ADA Presentation

Topic: Tackling Version Management in MLOps

DVC: https://dvc.org

Things to do:
Introduction to MLOps
Introduction to Version Controlling (and types)
Introduction to Data Version Controlling
Why do we need Data Version Controlling
Applications
Introduction to DVC library
Code Demo

We just need to follow: https://staff.fnwi.uva.nl/a.s.z.belloum/LiteratureStudies/Reports/2020-Internship_report-Yizhen.pdf

References:

- MLOps and data versioning in machine learning project: This is an industrial Internship Report. We are going to base most of our report on this.

https://staff.fnwi.uva.nl/a.s.z.belloum/LiteratureStudies/Reports/2020-Internship_report-Yizhen.pdf

- Tackling Version Management and Reproducibility in MLOps: This is a book on MLOPs and Version Management.

https://repositorio-aberto.up.pt/bitstream/10216/152181/2/636962.pdf

- MLOps Scaling ML in an Industrial Setting: Another industrial internship report, contains relevant info on MLOps

https://staff.fnwi.uva.nl/a.s.z.belloum/MSctheses/MScthesis_Yizhen_Zhao.pdf

References to include in the report:

- On the Co-evolution of ML Pipelines and Source Code - Empirical Study of DVC Projects: https://mcis.cs.queensu.ca/publications/2021/saner.pdf
- Data best practices and case studies: https://guides.library.stanford.edu/data-best-practices/version-files
- Creating reproducible data science workflows with DVC: https://medium.com/y-data-stories/creating-reproducible-data-science-workflows-with-dvc-3bf058e9797b
- Software Engineering for Machine Learning: A Case Study: https://www.microsoft.com/en-us/research/uploads/prod/2019/03/amershi-icse-2019_Software_Engineering_for_Machine_Learning.pdf
- Tackling Version Management and Reproducibility in MLOps: This is a book on MLOPs and Version Management: https://repositorio-

aberto.up.pt/bitstream/10216/152181/2/636962.pdf

Papers:

- On the Co-evolution of ML Pipelines and Source Code - Empirical Study of DVC Projects: https://mcis.cs.queensu.ca/publications/2021/saner.pdf
- Software Engineering for Machine Learning: A Case Study: https://www.microsoft.com/en-us/research/uploads/prod/2019/03/amershi-icse-2019_Software_Engineering_for_Machine_Learning.pdf
- Demystifying MLOps and Presenting a Recipe for the Selection of Open-Source Tools: https://www.mdpi.com/2076-3417/11/19/8861

GPT Prompts:

I want to prepare some slides (maybe 8-10 slides) on the topic "**Tackling Version Management in MLOps using DVC**".

I want the following content in the slides:
Introduction to MLOps (and why do we need MLOps), Introduction to Version controlling and types of version controls in Machine Learning Projects (project version control, data version control, model version control, deployment based version control etc.), Introduction to Data Version Controlling (and why do we need it), applications in industry, challenges faced, what is DVC, demo problem and solution offered using DVC (the demo problem is a toy demonstration of changing the dataset, by changing the random seed in train test split, and tracking the changes using DVC), Conclusion

Prepare a detailed blueprint for the slides on a per page and per topic basis based on the topic of "**Tackling Version Management in MLOps using DVC**" using the above described specifications (content).

Reply (BARD):

Slide 1: Title Slide
Title: Tackling Version Management in MLOps using DVC
Subtitle: Leveraging Data Version Control for Reproducible and Reliable Machine Learning Pipelines
Presenter: [Your Name]

Slide 2: Introduction to MLOps
What is MLOps?

- Machine Learning Operations (MLOps) is a set of practices that aim to

bridge the gap between machine learning development and deployment.
- It encompasses the entire lifecycle of machine learning models, from ideation to production, ensuring their continuous improvement and effectiveness.

Why do we need MLOps?
- Increasing complexity of ML models and pipelines
- Need for reproducibility and reliability
- Ensuring compliance and governance

Slide 3: Introduction to Version Controlling

What is version control?
- Version control is a system that tracks changes to code and other data over time, allowing users to revert to previous versions if necessary.

Types of version control in Machine Learning Projects
- Project version control: Tracks changes to code and configuration files
- Data version control: Tracks changes to data used for training and evaluation
- Model version control: Tracks changes to machine learning models
- Deployment-based version control: Tracks changes to deployment configurations

Slide 4: Introduction to Data Version Controlling

What is data version control?
- Data version control is a critical component of MLOps that ensures the reproducibility and traceability of machine learning models.
- It allows users to track changes to data used for training and evaluation, enabling them to compare different models and understand their impact on model performance.

Why do we need data version control?
- Ensures reproducibility of ML models
- Facilitates collaboration and experimentation
- Enables model lineage tracking and impact analysis
- Improves model performance and robustness

Slide 5: Applications of Data Version Controlling in Industry
- Fraud detection and risk assessment
- Customer segmentation and targeting
- Predictive maintenance and anomaly detection
- Supply chain optimization and demand forecasting
- Personalized recommendation and engagement

Slide 6: Challenges Faced in Data Version Controlling
- Managing large datasets

- Integrating with existing workflows
- Automating data versioning processes
- Ensuring data quality and integrity
- Addressing data privacy and security concerns

Slide 7: What is DVC?
DVC: Data Version Control

- DVC is an open-source data version control tool specifically designed for machine learning projects.
- It seamlessly integrates with Git, leveraging its powerful versioning capabilities.
- DVC provides a user-friendly interface and automates data versioning processes.

Slide 8: Demo Problem: Changing Dataset by Modifying Random Seed
Demo Problem:

- Create a simple classification model using a random seed for train-test split.
- Modify the random seed and observe the impact on model performance.
- Track changes to the dataset and model using DVC.

Slide 9: Solution Offered Using DVC
DVC Solution:

- Track changes to the dataset and random seed using DVC.
- Easily reproduce different versions of the dataset and model.
- Compare model performance across different data versions.

Slide 10: Conclusion
Key Takeaways:

- Data version control is essential for reproducible and reliable machine learning.
- DVC provides a powerful and easy-to-use tool for data version control in ML projects.
- Adopting DVC can significantly improve the efficiency and effectiveness of ML development and deployment.