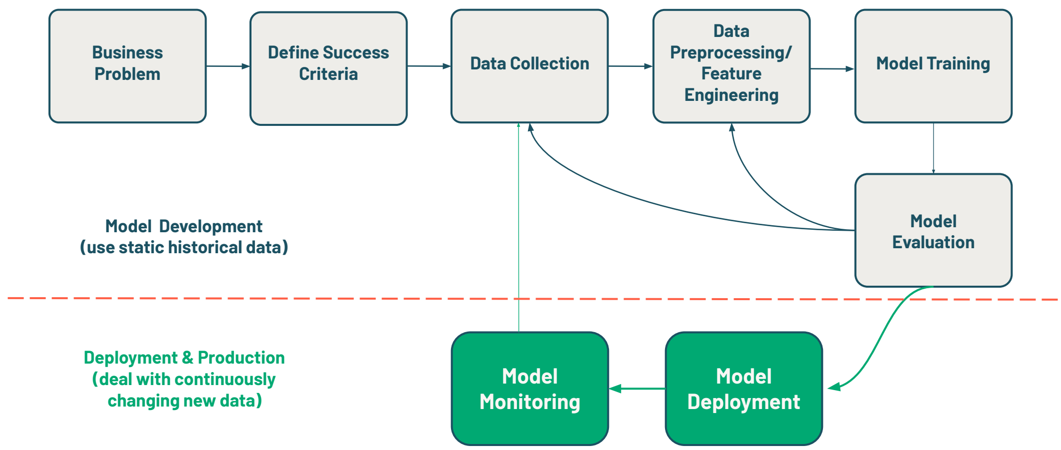
The process of integrating a machine learning model into a **production environment**, making it accessible for end-users or other systems to **generate predictions** or insights.

(Deployment Strategies: batch, streaming, real-time, or embedded/edge)

Model deployment is the process of taking a trained machine learning model and integrating it into a production environment, typically to support a business-critical application. This allows the model to generate predictions in batch, streaming, or real-time mode so that downstream systems or users can gain actionable insights.

The Machine Learning Full Lifecycle

Your Model is ready, then what?



When working with machine learning, there are several steps you need to go through to transition from your initial business problem to a deployed model capable of making predictions.

The first step is defining your business problem. What is the goal you wish to achieve with machine learning? Once you have a clear understanding of the problem, you can proceed to define success criteria. Next, you will collect data to train your model. This data must be cleaned and preprocessed to ensure effective use. Once you have preprocessed data, you can move on to training the model. After training, it is crucial to evaluate its performance using a validation set. This helps you gauge how well the model is performing and identify any issues that need addressing.

Once you have a well-performing model, the next challenge is deploying it into production. Deploying a machine learning model can be challenging, as it involves managing the infrastructure for deployment and ensuring continued performance over time. Ongoing monitoring of the model's performance and making necessary adjustments are essential to maintain its effectiveness.

Model Deployment Modes

Comparing methods

Deployment Method	Throughput	Latency	Example Application
Batch	High	High (*hours to days*)	Periodic customer churn prediction
Streaming	Moderate	Moderate (*seconds to minutes*)	Dynamic pricing application
Real-time	Low	Low (*milliseconds*)	Recommendation systems Chatbot
Edge/Embedded	Low	Low (Dependent on device processing power)	IoT Applications. Farm sensor for detecting humidity

When deciding on a deployment method, two key factors to consider are throughput—the number of data points the model needs to process—and latency—how quickly predictions are needed. Batch deployment is suitable for high throughput with low urgency, like customer churn predictions updated daily or weekly. Streaming deployment works well for moderate throughput and faster responses, such as dynamic pricing updates on e-commerce or travel platforms. Real-time deployment handles low throughput but requires immediate responses, like chatbots or recommendation systems reacting instantly to user actions.

In IoT or edge computing scenarios, models often run directly on devices to minimize data transfer and latency. Here, throughput is very low since each device handles its own sensor data, but response time must be extremely fast. Edge models are typically optimized to use minimal memory and power, ensuring efficient, quick predictions without relying on network connectivity.

Thank you for completing this lesson and continuing your journey to develop your skills with us.