

## Business Case for Machine Learning

This business case was prepared by the Machine Learning 2019 project manager following a sprint with the work package leaders, and is submitted to the HLG-MOS for their approval.

| Type of Activity   |                               |   |                                |
|--|-------------------------------|---|--------------------------------|
| <input type="checkbox"/>   | New project                   | <input type="checkbox"/>  | New activity                   |
| <input checked="" type="checkbox"/>  | Extension of existing project | <input type="checkbox"/>  | Extension of existing activity |
| The project is an extension of the 2019 Machine Learning project sponsored by the HLG-MOS.   |                               | Activities are undertaken by Modernisation Groups. These activities produce smaller, more detailed outputs to help achieve the HLG-MOS vision |                                |
| <p>For an overview of the progress of the current ML project, refer to the working documents (most of them were discussed at a sprint held in September):</p> <p><a href="https://statswiki.unece.org/display/MLP/Working+documents">https://statswiki.unece.org/display/MLP/Working+documents</a></p>   |                               |   |                                |
| Purpose  |                               |   |                                |
| <p>Based on mutual interest and building on existing national developments, the objective of the project is to advance the research, development and application of machine learning techniques (ML) to add value (relevance, timeliness, quality, efficiency) to the production of official statistics. Building on the engagement of 38 participants from 18 organisations in 14 countries, pilot studies aimed at demonstrating the value-added of ML will be completed; best practices in the implementation of ML techniques will be identified; and knowledge, tools and best practices will be shared. The latter will not focus solely on implementing the ML techniques, but how National Statistical Organisations (NSOs) are organized to move them quickly to the production processes. All this will be underpinned by a quality framework and supporting reference documents.</p>  |                               |   |                                |
| Description of the activity  |                               |   |                                |
| <p>One of the biggest threat to (NSOs) producing official statistics is that of quickly becoming irrelevant in today's fast-paced and ever increasingly complex society, economy and environment. NSOs are threatened by an increasing number of public and private organisations who produce and promote statistics in a more timely and accessible manner, attracting the attention of policy makers and other users. These data producers use approaches and methods beyond those traditionally used by NSOs, such as machine learning and artificial intelligence. In fact, these approaches and methods are no longer that new anymore and the importance of integrating them into the production of official statistics has now been widely recognized by NSOs. At its November 2018 workshop, the HLG-MOS further brought this point forward by supporting a proposal from the Blue-Sky Thinking Network to launch a Machine Learning Project.</p> <p>The following work packages comprise the project:</p> <p>Work package 1 – Demonstrating the value added of Machine Learning</p> <p>Pilot studies are being conducted on: (1) Coding (with examples on product descriptions, industry and occupation, sentiment); (2) Edit and Imputation (with examples on survey and register/administrative data); and, (3) the integration of satellite data with census, survey and administrative data. At the "local" level, the topics on which the studies are conducted are mostly relevant to the participating organisations. At a more global level, the studies are determining the value-added of ML in each of these three areas, identifying best practices in implementing ML solutions, raising challenges in integrating them in the production process and providing opportunities to test quality assurance practices. In spite the fact that the project was launched in mid-March and was fully functioning only in May, fourteen presentations, four working papers, 80 reference</p> |                               |   |                                |

documents, some software and ML code were being shared in September. Many of them were discussed a sprint.

#### Work package 2 – Quality

When it comes to trust in official statistics, NSOs hold a competitive advantage in being transparent and publishing details on data sources, methods and various indicators, as described in many quality frameworks. The goal of this work package is to propose quality framework components for evaluating ML processes and statistics produced using them, as well as to bridge the gap between these components and those in existing frameworks. This will allow NSOs to compare outputs from traditional and ML methods and to inform users of data quality when ML is used to produce outputs.

The emphasis has been placed on the accuracy dimension of the framework and ML uses were divided into uses for traditional processes and uses that produced direct outputs. Concepts in the proposed framework are being evaluated in the pilot studies being performed in work package 1 and will continue in year two. The pilot studies are raising issues and identifying best practices in quality assurance (QA). Members are realizing the quality challenges beyond demonstrating the value-added of ML, e.g. QA support in production. They are also identifying other dimensions where ML may add value, these will be included in the proposed framework and evaluated in the pilot studies.

#### Work package 3 – Integration of Machine Learning (more specific than Lessons Learned, originally proposed)

One of the recurring themes from the first iteration of this project was that integrating machine learning into official statistics requires more than simply building machine learning systems. In fact, a number of participants noted that they had already developed otherwise successful machine learning solutions, but had been unable to implement them into production processes because of a variety of organizational and structural impediments including uncertainty over who should be responsible for building, evaluating, and maintaining these highly interdisciplinary systems. The goal of this work package is to explore how different NSOs are organized to integrate ML in their production processes, and report on the different practices, sources of impediments and propose successful practices.

#### Alternatives considered

The project is likely the only one in the official statistics community solely dedicated to ML. It is also carried out by a large group with a very good mix of expertise in statistics, methodology, IT, data science and subject-matter. Its mandate could be passed on to another group, but given the size of the group and the familiarity that they have gained, a dedicated manager would be required to preserve their momentum. Its activities could be dispersed to other groups, even outside the HLG-MOS, working on broader topics (e.g. Big Data) or specific ones (e.g. use of satellite images), but the attention to ML and advancing its use to transform processes could be reduced.

#### How does it relate to the HLG-MOS vision and other activities under the HLG-MOS?

The addition and integration of ML solutions are needed to enable NSOs to remain relevant by producing more relevant information with more detail and in a timelier fashion. At the same time, they need to be developed and implemented in a responsible manner for NSOs to remain and further grow as a trusted data authority. The project is fortunate to have the engagement of 38 individuals collaborating on several pilot studies and other developments. The work of the project is also supported or followed by 30 other individuals from 14 national or international organisations who have requested access to the project's wiki pages. Once completed, all key information will be available online and through a workshop.

#### Proposed start and end dates

**Start:** January 2020

**End:** December 2020

- Complete and document pilot studies, including ML code: February 2020
- Complete summary of value-added of ML on each pilot study theme: April 2020
- Complete best practices in the implementation of ML with accompanying quality framework: June 2020
- Complete review and findings of ML integration practices: September 2020
- Complete the development of a workshop: November 2020
- Wrap-up and recommend where to move any next steps: December 2020

