DevOps for Machine Learning

Safir Mohammad Mustak Shaikh

Agenda

Motivation

- Knowledge Gap between Data Scientists and Operations Team
- Evolving Environment
- Model Drifting
  - Data Drifting
  - Concept Drifting
- Case Study [1]
  - A chatbot released by Microsoft for Twitter
  - Within a few hours, the bot learned not only the language from people but also their values
  - It started tweeting highly offensive things

Introduction

- Goal
- Why DevOps?
- What is MLOps?
- DevOps vs. MLOps
- Elements on an ML system
- Data Science steps in an ML system

MLOps Levels

- Level 0: Manual Process
- Level 1: Automating ML Pipeline
- Level 2: Automating CI/CD Pipeline

MLOps Level 0: Manual Process

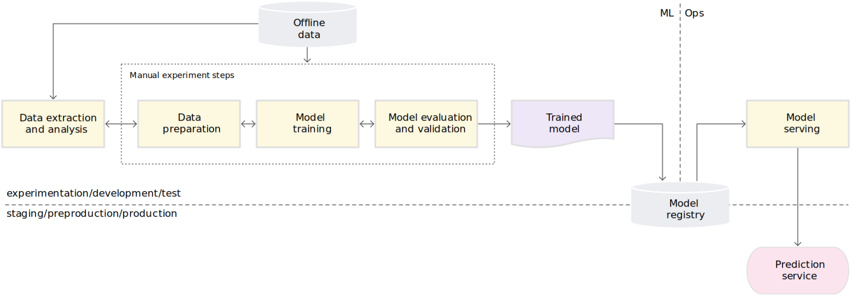


Figure: Level 0 Architecture [2]

MLOps Level 1: Automating ML Pipeline

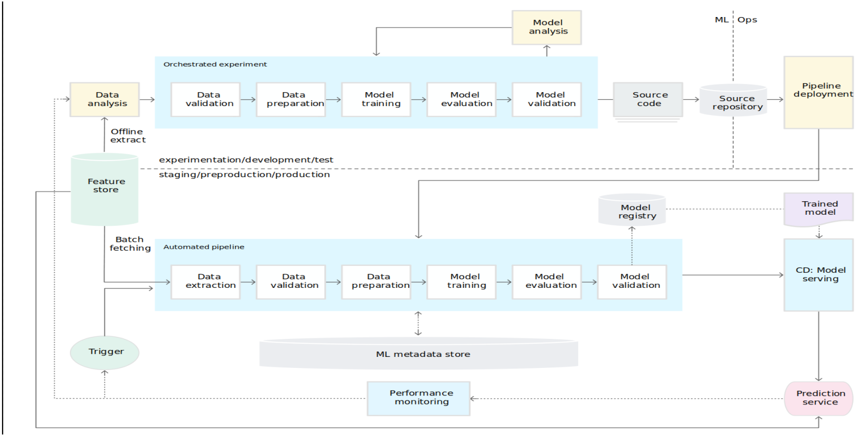


Figure: Level 1 Architecture [2]

MLOps Level 1 Properties

- Characteristics
- Added Components
  - Data & Model Validation
  - Feature Store
  - Metadata Management
  - Pipeline Triggers
    - On Demand
    - Schedule Based
    - Availability of New Data
    - Model Performance Degradation
    - Changes in Data Distribution
- Challenges

MLOps Level 2: Automating CI/CD Pipeline

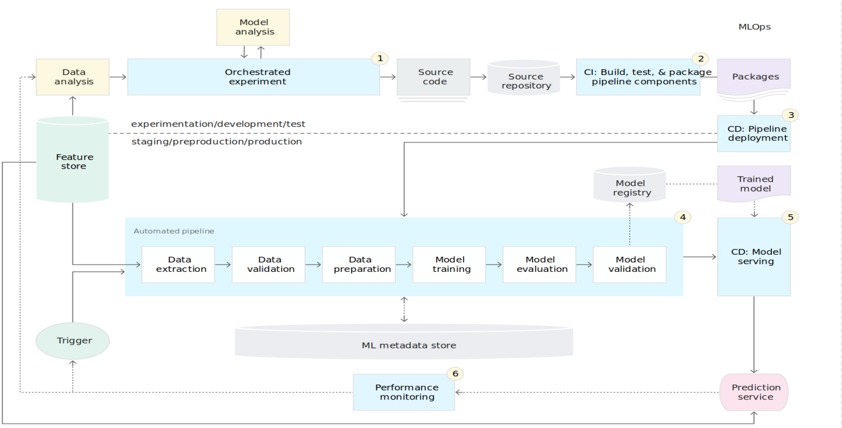


Figure: Level 2 Architecture [2]

Transparency

- Concept
- Methods
  - Local Methods
    - LIME [3]
    - Anchor [4]
    - SHAP [5]
    - ICE [6]
  - Global Methods
    - PDP [6]
    - Global Surrogate [7]

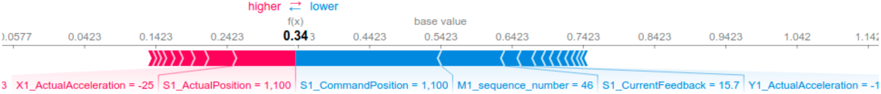
Local Transparency Methods

Anchor Explainator

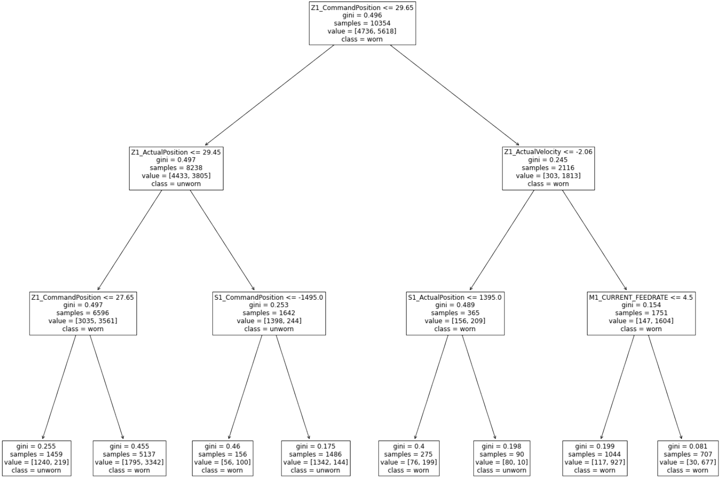
Anchor: M1_sequence_number > 86.00 AND M1_CURRENT_FEEDRATE <= 3.00 AND Y1_DCBusVoltage > 0.10 AND 326.00 < X1_OutputCurrent <= 327.00 AND Y1_OutputVoltage > 9.83 AND X1_DCBusVoltage > 0.09 AND clamp_pressure <= 3.00 AND X1_OutputVoltage > 10.40

Prediction	Precision	Coverage
1	0.9193899782135077	0.0

LIME Explainator



Global Surrogate



Implementation

CNC Mill Tool Wear App:

- The latest technology AutoML [8], for model development
- LIME [3], SHAP [5], Anchor [4], and Global Surrogate [7] as transparency methods
- ML-Flow [9] for tracking model experiments
- Prometheus [10] and Grafana [11] for monitoring
- Other endpoints for data distribution, model statistics, application metrics, and retraining

Benchmarks - MLOps

- MLOps Level 2
  - Automated Pipeline
  - Model Registry
  - Data and Model Analysis
  - Feature Store
  - Model Serving
  - Performance Monitoring
  - ML Metadata Store
  - Triggering

Benchmarks - Model Metrics

	Accuracy	Precision	Recall	F-1 Score	Correlation Coefficient
Decision Tree	99.00 %	0.9900	0.9915	0.9908	0.9799
Random Forest	99.34 %	0.9942	0.9937	0.9939	0.9868
Auto-Sklearn	**99.46 %**	**0.9958**	**0.9942**	**0.9950**	**0.9891**

Table: Performance Metrics

Conclusion

- MLOps
  - Standardized & Automated Model Deployment
  - Effective Model Performance
  - Lessened Failures
- Transparency
  - Better Understanding
  - Key to Analyze & Improve the System
- Cloud-Support Technologies
  - Faster Deployment
  - High Quality Operations

Future Scope

- Python Docker Image
- Under Development Tools

References I

- O. SCHWARTZ, “In 2016, Microsoft’s Racist Chatbot Revealed the Dangers of Online Conversation,” Nov 2019. [Online]. Available: https://spectrum.ieee.org/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation
- “MLOps: Continuous delivery and automation pipelines in machine learning,” 07 2020. [Online]. Available: https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning
- M. T. Ribeiro, S. Singh, and C. Guestrin, “"Why Should I Trust You?": Explaining the Predictions of Any Classifier,” in **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016**, 2016, pp. 1135–1144.
- Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin, “ Anchors: High-Precision Model-Agnostic Explanations ,” in **AAAI Conference on Artificial Intelligence (AAAI)**, 2018.

References II

- S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in **Advances in Neural Information Processing Systems 30**, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4765–4774. [Online]. Available: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," **Journal of Machine Learning Research**, vol. 12, pp. 2825–2830, 2011. [Online]. Available: https://scikit-learn.org/stable/modules/partial_dependence.html
- Christoph Molnar, "Global Surrogate," in **Interpretable Machine Learning - A Guide for Making Black Box Models Explainable**. BOOKDOWN, 10 2021, pp. 221–226. [Online]. Available: https://christophm.github.io/interpretable-ml-book/global.html

References III

- Frank Hutter, Marius Lindauer, “AutoML | Freiburg-Hannover,” 2018, Last Accessed: 2020-10-22. [Online]. Available: https://www.automl.org/automl/
- Databricks, “MLflow - An open source platform for the machine learning lifecycle,” 2018. [Online]. Available: https://mlflow.org/
- Bjorn Rabenstein and Julius Volz, “Prometheus: A Next-Generation Monitoring System (Talk).” Dublin: USENIX Association, May 2015. [Online]. Available: https://prometheus.io/docs/introduction/overview/
- Grafana Labs, “Grafana Documentation,” 2018. [Online]. Available: https://grafana.com/docs/