Optimizing Renewable Energy Forecasts:

An MLOps Conceptual Approach for Scalability

Riege, Raphael[1], Schütz, Johannes[2]

[1]Fraunhofer IEE, raphael.riege@iee.fraunhofer.de

[2]Fraunhofer IEE, johannes.schuetz@iee.fraunhofer.de

As the expansion of renewable energies grows, the need for accurate energy forecasts becomes crucial due to the dependency on volatile energy sources. Traditional forecasting systems, which utilize weather data and historical generation data, are challenged by the unique behaviors of individual power plants, lack of data and changing conditions (Yan et al. 2022). To address these challenges, we propose a highly scalable energy prediction system based on MLOps principles to ensure efficient model updates while adhering to appropriate quality criteria. This aims to improve the accuracy and efficiency of renewable energy forecasts, supporting a reliable energy supply in the dynamic energy sector.

The primary purpose of MLOps is to efficiently facilitate the deployment of ML models into production by eliminating bottlenecks in Development and Operations and automating the workflows (Subramanya et al. 2022). Building on these principles, we aim to develop a service mesh comprising multiple interacting microservices, which include a training service for both experimental and production environments, a forecasting service for operational forecasts, a centralized feature store that houses all necessary data including training, test, operational data, and master data, a model store for archiving models with their parameters and metrics, and a monitoring service to ensure prediction quality and service reliability.

The main challenge we address is the large number of assets, each requiring individual model predictions, necessitating a highly flexible and scalable ML pipeline. We employ dual training modes for initial and continuous model retraining within the same productive Kubernetes cluster used for inference. Furthermore, a separate model store for logging and tracking supports access at any stage of the ML pipeline (Alla and Adari 2021), complemented by a central feature store that enhances flexibility and adaptability by utilizing consistent data interfaces (Dowling 2023). For a high scalability of our inference service, we use Horizontal Pod Autoscaling provided by Kubernetes that automatically updates a workload resource to match request demand (The Kubernetes Authors 2024).

In summary, conventional static methods do not fulfill the growing requirements of increasingly frequent dynamic changes in modern energy systems. We therefore aim to use MLOps concepts to provide scalable services for the continuous improvement of forecasts under changing conditions.

Literature

Alla, Sridhar; Adari, Suman Kalyan (2021): Introduction to MLFlow. In Sridhar Alla, Suman Kalyan Adari (Eds.): Beginning MLOps with MLFlow. Berkeley, CA: Apress, pp. 125–227.

Dowling, John (2023): What is a Feature Store for Machine Learning? Available online at https://www.featurestore.org/what-is-a-feature-store, updated on 8/11/2023, checked on 5/17/2024.

Subramanya, Rakshith; Sierla, Seppo; Vyatkin, Valeriy (2022): From DevOps to MLOps: Overview and Application to Electricity Market Forecasting. In *Applied Sciences* 12 (19), p. 9851. DOI: 10.3390/app12199851.

The Kubernetes Authors (2024): Horizontal Pod Autoscaling. Available online at https://kubernetes.io/docs/tasks/run-application/horizontal-pod-autoscale/, updated on 2/18/2024, checked on 5/17/2024.

Yan, Jie; Möhrlen, Corinna; Göçmen, Tuhfe; Kelly, Mark; Wessel, Arne; Giebel, Gregor (2022): Uncovering wind power forecasting uncertainty sources and their propagation through the whole modelling chain. In *Renewable and Sustainable Energy Reviews* 165, p. 112519. DOI: 10.1016/j.rser.2022.112519.