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

An open source data science platform to foster innovative and production-ready machine learning systems

Romain Avouac (National Institute of Statistics and Economic Studies (INSEE), France)

romain.avouac@insee.fr

Abstract

Machine learning is becoming an increasingly important tool for the production of official statistics. Not incidentally, an increasing number of public statisticians trained as data scientists have joined NSIs in recent years. However, these new profiles often find themselves isolated in national statistical systems, and their ability to derive value from machine learning methods is limited by several challenges.

The first challenge is related to the lack of proper IT infrastructures. Training a machine learning model generally involves large amounts of data and high computational capacity. Similarly, emerging techniques often require specific hardware, such as GPUs, to perform computation in a massively parallelized way. Such resources are rarely found in personal computers or traditional IT infrastructures.

Another challenge is the transition from machine learning experiments to production-ready solutions. Production environments often differ from development environments, in such a way that the additional development costs needed to go from a proof of concept to a system working in production can limit the feasibility of this transition. Besides, in a production setting, a machine learning system needs both to be scaled to changing demand and to be properly monitored. Finally, it is generally the case that models need to be periodically or continuously updated, which require proper management of their lifecycle in order to ensure reproducibility. These various challenges highlight the need for both technical infrastructure and automation tools that can help statisticians and IT teams to implement the best practices advocated by the MLOps approach.

Against that background, we developed the SSP Cloud, an open-innovation data science platform built upon state-of-the-art IT components to provide statisticians with scalable and reproducible environments. The platform is based on three deeply structuring choices: cloud computing, object-storage and containerization, which enable to proivde extensive computing resources – the benefits of a centralized infrastructure – while managing concurrency in the access to these resources and services isolation. We provide an extensive catalog of services to cover the entire lifecycle of a machine learning project : interactive services (R, Python, Julia) for the development phase and automatization tools (MLFlow to industrialize models training, argo-workflow to orchestrate parallel jobs) to develop production-ready systems.

The building principles of this platform where further refined into an open-source project : Onyxia. As a result, public organizations can create their own internal instance of this modern data science platform and tailor it to the needs of their end users.