UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS
ModernStats World Workshop 2022
27-29 June 2022, Belgrade, Serbia

Use of GSIM for Modelling Machine Learning Processes

InKyung Choi (UNECE), choii@un.org

1. Introduction

Machine learning (ML) is a field of study that gives computer the ability to learn without explicitly being programmed[1]. It learns patterns in a large amount of data and applies it to make prediction for a given task without humans providing all rules explicitly.

ML holds a great potential to modernise the production process of the statistical organisations as well as to expand the range of services they can provide to society with ever increasing information needs. For example, classifying textual and image data is a crucial step in the statistical production that transforms the micro-level raw data into codes from statistical classification system. This task used to rely largely on time-consuming manual works or heavy rule-based system. With the advance of national language processing (NLP) and computer vision, ML models can be used to assist humans in classifying the textual responses and images. This can allow statistical organisations to process a large amount of data faster, thus making the production process more efficient. ML can also allow the organisations to better utilise the big data to support the survey operations (e.g., identification of residential area via satellite image) or produce new types of statistics (e.g., sentiment indicators based on social media posts).

While there is an increasing interest in using ML for production of official statistics, ML is still relatively new in statistical organisations and there are a number of organisational, technical and cultural challenges. Firstly, ML often requires a skill set that is different from traditional capabilities that statistical organisations have; hence needs to be built inside or acquired from outside. ML project needs a multi-disciplinary collaboration; it involves not only data science, but also subject matter expertise, IT support as well as sound statistical comparison; hence requires a wide extent of collaboration. Automating status-quo manual processes with ML inevitably impacts the regular work of staff which can make it hard to gain buy-in about the ML solution.

Among the challenges, this paper focuses on the lack of a common language to describe processes related to ML (processes that develop ML and processes that use ML) in the official statistics. By representing the process using the Generic Statistical Information Model (GSIM)[2], this paper aims to facilitate the development ML and its integration in the regular statistical business.

The remainder of this paper is structured as follow. Section 2 introduces the ML development process. In Section 3, GSIM is used to describe the information inputs/outputs of the process and how its final

[1] https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained

[2] https://statswiki.unece.org/display/GSIM

products (e.g., ML service) of the development process can be connected to the regular stat production processes. This paper concludes with summary in Section 4.

2. Machine learning development process

ML pilot studies conducted in the field of official statistics[3] show that ML can be used in many ways in the statistical organisations, the use cases include, but not limited to, automation of manual tasks, production of new types of statistics based on big data, assistance to survey operation, and so on.

Among these, this paper focuses on the first use case – automation of manual tasks in existing business processes such as record linkage (GSBPM sub-process 5.1), coding and classification (GSBPM sub-process 5.2), edit and imputation (GSBPM sub-process 5.3-5.4). Note that in such scenario, there already exists methods used in the production process (e.g., manual coding, rule-based coding, imputation based on linear model) and ML solution needs to be first developed to replace/assist this method. Thus, the use case involves two interconnected processes:

- Processes that develop the machine learning solution (P1); and
- Processes that use the developed machine learning solution (P2).

While these two processes can be considered as two sub-processes in the same production process (e.g., during a new cycle of a survey program, the survey team develops a ML model to carry out a part of micro editing and uses it for E&I process), they are considered separately in this paper for two reasons. First, model decay issue requires P1 and P2 run in different cycles. ML models are built based on data, hence, once the model is exposed to new data that it has not seen, its prediction accuracy starts decreasing (model decay) and the models need a continuous updating (model re-training). Hence process P1 can be repeated with its own cycle while P2 runs in parallel without being affected by changes from P1. Also, as a relatively new technique, ML is currently often developed and used by individuals in different times. However, with increasing demand for ML, the organisations may seek to centralise the ML development (as opposed to being carried out in an isolated manner), hence the two processes involving different capabilities and expertise will need to be managed differently.

2.1. Processes that develop the machine learning solution (P1)

At a high level, the ML development (P1) follows stages as below (each step is further explained in Chapter 3; for full description, see "Journey from Experiment to Production"[4]):

- **Understand business needs** and end users;
- **Assess preliminary feasibility** with respect to the business problem, data and technical resources (software and hardware) in the organisation;
- **Develop proof of concept (PoC)** to have concrete idea if machine learning solution is feasible for the given business problem or data, explore any constraints and determine if it is worth investing further resources;
- **Prepare business case** based on the preliminary feasibility assessment and findings from the proof of concept to get approval to develop the model for the production;
- **Develop Model** for the production; and

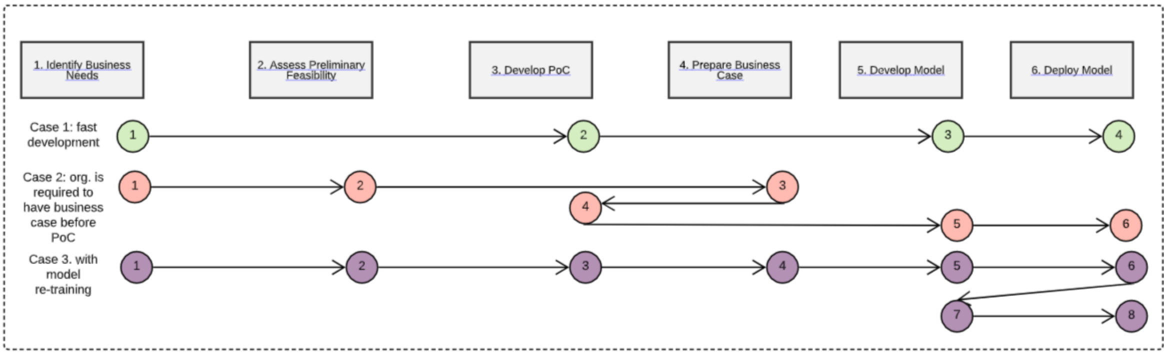
[3] For example, see https://statswiki.unece.org/display/ML/Studies+and+Codes

[4] Developed by ONS-UNECE ML Group 2021 Work Stream 2, see: https://statswiki.unece.org/pages/viewpageattachments.action?pageId=293535864&sortBy=date&startIndex=0&preview=/293535864/330370868/FromExperimentToProduction.pdf

- **Deploy Model** by making it available to the end-users to address the needs identified in the first stage.

Note that while the steps are in the logical order, they do not need to be followed in the sequential order. The steps can be conducted in parallel, repeated, skipped and re-visited depending on the situation. Also, each organisation is at a different level of ML maturity and has different policies and practices, hence activities undertaken and how they are carried out within each step may vary depending on the organisation. Some ML developments can be very quick and agile, while others can be lengthy (see Figure 1).

Figure 1. ML development process


1. Identify Business Needs
2. Assess Preliminary Feasibility
3. Develop PoC
4. Prepare Business Case
5. Develop Model
6. Deploy Model
Case 1: fast development
1
2
3
4
Case 2: org. is required to have business case before PoC
1
2
3
4
5
6
Case 3. with model re-training
1
2
3
4
5
6
7
8

2.2. Processes that use the machine learning solution

Under the use case of main interest for this paper, ML is used to automate the manual tasks or improve legacy methods. Hence, the process in which the ML solution would be used after the development already exist and have been run as a part of regular statistical programs which can be assumed, for example, to follow GSBPM (phase 4-7).

Note that, for ML to be considered as a long-term capability in the organisation, the ML model needs to be continuously monitored and updated (model re-training). This implies that the connection between development process (P1) and the processes that use the solution (P2) is not completed with ML solution hand-over, hence the re-training process should be aligned with the regular production processes and hand-over of the re-trained model should be scheduled in a way that it does not impact the regular production.

3. Representation of ML development processes and linkage to production process using GSIM

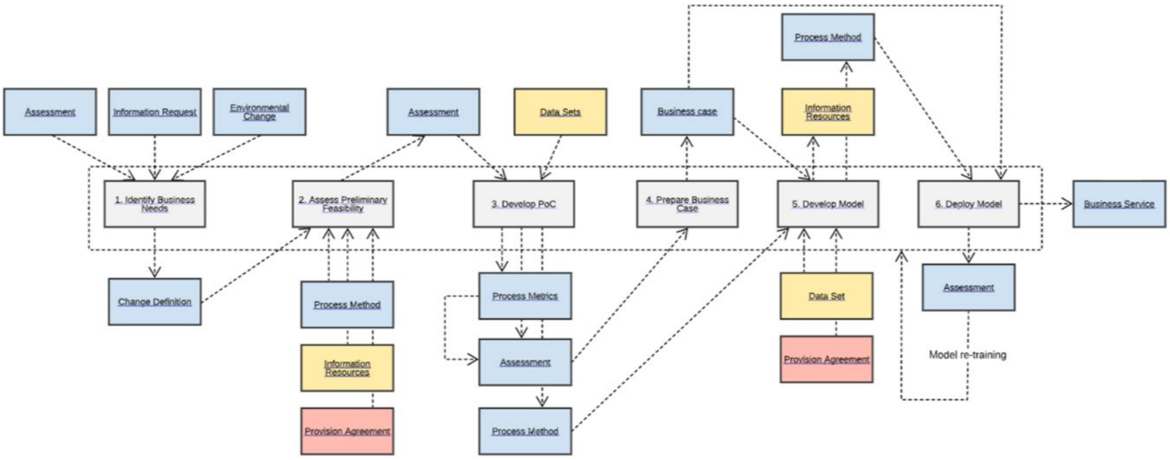
GSIM is a reference framework for statistical information that provides a set of standardised, consistently described information classes, which can be used as inputs and outputs in the design and the production of statistics. While the initial development and use of GSIM mainly had the traditional statistical productions context in mind, the generic nature of GSIM allows to use its information classes to represent the information inputs and outputs for the new type of processes such as ML development. In this section, GSIM is tested to represent the ML development process (Section 3.1) and its linkage to production process (Section 3.2).

In the rest of the paper, GSIM information classes are stylised in capital and italic. For the sake of brevity, full definitions of the individual GSIM classes are not provided in this paper, readers are referred to GSIM wiki page for further information.

3.1. Processes that develop the machine learning solution

Figure 2 provides the overview of the information inputs and outputs in the 6 steps of ML development process (P1). Note that they are not exhaustive list of inputs and outputs and that not all of them are needed for each step. The aim of this exercise is to test the extent with which GSIM can be used to capture core information needed to describe the ML development processes.

Figure 2. Representation of ML development process using GSIM



3.1.1. Understand business needs

Business needs affect various aspects around the ML solution that will be eventually put into production and decisions to be made along the journey to production. Therefore, understanding the business needs and end-users is an important first step in the development process. While the context is different, this step has a lot in common with sub-processes of GSBPM Phase 1 Identify User Needs. Based on the GSBPM Phase 1 modelling done in GSIM Information Flow in GSBPM[5], following GSIM classes can be used to represent information inputs in this stage:

- *Assessment*: report of analysis of the quality and effectiveness of any activity in the statistical organisation, for example, quality assessment revealing that a high level of error occurred from a rule-based coding system;
- *Environment Change*: requirement for change that originates from a change in the operating environment. For example, data collection strategy for price statistics might shift toward scanner and web-scrapping data to reduce survey cost, this leads to a requirement of a new way of processing the large amount of data;
- *Information Request*: need for new information that could require ML solution.

After the business needs are identified, the information output resulted from this stage can be presented as

- *Change Definition*: specification document defining proposed plan (e.g., changes to be made to existing methods).

3.1.2. Assess preliminary feasibility

In this stage, an initial evaluation of the suitability of machine learning solution is conducted with respect to the business problem, data and technical resources (software and hardware). High-level questions can help gauge the feasibility of machine learning, such as: are there large data sets, does

[5] For more information, see report from Linking GSBPM-GSIM task team under the HLG-MOS Supporting Standards Group

existing (status-quo) system require repetitive manual works that can be automated by ML to a certain extent, are there high-value works to which human resources that are saved from automation can be devoted.

Many ML models learn on (training) data and run on (new) data to make predictions, hence the ability to have a sustainable supply of data is crucial to ensure a long-term value of the machine learning solution.

Following GSIM classes can be used to presented information inputs in this stage:

- *Change Definition*: output from previous step;
- *Process Method*: methods that are currently used in the process or methods that are considered as candidate (both ML and non-ML);
- *Reference Document*[6]*:* any methodological handbook or guidelines that could inform the assessment of various candidate methods;
- *Information Resource*: *Data Set* on which ML model will be developed as well as its *Referential Metadata Set* to understand the usability of the *Data Set* for the given purpose. The real data set that ML is targeted may not be available for those who need to run the ML experiment for various reasons (e.g., data security, administrative hurdle, lack of hardware to accommodate the volume of the data) in which case, synthetic data or publicly available data might need to be used instead;
- *Provision Agreement*: provision agreements associated with *Data Set* to assess if it is possible to use them.

The result of the feasibility assessment can be captured in GSIM class:

- *Assessment*: results of preliminary feasibility assessment of ML solution.

3.1.3. Develop proof of concept

In this stage, a small-scale proof of concept (PoC) ML model development takes place to have concrete idea if ML solution is feasible for the given business problem and data, explore any constraints and determine if it is worth investing further resources. To measure the performance of the PoC model, detailed and quantifiable quality criteria to judge success (e.g., accuracy, time and cost) should be established. The development of the PoC ML model roughly follows:

i. Data collection and ingestion where data sets needed for building machine learning models are gathered together

ii. Data preparation and feature engineering where data are visualised, cleaned (e.g., outlier and error detection, treatment of missing values), transformed (e.g., box-cox transformation, re-scaling) before being fed into the machine learning algorithms

iii. Model training where the different machine learning models are trained on the data set prepared from the previous step

iv. Model testing where the final evaluation of the model is conducted on the test set.

Following GSIM classes can be used to represent information inputs in this stage:

- *Assessment*: output from previous step;
- *Data Set* to be used to develop the PoC model

The information output resulted from the development of PoC model can be captured with:

[6] GSIM is currently under revision, “Reference Document” is one of classes that are considered to be included in the new version of the model. See more about this class: https://github.com/UNECE/GSIMRevision/issues/3

- *Process Method*: resulting PoC model that includes all aspects involved (e.g., data transformation, ML algorithm, hyperparameter);
- *Process Metric*: any resources spent for PoC development (e.g., staff time, cost for acquisition of any hardware/software);
- *Assessment*: results of the PoC experiment summarising how the model performed against quality criteria of interest. All relevant findings and constraints should be documented so that they could be used for the next stages when deciding whether the ML models can be used in production or not.

3.1.4. Prepare business case

ML project often involves stakeholders with vastly different background (e.g., subject matter experts, data scientists, statisticians, IT specialists) and can also take long time to complete during which the team composition may change. Therefore, business case plays an important role to ensure that all those involved have a common understanding of objectives and requirements. Business case would typically include elements such as: problem statement, business value addition, cost, stakeholder, project plan, operational business process, data, governance and risk assessment. To maximise the return on investment, this stage can also explore the possibilities of expanding the application areas of the solution so that it can be used in other parts of the organisation with similar business needs.

Similar to GSBPM sub-process 1.6 “Prepare and submit business case”, this stage has an assessment as core input:

- *Assessment*: output from previous step;

The information output resulted from this stage can be presented as:

- *Business Case*: proposal for the development of the production-level ML solution. Note that in some organisations, business case is needed for PoC as well. In this case, this step would be carried out before PoC step (e.g., case 2 in Figure 2).

3.1.5. Develop Model

Once the business case is approved, the development of the production-level model is initiated. At a high level, the model development stage follows a similar process as the PoC model development (i.e., data collection, data preparation, model training, model testing), but there are several differences coming from data, model and IT environment. For example, the model developed in this stage uses the real-world data (as opposed to, e.g., a small portion of it used for PoC), the model that needs to be developed might have a different scope with extended the application area, the model might need to be developed using a different programming language to be compatible with production environment.

Following GSIM classes can be used to represent information inputs in this stage:

- *Business Case*: output from previous step;
- *Process Method*: output from PoC model development (3.1.3) as a basis or reference for the production-level model;
- *Data Set* on which the ML model will be developed

The development of production-level ML model resulted in:

- *Process Method*: ML model that includes all aspects involved (e.g., data transformation, ML algorithm, hyperparameter);
- *Information Resource*: collection of all relevant information (e.g., *Process Metric* and *Assessment* of the model against quality criteria set in the *Business Case*) which is important for reproducibility, monitoring and re-suability of the models.

3.1.6. Deploy model

The ML model is a tool designed to address a business problem identified (3.1.1). To provide its business value, therefore, the ML model developed (3.1.5), which may exist as programming script on the data scientist's computer, should be made available to the end-user (which can be either humans or another software in the bigger system). In this sense, model deployment can be considered as a process making the ML model available to the users. Depending on the problem and the end-users, deployment can take different paths (e.g., API, service application with user interface).

As described in Section 2.2, the ML model starts to decay over time, therefore the model needs to be continuously monitored and re-trained when needed. The monitoring can be done through tracking performance metrics (e.g., decrease of prediction accuracy) or comparing new data with the one used for model development (e.g., comparing covariates distribution in the data set used for the model training and the distribution in the new data). Depending on the assessment of the monitoring metrics, re-training of the model could lead to one of previous stages (e.g., the update of model parameters with a new data set leads to 3.1.5 while a bigger change in the model, such as change from classical ML model to deep learning model, might require a development of PoC model first).

In this stage, the core input is:

- *Process Method*: output from previous step

The output can be represented as:

- *Business Service*: ML model wrapped as a service;
- *Assessment*: analysis of the production model monitoring

3.2. Processes that use the machine learning solution

The ML development process can be considered as a part of a corporate activity that supports the statistical production (e.g., ML automatic coding application service can assist GSBPM sub-process 5.2. Classify and Code). Such support activity will impact the design of an existing production process as the ML solution becomes a new processing method or one of methods that can be chosen from.

GSIM Information Flow in GSBPM provides how GSIM classes can be used to represent the information input and outputs of all GSBPM sub-processes. Figure 3 shows how this production process can be linked with the ML development process. To break down the story (white boxes) depicted in the Figure from the bottom:

- Statistical organisation initiates a programme to modernise methods (*Statistical Support Program*), one of which is research looking at automating existing coding system (*Business Process*; note that this process is essentially ML development process described in Section 3.1.; see Figure 4)
- The ML research impacts the design of coding method (*Statistical Program Design*) of the labour force survey (*Statistical Program*) which used to use the legacy coding system
- New cycle of the labour force survey (*Statistical Program Cycle*) uses the newly developed ML solution (*Business Service*) in its coding task (*Process Step*) in process phase (*Business Process*)[7]

[7] Note that the level of granularity of activity assigned to GSIM *Business Process* or *Process Step* may depend on the choice of individual user. One can choose to call the entire production process as *Business Process* and GSBPM phase as *Process Step*. In this example, *Business Process* is chosen to represent GSBPM phase to use *Process Step* for its sub-process

Figure 3. Representation of the linkage between the production process and ML development process using GSIM

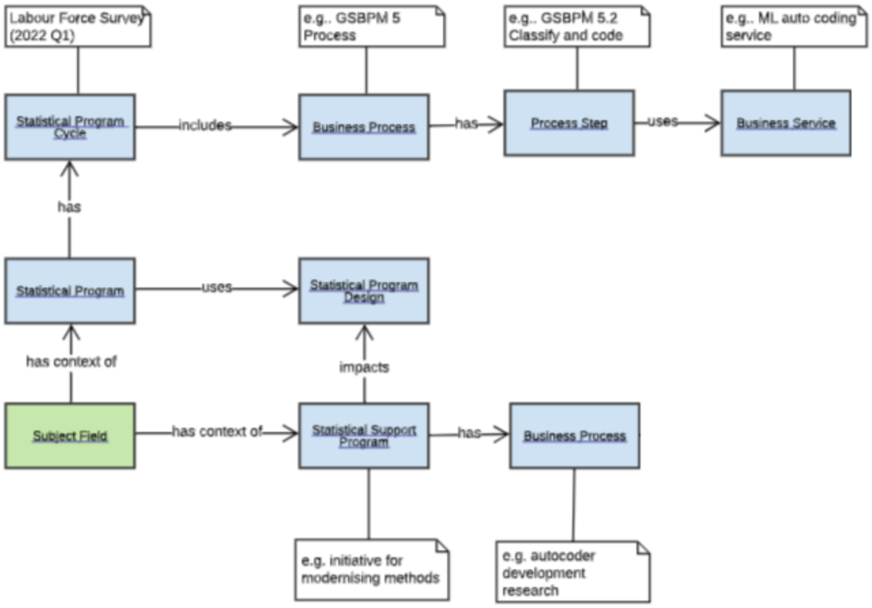
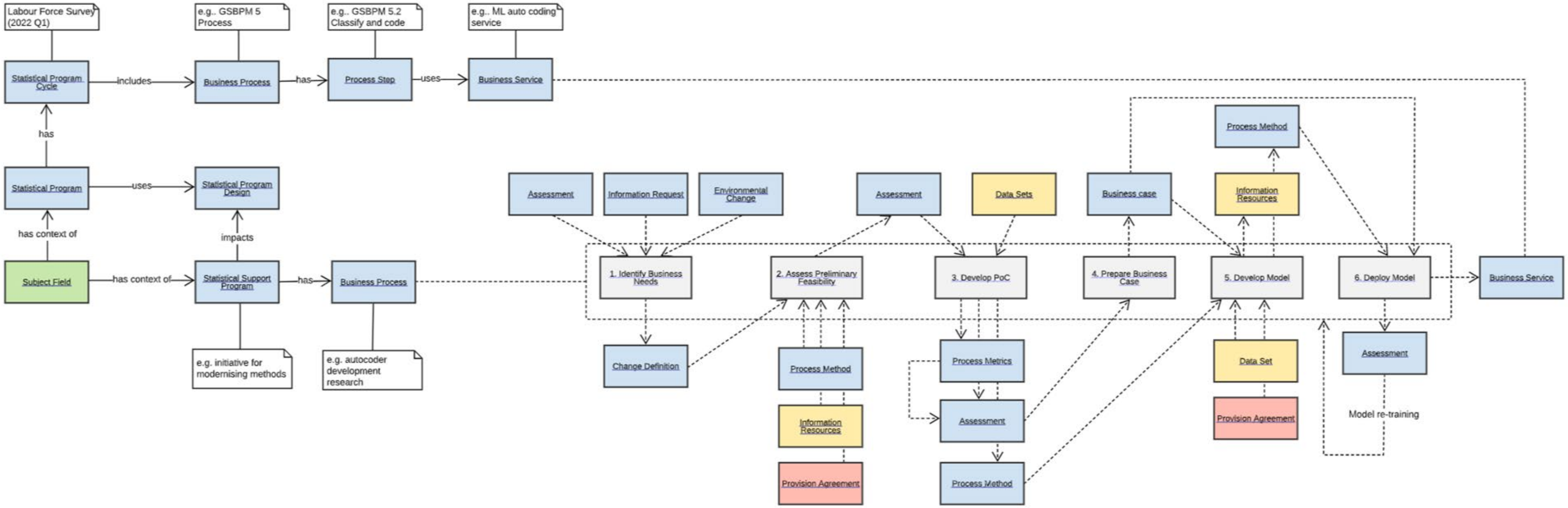

Labour Force Survey (2022 Q1)
e.g.. GSBPM 5 Process
e.g.. GSBPM 5.2 Classify and code
e.g.. ML auto coding service
Statistical Program Cycle
includes
Business Process
has
Process Step
uses
Business Service
has
Statistical Program
uses
Statistical Program Design
has context of
impacts
Subject Field
has context of
Statistical Support Program
has
Business Process
e.g. initiative for modernising methods
e.g. autocoder development research

Figure 4. Representation of the ML development process and its link to the production process using GSIM



4. Summary and conclusion

Statistical production process involves many resource-intensive tasks and machine learning holds a great potential to make them more efficient. ML can assist humans carry out the tasks faster or automate certain parts of the tasks which can also help better utilise the big data sources.

There are, however, several challenges in integrating ML as a regular capability in the statistical organisations. ML is a still relatively new technique and often developed and used in an isolated manner. Also, unlike typical software, even after deployed, ML requires monitoring as well as re-training, hence alignment with the regular production processes is needed.

To help facilitate the integration of ML into regular business of statistical organisations, this paper presented how GSIM can be used to represent the information needed for the ML development processes and linkage between ML development process and regular production processes that uses the resulting ML solution.

With standardisation of production processes, there is a greater need for centralised corporate support. While this paper focuses on the ML, there are many "non-production" processes (e.g., data integration, registry management) and much work would be needed to harmonise both processes and information.

Acknowledgement The development of this paper was greatly informed by discussions within the Infrastructure Theme Group of the ONS – UNECE Machine Learning Group 2022. I appreciate invaluable inputs from its members, in particular, Jakob Engdahl (Statistics Sweden), Joni Karanka (UK ONS), Jenny Pocknee (Australian Bureau of Statistics) and Anton Örn Karlsson (Statistics Iceland)