Journey from Machine Learning Experiment to Production

Authors: InKyung Choi (UNECE) and Claire Clarke (ABS, Australia)
Collaborators: Katja Loytynoja (Statistics Finland), Ayoub Mharzi (IMF), Issac Ross, Justin Evans and Kate Burnett-Isaacs (Statistics Canada), Abel Coronado, Jael Perez and Arturo Rubio (INEGI, Mexico), Luis Molina Martinez (UN OPS), Evyatar Kirshberg, Mark Feldman, David Wajnryt and Debby Sonino (CBS, Israel), Alex Measures (BLS, US), Eric Deeben and Alison Baily (UK)
Date: November 17, 2021
Version: 1.0

1. Introduction

The pilot studies in HLG-MOS Project demonstrated the value added of machine learning (ML) in improving the quality of official statistics, for example, by increasing accuracy, reducing processing time or making data more consistent. While these pilot studies can be helpful in convincing stakeholders about the potential of machine learning, integrating the machine learning solution, even with its proven effectiveness and validity, into production has often turned out to be very difficult and time-consuming. Unfortunately, many machine learning solutions from experiments could not complete this journey and end up being left on the shelf.

The difficulty of moving machine learning solutions to production is experienced widely across sectors and domains. For example, Venturebeat reported in 2019 that "*87% of data science projects never make it to production*"[1]. In its 2020 *State of Enterprise Machine Learning*, Gartner showed that "*18 percent of companies are taking longer than 90 days*" to deploy a machine learning model[2]. The situation is arguably more challenging for statistical organisations that are public organisations as well as primary producers of official statistics. The official statistics are required to provide not only accurate but also reliable and (temporally and spatially) comparable portraits of the society based on scientific standards[3]. As changes in the methods and data could impact these qualities that statistical organisations have maintained, the process of adopting new methods and data sources into production can be often slow and difficult.

For a machine learning solution to make it into production, one should examine what lays ahead and carefully plan accordingly to act pre-emptively and avoid unnecessary delays. To operationalise the machine learning solution, one needs to go beyond simply demonstrating that the solution works. There are organisational, technical and cultural challenges to overcome. Firstly, machine learning requires a multi-disciplinary collaboration; it involves not only data science, but also subject matter expertise, IT support as well as sound statistical comparison. The survey conducted in 2020 through the Machine Learning Project Work Package 3, for example, showed that "*coordination between internal stakeholders*" is the most significant factor that limits the organisation from using machine learning (Box 1). Also, while the "experiment environment" often

---

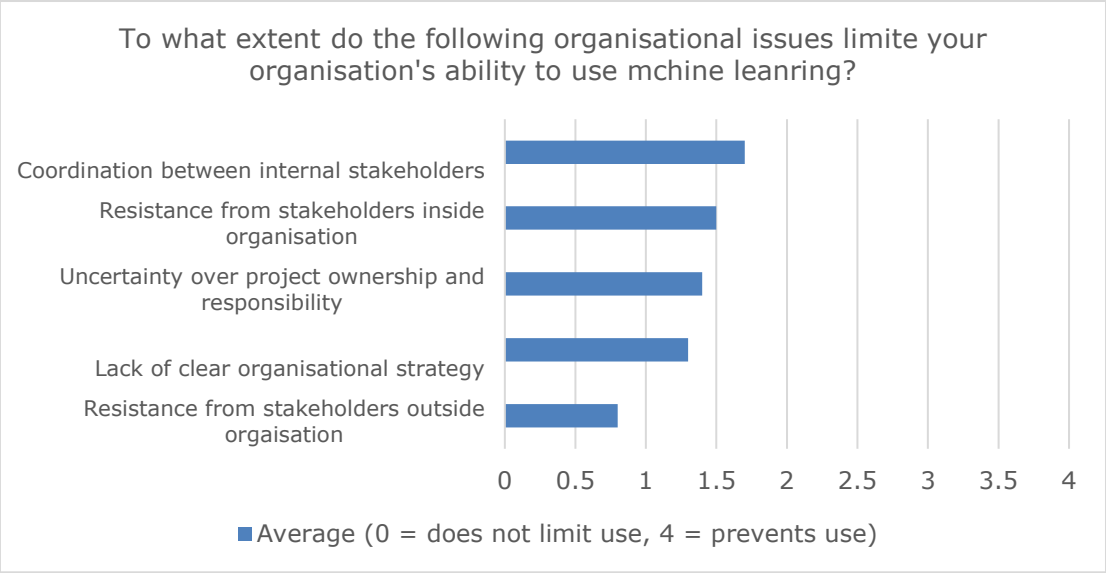
[1] https://venturebeat.com/2019/07/19/why-do-87-of-data-science-projects-never-make-it-into-production/
[2] https://algorithmia.com/state-of-ml
[3] Fundamental Principles of Official Statistics
https://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx

has more relaxed conditions, once the machine learning solution is to be moved to the "production environment", it needs to be embedded into software or system that is already used in the production. Obtaining the permission or security clearance for software or hardware needed for the machine learning solution is often a lengthy process which can stall the operationalisation. Also, automating status-quo manual processes by machine learning inevitably impacts the regular work of human staff and this makes it hard to obtain buy-in about the machine learning solution if consultation and communication with stakeholders did not take place in the early stage of the journey.

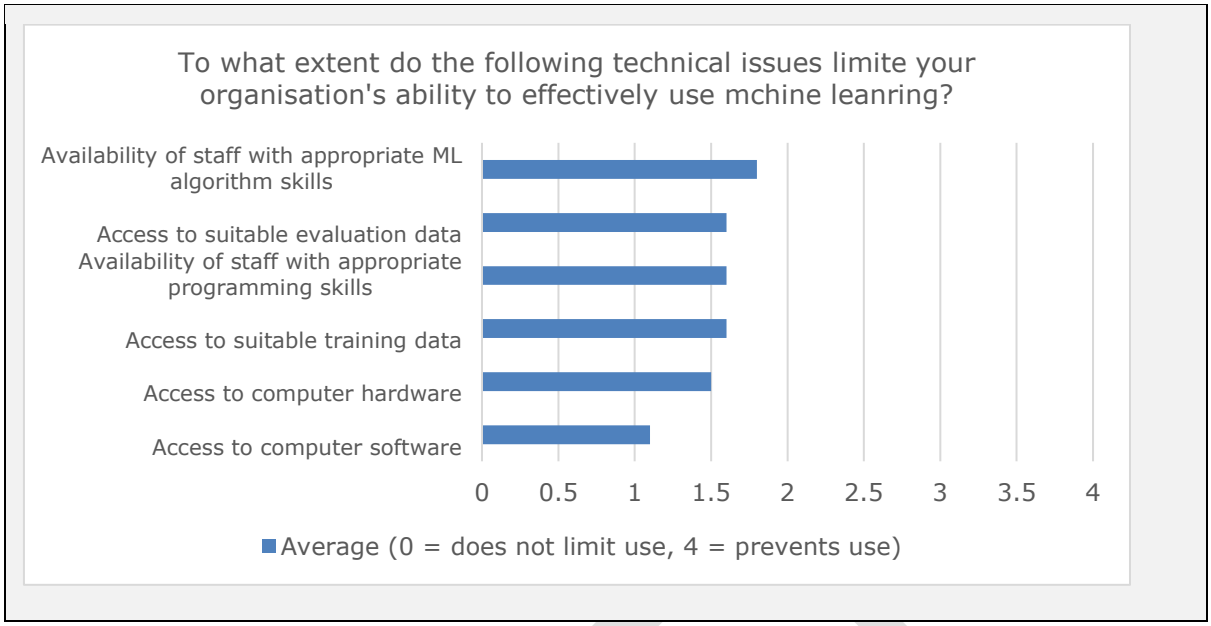
In this paper, typical steps that statistical organisations would take from the machine learning experiment to its deployment in production are described with some of technical and organisational issues and constraints often experienced in each stage. Note that, while the steps are in the logical order, 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.

Box 1. Findings from the Machine Learning Project Survey on Integration

The HLG-MOS Machine Learning Project Work Package 3 team aimed to identify and address the challenges to integration and production deployment. For this, a short online questionnaire designed to get a high-level overview of the key challenges and successes was conducted in 2020. Following charts summarise the results from the questions asking organisational and technical challenges[4].



[4] For the complete survey results and deeper investigation into key questions, see the full report on https://statswiki.unece.org/display/ML/WP3+-+Integration


To what extent do the following technical issues limite your organisation's ability to effectively use mchine leanring?
Availability of staff with appropriate ML algorithm skills
Access to suitable evaluation data
Availability of staff with appropriate programming skills
Access to suitable training data
Access to computer hardware
Access to computer software
0
0.5
1
1.5
2
2.5
3
3.5
4
Average (0 = does not limit use, 4 = prevents use)

2. Journey from ML Experiment to Production

2.1. Understand Business Needs

The machine learning journey may start in different ways. In some cases, it is initiated by curious and committed individuals who want to improve the status quo. It may start with directives from senior management or as a pure research project without a particular plan to put the ML solution in production. However, regardless of how it started in the beginning, a ML solution that does not address any business need in the organisation does not lend itself in production. Therefore, understanding business needs is a critical first step in the journey to production.

The business needs affect various aspects around the machine learning system that will be eventually put into production and decisions to be made along the journey to production. For example, decisions on which quality dimensions (i.e., accuracy, timeliness, cost-effectiveness, explainability, reproducibility) to focus may change depending on the specific business problem. For example, if the purpose of the machine learning solution is to assist human experts (e.g., ML proposing the top 3 likely building types for each building image in image classification tasks), the human experts might be more interested in getting accurate predictions than explainable predictions. On the other hand, if the machine learning solution is used for forecasting economic indicators that directly affect the policy and business decisions, one might prefer an explainable model than an accurate but completely black-box model.

It is also important to understand not only "what" is needed (i.e., business needs), but also "who" needs the machine learning solution (i.e., end-users, business owners) in this stage. Data scientists and engineers might be the ones who are mainly responsible for the development of the machine learning solution, but it is eventually the business owners who need to use the end-solution for their daily work. Therefore, the solution should be designed considering how it would be used by and interact with the end-users who are mostly not data scientists. Initiating a consultation with those in the business area at the early stage helps to better reflect the ultimate needs of users and set ground for their buy-in. The proper expectation management with users during the journey is also important as they may not be familiar with what ML can realistically accomplish [34].

The machine learning solution can heavily rely on non-technical and non-machine learning factors. For example, if the machine learning solution is for automating the coding process for statistical classification, the information about the classification is critical as its update can change the data set on which the machine learning model is to be trained. Hence, keeping contact with those who are responsible for maintaining the classification is needed so that the information could be taken into account during the development as well as the monitoring phase (see 2.6 Deploy the Model).

Machine learning is a relatively new area of work in the statistical organisation. Some organisations are equipped with a centralized place that is dedicated to coordinate machine learning-related works and projects, but many are in the process of determining the right organisational structure to accommodate this new work area. Given that a machine learning project needs expertise from various areas (e.g., IT, subject matter domain, methodology) with different work priority and schedule, coordinating and aligning the works of these different divisions might cause great difficulties and one should be aware of this challenge along the way. On the other hand, gaining and maintaining the enthusiastic support of potential end users can be an important factor in ensuring that a machine learning solution makes it all the way into production.

2.2. Assess Preliminary Feasibility

In this stage of the Preliminary Feasibility Study (PFS), an initial evaluation of the suitability of machine learning solution with respect to the business problem, data and technical resources (software and hardware) is conducted.

Machine learning is not a panacea and one should not expect it would resolve every business problem. While the Proof-of-Concept (PoC) experiment in the next stage would provide more concrete ideas on how machine learning would work for the given business problem, a few high-level questions can help gauge the feasibility of machine learning, such as: are there large data sets, does 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. Research on the growing body of machine learning use cases, particularly within the official statistics community[5] to see what types of machine learning methods were used and how they worked within constraints of statistical organisations, help avoid re-inventing wheels and save a significant amount of time and effort in advance. It also often happens that different teams within the same organisation work on the similar problems without knowing each other, hence scanning within the organisation is important to avoid duplication of efforts and potentially develop a common service that is applicable for different programmes within the organisation.

Many machine learning 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 for ensuring a long-term value of the machine learning solution. For example, if the solution is for the production of monthly urbanisation index based on satellite images between Census years, it is essential to have a secure and regular access to the data during this period. Just like traditional statistical methods, or perhaps even more, machine learning methods are subject to classical data issues. One might investigate the characteristics and quality of data by asking questions such as: how it is collected (e.g., web, survey, administrative), what population it covers, have there been any change how the data is prepared (e.g., change in editing and weighting methods).

The assessment of technical requirements and constraints is a crucial component of the evaluation in this stage. Many developments in the machine learning field have been occurring around the open-source software (e.g., python, R), which might not be supported by corporate IT systems. Also, some machine learning methods require large computational resources (e.g., GPU, TPU) that may not be available in the organisation. In this case, one might need to use the software temporarily for the PoC experiment or explore options for other environments (e.g., cloud). Either way, one should take into account time and resources into later stages when the machine learning model is moved into production for the appropriate tool to be acquired and/or the code re-worked (e.g., from python to a programming language that is supported by the corporate system).

Note that it can be difficult to convince a business area of the value of a machine learning solution without an example worked directly on the data in question, but at the same time, it often happens that the data may not be available for the immediate PoC experiment or accessible to those who need to run the experiment for various reasons (e.g., data security, administrative hurdle, lack of hardware to accommodate the volume of the data). If such constraints cause the experiment to be conducted in an environment where only public or synthetic data is available, this may shift elements of the PFS into the later stages. In such a case, the PFS would focus on demonstrating that the method or approach in question is capable of solving the type of problem at hand with the intention of acquiring an initial commitment of resources to address the technical constraints and apply the method in an environment where appropriate data is

[5] For example, https://statswiki.unece.org/display/ML/Studies+and+Codes, https://marketplace.officialstatistics.org/methods

available. The proof of concept or business cases development stages could then be used to show that the method in question works well for the particular problem using real data.

2.3. Develop Proof of Concept

The proof of concept (PoC) often proceeds the full-scale model development 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. PoC model can also provide opportunity to obtain quantitative results to be used to support the business case and discover issues unexpected from a desktop research and a preliminary feasibility assessment.

To measure the performance of the PoC model, detailed and quantifiable quality criteria to judge success such as accuracy, time and cost should be established. The choice of quality metric should take into account business needs and context. For example, when deciding accuracy measure, one might give more emphasis on precision metric than recall metric when false positive is costly, and vice versa.

Although this is a technical stage requiring data science and machine learning expertise, the domain experts, business owner and end-users play an important role, and sometimes, their involvement can be a prerequisite. In the case of supervised machine learning, for example, ML methods need labelled/annotated data set to train and test the model. Given that ML models "learn" from data, the quality of data set that one feeds into the algorithms is critical. As the old maxim "garbage in, garbage out" goes, a poor data set results in a poor model. The importance of high-quality training data was highlighted by several pilot studies in the ML project that "*successful pilot studies have shown that establishing a "ground truth" or "golden data set" that is created manually and is deemed to be accurate and free of errors is of prime importance*"[6]. This data set is created through the careful manual operation by human staff. Even when such data set already exists (e.g., manually edited data from the past surveys), the domain experts can provide important insights in the machine learning model development process during, for example, feature engineering and model diagnostic (see more below).

The development of the machine learning model roughly follows steps as below:

- **Data collection and ingestion** where data sets needed for building machine learning models are gathered together. Often, new needs for additional data arises during the model development and the data collection steps may need to be repeated. As discussed earlier, the data set at the PoC model stage may not be the real data set, but synthetic data, publicly available data or a small subset of the real data. In this case, PoC development team should be aware of the limitation caused by data (e.g., complexity, size) and reflect this when interpreting the results
- **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. New features (input variables) that are not in the raw data set but deemed important can be created through, for example, consultations with the subject-matter experts. For non-conventional form of data such as textual data, this is where the original form is converted into a numerical form (e.g., vectorization of text data).

[6] https://statswiki.unece.org/display/ML/WP1+-+Theme+1+Coding+and+Classification+Report

- **Model training** where the different machine learning models are trained on the data set prepared from the previous step. To avoid the overfitting problem, the data set is split into a training set and a testing set and only the former is used in this stage so that the model can be tested with an independent data set that it has not been exposed to. The hyperparameters of the models can be either set manually or determined by splitting the training set further or using cross-validation method[7].
- **Model testing** where the final evaluation of the model is conducted on the test set. Note that while accuracy is the most commonly used quality dimension for the evaluation of machine learning models, one should also pay attention to other quality dimensions such as time (e.g., how long does it take for training the models, how long does it take to make prediction), cost (e.g., was special computing hardware needed?). All relevant findings and constraints should be documented so that they could be used for the next stages when deciding whether the machine learning models can be moved into production or not.

2.4. Prepare a Comprehensive Business Case

Based on the preliminary feasibility assessment and findings from the proof of concept, a comprehensive business case is prepared to get approval to develop the model for the production. Machine learning 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. Business case plays an important role to ensure that all those involved have a common understanding of objectives and requirements. It is also vital to obtain the substantial resources often needed to move a solution into production. To maximise the return on investment, it is recommended to 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. Business case would typically include elements such as:

- **Problem statement**: description of "as-is" process and solution (e.g., manual coding by human coders, rule-based editing) including its cost, time, level of quality with highlights on any inefficiencies. This can include an assessment of alternative solutions other than ML (e.g., if manual coding can be replaced by rule-based coding, why ML?).
- **Business value addition**: description of how ML solutions can contribute to the business. The results from the PoC can provide a concrete idea on the added value in terms of accuracy, time and cost. Exploration of different areas where the ML solution can be expanded to (e.g., other business lines that use the same classification system) could help making a strong case. One should make sure that the value proposition is aligned with the corporate innovation strategy (e.g., transition to cloud, open-source software).
- **Cost**: description and estimation of cost involved such as purchase of new IT resource, staff working hours and cloud storage if needed. Unlike standard software, machine learning requires continuous maintenance (see stage 6), therefore estimated cost should include not only initial resource (time and cost) investment for the deployment, but also monitoring and maintenance.
- **Stakeholder**: identification of stakeholders (e.g., business owner, data science developer, data owner, subject matter expert, human coder) and analysis of their expectations and concerns which will help gaining buy-in.

[7] https://scikit-learn.org/stable/modules/cross_validation.html

- **Project plan**: identification of tasks and steps to follow from the development of the ML solution to its sign-off (deployment). The plan should include the estimation of resources required and timeline for each step, in particular, time needed for the acquisition and security vetting of software or data. Details for model development and strategy such as how to evaluate the model accuracy (e.g., establishing the gold standard data set) and how to find the threshold for ML-based prediction can also be included.
- **Operational business process**: description of the process steps and flow to be followed when ML solution is put in the production including how it would interact with existing business processes and components.
- **Data**: description of data needed for the model development and how to acquire it, assessment of its quality and impact on the model.
- **Governance**: description the roles of individuals (e.g., business owners, ML developers), maintenance plans (e.g., how to monitor the deployed model in the production, how to determine the re-training of the model, who should do these). Analysis of potential risk in terms of ethics (Box 2), privacy and security.
- **Risk assessment**: if PoC was done in the different environment than in the production (e.g., synthetic data), limitation and potential issues that could occur during the development can be described here

Note that depending on the organisation policy and practice, the business case might be required before the development of PoC model or prepared in parallel with the PoC experiment. In such cases, the weight given to different elements of the business case may vary from a business case developed after a PoC.

Box 2. The Ethics of Machine Learning

By the UK Statistics Authority's Centre for Applied Data Ethics

The use of machine learning provides substantial benefits for research and statistics. However, when embarking on any statistical or research project, using any method, it is important to consider any possible ethical issues relating to the collection, access, use and storage of data. This helps to both reduce potential harm to anyone involved in the research (i.e., data subjects and others who may be impacted by the work) and maintain public acceptability around the production of research and statistics using such methods. It is therefore important that National Statistical Offices (NSOs) take a lead role in considering the application of data ethics to their work and are seen to use data in ethically appropriate ways.

Following the identification of a need for further applied ethics guidance in the use of machine learning for the production of official statistics by the international research and statistical community, the UK Statistics Authority's Centre for Applied Data Ethics developed the ethics guidance on the application of machine learning to the research and official statistics context[8], as part of the ONS-UNECE Machine Learning Group 2021's Data Ethics Workstream. The guidance focuses on four main areas:

- The importance of minimising and mitigating social bias;
- The need to consider the transparency and explainability of machine learning research;
- The importance of maintaining accountability within all aspects of machine learning processes, and;
- The need to consider the confidentiality and privacy risks arising from the data use.

Minimising and mitigating social bias, which can creep into machine learning projects in a number of ways is imperative in ensuring that the research and statistics NSOs produce have accuracy and validity, and do not perpetuate negative (or positive) social discriminatory practices. Bias of course is not particular to machine learning, however there are a number of different ways it can be embedded into machine learning projects and can be particularly complex to eradicate.

Machine learning projects may also pose several risks to data protection and privacy. Not only does machine learning require the use of large, representative data sets for training the model, which may contain sensitive information (access to which may raise questions of data protection), but the models may also be able to identify nuanced differences between data points, thus enabling the correlation of certain characteristics to potentially sensitive information. Machine learning methods also raise questions relating to the confidentiality of data, the use of third party and linked data and the potential for re-identification. This means that it is important that stakeholders maintain accountability within all aspects of machine learning processes, ensuring that models are used only for their intended purposes, and that different stakeholders are aware of their responsibilities. Moreover, transparency is key - the decisions that are made about data, analysis, and methods,

[8] The full report is available on: https://uksa.statisticsauthority.gov.uk/publication/ethical-considerations-in-the-use-of-machine-learning-for-research-and-statistics/ The Centre proactively welcomes the views and comments of others on their guidance to ensure that it is supporting the broader international conversation

should be openly and honestly documented, and communicated in a way that allows others to evaluate them.

Whilst complex, these issues can all be mitigated if an "ethics by design" approach is taken when using machine learning and need not be a barrier for those embarking on machine learning projects. It is important that the official statistics community encourages researchers and statisticians to think about the ethics of their projects at the earliest opportunity, leading by example, and ensuring open, honest, and transparent communication between stakeholders.

Going forward, it would be beneficial for the official statistics community to continue to discuss these issues further in collaboration with other groups exploring this issue within a broader context (e.g., law, policy, data governance) as recommended by the UNECE HLG-MOS Machine Learning Project. The ethics guidance will provide an initial foundation from which to do this.

2.5. Develop the Model

Once the business case is approved, the development of the production-level model is initiated. At the 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). However, there are several differences coming from data, model and IT environment.

- **Data**: while the PoC experiment might have been conducted on smaller scale data (or even not real data), the model developed in this stage uses the real-world data. This can create complications such as data storage issues when the volume of data is large. Also, when the data needed for the model development come from different sources in different formats, pulling the data and preparing them for downstream consumption (let alone getting the data sets themselves) can be challenging and take a lengthy time. Some features may be available in a different format or as a slightly different concept (e.g., income for family instead of income for household). The production-level model at this stage requires a reliable supply of the data; some data sources used in the PoC experiment may need to be dropped if its supply is deemed unreliable.
- **Model**: the PoC model can be a basis for or a component of the production-level model. But the business problem after comprehensive business case might be different from those in the PoC stage (e.g., new classification system added as the target classifications, prediction frequency increased, higher accuracy requirement) hence may require a different set of evaluation criteria or different priority in choosing the final model. Legal and ethical considerations may play a greater role in deciding the model in this stage. Also, unlike PoC model that could be run in a stand-alone experiment, model developed in this stage needs to put in the existing process, hence there may be additional requirements in order for the outputs of the model to be fed into downstream systems (e.g., transformation of the outputs, format changes).
- **IT Environment**: the production environment might be different from the experiment environment (where PoC model is built and tested). The software used for ML experiments (e.g., Python, R) may not be supported in the production environment. Therefore, one may need to develop a wrapper for the model and connect to the existing system, or completely re-write the ML codes in a software language that is supported by the production system. Note that some ML algorithms have stochastic elements that might be difficult to reproduce from

one language/system to another (e.g., the same seed might produce a different outcome) making it harder to be sure if the model is producing the same results. The decision regarding what IT environment to use needs to be made in advance as certain ML algorithms may not be readily available in some software languages, hence affect the choice of the ML algorithms to try in this model development stage.

The ML model development is iterative process, one may need to repeat steps from data collection to model training/testing many times before the final model. Documenting this process and versioning of milestone models are critical in this stage for several reasons:

- For reproducibility: the original developer may not be able to complete the development or the model may need to be handed over to a different person or team
- For monitoring of the model: the changes in the distribution of data features and performance metrics can be used for detecting concept drift and model drift once the model is deployed (see stage 6).
- For re-usability: some of features and model components can be re-used for the development of other ML models in the organisation.

It is also important to have workflow around the ML solution established. For example, if the model is used to assist human staff for the data editing, it should be decided at what point the model interact with human staff during the editing process and, if needed, how the feedback from human staff (e.g., whether the ML prediction was correct or not) can be brought back to improve the model. If the model is used to make land cover prediction based on satellite data for regular statistics, the workflow should be set up to determine when and how the data is retrieved (e.g., manual batch download, automatic API pull).

The ML model is often packaged into an application tool to provide a user-friendly interface. This is main activity in the next stage (model deployment) but can be initiated in parallel with this stage and be connected to the model once it is finalized as the model development stage may take a long time.

As the use of ML spread and scaled up, statistical organisations would need systems that can support the ML development in a more systematic and efficient way (e.g., ML lifecycle management, repository of models and features).

2.6. Deploy the Model

What is the model deployment?

The ML model is a tool designed to address a business problem identified in the stage 1. To provide its business value, therefore, the ML model, which may exist as programming script on the data scientist's computer, should be made available to the end-user. In this sense, model deployment can be considered as a process of integrating the model in the existing system so that its results (e.g., predictions from the ML model) are available to the users.

How to deploy?

Depending on the problem and the users (which can be either humans or another software in the bigger system), deployment can take different paths. For example, when the model predictions are fed into another service in a fully automated manner, API built around the model may suffice in facilitating the interactions between the ML model and other connected services. If the model is used to semi-automate the coding and classification process by assisting human staffs (e.g., proposing top 5 most likely codes for a given text description), service application with user-friendly interface in

combination of API can help humans to interact with the model (Box 3). On the other hand, when the model is not used for intermediate process, but for the estimation of final statistics (e.g., forecasting economic indicators), the model may not need a front-end for the end-users (public), as they are mostly interested in the final data product rather than feeding data into the model directly and receiving the forecasting results.

In the deployment stage, the model should be packaged so that it can operate in any environment or system as it did in the local computer of the developers. ML model can depend on the combination of specific software libraries (versions) which may crash in system of different team. Advent of containerisation tools such as Docker has facilitated this process and simplified the complex dependency issues.

Given that the ML model is often handed over to a team different from the original development team after its deployment, it is also important that all relevant information regarding the model (e.g., training data used, hyperparameters, codes) is carefully documented. This will assist later users and support staff in understanding when the model is deviating from expected behaviour and how to address any issues. If the end users have little experience with ML models, it may be useful to consider training sessions as a part of the model handover process.

Monitoring plan after deployment

ML model is built based on patterns learned from data in the past, but after the deployment, the model needs to make predictions on the new data that it was not exposed before and these patterns can change over time. This happens due to change of data on which the model needs to make predictions (e.g., new products in market, new type of jobs) or change of relationship between input features and output (e.g., update of statistical classification system). Over time, therefore, the model starts to decay and it is important to have a governance plan in place before the deployment so that the model can 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. It will be helpful to have a clear plan for who will be responsible for monitoring the model performance and for adjusting or retraining it should that become necessary. Establishing communication channels with those who are maintaining the artefacts on which the model depends on (e.g., data owner, classification maintenance team) in the management plan could also help ensuring that information on any big updates to be shared in advance and acted on accordingly.

Box 3. Designing and Deploying a Machine Learning Solution for Official Statistics: The IMF Experience

Prepared by International Monetary Fund (IMF) Statistics Department (Ayoub Mharzi, Alberto Sanchez, Marco Marini, Alessandra Sozzi, Lamya Kejji and Yamil Vargas)

A successful Machine Learning (ML) solution for official statistics requires a careful design of the different stages of the data lifecycle. For example, data preparation and data ingestion are critical steps for an efficient upload of the input data. Furthermore, feedback from end users is essential to design the functionalities to be included in the solution interface – which is why it is important for end users to be involved in the design of the solution from the start of the project.

This box provides an overview of critical areas that should be considered when designing and deploying a ML solution for a data-producing organisation, drawing from our initial experience in the UNECE HLG-MOS ML Project to build an automated coding tool for economic and financial indicators collected from IMF member countries[9].

Who Will be Using the Tool?

Users should be involved throughout all stages of the implementation of an ML solution. Defining the target groups of users of the ML solution is a key step for shaping the final tool, as it helps to identify all user roles and their interaction with the ML solution, the data format to be used by different individuals, and the end-to-end workflow of the solution. Identifying the target audience will also impact how the data will be handled behind the scenes: data cleaning, formatting, feature selection, and other steps. We recommend spending the necessary time to clearly identify and engage with the target audience from the beginning of the project. Based on our experience, it is helpful to have a potential end user part of the project team.

Data Format

The data upload function is typically the first point of interaction between the end user and the solution. In our project, the first step for the user is to upload the description of indicators for which our teams need to generate codes for. In this regard, it is important to consider the possible formats of the data (Excel, CSV, XML, etc.) and the different data presentations (tables structures, headers, etc.). On the backend, data are extracted from the input files, processed, and prepared to feed into the ML models. There are many places where this process can break, hence it is important to find the right balance and try not to overengineer this step. It is advisable to develop a template to guide the user on how to prepare the input data for the tool.

Interactivity, Intuitiveness and Usability of the User Interface

A user interface should be developed to simplify the use and delivery of your ML solution. A well designed and functioning user interface will help your target audience to be on board with your solution. An important aspect to consider is efficiency, as the user will be more inclined to switch to the solution if it takes little time to run the full process. Other factors to consider are as follow:

- **Explainable ML and transparency**: your user interface should provide a certain level of interactivity that allows users to see what is happening in the backend and how your overall solution is operating. The main argument

[9] The full report on the designing and deploying a ML solution for Official Statistics from IMF will be made available on: https://statswiki.unece.org/display/ML/Machine+Learning+Group+2021

against the use of ML solutions is that they are black boxes and very difficult to explain and interpret. However, one can incorporate functionalities allowing the user to get some details on the predictions, models, feature extraction techniques used and performance measures. Although this might not be useful to all users, it can help those more familiar with ML to have more understanding on what's happening in the background and potentially provide feedback on it, and propose new ideas.

- **Intuitiveness and usability**: your ML solution will be driven by two goals: (i) provide a solution to an unanswered problem; or (ii) improve an existing process. In both cases, intuitiveness and usability are key aspects to get buy-in from users. The end-to-end process behind the interface should be streamlined as much as possible, to eliminate unnecessary steps that may impact the intuitiveness of the tool. It may be useful to let users outside of the project group run the tool and gather feedback. When the proposed ML solution (and user interface) aims to replace an existing process, the transition for the users should be as smooth as possible. Because introducing changes to existing processes is always challenging, it is important to help the end user to better transition to the new solution. In our project, users manually assign and review codes for the indicators they are presented with. Our goal is to automate the code generation by using ML techniques. However, we do not want to force users to review the predicted codes directly on the solution interface. Instead, we will allow them to download the ML predicted results in a more familiar file format (e.g., Excel) to complete their review process in their preferred environment.

User Feedback and Retraining the Model

A ML solution should always incorporate feedback from its own mistakes. Two ways to learn is through user feedback and model retraining with the new inputs. Tackling this task will be on a case-by-case basis. For our solution, we are planning to implement the following steps:

- Identify subject matter experts to review, in an initial stage, both the manual and ML-based assigned codes. This will help provide an accurate assessment of the predictions and improve the review process moving forward. Subject matter experts should be staff having the needed level of expertise to accurately review the predictions;
- Split the review tasks among different users, either to reduce the burden of the review process and to double check predictions by the group of subject matter experts;
- Identify data domains with higher-quality predictions. For these domains, predictions should be automatically fed into the training dataset. For domains with lower-than-average accuracy, subject matter experts' review is needed to add these predictions.
- Retrain the model using inputs adjusted by the subject matter experts.

3. Conclusion

Machine learning holds a great potential for statistical organisations, it can make the existing processes more efficient and allow the production of new statistics and services that could meet the growing needs of society. While there is increasing evidence demonstrating its potential, moving the machine learning solutions from experiments to production is often a very challenging task. The development of machine learning solutions requires a close collaboration among multidisciplinary stakeholders, the buy-in from end users and establishing a system to monitor and maintain the deployed ML solution. Machine learning also involves technical challenges as it often requires software and hardware that are not often readily available or supported in the organisation.

This paper described the six stages toward the operationalisation of ML solution, from the business needs identification stage to the model deployment stage. Several factors play important roles in this journey:

- **Business needs** affect many decisions to be made along the journey, such as prioritisation of quality dimensions and workflows around the solution. They should be identified at the beginning with a broad consultation with stakeholders.
- Design of ML model and interface should take into account the needs and profile of **end users** to increase the usability of the solution and buy-in.
- **IT requirements** (software, hardware) can affect the journey to production significantly. The difference between "experiment environment" and "production environment" and the constraints that arise from this should be identified at the early stage and incorporated in the planning.
- ML models are built based on the **data**, hence ensuring the quality of data, obtaining access to data as well as addressing any privacy and ethical issues involved are important.
- Even a high-performing model can quickly decay once it is deployed. A **maintenance** system should be in place to monitor the model as well as data before sign-off.