**UNECE High-level Group for the  
Modernisation of Official Statistics**

**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** | | | | |
|  | New project | |  | New activity |
|  | Extension of existing project | |  | 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 | |
| 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): [https://statswiki.unece.org/display/MLP/Working+documents](about:blank) | | | | |
| **Purpose** | | | | |
| 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, the project will:   * Summarize the numerous pilot studies that are being conducted to demonstrate the value-added of ML on demand-driven needs, as well as share the ML code used to conduct these demonstrations; * synthesize the results and experiences into recommendations and best practices in the implementation of ML techniques in the areas under study and in general, supported by additions to traditional quality assurance frameworks; * conduct a review on how NSOs are organized to bring together the multi-disciplinary expertise required to integrate ML techniques from the definition of the business need their implementation in the production processes; * make numerous documents available as they are finalized to the statistical community during the year (see below the proposed schedule on the delivery of study reports, synthesis and best practices, quality assurance practices, organisational review), as well as combine this information into a workshop at the end of the project. | | | | |
| **Description of the activity** | | | | |
| One of the biggest threats 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 information and 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.  As a second part of the project, the following work packages comprise the objectives for 2020:  Work package 1 – Demonstrating the value added of Machine Learning  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 documents, some software and ML code have been shared. Many of them were discussed a sprint held in September.  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 potentially 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 an additional 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 | | |
| The following outputs will be shared with the statistical community as they are completed during the year:   1. Report on each pilot study and related ML code; demonstration of added value of ML to demand-driven needs of the participating organisations along with documented ML code to share with other organisations; **delivery February 2020** 2. Synthesis of demonstrations, recommendations on use of ML and supporting reference documents; synthesis of added value of ML to the statistical process (C&C, E&I, Imagery) and experiences/recommendations on how to implement. In addition to the outcomes of the pilot studies, we would like to add to this synthesis other proofs of concept already conducted by other organisations; **delivery April 2020** 3. Best practices in the development of ML, supporting quality framework and review of integration practices in organisations; synthesis on the general use of ML that would apply beyond the areas covered in the project supported by related quality practices. In spite of these demonstrations, best practices, etc., for various reasons, some organisations seem to be able to bring the ML solutions further along than others. Some organisations seem to have ML solutions waiting on shelves. We propose to conduct a review of organisational practices that facilitate or impede on the progress in integrating ML solutions and share findings with others; **delivery September 2020** 4. Workshop/training material to share the work of the project, including a hands-on learning exercise; this would be a live means of sharing the knowledge and best practices accumulated by the project. It would include a hands-on ML exercise with a dataset and code on which to apply it; **delivery November 2020** 5. Project wrap-up and recommended next steps; **delivery December 2020** 6. Workshop/training delivery; **near the end of the project** (depending on the progress of the project early in the year, the project may deliver a portion of the workshop/training material in November 2020). | | | | |