Menu

Upload your PDFs here and click on 'Process'

Drag and drop files here

Limit 200MB per file

Browse files

Submit and Process

Chat with Multiple PDFs

Ask a question about your documents:

give me the detailed summary of ML deployment patterns

Replay: There are several deployment patterns for real-time ML models:

- **A/B testing:** Deploy multiple model versions simultaneously, distributing traffic among them to evaluate performance. The best-performing model is then selected based on predefined criteria (e.g., accuracy, conversion rates). A/B testing typically uses fixed traffic splits, unlike gradual rollouts.
- **Gradual rollout (Canary deployment):** A new model version is initially exposed to a small segment of traffic. Performance is closely monitored, and if successful, traffic to the new model is gradually increased. This allows for continuous exposure to real traffic while providing a safety net to scale back if issues arise. Gradual rollout adjusts traffic dynamically based on performance, unlike A/B testing.
- **Shadow deployment:** A new model version runs alongside the existing one but doesn't actively serve traffic. It receives a copy of incoming traffic and generates predictions, which are compared to the current version for evaluation without affecting users. This offers a risk-free comparison but involves extra work as multiple predictions are generated for the same input.

These patterns are not mutually exclusive and can be combined. For example, start with shadow deployment for risk-free evaluation and then proceed with a gradual rollout based on the findings. Using model aliases like "Champion" and "Challenger" helps manage different versions and switch traffic easily. Databricks Model Serving facilitates these patterns by allowing traffic splitting and monitoring performance metrics of different model versions.