Training Course on Data Versioning and Experiment Tracking for ML

Professional Development Training Course Outline

Category	Data Science	**Duration**	10 Days
Base Fee	$3,000 USD	**Delivery Mode**	Online / On-site Hybrid

Course Introduction

In the rapidly evolving landscape of Machine Learning (ML) and Artificial Intelligence (AI), achieving **reproducibility** and **traceability** is paramount for robust and reliable model development. Training Course on Data Versioning & Experiment Tracking for ML: Tools like DVC, MLflow for Reproducible Research delves into the critical practices of **data versioning** and **experiment tracking**, empowering data scientists, ML engineers, and researchers to effectively manage their ML workflows. We will explore industry-standard tools like **DVC (Data Version Control)** and **MLflow** to streamline operations, enhance collaboration, and ensure the integrity of your ML projects, from data ingress to model deployment.

The increasing complexity of ML models, coupled with the iterative nature of experimentation, necessitates systematic approaches to version control for datasets, code, configurations, and trained models. This course will equip participants with the practical skills to implement **MLOps best practices**, mitigate **model drift**, and establish a foundation for **responsible AI** development. By mastering these techniques, organizations can significantly improve the efficiency, transparency, and trustworthiness of their ML initiatives, accelerating innovation and delivering production-ready models with confidence.

Programme Curriculum

Training Course on Data Versioning & Experiment Tracking for ML: Tools like DVC, MLflow for Reproducible Research

Introduction

In the rapidly evolving landscape of Machine Learning (ML) and Artificial Intelligence (AI), achieving **reproducibility** and **traceability** is paramount for robust and reliable model development. Training Course on Data Versioning & Experiment Tracking for ML: Tools like

DVC, MLflow for Reproducible Research delves into the critical practices of **data versioning** and **experiment tracking**, empowering data scientists, ML engineers, and researchers to effectively manage their ML workflows. We will explore industry-standard tools like **DVC (Data Version Control)** and **MLflow** to streamline operations, enhance collaboration, and ensure the integrity of your ML projects, from data ingress to model deployment.

The increasing complexity of ML models, coupled with the iterative nature of experimentation, necessitates systematic approaches to version control for datasets, code, configurations, and trained models. This course will equip participants with the practical skills to implement **MLOps best practices**, mitigate **model drift**, and establish a foundation for **responsible AI** development. By mastering these techniques, organizations can significantly improve the efficiency, transparency, and trustworthiness of their ML initiatives, accelerating innovation and delivering production-ready models with confidence.

Course Duration

10 days

Course Objectives

Upon completion of this course, participants will be able to:

1. Grasp the fundamental concepts and importance of **reproducible machine learning** and **traceable AI workflows**.
2. Implement robust strategies for **data version control** using **DVC** to manage large datasets and track changes effectively.
3. Utilize **MLflow** for comprehensive **ML experiment tracking**, including parameters, metrics, code versions, and artifacts.
4. Seamlessly combine **DVC for data management** and **MLflow for experiment lifecycle management** in a unified MLOps pipeline.
5. Ensure that **ML models** can be consistently **reproduced** and validated across different environments and team members.
6. Leverage MLflow's Model Registry for **model versioning**, staging, and production deployment, facilitating **model governance**.
7. Track and compare the results of **hyperparameter optimization** experiments to identify optimal model configurations.
8. Design and implement **automated ML pipelines** incorporating data versioning and experiment tracking for efficiency.
9. Enhance **team collaboration** by establishing shared, versioned data and reproducible experiment results.
10. Efficiently **debug ML models** and **rollback to previous versions** of data, code, or models when issues arise.
11. Apply best practices for **data integrity** and **data quality management** within versioned datasets.
12. Understand how data versioning and experiment tracking contribute to **model drift detection** and mitigation.
13. Contribute to the development of **ethical AI** and **transparent machine learning** by ensuring full traceability and accountability.

Organizational Benefits

- Streamlined workflows lead to faster iteration and deployment of ML models.

- Enhanced reproducibility reduces errors and ensures consistent model performance.
- Efficient experiment tracking and data management minimize wasted resources and redundant efforts.
- A unified approach to versioning and tracking fosters seamless teamwork.
- Comprehensive record-keeping supports regulatory requirements and audit trails.
- Transparency and reproducibility build confidence in ML-driven decisions.
- Clear historical records enable quicker understanding of past experiments and data.
- Documented experiments and versioned data facilitate knowledge sharing across the organization.
- A robust MLOps foundation enables faster adoption of new ML techniques and technologies.

Target Audience

1. Data Scientists.
2. Machine Learning Engineers.
3. AI/ML Researchers.
4. MLOps Practitioners.
5. Data Engineers.
6. Team Leads and Project Managers (ML/AI)
7. Software Engineers working with ML.
8. Anyone interested in building scalable and reliable ML solutions.

Course Outline

Module 1: Introduction to Reproducible Machine Learning

- Understanding the "Reproducibility Crisis" in ML/AI.
- Defining reproducibility, replicability, and generalizability in ML.
- Challenges in achieving reproducibility: data, code, environment, randomness.
- The importance of MLOps for reliable ML lifecycle management.
- Case Study: A pharmaceutical company facing challenges in replicating drug discovery ML models due to undocumented data changes.

Module 2: Fundamentals of Version Control (Git Refresher for ML)

- Brief overview of Git for code versioning.
- Common Git commands essential for ML projects (commit, branch, merge).
- Why Git alone is insufficient for large datasets and model artifacts.
- Strategies for organizing ML projects with Git.
- Case Study: A data science team struggling with multiple versions of feature engineering scripts leading to inconsistent model performance.

Module 3: Introduction to Data Version Control (DVC)

- What is DVC? Key features and architecture.
- DVC for managing large files and datasets outside of Git.
- Installing and initializing DVC in an ML project.
- dvc add, dvc push, dvc pull for data versioning.
- Case Study: A financial institution needing to track changes in customer transaction data used for fraud detection models.

Module 4: Advanced DVC for Data Pipelines

- DVC stages and pipelines for reproducible data transformations.
- Defining dependencies and outputs in DVC pipelines.
- Automating data preprocessing and feature engineering with DVC.
- Managing DVC remotes (S3, GCS, Azure Blob, local).
- Case Study: An e-commerce company building a recommendation system, requiring reproducible data pipelines from raw clicks to aggregated features.

Module 5: DVC for Model and Experiment Artifacts

- Versioning trained models and evaluation metrics with DVC.
- Tracking configuration files (params.yaml) for experiments.
- dvc metrics and dvc plots for comparing experiment results.
- Branching and merging DVC experiments.
- Case Study: An autonomous vehicle company needing to version different iterations of their perception models and associated sensor data.

Module 6: Introduction to MLflow Tracking

- What is MLflow? Overview of its components (Tracking, Projects, Models, Registry).
- Setting up an MLflow tracking server (local and remote).
- Logging parameters, metrics, and artifacts with mlflow.log_param, mlflow.log_metric, mlflow.log_artifact.
- Organizing runs into experiments.
- Case Study: A marketing analytics firm wanting to compare the performance of various customer segmentation models across different parameter sets.

Module 7: Advanced MLflow Tracking and UI

- Programmatic access to MLflow runs and experiments.
- Comparing multiple MLflow runs in the UI.
- Tagging runs for better organization and searchability.
- Visualizing experiment results using MLflow UI.
- Case Study: A healthcare provider tracking hundreds of machine learning experiments for disease prediction, needing a unified view for analysis.

Module 8: MLflow Projects for Reproducible Code

- Defining MLflow Projects for packaging ML code.
- Reproducible execution of ML projects using mlflow run.
- Environment management for MLflow Projects (Conda, Docker).
- Integrating external dependencies within MLflow Projects.
- Case Study: A research institution sharing ML code with collaborators, ensuring consistent execution environments and dependencies.

Module 9: MLflow Model Registry for Model Management

- Concept of Model Registry and its importance in MLOps.
- Registering, versioning, and transitioning models between stages (Staging, Production).
- Model governance and approval workflows with MLflow Registry.
- Loading and deploying models from the Model Registry.
- Case Study: A tech company managing multiple versions of a fraud detection model in production, requiring clear staging and approval processes.

Module 10: Combining DVC and MLflow for End-to-End Reproducibility

- Strategies for integrating DVC and MLflow in a single ML workflow.
- Logging DVC-versioned data references in MLflow runs.
- Ensuring traceability from raw data to deployed model.
- Building an end-to-end reproducible ML pipeline with both tools.
- Case Study: An AI startup building a new language model, requiring full traceability of training data, model architectures, and experiment results.

Module 11: Best Practices for Reproducible ML Workflows

- Structuring ML projects for reproducibility.
- Documenting experiments and decision-making processes.
- Managing random seeds for deterministic results.
- Strategies for handling large-scale data and compute.
- Case Study: A data science team struggling with "tribal knowledge" regarding their ML models, leading to difficulties in onboarding new members and maintaining existing systems.

Module 12: Advanced Topics: Model Drift and Data Shifts

- Understanding model drift and its impact on performance.
- Detecting data shifts and concept drift using versioned data.
- Leveraging DVC and MLflow for monitoring and alerting on data/model issues.
- Strategies for retraining and redeploying models to mitigate drift.
- Case Study: A retail company whose sales forecasting model performance degrades over time due to changing consumer behavior, necessitating proactive drift detection.

Module 13: Ethical AI and Responsible ML with Versioning

- The role of reproducibility in ethical AI.
- Tracking data provenance and bias in datasets.
- Documenting model fairness and transparency experiments.
- Auditing ML systems for compliance and accountability.
- Case Study: A lending institution facing scrutiny over potential bias in their loan approval ML model, requiring robust documentation of data sources and model decisions.

Module 14: Deployment and Monitoring Considerations

- Brief overview of model deployment strategies (e.g., Docker, Kubernetes).
- Integrating DVC and MLflow into CI/CD pipelines.
- Monitoring deployed models with MLflow.
- Rollback strategies for problematic deployments.
- Case Study: A SaaS company deploying a new ML-powered feature, needing a reliable way to deploy, monitor, and quickly revert if issues arise in production.

Module 15: Future Trends and Ecosystem

- Emerging tools and technologies in MLOps, data versioning, and experiment tracking.
- The role of cloud platforms (AWS Sagemaker, GCP AI Platform, Azure ML) in ML lifecycle management.
- Open-source vs. commercial solutions.
- Continuous learning and staying updated in the MLOps landscape.

- Case Study: Exploring how a fast-growing tech company evaluates and adopts new MLOps tools to scale their ML operations.

Training Methodology

This training course will adopt a highly interactive and practical methodology, combining theoretical concepts with extensive hands-on exercises and real-world case studies.

- **Interactive Lectures:** Engaging presentations introducing core concepts, tools, and best practices.
- **Live Demonstrations:** Step-by-step walkthroughs of DVC and MLflow functionalities.
- **Hands-on Labs:** Practical exercises and coding sessions to reinforce learning and build proficiency. Participants will work with realistic ML projects to apply learned concepts.
- **Case Study Analysis:** Discussions and problem-solving based on industry-relevant scenarios to illustrate the application of data versioning and experiment tracking in diverse contexts.
- **Group Discussions and Q&A:** Facilitating peer learning, sharing of experiences, and addressing specific challenges.
- **Best Practices and Pitfalls:** Highlighting common mistakes and strategies for successful implementation.

Register as a group from 3 participants for a Discount

Send us an email: info@fineskilltrainingcenter.org or call *+254769199797*

Certification

Upon successful completion of this training, participants will be issued with a globally- recognized certificate.

Tailor-Made Course

We also offer tailor-made courses based on your needs.

Key Notes

a. The participant must be conversant with English.

b. Upon completion of training the participant will be issued with an Authorized Training Certificate

c. Course duration is flexible and the contents can be modified to fit any number of days.

d. The course fee includes facilitation training materials, 2 coffee breaks, buffet lunch and A Certificate upon successful completion of Training.

e. One-year post-training support Consultation and Coaching provided after the course.

f. Payment should be done at least a week before commence of the training, to Fineskill Training Center account, as indicated in the invoice so as to enable us prepare better for you.

Upcoming Scheduled Sessions

Start Date	End Date	Location	Price
23 Nov 2026	04 Dec 2026	Online	$2,000
23 Nov 2026	04 Dec 2026	Physical	$3,000
23 Nov 2026	04 Dec 2026	Physical	$6,000
23 Nov 2026	04 Dec 2026	Physical	$6,000
23 Nov 2026	04 Dec 2026	Physical	$12,000
23 Nov 2026	04 Dec 2026	Physical	$9,000
23 Nov 2026	04 Dec 2026	Physical	$12,000
23 Nov 2026	04 Dec 2026	Physical	$12,000

Ready to enroll or request a customized program? Book online or contact us:

Register Online Now

Email: info@fineskilltrainingcenter.com | Website: www.fineskilltrainingcenter.com