UNECE

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

UNECE Machine Learning for Official Statistics Workshop 2023

10 May 2023

5-7 June 2023, Geneve Switzerland

ML training : Who? What? How? and… What for?

Christophe Bontemps (UN SIAP, ESCAP)

christophe.bontemps@un.org

Abstract

Designing a Machine Learning (ML) course for Official statistics (OS) is not an easy task as the question on training remains mostly unaddressed in the literature that focuses on the benefits of ML for OS to improve data analysis and statistical production (Sanchez *et al.* 2018, Puts & Daas. 2021, Beck *et al.* 2018).

Based on the experience of the e-learning courses developed from 2021 at the SIAP, we focus here on the questions that ultimately prevail when designing any training: Who? What? How?… and what for?

[Who?] NSO learners interested in ML come from different backgrounds (IT, Statistics, Analytics, Data management, …) and are heterogenous in their profiles and positions, as well as in their expectations. They also have different time constraints, even within the same NSO. This is particularly true for online training, where one has to expect some attrition in particular if the course include mathematical or theoretical components, which was the case in our courses

[What?] We thought that the course should be statistically oriented and include pedagogical activities aimed at interpreting current algorithms and opening the ML *black boxes.* The change of paradigm for the one used in the ML world, focusing on predictions accuracy is thus at the core of our courses. We also stressed how important it is to visualize results and (hyper) parameters choices in data science in general and in ML in particular. Importantly, ethical considerations when using ML proved to be among remarked and most viewed activities.

[How?] When it comes to teaching, there is no universal approach and one has to propose different types of activities to different audiences (or *personas).* The choice between *teaching* ML, where a software agnostic course is possible vs *training* in the use of ML which ultimately would rely on some software was certainly a tough one. We'll discuss that choice and the motivation for a balanced approach.

Also, we decided to use a diversity of pedagogical activities and to strictly align each activity with learning objectives We used, short interactive videos for ML concepts and principles, interactive widgets to play with some hyper-parameters in simple setting (trees, k-NN, classification), case studies & interviews to highlight implementation and technical issues, ad other interactive elements available[1]. In addition, we proposed reproducible, but optional, hands-on in R/Python for those already familiar with programming in an attempt to balance *teaching* and *training*. Finally, allowing and encouraging discussions within forums, webinars as well as experience sharing proved to be efficient tools for social learning, a goal often neglected.

[What for?] This is the main question and it was addressed first with the definition of the learning objectives (LOs). Our courses are "*for Official Statistics"* which is a specificity. Moreover, we made a clear distinction of LM applications with big data applications. With these constrains in mind, we followed simple, practical ideas.

[1] And no interactive PowerPoint-like slides!

First, we had to align the assessments and assessment capabilities with the LOs. Some objectives are of practical nature (*apply* a specific ML technique) and required specific features for assessments. Also, our potential learners may belong to small NSOs with limited computational capabilities which implies that examples should be usable online (*e.g.* interactive shiny apps) or based on small data sets (*e.g.* imputation exercise). Another leading idea was to showcase the diversity of applications while highlighting the similarity of the principles. A third constructive idea was to highlight that while ML proposes a new approach to prediction, some of its underlining techniques were not new (classification, regression, decision trees). Also, we wanted to explain the mathematics behind ML in a simple way, and to demystify the apparent complexity of ML, in particular through small-sized and interactive applications. Finally, designing courses “for official statistics”, implies discussing the challenges, and ethical considerations to consider before implementing any ML-based solution in an NSO.

Details on SIAP’s ML courses

- Facilitated Courses: “*Machine Learning for Official Statistics and the SDGs*”
  - First run: November-December 2021 (6 weeks)
    69 different activities, 7 Modules, 6 Webinars
    470 rticipants (51% women), 73% completion ,100 (optional) data-based projects,

  - Second run: November 2022- January 2023 (8 weeks)
    74 dfferent activities, 7 Modules, 6 Webinars
    383 participants (55% women), 62% completion, 97 (optional) data-based projects,

- Self-paced course: “Principles of Machine Learning for Official Statistics and SDGs”
  - June 2022 – ongoing (link to self-paced course, see also the flyer)
    47 different activities, 6 Modules, no webinars
    247 participants, 9% completion

References

Puts, M. & Daas, P. (2021). “Machine Learning from the Perspective of Official Statistic”, The Survey Statistician, 2021, Vol. 84, 12–17

Beck, M., Dumpert, F., & Feuerhake, J. (2018). “Machine Learning in Official Statistics”. *ArXiv, abs/1812.10422*.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). “An introduction to statistical learning” (Second Edition). New York: springer.

Sanchez, J.A. et al. (2018). “The use of machine learning in official statistics”, UNECE Machine Learning Team, November 2018

UNECE ML Group (2021) “Machine Learning for Official Statistics”, Geneva 2021 ISSN: 0069-8458