UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS
UNECE Machine Learning for Official Statistics Workshop 2023
5-7 June 2023, Geneva Switzerland
04 April 2023

Hands-on Lab : An introduction to MLOps with Mlflow

Tom Seimandi, Romain Avouac, Thomas Faria (National Institute of Statistics and Economic Studies (INSEE), France)
tom.seimandi@insee.fr

Abstract

Machine learning is becoming an increasingly important tool for the production of official statistics. It is used by many statistical organizations in a wide range of application, from text classification to anomaly detection in time series or imputation of missing values in surveys.

The widespread adoption of machine learning methods requires the adoption of best practices in order to industrialize their use. Developing, training and evaluating a model is a complex process. In particular, the transition from the development phase of a model to a production environment poses multiple challenges. How to deploy a model in a production environment ? In what format ? How to monitor the model over time ? These are precisely the challenges addressed by MLOps, a relatively new field that aims to establish a set of practices necessary to manage the entire lifecycle of machine learning models.

Several frameworks have been developed to implement the principles of MLOps. MLflow is an open-source software that allows data scientists and statisticians to manage their end-to-end machine learning workflows. MLflow facilitates model development as it allows to track experiments in a reproducible way and compare training runs according to various evaluation metrics. It can also be used as a model and model meta-data store. Finally, MLflow offers tools to rapidly serve machine learning models to end users.

In this hands-on lab, we will develop a standard machine learning pipeline from scratch, comparing the traditional way of training and deploying models with the MLOps way using MLflow. The lab will be based on open data from the French National Business Registry Sirene. The training will take place on the SSP Cloud, a data science platform developed by the French National Statistical Institute which is suited for such innovative use cases both in terms of available services and computing resources. This 3-hour session is designed for practitioners with some experience in machine learning and Python. The number of participants is limited to 30. All you need to bring is your laptop with a current web browser installed!