Politecnico di Milano – Master Universitario di II livello

Artificial Intelligence & Data Engineering

PROVA D'ESAME MODULO 5

aws

MLOps Engineering on AWS
Workbook

Predicting House Prices with Advanced Regression Techniques and Simulating MLOps Practices

Overview:

In this assignment, your team will collaborate to solve the Kaggle competition "House Prices - Advanced Regression Techniques." The objective is to predict house prices based on a variety of features using machine learning models. As part of this project, you will simulate key MLOps (Machine Learning Operations) practices, including data and model versioning, pipeline automation, and local deployment testing using containers.

https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques

You are required to set up an automated workflow with GitHub Actions that will build and push the container to Docker Hub upon any changes to the repository. The model's container can be tested locally before final deployment. This assignment will give you hands-on experience in deploying machine learning models within a production-like environment.

Team Roles and Responsibilities:

Each member of your team will assume specific roles that align with an MLOps workflow, based on the MLOps workbook framework. These roles will ensure that different parts of the machine learning lifecycle are covered, from data preparation to deployment and monitoring:

- Data Engineer: Manages data preparation and ensures that different versions of the dataset are tracked using DVC.
- Data Scientist: Selects the best machine learning models, performs feature engineering, and leads the experimentation phase.

- ML Engineer: Automates the pipeline to ensure that updates in data or code trigger the necessary processes like retraining and model evaluation.
- MLOps Engineer: Focuses on containerizing the final model, setting up a local deployment, and ensuring integration with GitHub Actions for automated builds and container publishing.
- DevOps Engineer: Assists with infrastructure setup, managing the container workflow, and integrating Docker Hub for deployment.

Each team member must clearly document their responsibilities, outlining how they contributed to each stage of the machine learning lifecycle. This will ensure accountability and streamline collaboration within the group.

Simulating MLOps Practices:

As part of the MLOps simulation, your team will use DVC for version control of datasets and models, while Git will be used for code collaboration. Key MLOps practices that will be implemented include:

- Data Versioning: The Data Engineer will manage dataset versioning using DVC. Any updates in data preprocessing or feature engineering should be tracked with DVC.

Local Deployment and Testing:

The final model will be deployed locally using a containerized approach. The MLOps Engineer will containerize the selected model based on the approach outlined in the Amazon SageMaker scikit_bring_your_own/container example:

https://github.com/aws/amazon-sagemaker-examples/tree/main/advanced_functionality/scikit_bring_your_own/container

You will prepare a container that serves predictions and can be tested locally. The container will simulate a production environment and allow for end-to-end testing.

Testing the container can be performed locally using Docker, where the containerized model can serve predictions. This will allow you to ensure the model functions correctly in a simulated production-like setting before finalizing the deployment process.

Action Plan for Monitoring and Continuous Improvement:

Throughout the project, you will follow an action plan to monitor model performance and ensure continuous improvement. The team should identify areas for improvement in the workflow and simulate monitoring to track model performance over time. This could include tracking metrics related to model accuracy or latency, as well as identifying any data drift or changes in model performance that might indicate the need for retraining.

You will also need to ensure that the automated workflows in GitHub Actions and Docker Hub integration are running smoothly, so any potential errors in building or publishing containers can be caught early and addressed.

Submission Requirements:

Your final submission will include the following deliverables:

1. GitHub Repository: A link to your team's GitHub repository, which should include:
   - All relevant scripts for data preprocessing, model training, and evaluation.
   - The DVC pipelines for data and model versioning.

2. Docker Hub Repository or ECR Repository: A link to your team's docker registry account, where the containerized model is hosted.

4. Reflection Document: A 1–2-page document reflecting on your workflow, the challenges you encountered, and how you managed collaboration using Git, DVC, and CI/CD Pipeline. Discuss areas where you identified gaps in the workflow and how you improved them.

Assessment Criteria:

You will be assessed based on the following criteria:

- MLOps Practices: How effectively your team simulated the full MLOps lifecycle, including data and model versioning, pipeline automation, and containerized deployment using GitHub Actions and Docker Hub.
- Automation: Your ability to successfully set up GitHub Actions to automate the Docker build and push process, ensuring continuous integration and delivery of your model container.
- Model Selection and Performance: The rigor of your model selection process, including how well you experimented with different models, tuned hyperparameters, and chose the best-performing model.
- Collaboration and Workflow: The effectiveness of your team's collaboration using Git, DVC, and other tools. Each team member's role and contribution to the workflow should be clearly documented.
- Reflection and Improvements: A thoughtful reflection on the challenges you faced during the project and the steps your team took to improve the workflow and ensure smooth collaboration.

By completing this assignment, your team will gain experience in solving a real-world machine learning problem while simulating a full MLOps pipeline. This will include the integration of data and model versioning, pipeline automation, and containerized deployment using modern tools such as GitHub Actions and Docker Hub.

Introduction to MLOps

1. Activity – Communication

The purpose of this activity is to strengthen the machine learning operations (MLOps) communications in your environment. You will create a table to document your MLOps stakeholders.

Fill out the following table. For each of the idealized roles on the left (rows), look across the ML lifecycle stages (columns). Write the job title or the name of the person or people in your organization that are responsible for each kind of work.

If you are not sure who has these responsibilities in your environment, you might want to investigate and continue to completing the table after this course. You might discover other roles that are important in your environment, and you can use the "other" row for those.

Example:

Roles	Data preparation	Model build	Model evaluation and selection	Deployment	Monitoring
Security engineer	*Mateo Jackson* *Data Security Analyst*			*Terry Whitlock* *DevSecOps Manager*	

Roles	Data preparation	Model build	Model evaluation and selection	Deployment	Monitoring
Business stakeholder					
Data engineer	Nicola Zambelli Tommaso Pagliari				
Data scientist	Jacopo Di Prima Emanuele Eusepi	Jacopo Di Prima Emanuele Eusepi	Jacopo Di Prima Emanuele Eusepi		
ML engineer		Jacopo Di Prima Emanuele Eusepi	Jacopo Di Prima Emanuele Eusepi	Jacopo Di Prima Emanuele Eusepi	
DevOps engineer				Nicola Zambelli Tommaso Pagliari	Nicola Zambelli Tommaso Pagliari
MLOps engineer				Nicola Zambelli Tommaso Pagliari	Nicola Zambelli Tommaso Pagliari
Governance officer					
Model approver					
Security engineer					
Other					

2. **Where do you see overlapping interests and responsibilities? What actions should you take to balance the interests and coordinate efforts?**

   Do you see any gaps in the table? What problems could those gaps create? What actions should you take to mitigate these risks?

   The Data Engineer and Data Scientist both work on data preparation and feature engineering. To avoid overlap, the Data Engineer can focus on data cleaning and versioning, while the Data Scientist handles feature selection and engineering.

Action plan:

Consider the activities in this section of the workbook. List any action items you want to complete to advance MLOPs practices within your organization.

Invented Scenario

Action / outcome description	Stakeholders	Considerations
Automate Pipeline with GIT actions	MLOps Engineer	
Data versioning with DVC	Data eng	
Monitor Model Performance	Data Scientist, ML Eng	
Documentation and Communic.	all	

Invented Scenario

MLOps maturity level - initial

During module two, you learned about providing experimentation environments for data scientists. Data scientists use these environments for exploring data and creating new ML models. Providing standard experimentation environments promotes consistency in the data analysis and model creation steps of the ML lifecycle.

This section of the workbook helps you identify opportunities to implement what you've learned to set up experimentation environments in your organization.

3. **How are experimentation environments created in your organization? How could this practice be optimized?**

   In the past, our organization set up experimentation using local virtual machines or individual containers for each data scientist. This method often caused inconsistencies and took a lot of time to set up.

   To improve this, we can switch to using Amazon SageMaker on AWS. By creating standardized, pre-configured SageMaker notebooks, we can ensure consistency and cut down on setup time

4. **To standardize experimentation environments, you might need to centralize approved model building resources. Which algorithms, frameworks, and libraries does your organization use? Note any custom-built components.**

Training algorithms	ML frameworks	Libraries

• Linear Regression • Decision Trees • Random Forest • Neural Networks	• Scikit-learn • TensorFlow	• NumPy • Pandas • Matplotlib • Seaborn

5. **Consider the components listed in the previous question. Which methods of using Amazon SageMaker to implement training and inference support your existing tools?**

- ☐ Use container images managed by AWS with built-in algorithm
- ☑ Use container images managed by AWS with your own algorithm in script mode
- ☐ Bring your own container (BYOC) with a custom ML framework
- ☐ Extend pre-built container with custom dependencies
- ☐ Extend pre-built container with requirements.txt
- ☑ Bring your own model (The model is built on premises and then brought to the cloud for hosting.)
- ☐ Bring a third-party model (such as from AWS Marketplace.)

Action plan:

Consider the activities in this section of the workbook. List any action items you want to complete to advance MLOPs practices within your organization.

Action / outcome description	Stakeholders	Considerations
Standardize Exp Environment	Data Scientist	
Implement Version Control	Data Eng	Git
Documentation Practices		

MLOps maturity level – repeatable

After the data scientists are familiarized with SageMaker experimentation, the next step is to automate processes for model training and deployment. At the Repeatable level of MLOps maturity, you create ML pipelines for each step of the ML lifecycle.

6. What DevOps practices that your organization already uses can be adopted for MLOps? What MLOps practices will be new or different?

For MLOps, we'll need to adopt some new practices. These include using tools like DVC for data versioning, setting up automated model training and deployment pipelines with Amazon SageMaker, and continuously monitoring model performance to catch data drift and other issues. Additionally, we'll need to implement strong model governance practices to track model lineage and ensure reproducibility

7. Activity – From DevOps to MLOps

Fill out the following table. The purpose of this activity is to familiarize yourself with existing conditions in your environment, from an MLOps perspective. This table will help you verify the tools and processes that you are using, who is responsible, and potential opportunities to further improve MLOps in your environment.

Operations tools and processes	Do you have this? [Y/N]	Who is responsible for it?	If you do not have it, what are your next steps (action plan)?
Code building pipeline	Y	DevOps Eng	
Code version control	Y	DevOps Eng	Implement Git
Model version control	N	Data Scientist	Implement Sagemakers Registry
Data version control	N	Data Engineer	Set up DVC
Model building pipeline	N	ML Engineer	Automated pipeline with github actions and sagemaker
Data pipeline	Y	Data Eng	
Approval process	N	Leader (?)	

8. How does your team validate that a new or updated model meets quality standards and functions properly before making it available for deployment?

We start by running unit tests on individual components of the model to ensure they function correctly. Next, we perform integration testing to verify that all components work together, including data pipelines, feature engineering steps, and the model itself. We evaluate the model's performance using predefined metrics (R-squared, RSS) to determine if it meets the required standards.

We also test the model with stakeholders to make sure it meets business needs. Before using the model, we set up tools to monitor its performance and catch any problems early.

9. Of the different ways to implement inference that were discussed, which are most like those you use in production?

- ☐ Real-time
- ☐ Serverless
- ☑ Asynchronous
- ☐ Batch transform

Action plan:

Consider the activities in this section of the workbook. List any action items you want to complete to advance MLOPs practices within your organization.

Action / outcome description	Stakeholders	Considerations
Automate Model training and deployment	MLOPS Engineers	
Implement CI/CD	DevOps Eng, MLOps Eng	
Implement Monitoring	ML Engineers	

MLOps maturity level – reliable

After your repeatable processes and pipelines are in place, the next level of MLOps maturity focuses on reliability.

10. Which deployment strategies are relevant for your organization?

- ☑ A/B testing
- ☐ Shadow testing
- ☐ In-place deployment
- ☑ Blue/green deployment: all-at-once, canary, linear
- ☐ Rolling deployment

11. Monitoring solutions

Fill out the following table to illustrate your organization's monitoring solutions. Consider how your infrastructure and model metrics impact your business KPIs. Also, consider how these metrics can alert you to the need for remediation.

Metric types	Owner	Monitoring tools	Types of alerts
Business KPIs	Business Analyst	PowerBI	Revenue impact, Coustomer satisfaction
Infrastructure performance and health	Dev Ops Eng	AWS Cloudwatch	CPU/memory usage, Network latency
Data and model performance	Data Scientists, ML Eng	AWS Cloudwatch	Data drift, accuracy

12. **When the monitoring solutions in the previous table detect an issue, an automated remediation or notification should start. Use the space to follow to design one or more remediation paths.**

If model accuracy drops below a certain threshold, an alert is sent to the Data Scientist and ML Engineer. This triggers an automatic retraining job using the latest data in Amazon SageMaker. The new model's performance is then validated the baseline. If it performs better, it's deployed to production. If not, the current model stays in place, and the team is notified for further investigation

When customer satisfaction scores drop significantly, an alert is sent to the Business Analyst and Customer Support team. Power BI automatically generates a report to highlight potential causes. A possible development could be a meeting with stakeholders to discuss the findings and develop an action plan. The impact of the changes is monitored, and strategies are modified as needed.

Action plan:

Consider the activities in this section of the workbook. List any action items you want to complete to advance MLOPs practices within your organization.

Action / outcome description	Stakeholders	Considerations
Advanced Monitoring & Alerting	ML Engineers	Detect data drifts
Audits and reviews	Leads	
Disaster recovery plans	DevOps Engineers	Develop and test

MLOps maturity level – scalable

As your organization's use of ML matures, you will need a more scalable way to onboard additional ML use cases and teams to your ML environment. SageMaker supports options such as customized SageMaker project templates, separate MLOps accounts per team and environment, and shared accounts for centralized, reusable resources.

You can learn more about these options at https://aws.amazon.com/blogs/machine-learning/mlops-foundation-roadmap-for-enterprises-with-amazon-sagemaker/