UNECE

Machine Learning for Official Statistics

UNITED NATIONS

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

Machine Learning for Official Statistics

UNITED NATIONS

Geneva, 2021

This publication is issued in English.

Preface

Machine Learning holds a great potential for statistical organisations. It can make the production of statistics more efficient by automating certain processes or assisting humans to carry out the processes. It also allows statistical organisations to use new types of data such as social media data and imagery.

Many national and international statistical organisations are exploring how machine learning can be used to increase the relevance and the quality of official statistics in an environment of growing demands for trusted information, rapidly developing and accessible technologies, and numerous competitors. While the specific business environments may vary depending on the country, these statistical organisations face similar types of challenges which can benefit from sharing knowledge, experiences and collaborating on developing common solutions within the broad official statistical community.

This publication presents the practical applications of machine learning in three working areas within statistical organisations and discusses their value added, challenges and lessons learned. It also includes a quality framework that could help guiding the choice of methods, challenges that arise when integrating machine learning into statistical production, and key steps for moving machine learning from the experimental stage to the production stage and concludes with key messages on advancing the use of machine learning for the production of official statistics.

This publication is based on the results from two international initiatives: the UNECE High-Level Group on Modernisation of Official Statistics (HLG-MOS) Machine Learning Project (2019-2020) and the United Kingdom's Office for National Statistics (ONS) – UNECE Machine Learning Group 2021, and approved by the HLG-MOS.

Acknowledgements

This publication is prepared based on the reports from the UNECE HLG-MOS Machine Learning Project (ML Project) and the United Kingdom's Office for National Statistics (ONS) – UNECE Machine Learning Group 2021 (ML Group 2021) as below:

- Chapter 1 and Chapter 6 – ML Project *Final Report* by Claude Julien (UNECE Project Manager)
- Chapter 2 – ML Project *Coding and Classification Theme Report* by Claus Sthamer (United Kingdom)
- Chapter 3 – ML Project *Coding and Classification Theme Report* by Claus Sthamer (United Kingdom)*, Edit and Imputation Theme Report* by Florian Dumpert (Germany) and *Imagery Theme Report* by Abel Coronado and Jimena Juárez (Mexico)
- Chapter 4 – ML Project *Quality Framework for Statistical Algorithms* by Siu-Ming Tam (Australia), Bart Buelens (Belgium), Wesley Yung (Canada), Florian Dumpert (Germany), Gabriele Ascari, Fabiana Rocci (Italy), Joep Burger (Netherlands), Hugh Chipman (Acadia University) and InKyung Choi (UNECE)
- Chapter 5 – ML Group 2021 *Journey from Machine Learning Experiment to Production* by Claire Clarke (Australia) and InKyung Choi (UNECE)

The original reports can be found on the UNECE Wiki Space *Machine Learning for Official Statistics* (https://statswiki.unece.org/display/ML). We appreciate permission of the authors to use the reports for this publication. Further adaption and editorial changes were made by the UNECE Secretariat (InKyung Choi, Taeke Anton Gjaltema, Christophe Jones and Wai Kit Si Tou).

ML Project and ML Group 2021, on which this publication is based, would have not been possible without many individuals around the world in the official statistics community. The contributions from the members, in particular the ML Project leads, Wesley Yung (Canada), Eric Deeben (United Kingdom), Alexander Measure (United States of America) and Claude Julien (UNECE Project Manager), are greatly appreciated.

List of Chapters

List of Figures

List of Tables

List of Box

1. Background

Modernisation of Statistical Organisations

National statistical organisations (NSOs) are being challenged to be more responsive to the increasing need for more relevant, timely, detailed and accessible statistical information and data services. NSOs also need to distil the ever-increasing amount of data available from a wide variety of sources, in various formats and levels of quality to produce information that can be trusted and used to make data-driven policy decisions. At the same time, they are under pressure to meet these expectations within existing budget levels.

NSOs also face competition from a growing number of private companies who produce and communicate statistics in a more timely and accessible manner that attracts the attention of policy makers and many other users. These companies could produce these statistics for several reasons such as a quick access to alternative data sources and cutting-edge technologies as well as fewer constraints on quality and transparency compared to NSOs.

However, NSOs also hold a competitive advantage in several aspects. They have considerable collective expertise in efficiently integrating diverse sources of data. They are also more transparent by publishing details on data sources, methods and various indicators, and their legal obligation to respect privacy and protect against disclosure help NSOs to gain public trust. Furthermore, the capacity to do so not only lies within each NSO, but increasingly through a network of professionals around the world that are brought together through collaborative initiatives in the broad official statistics community.

In addition to counting on their individual and collective expertise, statistical organisations must have an adaptive culture to remain relevant, by responding to the timely needs of stakeholders in a continuously responsible manner. The pandemic crisis has also "*changed the relative importance of the different components of quality, with a much greater focus on timeliness*"[1]. To make statistical information and services relevant to the growing needs of users, and to produce them efficiently in cost and time, NSOs have to adapt to and embrace new technologies and data sources.

Machine Learning for Official Statistics

With the increased computing power, methodological advances and an unprecedented amount of data arising from the digitalisation of society and business, machine learning has been making breakthroughs across many disciplines. Computers have learned to draw a painting in the style of Rembrandt[2], to write an article just like humans[3] and to determine the 3D shape of proteins[4]. Indeed, "*any industry with very large amounts of data — so much that humans can't possibly analyze or understand it on their own — can utilize artificial intelligence*"[5] (machine learning) and many private companies are now utilising the technology for providing personalised recommendation and tailored

[1] Conference of European Statisticians (2021) Summary of key points from the Chief Statisticians' sprint on "Innovation, business continuity and staff motivation during the pandemic" (https://unece.org/sites/default/files/2021-06/2107622E.pdf)
[2] https://www.nextrembrandt.com/
[3] https://amp.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3
[4] https://www.nature.com/articles/d41586-020-03348-4
[5] https://www.gartner.com/smarterwithgartner/the-disruptive-power-of-artificial-intelligence/

customer services. Knowingly or unknowing, machine learning has slipped into daily lives of people in the current society.

The interest in machine learning in the official statistics community has been growing rapidly. Many national and international organisations are investigating how it can be used to increase the relevance and quality of official statistics in an environment of growing demands for trusted information, rapidly developing and accessible technologies, and numerous competitors. The position paper by the Blue Sky Thinking Network of the UNECE High-Level Group for the Modernisation of Official Statistics (HLG-MOS) postulated in 2018 that: *"for the processing of some secondary data sources (e.g., administrative sources, big data, Internet of Things), it seems essential to look into opportunities offered by machine learning, while also for primary data, the technique might offer added value"*[6].

However, the use of machine learning for official statistics requires a more cautious approach as NSOs operate in a different way than companies in the private sector. There is a great weight of responsibility associated with each number they produce. NSOs cannot simply use whichever technique that appears to work but then change because the results start going astray[7]. Credibility of official statistics are based on the Fundamental Principles of Official Statistics[8] to ensure that they are produced in a sound, reliable and transparent manner. The official statistics community needs to make sure that new technology and methods are used in a responsible way, so as to maintain the public trust bestowed on them.

The UNECE HLG-MOS Machine Learning Project

To facilitate the investigation of the use of machine learning for official statistics, and to consolidate the lessons learned by statistical organisations, the UNECE HLG-MOS launched a Machine Learning Project in March 2019. The Project aimed to demonstrate the added value of machine learning, i.e., whether it can help in the production of more relevant, timely, accurate and trusted data in an efficient manner. The Project also aimed at identifying and addressing some common challenges encountered when incorporating machine learning in organisations and their production processes. The work of the Project was organised around following three work packages (WPs):

- Work Package (WP) 1. Pilot Studies;
- Work Package (WP) 2. Quality; and
- Work Package (WP) 3. Integration Challenges.

One can combine these work packages and their goals into a single sentence: to integrate demonstrated machine learning solutions (WP1) within production processes (WP3) in a sound and efficient manner (WP2).

The Project started with a small group of about 10 participants, but grew to more than 120 participants from 23 countries across the world when the Project was completed in 2020[9]. The Project produced 21 pilot studies, thematic reports that summarised and analysed the pilot studies, a quality framework, and a report that identified the common

---

6 https://statswiki.unece.org/download/attachments/223150364/The%20use%20of%20machine%20learning%20in%20official%20statistics.pdf?version=2

7 https://en.wikipedia.org/wiki/Google_Flu_Trends

8 https://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx

9 The international collaborative initiative on machine learning is succeeded by the U.K. Office for National Statistics (ONS) – UNECE Machine Learning Group 2021. For more information, see ML Group 2021 wiki page (https://statswiki.unece.org/display/ML/Machine+Learning+Group+2021)

challenges in integrating machine learning into production processes[10]. This publication is based on the main findings and lessons learned from the Project.

Structure of the Publication

The remainder of this publication consists of a further five Chapters. Chapter 2 introduces some commonly used machine learning algorithms and accuracy metrics that are used to assess the performance of machine learning models. The practical applications of machine learning in three working areas within statistical organisations are examined in Chapter 3. Chapter 4 discusses quality dimensions that can provide guidance on the choice of algorithm for the production process. Challenges that arise when integrating machine learning into statistical production, and key steps for moving machine learning from the experimental stage to the production stage are described in Chapter 5. Lastly, this publication concludes with key messages on advancing the use of machine learning for the production of official statistics and recommendations for future work in Chapter 6.

Chapter	Target Audience
Chapter 2. Machine Learning Algorithms	(reference material for later chapters)
Chapter 3. Machine Learning Application Areas	Methodologists, Statisticians, Data Scientists
Chapter 4. Quality Framework for Statistical Algorithms	Methodologists, Statisticians, Data Scientists
Chapter 5. Journey from Machine Learning Experiment to Production	Data Scientists, Project Managers, Line Managers
Chapter 6. Key Messages and Conclusion	Senior Managers

[10] All project materials (e.g., reports, codes, data, presentation, papers) are available on the UNECE Wiki (https://statswiki.unece.org/display/ML/HLG-MOS+Machine+Learning+Project)

2. Machine Learning

Machine learning is *a "field of study that gives computer the ability to learn without explicitly being programmed"*[11]. It is closely related to, and uses methods from, other fields such as statistics, computer science and artificial intelligence.

In Chapter 2.1, some commonly used machine learning algorithms are introduced. The Chapter does not aim to provide an exhaustive list of machine learning algorithms nor technical details, but rather to provide a brief overview of some of the machine learning algorithms that are referred in Chapter 3[12]. Chapter 2.2 describes accuracy metrics that are used to assess the performance of machine learning models.

Given that machine learning is comprised of and influenced by several disciplines, there are different terms that are used interchangeably as well as concepts that are not commonly used in statistics. Below is a working definition of some of key terms and concepts that appear in the rest of this Chapter.

- **Algorithm** is *a "finite sequence of well-defined, computer-implementable instructions, typically to solve a class of specific problems or to perform a computation"*[13] (synonym: method);
- **Feature** is "*an input variable used in making predictions*"[14] (synonym: input, predictor, explanatory variable, independent variable). Machine learning is often used for data with a large number of features. In this case, feature engineering can be conducted before applying the machine learning algorithm to extract or find a smaller set of features that are more useful for the prediction;
- **Target variable** is a variable that needs to be predicted, such as occupation of a person, type of land use (synonym: output, response variable, dependent variable). Machine learning algorithms can be grouped into two different types:
    - **Supervised machine learning** where a data set has known target values and machines are instructed to learn the relationships between features and the target; and
    - **Unsupervised machine learning** where there is no target in the data. set, and the algorithm needs to figure out any patterns on its own.
- **Model** is an output of a machine learning algorithm that is run on the data set. Note that while an algorithm, as a set of instructions to be applied to a data set, exists prior to data, the model is obtained after applying the algorithm to the data set;
- **Training** is a process of determining the model. This is where "learning" takes place in machine learning. Once the model is established it can be tested to measure its prediction accuracy, with respect to a test data set (a part of the data set that is set aside and not used for training). Evaluating the accuracy of a model with the same data set that was used to build the model often leads to the overestimation of accuracy. Hence, the partitioning of data sets into separate parts used respectively to training and testing of machine learning models is a common practice in the field of machine learning; and

[11] https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained
[12] Readers who are interested in more resources with technical details are referred to UNECE Machine Learning for Official Statistics – Learning and Training wiki (https://statswiki.unece.org/display/ML/Learning+and+Training) or other resources online
[13] https://en.wikipedia.org/wiki/Algorithm
[14] https://developers.google.com/machine-learning/glossary#feature

- **Hyperparameter** is "*a parameter whose value is used to control the learning process*"[15]. Compared to the (model) parameter, the hyperparameter does not contribute to the prediction directly, and can be set manually at a specific value or "tuned" by searching through a pre-defined set of values. Chapter 2.1 provides few examples of hyperparameter.

[15] https://en.wikipedia.org/wiki/Hyperparameter_(machine_learning)

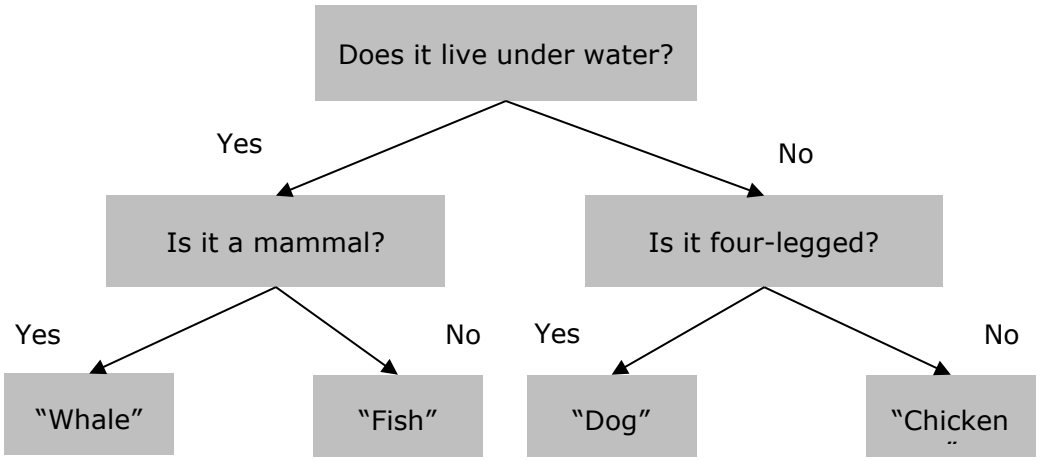
2.1. Machine Learning Algorithms

Decision Tree

The Decision Tree algorithm builds a sequence of hierarchical decision rules. For example, assume that a set of animals needs to be classified into one of 4 categories (dog, chicken, whale, fish) based on their features. Decision Tree starts from the root decision point (e.g., "does it live under water?") then splits into branches (e.g., if answer to the root is "yes", then "is it mammal?"; if "no", then "is it four-legged?") repeatedly until the end nodes (called "leaves") where the prediction is made (e.g., "whale" if it lives under water and mammal; "fish" if it lives under water but not a mammal) as in Figure 2.1. As the prediction is made based on a set of rules, it is often straightforward to understand how the machine learning model reaches a certain prediction for each example encountered.

Decision Trees can either be used to classify data items into categories (as illustrated by the above example), or they can be used to make numerical predictions when the target variable is numeric. For example, if the model was used to estimate the weight of the animal based on the features described, the output would be a numerical value rather than an assigned category in each case. This application is described as a regression task. Decision Trees can only generate a discrete number of values, and so can only approximate a continuous target variable.

Figure 2.1. Decision Tree Example



Random Forest

The Random Forest algorithm, as its name implies, creates a large number of separate Decision Trees that operate as an ensemble. In categorising items of data, each individual tree in Random Forest produces a prediction score and the category with the most votes becomes the prediction of Random Forest.

The Random Forest algorithm is initialised with a set of hyperparameters. One of them is the number of trees that it builds for the prediction task. For example, when set as 1000, the algorithm will build 1000 trees out of a random selection of data set features and data records.

When the model is used to predict into which category a record falls, a score is given for each category based on the results from all trees, e.g., [0.15 (dog), 0.30 (chicken), 0.45 (whale), 0.10 (fish)]. In this case, the highest score is 0.45, meaning that 450 out of the 1000 trees voted for the category "whale". Random Forest model, therefore, predicts that "whale" is the most likely category for this record.

Random Forest is a flexible and easy to use, and it can produce, even without hyper-parameter tuning, a great result. It requires relatively little computation power and is also one of the most commonly used algorithms because of its simplicity and diversity. Similar to Decision Trees, Random Forest can be used for either classification tasks (for categorical target variable) or regression tasks (to approximate a continuous target variable).

Logistic Regression

In contrast to Linear Regression, which uses features to predict a continuous target value (e.g., age, price), the Logistic Regression algorithm uses the features to predict a binary categorical target variable. It is used to predict which binary values (e.g., pass vs fail) a data record falls into, by fitting a function of the odds of that outcome to a linear combination of the features (as shown in the equation below). The function that relates the odds to the features is referred to as a link function, which often to be a logarithmic or logit function. For this reason, the algorithm is also called Logit Regression.

$$\log Odds = \log \frac{P\ (Y = "pass"\ |x_1, \dots, x_p)}{1 - P\ (Y = "pass"|x_1, \dots, x_p)} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

where Y is the binary target variable, $x_1, \dots, x_p$ are the features and $\beta_0, .., \beta_p$ are regression coefficients.

K-Nearest Neighbour (k-NN)

The k-Nearest Neighbour (k-NN) algorithm categorises records based on a set of features, which it treats as coordinates in a multidimensional space. If there are already records for which the categories are assigned, then for a new (without assigned categories) record the k-NN algorithm will predict its category by examining the categories assigned to its "Nearest Neighbours" within the multidimensional space defined by its feature variables. For such an additional data record, the algorithm locates a pre-specified number k of records in the training data set that are the "nearest" based on a measure of distance (e.g., Euclidean distance). The categories that these neighbours belong to are counted and the record is assigned the category representing the majority of the categories of these k nearest neighbours.

The scale of features (or their units of measurement) can give a greater importance to certain feature variables than others, unless standardisation of the feature variables is performed in advance. Alternatively to the standardisation, a weighting can be incorporated into the distance measure to adjust the importance of certain features. Similarly, where a large number of features exist, it may be desirable to reduce the number of features or dimensions prior to using the k-NN algorithm. Non-Euclidian distance measures may also be desirable when there are a large number of features.

K-NN is known as a "lazy" machine learning algorithm as it does not produce a model and its understanding about the relationships of the features to the target variable is limited. It simply stores the training data and then calculates the distance measure between a record to the training data and then picks the most common categories in its k-nearest neighbours to predict the category of the record.

Least Absolute Shrinkage and Selection Operator (LASSO) Regression

In a traditional least squares regression, a linear equation relates a target variable to a set of features, each having its own coefficient. These coefficients are fitted by minimising the sum of the squared residuals, where each residual represents the difference between the values of the target variable and their predicted values. The sum of the squared residuals takes the following form:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_j\right)^2$$

where n is the number of records in the dataset, p is the number of features, y is the target variable, and $\beta_0,..,\beta_p$ are regression coefficients.

The Least Absolute Shrinkage and Selection Operator (LASSO) Regression is a modification of traditional least squares regression that employs regularisation to prevent overfitting. It constrains the minimisation of the regular regression mean-squared loss function by adding a penalty term (the sum of absolute values of coefficients) as follows:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

where λ is a hyperparameter that controls the strength of the penalty[16]. Note that although the hyperparameter affects the estimation of parameters β, once parameter values are obtained, it does not directly affect the predicted value which is determined by the coefficients only (i.e., $\hat{y} = \hat{\beta}_0 + \sum_{j=1}^{p}\hat{\beta}_j x_j$).

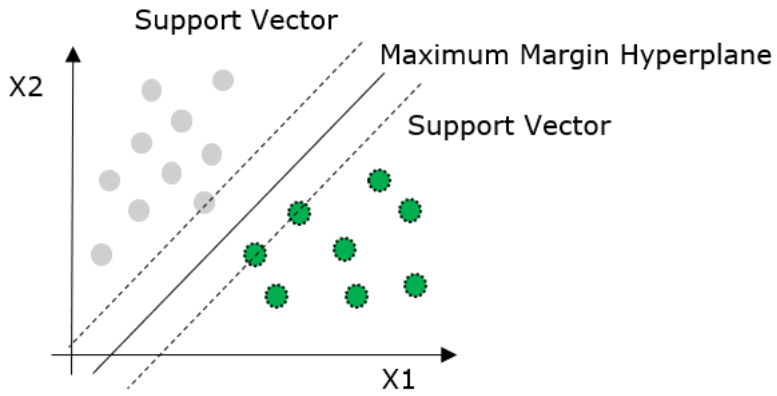
The aim of constraining the minimisation in this way is to penalise higher absolute values of the fitted coefficients, since these would imply stronger relationships between a given feature and the dependent target variable. Although constraining the way in which these coefficients are fitted introduces bias to predicted values, the idea is that when linear models are fitted to a relatively small training data set, reducing the values of fitted coefficients can reduce the occurrence of spurious correlations being incorporated into the model. Depending on the value selected for the hyperparameter λ, fitted values for some of the coefficients can be zero, meaning that such features are eliminated from the model.

Support Vector Machine (SVM)

With the Support Vector Machine (SVM) algorithm, significant results can be achieved with relatively little computation power. In this algorithm, each data record is a point in a p-dimensional feature space, where p is the number of features. When p is 2 (i.e., 2-dimensional feature space), the Support Vector Machine algorithm finds a line that is the maximum distance away from the nearest point of each category to this line. When p is larger than 2, a hyperplane (instead of a line) is found that separates best the data points belonging to each category, this is called the Maximum Margin Hyperplane (MMH). The points from each category that are the closest to the MMH are called Support Vectors (SV) (See Figure 2.2). Each category must have at least one SV but may have more than one. These SVs define MMH and thus provide a very compact way to store a model.

[16] Note that when λ is equal to 0, LASSO is equivalent to the ordinary linear regression

Figure 2.2. Support Vector Machine (p=2)



For a new data record, that data record is drawn and the area it falls into is the predicted category. In this way, the Support Vector Machine combines aspects of both the instance based Nearest Neighbour and regression methods. As this combination is very powerful, Support Vector Machines can model highly complex data relationships.

Naive Bayes

Naive Bayes is a classification algorithm based on Bayes' Theorem with an assumption of independence among features. In simple terms, the Naive Bayes algorithm assumes that, for a given value of the target variable, the presence of a particular feature is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features may depend on each other or upon the existence of the other features, in Naive Bayes, all of these properties are assumed to independently contribute to the probability that this fruit is an apple and that is why it is called as "Naive". So if $P(Apple|red, round, 3\ inches)$ denotes the probability of a fruit being an apple given that it is red, round and 3 inches, by Bayes Theorem, it becomes

$$\frac{P(Apple)P(red, round, 3\ inches|Apple)}{P(red, round, 3\ inches)}$$

which, by the independence assumption, becomes

$$\frac{P(Apple)P(red|Apple)P(\ round|Apple)P(\ 3\ inches|Apple)}{P(red, round, 3\ inches)}$$

The above quantity is computed using observed frequencies from the training data set, and compared to the other possible classifications, such as $P(Pear|red, round, 3\ inches)$ and $P(Plum|red, round, 3\ inches)$, etc., and is classified according to the highest value among these probabilities.

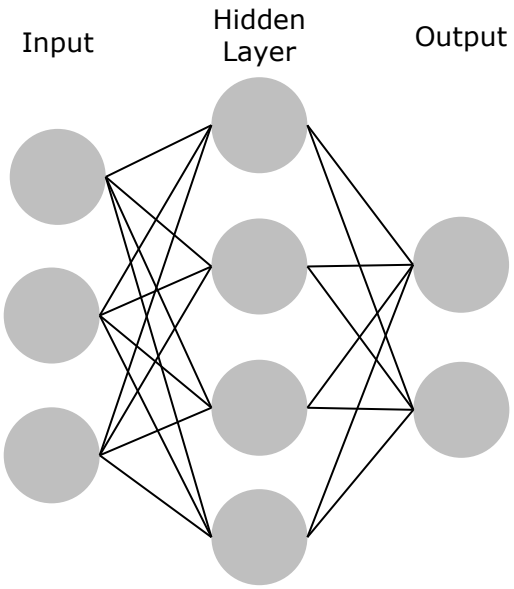
The Naive Bayes models are easy to build and particularly useful for very large data sets. Along with simplicity, it can outperform even highly sophisticated algorithms. Naive Bayes is mostly used for classification, and is a computationally cheap algorithm.

As long as the conditional independence requirement is fulfilled, the training set may be modest in size, although if training data does not contain all possible classes, then missing categories will be assigned zero probability.

Artificial Neural Network (ANN)

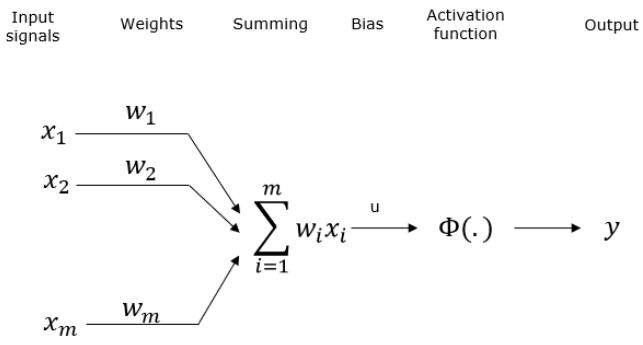
"*An Artificial Neural Network (ANN) is a collection of connected nodes called "artificial neurons" which loosely model the neurons in a biological brain. Typically, neurons are aggregated into layers. Different layers may perform different transformations on their inputs. Signals travel from the first layer (the input layer), to the last layer (the output layer), possibly after traversing the layers multiple times*"[17] (see Figure 2.3).

Figure 2.3. Artificial Neural Network Diagram



Artificial neurons (nodes) are elementary units in an artificial neural network, that process a set of input signals to generate a single output as in Figure 2.4.

Figure 2.4. Artificial Neuron



Each node receives a set of *m* inputs x_i, which are processed by weighting them with weights w_i and adding a bias term as follows:

[17] https://en.wikipedia.org/wiki/Artificial_neural_network

$$s = \left(\sum_{i=1}^{m} w_i x_i\right) + bias$$

An activation function Φ is then applied to the quantity s, such that the output from the neuron y is:

$$y = \Phi(s)$$

The result of this is then passed on to other nodes in the network. Various types of activation functions can be used depending on the application, producing outputs that are discrete or continuous, and bounded or unbounded.

Several powerful state-the-of-art machine learning algorithms are based on the concept of an ANN, some of which are described below. ANNs are also powerful in learning complex patterns in the data set, but a large ANN requires considerable computation power, hence investment in hardware (e.g., CPU, GPU) might be needed.

Multi-Layer Perceptron (MLP)

A Multi-Layer Perceptron (MLP) is a type of feedforward ANN which consists of one or more hidden layers. Signals travel from the input, through the hidden layers, to the output (hence signals "feedforward"). Each node (apart from the inputs) is a neuron and each input node represents a single feature of the data set. The value of this feature is then passed forward to the first node, which combines and processes the inputs it receives and transforms the result via an activation function. The result of this is then passed on to the next layer.

MLPs have to be trained on a training set where the outputs are known, in order to adjust the weights used by the artificial neurons. This is done iteratively, using each member of the training data set in turn, via a method called "backpropogation".

Backpropogation minimises the value of a loss function (such as the average squared difference between predicted and actual output), with respect to the values of the weights. In order to do this computationally efficient, it typically uses a gradient decent method, to iteratively adjusts the weights based on the results from using each successive data point from the training data set. The speed of decent (the size of steps) is controlled by a set learning rate (which is hyperparameter of MLP).

The more complex the MLP, the more complex relationships in the data can be recognised. The processing expense grows rapidly with the number of layers and neurons in each layer.

Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN)

Convolutional and Recurrent Neural Networks are types of ANN designed to efficiently learn specific relationships in the data. Convolutional Neural Networks (CNN) were originally designed for image processing and focus on efficiently learning spatial patterns in images.

In Recurrent Neural Networks (RNN), signals are allowed to travel backwards using loops. This ability mimics more closely how a biological neural network operates. This allows very complex patterns to be learned. The addition of a short-term memory or delay increases the power of RNNs, including the ability to learn sequential patterns.

Both approaches have been found to be useful for a variety of language processing tasks and are therefore powerful text classification tools.

FastText

FastText was created by Facebook Artificial Intelligence (AI) lab, created for a text classification task. It is a library for efficient learning of word representations and sentence classification. The model allows the creation of an unsupervised learning or supervised learning algorithm for obtaining vector representations for words. Facebook makes available pre-trained models for 157 languages[18].

eXtreme Gradient Boosting (XGBoost)

Boosting is an ensemble technique where new models are added to correct the errors made by existing models. Models are added sequentially until no further improvements can be made.

Gradient Boosting is an approach where new models are created that predict the residuals or errors of prior models, and are then added together to make the final prediction. It is called gradient boosting because it uses a gradient descent optimisation algorithm to minimise the loss when adding new models.

eXtreme Gradient Boosting (XGBoost)[19] is a library that implements the gradient boosting decision tree algorithm, using a sequence of fairly small decision trees. It is one of the fastest implementations for gradient boosting, and can be applied to both regression and classification tasks.

[18] As of June 2021
[19] https://xgboost.readthedocs.io/en/latest/

2.2. Prediction Accuracy

Prediction accuracy evaluates the performance of a machine learning model. Using the same data set to calculate the accuracy metrics could potentially lead to overestimation of its accuracy as the model was developed based on the training data set, hence different data sets are used for developing the model (training data set) and estimating the accuracy of the model (testing data set).

One approach that is often used to evaluate the machine learning models is k-fold cross-validation, where the data set is partitioned into *k* subsets, the model is trained on a training set comprising all but one of them, and the model is tested using the remaining subset. This model fitting is performed a total of *k* times, each time leaving out a different one of the *k* subsets, and the results of the *k* models are combined to evaluate accuracy. This approach allows all of the data to be used for training models, with each data point being used once also for evaluation/testing purposes. Other approaches to assessing the accuracy of models include bootstrap methods which sample data points from the data set to make inferences about model accuracy.

The prediction accuracy is often used as one of the quality dimensions with which a machine learning model is compared with other machine learning models or existing methods, such as manual classification, rule-based edit, traditional statistical methods. For more about quality dimensions, see Chapter 4.

2.2.1. Continuous Target Variable

When the target is a continuous variable (e.g., income, age), Mean Squared Error (MSE) and Mean Absolute Error (MAE) are commonly used as accuracy metrics.

For i-th record in the data set, let y_i be the value of target variable and $f(\vec{x}_i)$ be the predicted value based on the feature vector $\vec{x}_i$.

- **Mean Squared Error (MSE)** is the average of the square of the difference between the original value and the predicted value

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(\vec{x}_i))^2$$

- **Mean Absolute Error (MAE)** is the average of the absolute of the difference between the original value and the predicted value

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - f(\vec{x}_i)|$$

2.2.2. Categorical Target Variable

Binary Classification

When the target variable belongs to either one of two categories (e.g., positive, negative), machine learning model predictions fall one of four cases as below:

- True Positives (TP): the model predicted the positive category correctly
- False Positive (FP): the model incorrectly predicted the negative category as a positive category
- False Negative (FN): the model incorrectly predicted the positive category as a negative category

- True Negative (TN): the model predicted the negative category correctly

For example, in a classification task where a bank wants to predict fraudulent and non-fraudulent transactions, the positive category would be fraud, as these are the ones the bank wants to predict and find. Cases belonging to the True Positives are the ones that are truly fraud, and where the model correctly predicted fraud. False Positives would be non-fraud cases, which the model wrongly predicted as fraud. These possibilities are laid out in the following table, which if populated with frequencies is referred to as a "confusion matrix".

		Predicted category	
		Positive	Negative
Actual category	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

For the binary classification, the following accuracy metrics are commonly used:

- **Accuracy** is the proportion of correct predictions among all predictions made. This indicates the ability of the model to make correct predictions for positive and negative categories:

$$\frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions\ made} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **Precision** is the proportion of correct positive predictions among all cases that were predicted as positive. This indicates the ability of the model to predict True Positives and to avoid False Positives:

$$\frac{True\ Positives}{True\ Positives + False\ Positives} = \frac{TP}{TP + FP}$$

- **Recall** (also known as Sensitivity) is the proportion of correct positive predictions among all positive cases. This indicates the ability to predict the records we want to find:

$$\frac{True\ Positives}{True\ Positives + False\ Negatives} = \frac{TP}{TP + FN}$$

- **F1 score** is the harmonic mean between Precision and Recall. The range for the F1 score is [0, 1]. The F1-score shows how precise the model is (how many cases were predicted correctly) as well as how robust it is (it does not miss a significant number of records). Note that the F1-score will be lower than the simple arithmetic means if one of the two metrics is much lower than the other. And this is the reason why the F1-score can be more useful than Accuracy or Precision alone:

$$2 * \frac{Precision\ *\ Recall}{Precision\ +\ Recall}$$

For a binary classification, one typically focuses on accuracy metrics for the positive category which is the target we want to find.

Obtaining good Precision *or* Recall is usually easy, but getting good Precision *and* Recall is often difficult. In general, models with high Precision and Recall scores are preferred,

but there is a trade-off between Precision and Recall - improving the Precision score often results in lowering the Recall score and vice versa.

If the number of cases per category is not evenly distributed and un-balanced (e.g., if the number of cases labelled as fraud is much smaller than the cases labelled as no-fraud), Accuracy, Precision and Recall can be misleading if seen in isolation. Assume that the non-fraud category is the majority category where, say, 90% of all cases are non-fraud and the model predicts all as being non-fraud which means the model predicts correctly in 90% of all records. This might look like a good result by simply considering Accuracy, as it is a measure for correct prediction. Recall for the positive category (i.e., fraud) would be 0% but Recall for the non-fraud category would be 100%. Precision would be 0% for our positive fraud category, but 90% for the non-fraud class. For a bank, it would mean that no fraudulent transactions are detected, and this model would be rather useless. A better approach is to combine these metrics into composite metrics such as the F1-score and Macro F1-score (see below).

Multi-Class Classification

When the target variable belongs to one of more than two categories (e.g., classifying a building into one of four categories: commercial, residential, under construction, others), the overall prediction accuracy of the model can be assessed using the same classification metrics as for the binary case above for each class, and then combining them in either Macro or Micro metrics. This provides a balanced view of the ability of the model to predict all categories.

Assume that the target variable has 3 categories: C_1, C_2 and C_3 and $N_{i,j}$ denotes the number of cases that belong to category C_i and predicted as category C_j, as laid out in the confusion matrix below.

		Predicted category		
		C_1	C_2	C_3
Actual category	C_1	$N_{1,1}$	$N_{1,2}$	$N_{1,3}$
	C_2	$N_{2,1}$	$N_{2,2}$	$N_{2,3}$
	C_3	$N_{3,1}$	$N_{3,2}$	$N_{3,3}$

For each category C_i, a single category accuracy metrics can be obtained by treating the category as positive and other categories as negative. For example, for category C_1:

$$Accuracy_1 = \frac{TP_1 + TN_1}{TP_1 + TN_1 + FP_1 + FN_1}$$
$$= \frac{N_{1,1} + (N_{2,2} + N_{3,3})}{N_{1,1} + (N_{2,2} + N_{2,3} + N_{3,2} + N_{3,3}) + (N_{2,1} + N_{3,1}) + (N_{1,2} + N_{1,3})}$$

(In fact, the accuracy measure is identical for all categories)

$$Precision_1 = \frac{TP_1}{TP_1 + FP_1} = \frac{N_{1,1}}{N_{1,1} + (N_{2,1} + N_{3,1})}$$

$$Recall_1 = \frac{TP_1}{TP_1 + FN_1} = \frac{N_{1,1}}{N_{1,1} + (N_{1,2} + N_{1,3})}$$

$$F1\ Score_1 = 2 * \frac{Precision_1 * Recall_1}{Precision_1 + Recall_1}$$

Macro Averages

The simplest way to combine individual metrics into the Macro metric is to calculate the arithmetic mean for each of them. Macro averages do this in a way that gives equal weight to each category (rather than giving equal weight to each observation).

- **Macro Accuracy** is the mean of all individual Accuracy
$$\frac{1}{C}\sum_{i=1}^{C} Accuracy_i$$
- **Macro Precision** is the mean of all individual Precision
$$\frac{1}{C}\sum_{i=1}^{C} Precision_i$$
- **Macro Recall** is the mean of all individual Recall
$$\frac{1}{C}\sum_{i=1}^{C} Recall_i$$
- **Macro F1 Score** is the mean of all individual Accuracy
$$\frac{1}{C}\sum_{i=1}^{C} F1\ Score_i$$

where C is the number of categories. Note that by taking the simple arithmetic average across categories, the accuracy of each category contributes equally to the macro metric, which may not be desirable when the cases are highly unbalanced. The averages can be weighted by the numbers of records for each category which are then called Weighted-Average or simply Weighted-Accuracy and so on.

Micro Averages

Micro averages dispense with the need to calculate individual accuracy metrics for each category before averaging across categories, by calculating the average in a single step.

- **Micro Accuracy** is the proportion of correct predictions among all predictions made:
$$\frac{N_{1,1} + N_{2,2} + N_{3,3}}{N_{1,1} + N_{2,2} + N_{2,3} + N_{3,2} + N_{3,3} + N_{2,1} + N_{3,1} + N_{1,2} + N_{1,3}}$$
- **Micro Precision** is the proportion of correct predictions among all cases that were predicted as the category (e.g., for C_1, cases that are predicted as 1 are $N_{1,1}, N_{2,1}, N_{3,1}$), which becomes identical to the Accuracy:
$$\frac{N_{1,1} + N_{2,2} + N_{3,3}}{(N_{1,1} + N_{2,1} + N_{3,1}) + (N_{1,2} + N_{2,2} + N_{3,2}) + (N_{1,3} + N_{2,3} + N_{3,3})}$$
- **Micro Recall** is the proportion of correct predictions among all cases that actually belong the classes (e.g., for C_1, cases that actually belong to the category are $N_{1,1}, N_{1,2}, N_{1,3}$), which becomes identical to the Accuracy:
$$\frac{N_{1,1} + N_{2,2} + N_{3,3}}{(N_{1,1} + N_{1,2} + N_{1,3}) + (N_{2,1} + N_{2,2} + N_{2,3}) + (N_{3,1} + N_{3,2} + N_{3,3})}$$
- **Micro F1 Score** is the harmonic mean between Micro Precision and Micro Recall

$$2 * \frac{Micro\ Precision\ *\ Micro\ Recall}{Micro\ Precision\ +\ Micro\ Recall}$$

Because the Micro-Precision and Micro-Recall terms are identical, the Micro-F1-Score simplifies to also become identical to the Micro Precision (or Micro Precision).

3. Machine Learning Application Areas

Machine learning holds a great potential to contribute the work of statistical organisations in various ways. It can automate the process that was used to be largely done by humans, assist humans do the work more efficiently, and allow the organisations to make use of new data sources, which can ultimately increase the relevance and timeliness of statistics produced. As mentioned in the Chapter 1, however, machine learning should not be used for the sake of machine learning, it should be aligned with business needs in the organisations.

This Chapter introduces three application areas (namely, classification and coding of textual data, editing and imputation, and imagery analysis) that were investigated by the HLG-MOS Machine Learning Project for their potential added value and lessons learned from the experiences.

3.1. Classification and Coding of Textual Data

3.1.1. Introduction

"*Classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known*. *Examples are assigning a given email to the "spam" or "non-spam" and assigning a diagnosis to a given patient based on observed characteristics of the patient (sex, blood pressure, presence or absence of certain symptoms, etc.)*"[20].

In the context of the coding and classification work in the statistical organisations, the given set of data referred to in the above quote is typically a text or narrative provided by the respondent from a survey or administrative data source[21]. For example, it could describe an individual's occupation or the economic activity of the company described in an administrative business register. With the increasing use of new data sources, the text data that statistical organisation might work with could also include product descriptions scraped from the internet or text posts obtained from social media platforms such as Twitter.

The aim of classification and coding in this scenario is to classify the descriptions into international or corporate statistical classification system, such as Standard Industrial Classification (SIC), Standard Occupational Classification (SOC) (see Table 3.1 for examples). Beyond the textual variable, other variables that exist within a data source that relate to other attributes of the data item (e.g., age, net pay) could also be used to perform the classification.

[20] https://en.wikipedia.org/wiki/Statistical_classification

[21] Although classification can be applied to various types of data, this Chapter concerns its application to textual data

Table 3.1. Examples of Statistical Classification System

Classification System	Description
NAICS	North American Industry Classification
SCIAN	Sistema de Clasificación Industrial de América du Nord – Spanish version of NAICS
NOC	National Occupational Classification – Canada's national system for describing occupations
SINCO	National Classification System for Occupations (Sistema Nacional de Clasificación de Ocupaciones)
NACE	European Classification of Economic Activities (Nomenclature statistique des Activités économiques dans la Communauté Européenne)
SIC	Standard Industrial Classification – Established by the USA in 1937, replaced by NAICS in 1997
SOC	Standard Occupational Classification
OIICS	Occupational Injury and Illness Classification System – Developed by the Bureau of Labour Statistics of the United States of America
ECOICOP	European Classification of Individual Consumption by Purpose
CTS	Catalogue of Time Series by the International Monetary Fund (IMF)

Table 3.2 shows example of entries in the SOC 2010 used by the United Kingdom[22], for which the classification and coding task is, in this case, to assign a category (code) from the table to the textual narrative given by the respondent (e.g., "I drive a fork-lift truck" or "I use fork-lift trucks in my job to load up lorries").

Table 3.2. Examples of Entries from the United Kingdom SOC 2010

Code	Description
8221	Crane drivers
8222	Fork-lift truck drivers
8223	Agricultural machinery drivers
8229	Mobile machine drivers and operatives n.e.c.

[22] https://www.ons.gov.uk/methodology/classificationsandstandards/standardoccupationalclassificationsoc/soc2010/soc2010volume2thestructureandcodingindex#electronic-version-of-the-index

3.1.2. Data Preparation

In the text classification and coding, raw data consist of textual responses to open-ended questions that have been written by individual respondents. Texts written in such natural language can contain mistakes and errors, as well as words that serve grammatical and syntactical purposes which may not necessarily be useful for the classification task. Therefore, **text pre-processing** may be performed to some extent in order to remove these elements that do not add information and/or produce noise in the data, hence reducing the ability of the machine learning algorithm to recognise words of the same meaning. Some commonly used text pre-processing methods include:

- **Removal of noise** such as punctuation marks (e.g., commas, period points, exclamation marks) and special characters. However, note that what is considered as noise depends on the domain and task (e.g., a hash symbol (#) could be noise for data collected from survey questionnaire, but could be important for twitter data analysis as they are used to designate "hashtags");
- **Normalisation to lower case** as most programming languages are case-sensitive (e.g., the word "bus" is differentiated from "Bus" even though they are essentially the same word);
- **Removal of stop words** (words that commonly appear in texts but do not add much meaning to the texts to be analysed, and hence considered unimportant) such as "the", "a", "an" and "in". By removing these words, the algorithm can focus on more important words;
- **Stemming and Lemmatisation** are two approaches to handle the inflections or syntactic differences between word forms. Both stemming and lemmatisation can be achieved with commercial or open-source tools and libraries:
  - **Stemming** is a process where words are reduced to their stem or root by chopping off the end of the word to reduce it to its stem (e.g., the word "flying" has the suffix "ing" and stem "fly")[23]. The aim is to reduce the inflectional forms of each word into a common base. This increases the frequency of the word's occurrence and gives the algorithm more identical instances of that word to learn from; and
  - **Lemmatisation** also tries to remove inflections, but it does not simply chop off these inflections. It uses lookup tables (e.g., WordNet[24]) that contain all inflected forms of the word to find the base or dictionary form of the word, which is known as the lemma (e.g., "geese" is lemmatised by Wordnet to "goose"). If the word is not included in the table, it is passed as the lemma.
- **Tokenisation** is a process of splitting text into smaller pieces (called "tokens") such as words or character sequences. When n consecutive words are used, the set of tokens created are called "n-grams"[25]. For example, for the text "the fox jumps over the fence":
  - Word 1-gram: the, fox, jumps, over, the, fence;
  - Word 2-gram: the fox, fox jumps, jumps over, over the, the fence; and
  - Character 3-gram: 'the', 'he_', e_f', '_fo', 'fox', 'ox_', 'x_j', '_ju' …...[26]

[23] The Porter Stemmer algorithm is very popular for the English language, which chops both "apple" and "apples" down to "appl". This shows that stemming might produce something that is not a real word. Nevertheless, doing this to all the narratives to be classified and all target documents helps the algorithm can find matches

[24] A lexical database for English (https://wordnet.princeton.edu/)

[25] The Mexican pilot study on occupation and economic activity classification in Chapter 3.1.3 describes in detail their work on n-grams, and what impact changes in the n-grams have on their prediction results

[26] for the space character the underline '_' is used

With tokenisation, a set of unique words used in the data (called **"Bag of Words"**) is created which then can be used to classify individual text in the data.

Table 3.3 shows how four text descriptions of occupation change as they go through pre-processing described above.

Table 3.3. Text Pre-Processing Applied to Text Examples

Original text	"I drive a bus."	"Driving Bus"	"restaurant chef"	"I cook at a restaurant"
After tokenisation	"I", "drive", "a", "bus", "."	"Driving", "Bus"	"restaurant", "chef"	"I", "cook", "at", "a", "restaurant"
After removing noise	"I", "drive", "a", "bus"	"Driving", "Bus"	"restaurant", "chef"	"I", "cook", "at", "a", "restaurant"
After normalising to lower case	"i", "drive", "a", "bus"	"driving", "bus"	"restaurant", "chef"	"i", "cook", "at", "a", "restaurant"
After stemming	"i", "driv", "a", " bus"	"driv", "bus"	"restaurant", "chef"	"i", "cook", "at", "a", "restaurant"
After removing stop words	"driv", "bus"	"driv", "bus"	"restaurant", "chef"	"cook", "restaurant"

Text, in its raw form, is simply a collection of strings for machines. Humans understand that both "chef" and "cook" refer to an occupation and are synonymous, but for machines, the two words are simply 4-character strings with common character "c" in them, which is not helpful information for classifying texts such as "I am a chef" and "I am a cook". **Vectorisation** is the process of converting a text description into a real-valued list with fixed length (i.e., vector) that machine learning algorithms can process and extract meaningful information[27].

Term Frequency-Inverse Document Frequency (TF-IDF) is a commonly used vectorisation method, and is defined as:

TF-IDF = Term frequency x Inverse document frequency

where the term frequency of a word in a given text is defined as the number of times the word appears in the text divided by the total number of words in that text. The inverse document frequency term adjusts for how rare or common that particular word is and represents the reciprocal of how often the word appears across all texts in the data set, and can be regarded as a weighting.

[27] This process is also called feature engineering because the results from the vectorization (e.g., set of words) are often used as input features for machine learning algorithms

Natural language processing techniques are advancing rapidly in recent years. Word embedding techniques such as Wor2Vec[28] and GloVe[29] are gaining an increasing prominence, as they allow encoding of the semantic meaning of the words in the vector space so that, for example, vectors representing the word "chef" and "cook" are close to each other.

3.1.3. Pilot Studies

This Chapter provides a brief summary of the pilot studies conducted by members of the HLG-MOS Machine Learning Project Work Package 1 - Coding and Classification theme as below:

- National Institute of Statistics and Geography (INEGI), Mexico – Occupation and Economic Activity Coding Using Natural Language Processing;
- Statistics Canada – Industry and Occupation Coding;
- Statistics Flanders, Belgium – Sentiment Analysis of Twitter Data;
- Statistical Office of the Republic of Serbia – Coding Economic Activity;
- Statistics Norway – Standard Industrial Code Classification by Using Machine Learning;
- Bureau of Labour Statistics (BLS), the United States of America – Coding Workplace Injury and Illness;
- Statistics Poland – Production description to ECOICOP; and
- International Monetary Fund (IMF) – Automated Coding of IMF's Catalogue of Time Series.

Complete reports of all pilot studies are available on the UNECE Machine Learning for Official Statistics wiki page (https://statswiki.unece.org/display/ML/WP1+-+Pilot+Studies)[30]. One of the pilot studies from Statistics Poland is highlighted in Box 3.1.

An overview of the eight statistical organisations participating in this Project, their existing methods of classification (legacy systems) and data used for the pilot studies is provided in Table 3.4. At the time of the Project (2019-20), three of these pilot studies from Statistics Canada, Statistics Norway and the Bureau of Labour Statistics either went into production or had already done so by that time.

[28] https://en.wikipedia.org/wiki/Word2vec
[29] https://nlp.stanford.edu/projects/glove/
[30] Codes and data from some of the pilot studies are available on the UNECE wiki (https://statswiki.unece.org/display/ML/Studies+and+Codes)

Table 3.4. List of Pilot Studies, Legacy System and Data Used

Organisation	Classification	Legacy System	Data
INEGI	SCIAN, SINCO	Deterministic coding system assisted with manual coding with accuracy > 95%	Household Income and Expenditure - 74K households, 158K persons
Statistics Canada	NAICS, NOC	G-Code word matching with accuracy level > 95% (which is higher than human coders)	Canadian Community Household Survey (CCHS) – 89K records Labour Force Survey – 440K records Canadian Health Measures Survey (CHMS) – 114K NOC Index Entries, 38K NAICS Index Entries
Statistics Flanders	N/A	Life Statistics via surveys	Twitter data (tweets that contain either a positive or a negative emoticon as a label to avoid manual production of training data)
Statistical Office of the Republic of Serbia	NACE (at 2 & 3 digits)	Manual coding	Labour Force Survey – 20K cases
Statistics Norway	SIC (821 labels)	SIC classification of new companies for the Central Coordination Register are made manually from the description provided	Description of economic activities and "official" descriptions of codes and keywords – 1.5 million historical records
BLS	SOC, OIICS	Manual coding	Survey of Occupational Injuries and Illnesses – initially 261K records, later grew to > 2 million
Statistics Poland	ECOICOP (1st group of products: Food and non-alcoholic beverages 61 codes, all 5 digits)	N/A	Web scraped product names – 17K cases manually coded to ECOICOP
IMF	CTS with 28,886 codes	Manual coding	Time series data sets from member countries

Technical details (e.g., text pre-processing methods, machine learning algorithms[31]) and results are provided in Table 3.5 and Table 3.6 respectively.

[31] See Chapter 2.1 for description of ML algorithms

Table 3.5. Data Preparation Methods, Machine Learning Algorithms, Software and Hardware Used in the Pilot Studies

Organisation	Data preparation	ML algorithms	Software and hardware
INEGI	Article suppression, stemming, lemmatisation, uppercase, synonyms, TF-IDF	Assembly of algorithms: SVM, Logistic Regression, Random Forest, Neural Networks, XGBoost, K-NN, Naive Bayes, Decision Trees	• Python, scikit-learn, keras • 20 cores, 256 GB RAM, 4 TB drives
Statistics Canada	Removal of stop words, lowercasing character conversion, merging of variables, Caesar Cipher, addition of LFS 440K records to CCHS's training datasets (89K records)	Mandated to use FastText or XGBoost as they are already in G-Code[32]	• G-Code • 3 GHz Intel i5-3570, 16 GB RAM
Statistics Flanders	Lower casing, stemming, removing stop words, lemmatisation, removing special characters, n-gramming; count vectorization, TF-IDF vectorisation, autoencoder neural network embedding, retrained Neural Network	Penalised Logistic Regression, Random Forest, Gradient Boosting Trees, MLP	• Python • 4 GHz Intel i7 6700K, 16 GB RAM
Statistical Office of the Republic of Serbia	N/A	Random Forest, SVM, Logistic Regression	• Python, scikit-learn, Pandas, Pyzo IDE
Statistics Norway	Removal of obvious unreliable activities/code, removal of digits and punctuation, removal of stop words, lowercasing	FastText, Logic Regression, Random Forest, Naive Bayes, SVM, CNN	• Python • Google Cloud

[32] Automated and Interactive Coding Generalized System used in Statistics Canada

BLS	Very little data cleaning or normalisation[33]	Logistic Regression, SVM, Naive Bayes, Random Forest, MLP, CNN, RNN	• Python, scikit-learn • Initially 2-4 cores 8-16 GB RAM, later 4 Titan X Pascal GPUs each with 12 GB and 3584 cores
Statistics Poland	Vectorisation, normalisation	Naive Bayes, Logistic Regression, Random Forest, SVM, Neural Networks (Ludwig Library)	• Python, scikit-learn • Office PCs
IMF	Standardising the different country file formats; TF-IDF, Word2Vec	Logistic Regression, K-NN	• Python • 2.4 GHz Intel Core i5-6300U

The above Table 3.5 shows that the processing power used for machine learning training and prediction is typically that of a desktop or laptop computer. The exceptions are the two already operationalised solutions in the production from Statistics Norway and the Bureau of Labour Statistics that used Neural Networks. Statistics Norway used Google Cloud and the Bureau of Labour Statistics used 4 GPUs with 3584 cores each.

[33] The Bureau of Labour Statistics stopped using stop-word removal or stemming after they found out in early experiments that these techniques proved to be unhelpful due to the nature of the text narratives they need to classify. But they used CountVectorizer as a tool to create a "bag of features" representing the input. This shows which words (or sequences of words, or sequences of characters) occur in the input, but not the order in which those words or sequences appear. When they moved over to Neural Networks, they stopped doing this also. Preserving the original ordering of the sequence of letters allows the Neural Network to gain more insight into the text, while simpler algorithms are not capable of learning the intermediate representations, i.e., that letters form words, and words form phrases and sentences, and sentences form paragraphs and so on

Table 3.6. Results and Status

Organisation	Results	Status
INEGI	For Economic Activity / For Occupation Accuracy: 87.7 % / 83.1 % Precision: 66 % / 57.8 % Recall: 64.5 % / 57.3 %	Proof of concept
Statistics Canada	Accuracy rate > 95% when combined with clerical classification and up to 100% Recall and precision on quality control sample	In production for two surveys (CCHS and CHMS)
Statistics Flanders	• Precision: 80% • Recall: 81%	Proof of concept
Statistical Office of the Republic of Serbia	Accuracy • Random Forest: 69% • SVM: 75% • Logistic Regression: 69%	Proof of concept (investigation carries on to achieve > 90% accuracy)
Statistics Norway	FastText, SVM and CNN produce similar results; FastText faster in training; 22% of units predicted	In production as a supporting tool (5 best codes with probability is offered which allows for humans making a faster choice)
BLS	From comparison with the 'Gold Standard' data: ML is more accurate that manual; Neural Network coding is better in any of the 6 codes to be assigned than humans, with Accuracy between 69.8% and 91.9%.	In production (ML auto coding only above set threshold to maximise the overall macro-F1-score for the human/ML coding; > 85% of codes are assigned by a neural network)
Statistics Poland	Naive Bayes / SVM / Logistic Regression / Random Forest Accuracy: 90.5 % / 92 % / 91.6% / 92.2% Precision: 90 % / 92 % / 92 % / 93 % Recall: 90 % / 92 % / 92 % / 92 % F1-Score: 90 % / 92 % / 92 % / 92 % MCC: 90 % / 92 % / 91 % / 92 %	Proof of concept
IMF	Accuracy about 80%	Proof of concept

Box 3.1. Pilot Study from Statistics Poland

The objective of this pilot study was to test if it is possible to automate the manual ECOICOP product classification process using machine learning methods.

Given the lack of access to the actual data used to produce the Harmonised Index of Consumer Prices (HICP), the author of the pilot study collected their own data set through web scrapping, resulting in a data set of around 16,700 products names from online shops corresponding to about 60 ECOICOP categories. The data was reindexed randomly and divided into three groups, (i) training data set, (ii) validation data set, and (iii) test data set. In addition, vectorisation and normalisation were carried out by CountVectorizer or TfidVectorizer[34] depending on the algorithm.

Initially, five algorithms were considered, namely, Naive Bayes, Logistic Regression, Random Forest, Support Vector Machine, and Neural Networks, but Neural Network was not further experimented due to hardware and software restrictions.

All the four algorithms tried have accuracy above 90% (see Table 3.6). The hyperparameters were chosen by grid search or manually written selection rules. The prediction with very low confidence could be classified manually to improve the accuracy. The impact of the chosen random seed (during re-indexing) on accuracy and other results was also tested and the results remained robust.

This pilot study proved that the ECOICOP classification process can be supported by machine learning methods with high accuracy. The results were presented to the management of the office and the President of Statistics Poland and were well received.

Looking forward, more cooperation between all the departments in Statistics Poland is needed to share knowledge and experience in order to introduce modern methods more efficiently.

Technical details and python codes used for the pilot study are available at https://github.com/statisticspoland/ecoicop_classification. To help machine learning beginners to kick start with text classification, the authors of the pilot study has prepared tutorial on Google colab: https://colab.research.google.com/drive/1Epn2NeFRuFC_XyXtQ4qezGVBA5aAzqIh.

3.1.4. The Value Added from Machine Learning and Lessons Learned

The traditional and manual classification process is often lengthy, resource intensive and can be prone to errors. Even experienced human coders can assign different categories to the same text narrative, which leads to inconsistency issues in the results.

Machine learning can **automate the classification process, and this can be done to varying degrees with human supervision and collaboration**. For example, in the pilot studies from Statistics Canada and the Bureau of Labour Statistics, the machine learning solutions are only used when the prediction is made with a confidence level above a certain threshold. Predictions below this threshold are ignored and completed manually by human coders instead, thus only predictions with a high level of confidence are allowed to be auto-coded.

Machine learning can also assist humans. The machine learning solution for Statistics Norway acts as an advisory to the human coding process. The human coder is given the

[34] Python library scikit-learn modules for vectorisation

5 best machine learning predictions with their confidence levels and the option to either accept one of them or reject all. As the machine learning solution auto-codes big parts of these data sets and expediates the manual coding, classification processing can be done more rapidly. With a faster classification process, financial gain can be gained. For example, Statistics Norway reported that the gain is expected to be over a 10-year period between 7 and 17 million Norwegian Krone (equivalent to 0.65 to 1.6 million Euro).

The more mature and advanced a machine learning model is, the more confidence can be put into it to let the model take on a greater share of the predictions. As manual coding resources are freed up, possibly on an increasing scale over time, the cost of building, monitoring and maintaining the machine learning solution has to be accounted for as a possibly expensive IT infrastructure. Data consistency can also increase as the manual process is reduced.

Note that machine learning models may not fully replace manual classification and coding processes. Difficult and rare cases may still have to be coded by humans, especially ones that were not already encountered within the training data set. However, these manually classified cases can then be added as new training cases for the machine learning model on a continuous basis. This allows the machine learning solution to mature over time, which in turn should lead to a rise in the proportion of auto-coded cases. This will help to balance the presence of labelled cases for the minority or difficult classes, but care has to be exercised not to create a bias in the training data. Such an approach achieves ever-increasing resource savings.

3.1.5. Best Practices

Best practices can be subjective as they depend on the expectations of the organisation, and the context in which machine learning is to be used. The 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**. A comparison between machine learning predictions, manual process and other legacy systems, such as rule-based systems, can only be clearly and credibly established when each is compared to a golden data set in a statistically sound manner.

Sophisticated models such as Neural Networks (used by the Bureau of Labour Statistics) can provide better performance than the bag-of-words approaches that many others are using for text classification, but there are two big caveats:

- Firstly, as knowledge and techniques in the Neural Networks field advance rapidly, the best approach today may not be the best approach tomorrow; and
- Secondly, Neural Networks can be difficult to use effectively. Simply plugging in a generic implementation might not produce good results, the structure of the network needs to be adapted to the task (e.g., using tools like Tensorflow and Pytorch).

Depending on the task, computationally expensive approaches such as the Recurrent Neural Networks may be needed, leading to the requirement for specialised computing resources (specifically powerful GPUs). Most organisations do not have such IT resources, and cannot easily acquire them. Neural Networks should be considered in more complex contexts (use case) and, preferably, after experimenting first with less complex algorithms.

For many organisations, especially ones that are at the starting point of their machine learning journey, **the most suitable approach may not be the most advanced approach available, but rather a good approach that they have the resources to easily implement**. That could mean some variation of bag-of-words, or even no or little

pre-processing at all, in order to try and experiment with machine learning. Good results can often be achieved quickly using simple methods. To improve on these simpler methods can become more and more costly and time consuming. The most appropriate machine learning method depends on expectations, use cases and available resources of the organisation.

Cloud-based machine learning resources may help statistical organisations to avoid the initial investment that might be required (in some cases) in on-premises IT resources. However, these facilities come with other challenges, such as governance, security and the risk of disclosure of sensitive data, as well as often requiring ongoing payments to the provider for continued service.

The collaboration and code sharing within the HLG-MOS Machine Learning Project demonstrated that, to some extent, proven solutions developed by one organisation can be used in different settings (data, IT infrastructure, knowledge) in other organisations. Therefore, an organisation that has basic machine learning knowledge can quickly achieve good prediction results, even on small data sets, through a collaboration with other organisation.

3.1.6. Conclusion

Classification and coding processes are a prime example of where machine learning can play an important role in expediating the production of official statistics. Achieving this objective often requires investment of efforts to:

- Build and accumulate suitable training data to enable the machine learning models to learn adequately;
- Monitor its accuracy along the way;
- Make appropriate use of machine learning predictions in combination with manual coding; and/or
- Allow data users time to gradually build their confidence in data made largely from or with assistance by machine learning models to make data more accurate or consistent.

Machine learning prerequisites may include the availability of a training data set of sufficient quality, computation capacity and the machine learning expertise. There also has to be a predicted benefit of automation that outweighs the investment in the machine learning model development, maintenance over time and quality validation.

3.2. Editing and Imputation

3.2.1. Editing and Imputation in Statistical Organisations

The data that statistical organisations collect, through surveys, administrative data sources or web scraping, need to go through editing and imputation (E&I) processes to identify and treat problematic and missing values. This is a critical stage in the production process to ensure the data quality. To carry out this task, statistical organisations have employed various methods. For example, detection of suspicious values can be done in a rule-based way where data records are checked if they fulfil conditions on their values or through the comparison with the distribution of the data set. The data points that are not plausible should be treated or omitted from the data set where needed. Domain knowledge of subject matter experts is usually a necessary component of this editing and imputation process.

To clarify the distinction between editing and imputation, the following working definitions[35] are used in the rest of this Chapter:

- **Editing**: a task of identifying missing and problematic data (e.g., implausible values, contradictions in records) in data sets; and
- **Imputation:** a task of altering values that have been identified as incorrect and inserting missing values.

3.2.2. Expectations on Machine Learning

Often, machine learning is not used exclusively but in addition to or at least compared to an existing process (i.e., statistical methods, manual interactive work). This is true for both the exploratory and production phases and for different types of data (e.g., survey, administrative source). Machine learning could help increase the proportion of records in a data set that can be treated in a more automated way and improve the statistical production process by delivering better (e.g., more accurate, faster) results. This Chapter describes the expectation on machine learning from the members of the HLG-MOS Machine Learning Project Work Package 1 – Editing and Imputation Theme (more details about the pilot studies can be found in Chapter 3.2.3).

Editing

Broadly speaking, the editing methods can be classified as (i) rule-based methods (e.g., hard edit rules and soft edit rules that represent constraints on data, expected values or relationships between variables), or (ii) explorative methods that aim to identify potential anomalous data or with respect to some models that are deemed to represent the data properly.

Machine learning may discover rules that have only been "known" by intuition of the subject matter experts, through learning from a data set previously done by the experts. The supervised machine learning algorithm could learn, from these former editing results, which units, records or cells in a data set are problematic. This means that:

- The eventual goal of the machine learning model is to classify every unit of an incoming data set as "plausible" or "not plausible"; and

[35] Some definitions, which are not used in this Chapter, treat the process of altering incorrect values as a part of the editing process

- If such a model is sufficiently interpretable (explainable), rules that represent possible ways to classify a unit as "plausible" or "not plausible" can be extracted.

which would help:

- To conserve knowledge over time and changes in editing teams;
- To formalise the knowledge and to improve the automated detection of "problematic cells" in data sets; and
- To allow human editing staff to focus on validating "important", or in some sense, "influential" records.

Also, machine learning (as well as model-based approaches) may offer a valid and efficient new instrument for non-rule-based editing. The unsupervised machine learning model could be used to analyse data with respect to its "hidden structure" with less need for an a priori model for the data. This can help to gain efficiency to:

- Find outlier candidates or typical subgroups in an incoming data set; and
- Identify possible (soft) edit rules to classify a specific group of data as being problematic, to be further analysed.

which would help

- To detect "problematic cells" that are difficult to find by intuition or rules; and
- To use not only logical but also statistical aspects in the editing process.

It is also expected that, given a suitable amount of data, machine learning has the capacity to exploit a vast amount of information in the data to support the design and the maintenance of the editing process features.

Imputation

For the imputation, machine learning may improve prediction accuracy within already existing imputation schemes (e.g., regression imputation, predictive mean matching), which would possibly result in better imputation results. This leads to the question of how to determine if an imputation job is done satisfactorily. Indeed, there are different goals of imputation which can be summarised as following[36]:

- Predictive accuracy: the imputation procedure should maximise the preservation of true values. That is, it should result in imputed values that are as "close" as possible to the unknown true values;
- Ranking accuracy: the imputation procedure should maximise the preservation of order in the imputed values. That is, it should result in ordering relationships between imputed values that are the same (or very similar) to those that hold in the true values;
- Distributional accuracy: the imputation procedure should preserve the distribution of the true data values. That is, marginal and higher order distributions of the imputed data values should be essentially the same as the corresponding distributions of the true values;
- Estimation accuracy: the imputation procedure should reproduce the lower order moments of the distributions of the true values. In particular, it should lead to unbiased and efficient inferences for parameters of the distribution of the true values including a correct estimation of the additional uncertainty caused by the imputation (inferential accuracy); and

[36] EUREDIT project (Chambers 2001)

- Imputation plausibility: the imputation procedure should lead to imputed values that are plausible. In addition, they should be acceptable values as far as the editing procedure is concerned. This criterion should be applied in addition to the above four criteria.

These different goals have to use different metrics to measure their success. Machine learning may offer an additional value when there is either a regression or a classification step within the imputation process. If the focus is on predictive or ranking accuracy, this is apparent because machine learning is known to yield good predictions. If the focus is on distributional or estimation accuracy, very often, a "prediction step" is involved such as in (stochastic) regression imputation or predictive mean matching. There may also be value added by machine learning on the task of building imputation classes. Clustering and tree-based algorithms might be useful in this situation.

Machine learning is also expected to be faster in doing imputation compared to other methods once the model is established.

3.2.3. Pilot Studies

The expectations above have been checked against the results of pilot studies conducted by members of the HLG-MOS Machine Learning Project Work Package 1 – Editing and Imputation Theme as below:

Editing

- Istat, Italy – Machine Learning Tool for Editing in the Italian Register of the Public Administration; and
- Office for National Statistics (ONS), the United Kingdom – Classification of Records of Living Cost and Food (LCF) Survey Income Data that Need Editing.

Imputation

- VITO, Belgium – Early Estimates of Energy Balance Statistics using Machine Learning;
- Federal Statistics Office of Germany – Machine Learning Methods for Imputation;
- Istat, Italy – Imputation of the Variable "Attained Level of Education" in Base Register of Individuals; and
- Statistics Poland – Imputation in the Sample Survey on Participation of Polish Residents in Trips.

Complete reports of all pilot studies are available on the UNECE Machine Learning for Official Statistics wiki page (https://statswiki.unece.org/display/ML/WP1+-+Pilot+Studies)[37]. One of the pilot studies from the Office for National Statistics is highlighted in Box 3.2.

Table 3.7 offers insights into the motivation why the pilot studies have been conducted.

[37] Codes from some of the pilot studies are available on the UNECE wiki (https://statswiki.unece.org/display/ML/Studies+and+Codes)

Table 3.7. Legacy System and Aims

Organisation	Legacy System and Aims of Pilot Studies
Editing	
Istat	No legacy system, the task is new. Edit rules are of the main focus, but there are also investigations whether the application of ML can add value to the traditional editing approaches.
ONS	So far, there is only manual detection of spurious records. The goal was to replace the need for manual detection by learning a supervised model from former editing steps.
Imputation	
VITO	Old-fashioned working methods, such as large and complex Excel sheets should be replaced.
Federal Statistics Office of Germany	No legacy system. The study should show the principal behaviour of several ML methods in an imputation task. The aim was to investigate whether ML can replace other approaches in regression imputation.
Istat	No legacy system. The task is new. Goal of the investigation was to determine how and where ML can give greater benefits in solving the imputation problems compared with classic statistical models.
Statistics Poland	No legacy system. The goal was to achieve high predictive accuracy by imputation to avoid additional surveys.

Table 3.8 below gives an overview of the data used, important steps conducted, and machine learning algorithms compared in the pilot studies. Table 3.9 shows some details on the software and hardware as well as on the metrics used to assess the performance of the machine learning models.

Table 3.8. Data Used in Pilot Studies, Data Preparation Steps and Algorithms

Organisation	Data	Steps	Algorithms
Editing			
Istat	Public Administration Database (BDAP) and the Information System on the Operations of Public Bodies (SIOPE)	Comparing several variables from the two sources, identifying different types of inconsistent data, list of units regarded as important to be analysed deeper delivered by subject matter experts, identifying edit rules behind such units	Decision Trees, Random Forests
ONS	2018 Q2 and Q3 Living Cost and Food (LCF) survey data	Data preparation, calculation of the change vector, learning models to predict the change vector	Decision Trees, Random Forests, Neural Network
Imputation			
VITO	Quarterly data, ranging from Q1 2000 through Q1 2019	Z-standardisation of the data, feature selection for linear regression, calculating and comparing predictions	Linear Regression, Ridge Regression, LASSO, Random Forest, Neural Network, Ensemble Prediction
Federal Statistics Office of Germany	German cost structure survey of enterprises in manufacturing, mining and quarrying	Creating missing values (several proportions, several missing mechanisms), calculating and comparing predictions	K-NN (weighted and non-weighted), Bayesian Networks, Random Forests, SVM
Istat	Administrative information from the ministry of education, university and research, 2011 census data, sample survey data	Focusing on one region and on incomplete records, some manual feature selection, calculating and comparing predictions	MLP, Random Forests, Log-Linear Model
Statistics Poland	Quarterly sample survey on participation of Polish residents in trips for 2016 to 2018 and some big data sources	Learning different models for estimation and comparing their predictions by several measures	Different kinds of (generalised) linear models, Regression Tree, Random Forest, K-NN, different kinds of SVM

Table 3.9. Software, Hardware and Accuracy Measures

Organisation	Software/Hardware	Accuracy measures
Editing		
Istat	• R • No special hardware	Usefulness of the results indicating whether a variable determines the presence of a dangerous error in data; accuracy for model selection
ONS	• Python • Intel Core i5-8365U, 1.60GHz, 8 GB RAM	Recall, Precision, F1-score
Imputation		
VITO	• Python • Intel i7 CPU with 6 cores, and 32 GB of RAM	Root mean squared error, mean error, mean absolute error, mean absolute percentage error
Federal Statistics Office of Germany	• R • Intel Core i5-6500, 3.2 GHz, 8 GB RAM	Mean, standard deviation, skewness, kurtosis, minimum, maximum, 25 %-quantile, median, 75 %-quantile of the imputed variables, correlations between the variables
Istat	• Python • Azure cloud platform with Tesla V100-PCIE-16GB GPU	Micro-level accuracy, macro-level accuracy
Statistics Poland	• R • Intel Core i7-4770, 2x3.40 GHz, 64bit, 16 GB RAM	Mean absolute error, mean absolute percentage error, root mean squared error, R-square

Table 3.10 below summarises the most important aspects of the conclusions drawn from the pilot studies.

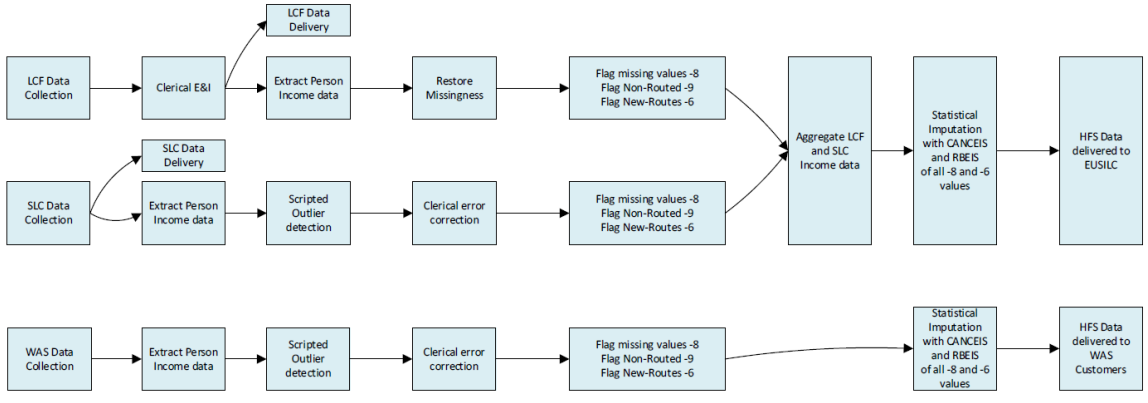
Table 3.10. Conclusion from Pilot Studies

Organisation	Conclusion
Editing	
Istat	• The first application of ML methods in this context has shown the possibility to use ML to support the design of an E&I scheme to make it more efficient • Exploring hidden patterns in the data with ML tools can help to understand how to classify units in a more efficient way in erroneous/not erroneous in terms of different error types and, therefore, how to combine the different E&I process steps

ONS	ML can be used for editing, but some points have to be borne in mind: • A ground truth/gold standard data set for retraining the model has to be created and enhanced periodically • ML expertise should be within the survey team to monitor and retrain the model when required • Editing will be far more efficient and faster with the ML solution compared to existing processes • Survey data will be available sooner for further processing and this will allow for more timely data and faster release • It remains open if ML can save cost here, because some clerical editing resources have to be maintained as well as technical expertise to build, analyse and keep the ML solution in operation
Imputation	
VITO	• Think of a baseline method that is simple, common-sensical and reasonably performing • No single ML method worked best, or even better than a very simple method • In this study the ensemble method, averaging results from several ML methods, seems promising • Manage expectations well; some people expect great results without effort or investment • Substantial effort is needed to conduct a proper investigation into the usability of ML methods • Making data and code publicly available has been well received by the community and can stimulate future joint work
Federal Statistics Office of Germany	• It is too early to give a general (not survey specific) advice to use one of the investigated methods for imputation • Random Forest does the imputation faster than the other tested methods in the study • The usage of weighted K-NN and Random Forest lead to more stable and "correct" estimations of the moments and quantiles; furthermore, the boxplots of these two methods are more symmetric than the other ones
Istat	• The results of estimation with the two approaches (MLP vs. log-linear model) are completely comparable • For particular sub-populations, such as extreme items (PhD), log-linear imputation is better • MLP micro accuracy is a bit better with respect to the log-linear model • MLP approach does not require variables pre-treatment
Statistics Poland	• Machine learning is much more powerful than traditional models and can easily overfit the data • Estimating the out-of-bag error is important to compare various methods by bootstrapping or cross validation • When k-fold cross validation was run several times, it led to confusion about which model is the optimal model; bootstrapping seems to be a more reliable method for model selection but at the same time it is more time-consuming • Model selection cannot be based just on the accuracy measures like MAPE, RMSE, etc. without checking distributional accuracy including biasedness • When data is imputed, it is hard to expect to impute data perfectly on the individual level; it may be expected to retrieve a true mean level of imputed data with respect to some strata; then, on average, totals can be calculated correctly

Box 3.2. Pilot Study from the Office for National Statistics, United Kingdom

The aim of this pilot study was to investigate if machine learning can be applied to identify suspicious personal income data records of the Living Cost and Food (LCF) survey that require manual clerical error correction, thereby building an efficient and accurate machine learning solution for the Household Financial Survey (HFS) that comprises of the LCF, Survey of Living Conditions (SLC), and Wealth and Asset Survey (WAS) (see below diagram for survey pipeline).



With the availability of both the raw survey data and the high-quality edited and imputed data of the LCF survey, it allows for changes of the data made during the clerical editing and imputation process to be labelled and used for supervised machine learning. The LCF data of Q2 2018 was used as the test data, whereas Q3 2018 was used as the training data. Several data preparation techniques were used to increase model performance such as feature selection, one-hot-encoding of the categorical features, normalisation and calculation of the change vector.

Machine learning algorithms from the python scikit-learn library for supervised learning were tried, including Decision Tree, Neural Network, and Random Forest, with Random Forest being selected as the final model given its better performance.

This pilot study is still a proof of concept, but it has shown that data records can be predicted to a high level of confidence for clerical error correction. Nevertheless, the question of what is accurate enough needs to be answered given the tension between recall and precision. While recall looks to minimise false negatives, precision is about minimising false positives. The F1-score, which is the harmonic mean of recall and precision, could be used as a quality measure. Yet, one still has to decide the appropriate recall and precision thresholds based on the end user's priorities.

As discussions with the survey teams progressed along with early promising results, the survey teams expressed interest in this pilot study and the need to find a new way of identifying data records that require error correction for the HFS. Detailed discussions about the scope and timetables of the software uplift of all social survey systems is under way.

3.2.4. The Value Added from Machine Learning and Lessons Learned

Editing

Traditionally, methods applied for data editing (i.e., task of finding missing and problematic data) include rule-based comparisons of observed values with (weak or strong) plausibility constraints, distributional investigations (e.g., for outlier detections),

and comparisons with external and/or former data sets. Every editing procedure can be designed in different flows according to the process features. It also involves several steps in which both automation (through edit rules) and subject matter experts (through interactive editing) play an important role in detecting problematic data. The degree of automation usually depends on the type of errors identified to be most common and on the possibility to define edit rules that characterise them. However, complete automation should not be the most important goal of machine learning in editing. For example, the pilot study from the United Kingdom, which aimed to analyse the capacity of the use of machine learning to increase the automation of the editing phase as much as possible (i.e., to reduce interactive editing in favour of automation), showed that:

- Learning from former editing results was possible (i.e., it is possible to predict whether a unit needs special attention); and
- The extraction of rules suffers from the trade-off that good predictions were only achievable with very detailed (i.e., long and complex) rules.

According to the pilot studies, **the editing process can be completed much faster and more consistently (compared to manual editing) with machine learning**. It may possibly even lead to a higher quality of the data and allow to release the final statistical products more quickly. Still, the effort required to maintain training data, the machine learning model and the analysis of the results might not prove to be a cost saver in the short term. Hence, the gain until now seems to be not so much in the efficiency of the results but the efficiency of the statistical process: machine learning allows using a huge amount of data with much less a priori knowledge, hypotheses and data preparation (e.g., general underlying structure of the data, stratification).

Imputation

For the imputation (i.e., task of altering incorrect values and inserting missing values), the pilot studies observed that:

- Machine learning delivered comparable (compared to traditional methods) results in a more automated way;
- Machine learning often produced plausible predictions. Nevertheless, in some cases, unplausible predictions appeared;
- Machine learning could produce more timely statistics by skipping some pre-treatment of variables (e.g., statistical transformations of the values such as logarithm transformation, grouping of variables, treatment of ordinal and nominal variables), but a successful use of machine learning in production would be possible only after a lot of (successful) experimentation on the topic;
- Machine learning could reduce human intervention (e.g., automatic variable selection);
- Imputation projects with time dependencies in the data (e.g., with time series data set) could be successful;
- It may happen that no single machine learning method works best for a given problem; and
- Some machine learning methods (or approaches within them) performed better in terms of distributional aspects than other ones.

Machine learning can be more powerful because they require fewer assumptions (compared to the fully parametric models). It is flexible enough to work very well on the training data set, but often perform poorly on a new data set. To avoid this, it is highly recommended to assess the performance of a machine learning model on a separate test data set, (e.g., to estimate population parameters based on a test set). Using machine learning successfully in production is possible only after a lot of (successful) experimentations on the topic of interest, substantial effort is needed to conduct a

proper investigation into the usability of machine learning methods. Parametric models are preferred, from every point of view, if the hypothesis of the model is satisfactorily met. Unfortunately, mistakes are often made in specifying the underlying hypothesis (i.e., in modelling the phenomena), in which case, the parametric model is not able to provide good predictions. Non-parametric models run a lower risk from this point of view, but fit (in the finite data situation) less well than the "true" parametric model. Furthermore, there is a need to shift the interest of stakeholders to accuracy and timeliness of results rather than to the interpretation of the parameters. There are no obvious quick wins to be made, and the uptake of machine learning methods in standard procedures requires substantial and continued effort and commitment. One should also always consider and check against a baseline method that is simpler, well-accepted, and reasonably performing to avoid drowning in complexities with only marginal effects.

3.2.5. Conclusion and Further Recommendations

Machine learning and statistical methods can assist the subject matter experts and the management in their decisions. For example:

- They can flag an observation as suspicious. The decision whether it needs to be corrected has to be made and to be accounted for by a subject matter expert; and
- Machine learning can provide a classification model but the choice of the threshold that should be used in the corresponding classification task has to be made and to be accounted for by a subject matter expert.

In addition to this, as pointed out during the Machine Learning Project Webinar 2020, machine learning can be used to tap into "fuzzy" forms of information (e.g., financial statements, articles in trade and financial magazines, company websites) that supports domain experts conducting E&I. This is an area that has largely been untouched by official statistics community and machine learning may offer a way forward as traditional methods (non-machine learning methods) cannot be used[38].

Applying machine learning needs a bit more data science skills (e.g., programming, coding, training/testing principles) than using traditional statistical methods (that are typically taught at the university in statistics courses).

It is also important that subject matter experts should always be involved. Programmers, statisticians, subject matter experts have to work together intensively, and all of them need some data wrangling skills. This has already been expressed, for example, by [1], who wrote: *"data science is a new interdisciplinary field that synthesizes and builds on statistics, informatics, computing, communication, management, and sociology to study data and its environments (including domains and other contextual aspects, such as organizational and social aspects) in order to transform data to insights and decisions by following a data-to-knowledge-to-wisdom thinking and methodology."*

The usage of machine learning is only useful if it is better (for quality dimensions, see Chapter 4) than the currently used baseline method and more simple statistical methods.

[38] Discussion paper by Mark van der Loo from ML Project Webinar 2020 (https://statswiki.unece.org/display/ML/WP1+-+Pilot+Studies?preview=/285216428/295633149/E%26I%20discussion%20paper.pdf)

3.3. Imagery Analysis

3.3.1. Introduction

Satellite information is becoming more and more available from a range of sources [2]. For example, the Landsat satellite of the United States of America's National Aeronautics and Space Administration (NASA) generates images with a 30-meter resolution for the whole globe in periods of 16 days. The complete collection of these Landsat 8 satellite images, which amounts to approximately 8-terabyte per year, is available in Amazon's cloud service, facilitating access to large volumes of information for non-Earth Observation (EO) experts. NASA also offers access to their MODIS satellite images with a resolution of 500 meters that generate a complete image of the Earth on a daily basis. It is also possible to access Sentinel-2 images from the European Space Agency (ESA) with a 10-meter spatial resolution and 5-day temporal frequency. In addition, there are also private companies with constellations of nanosatellites that are capable of generating an image at a resolution of 3-5 meters of the entire Earth daily [3][4][39].

This increasing availability of satellite imagery opens opportunities for official statistics. With the ever-fast-changing world, many of the issues that a current society faces often require more frequent monitoring at a more disaggregated level. The image data can be utilised to meet the growing demand for information to monitor various environmental and social-economic phenomena, such as monitoring the Sustainable Development Goals (SDGs) through computer vision and machine learning techniques [5].

3.3.2. Pilot Studies

Image data is still a relatively new type of data for statistical organisations but there is a growing body of works exploring how image data can be used for the production of statistics, such as the UN Global Working Group on Big Data (2017) Satellite Imagery and Geospatial Data Task Team Report[40] and the Conference of European Statisticians (CES) In-Depth Review on Satellite Imagery and Earth Observation Technology in Official Statistics[41]. The HLG-MOS Machine Learning Project Work Package 3 – Imagery Theme focused on the use of machine learning for image (both satellite and aerial) analysis, and this Chapter summarises pilot studies conducted by its members as below:

- Australian Bureau of Statistics (ABS) – Address Register Automated Image Recognition (AIR) Model;
- Statistics Netherlands – Learning Statistical Information from Images: a Proof of Concept;
- Federal Statistics Office (FSO), Switzerland – Arealstatistik Deep Learning (ADELE); and
- National Institute of Statistics and Geography (INEGI), Mexico – Use of Landsat Satellite Data for the Mapping of Urban Areas in Non-census Years.

[39] For more discussion on the different types of satellite data and their characteristics, in context of official statistics, see United Nations Global Working Group on Big Data (2017) Satellite Imagery and Geospatial Data Task Team Report (https://unstats.un.org/bigdata/task-teams/earth-observation/UNGWG_Satellite_Task_Team_Report_WhiteCover.pdf)
[40] Ibid.
[41] https://unece.org/fileadmin/DAM/stats/documents/ece/ces/2019/ECE_CES_2019_16-1906490E.pdf

Complete reports of all pilot studies are available on the UNECE Machine Learning for Official Statistics wiki page[42]. One of the pilot studies from INEGI is highlighted in Box 3.3.

Motivation and Organisational Context

The primary motivation of the pilot study organisation for using image data and machine learning was to reduce the cost and time required to conduct existing business processes or examine their suitability for producing statistics (see Table 3.11).

For example, to produce the Land Cover and Land Use (LCLU) statistics, Federal Statistics Office used to rely on a manual inspection process where human experts examined satellite images to determine the type of land use/cover of the areas shown in those images. This process is resource-intensive in terms of time and money, and machine learning is expected to facilitate the process and improve the detection of LCLU changes. The pilot study of INEGI aimed to detect the extension of urban areas across its vast national territory using satellite data and machine learning to help generate information products more rapidly.

Table 3.11. Motivations and Objectives

Organisation	Problem to Solve	Contribution	Value Assessment
ABS	Use an ML model to reduce the amount of manual intervention required during regular Address Register (AR) maintenance	Reduce costs (time) by making the process less resource-intensive	The number of automatically classified addresses
Statistics Netherlands	Explore the potential of ML for detecting poverty and population distribution from aerial or satellite imagery	Learn how to use ML to exploit imagery as a new data source in the production of official statistics and to assist other countries who do not have income data in measuring poverty from imagery	A working computer prototype
FSO	Facilitating land use and cover classification and improving change detection	Improve existing process to reduce costs (time). At present, internal resources are almost entirely allocated to visual interpretation, at the expense of other activities	A working computer prototype that demonstrates the innovative potential of the FSO in the use of ML/AI to process images
INEGI	Detect the extension of urban areas nationwide using ML	Reduce time and cost. Generate information products that contribute to cartographic updates. It will also be possible to incorporate urban	Clear objectives with links to potential impacts on existing and

[42] https://statswiki.unece.org/display/ML/WP1+-+Pilot+Studies

		growth data into the population estimation models. Finally, it will be possible to generate new types of statistics for monitoring the evolution and the extension of the cities of Mexico	future data products

The institutional priorities that led to the pilot studies and the stakeholders are summarised in Table 3.12.

Table 3.12. Organisational Context

Organisation	Relevance to the Organisation	Stakeholders Involved
ABS	Freeing manual classification experts from the simple work that can be performed by automatic algorithms, allowing them to focus their efforts on more complex address that cannot be classified automatically. Results from Automated Image Recognition (AIR) can be used in conjunction with other administrative data sources to strengthen confidence and quality in the Address Register	ABS officers. Since the Address Register forms the population frame for survey sampling it is important that it is of the highest quality and truly reflects the Australian population and its housing stock
Statistics Netherlands	The Centre for Big Data Statistics was launched in 2016 and has attracted data scientists with strong expertise in machine learning and computer science. This project has greatly stimulated the collaboration between the two groups. The study has also stressed the importance of specialised hardware and IT skills needed to be able to apply deep learning	Countries without income registers but with access to aerial or satellite images, and departments within Statistical Netherlands responsible for measuring SDGs or producing regional statistics
FSO	The FSO's land use statistics are an invaluable tool for long-term spatial observation, with an acquisition period that has been gradually reduced from 12 years (in 1979) to 6 years today. At present, internal resources are almost entirely allocated to visual interpretation, at the expense of other activities. Therefore, having a tool that simplifies the task of visual interpretation experts will allow them to generate information more quickly and allow them to contribute to other activities	A non-exhaustive list for stakeholders in Switzerland is as follows: federal administration, regional statistics, regional geoinformation centres, regional spatial planning offices
INEGI	Massive sources of information, such as satellite images, require a great amount of work to analyse manually, given the nearly 2 million square kilometres that Mexico covers. ML can be a key differentiator especially in the recognition of easily separable categories. In the most complex cases, human intervention is required to train the algorithm by performing a continuous and incremental update of training sets. ML does not replace field work nor manual validation, but it can complement and cover those aspects that have reached enough maturity to be automated	The General Directorate of Sociodemographic Statistics is interested in incorporating quarterly predictions that detect the change in growth in cities to incorporate their values in the population estimation models. Additionally, the cartographic update areas could also take advantage of the quarterly estimates.

Note that problems to be tackled through all pilot studies can be considered as a classification task where the goal is to predict to which class a certain image (or part of the image) belongs (e.g., land cover type, building type, urban vs. rural).

Data

Each pilot study organisation determined the study region and proceeded to acquire the necessary satellite and/or aerial imagery data for that area. Note that to work with the image data, raster information handling capabilities such as Geographic Information Systems software (e.g., ArcGIS, QGIS) or processing algorithms through specialised libraries in programming languages (e.g., Rasterio, RasterFrames of Python, Raster of R) are required.

The image data used in the pilot studies included open source Landsat images with a resolution of 30 meters per pixel and aerial images with sub-metric resolution (~ 25 cm per pixel) for which the organisations have developed infrastructure and invested in specialised flights to acquire them.

The image data can be considered as a set of signal strengths captured at different wavelength bands for different x-y coordinates. Aerial image captures signals from visible spectrum (red, green and blue) while satellite image captures more than 3 bands. Table 3.13 shows a summary of the characteristics of the image data used in the pilot studies.

Table 3.13. Image Data Used

Organisation	Images Used	Pixel Resolution	Bands/Channels
ABS	Aerial	~23 cm	3 (red, green, blue)
Statistics Netherlands	Aerial	25 cm	3 (red, green, blue)
	Satellite (Landsat 8)	30 m	11 (aerosol, blue, green, red, nir, swir1, swir2, pan, cirrus, tirs1, tirs2)[43]
FSO	Aerial	25 cm	3 (red, green, blue)
	Satellite (Landsat 8)	30 m	11 (aerosol, blue, green, red, nir, swir1, swir2, pan, cirrus, tirs1, tirs2)
INEGI	Satellite (Landsat 5, 7)	30 m	6 (blue, green, red, nir, swir1, swir2)

[43] Landsat 8 consists of 9 spectral bands (costal aerosol, blue, green, red, NIR (near-infrared), SWIR (short-wave infrared) 1, SWIR 2, Panchromatic band, Cirrus (bands used for detection of high-altitude cloud contamination) and 2 thermal bands (thermal infrared (TIRS) 1 and TIRS 2) (source: https://www.usgs.gov/faqs/what-are-band-designations-landsat-satellites?qt-news_science_products=0#qt-news_science_products)

Data Labelling and Use of Complementary Information

After the images were obtained, a labelling procedure was carried out to be used for the training and testing of machine learning models[44]. This is often done based on the manual work of experts through visual interpretation and/or in fieldwork activities. The cost of this manual labelling process is prohibitive if it has to be done for the entire set of vast image data. Machine learning models can learn patterns (e.g., characteristics associated with a certain class) from training data which is often a small subset of the data, and then be used to automating the labelling processes (i.e., classification) to ease the manual workload.

In addition to the images, complementary information from the study area can enrich the characterisation processes. For example, georeferenced information in vector format (e.g., ESRI Shapefiles, open GeoPackage format), which contain statistical or geographic information, can be used as the basis for new labels or contribute to the classification processes. It is also possible to incorporate digital elevation models (reticular information where the values of the pixels represent the elevation with respect to sea level) from which additional information can be generated (e.g., calculation of slopes).

Data Preparation and Feature Extraction

Image data typically undergo several data processings before being fed into machine learning algorithm. For deep learning algorithms (e.g., Convolutional Neural Networks), which are commonly used for image processing and recognition, data augmentation is often conducted. Data augmentation consists of carrying out systematic variations of the original images to expand the number of labelled examples available, for example, by rotating the images, changing the scale, etc. This is done in order to prevent or reduce the chances for the algorithm to overfit when using very small data sets. All the pilot studies carried out data augmentation to increase the amount of information used to train the algorithms.

Feature extraction is a procedure that derives or defines variables (features) that are helpful to characterise the image, such as texture, shape, or spectral indices. Sometimes, as in the case of the pilot study of INEGI, feature extraction is performed manually, which means that experts determined the characterisation strategy. The other pilot studies relied on the capabilities offered by the convolutional algorithms of deep neural networks for the automatic extraction of characteristics.

Machine Learning Algorithms

There are a wide variety of machine learning algorithms [6][7] applied to EO data. In the pilot studies, two types of algorithms were used: the state-of-the-art algorithms based on the Convolutional Neural Networks and more "traditional" machine learning algorithms such as Extremely Randomised Trees, Random Forest and Support Vector Machines. Convolutional Neural Networks-based algorithms use basic building blocks such as convolution filters and pooling layers, and organise them in stacks. One can use architecture (structure and composition of stacks and layers) used by the state-of-the-art Convolutional Neural Networks models[45] or build own architecture according to the

[44] For more about steps involved in satellite image analysis using machine learning, see Generic Pipeline for Production of Official Statistics Using Satellite Data and Machine Learning (https://statswiki.unece.org/display/ML/Studies+and+Codes?preview=/285216428/290358694/ML_WP1_Imagery_UNECE.pdf)

[45] Convolutional Neural Network architecture is formed by a stack of distinct layers (e.g., convolutional layers, pooling layers) that transform the input volume into an output volume (e.g.,

needs of each project. Due to the complexity in the training of these algorithms, some tools have been developed (e.g., Tensorflow, CNTK, PyThorch, Keras) that take advantage of the computational power of specialised hardware such as Graphics Processor Unit (GPU) and Tensor Processing Unit (TPU). The machine learning algorithm and software used for the pilot studies are summarised in Table 3.14 below.

Table 3.14. Algorithm and Software

Organisation	Algorithm	Python Library
ABS	Custom 12 layers CNN Architecture	Tensorflow (CPU)
Statistics Netherlands	CNN Architecture based on VGG16 and ResNet50 Random Forest and Support Vector Machine	Tensorflow (GPU) Scikit-learn
FSO	CNN Architecture based on Xception Random Forest	Tensorflow (GPU) Scikit-learn
INEGI	CNN Architecture based on LeNet Extremely Randomised Trees	Tensorflow (CPU) Scikit-learn

Results

Table 3.15 summarises the machine learning models selected as the final model in each pilot study and its accuracy which range from about 74% to 97%.

The pilot studies were in a proof-of-concept stage, with the exception of ABS which has moved its project to production. However, all organisations in the proof of concept stage were in the process of moving further towards production. FSO was in the validation and integration stage of established methodologies. INEGI is in discussion with key stakeholders to use the results of their pilot study in the production; the results of the 2020 Population Census are expected to help validating the results with field data. Statistics Netherlands that had to use open data for the pilot study to avoid privacy issues was in the process of implementing the approach on confidential data (income-related poverty label data) within a closed environment to ensure its security.

holding the class scores) through a differentiable function (source: https://en.wikipedia.org/wiki/Convolutional_neural_network). Several CNN architectures demonstrated competitive performance in image recognition by winning annual ImageNet Large Scale Visual Recognition Challenge (ILSVRC) such as VGG16, ResNet50 and Xception.

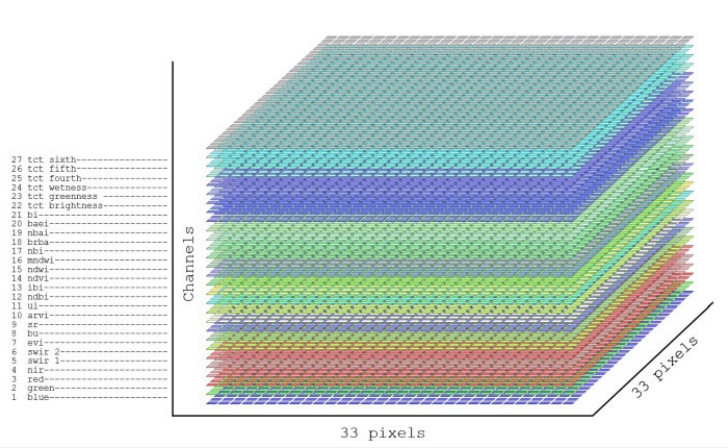
Table 3.15. Accuracy Results and Status of Pilot Studies

Organisation	Best Model	Overall Accuracy	Status
ABS	Custom CNN	96.9 %	Moved to production
Statistics Netherlands	ResNet CNN	74.0 %	Proof of concept
FSO	Xception CNN	~ 90.0 %	Proof of concept (fine-tuning the algorithm)
INEGI	Extremely Randomised Trees	93.9 %	Proof of concept

Box 3.3. Pilot Study from the National Institute of Statistics and Geography, Mexico

The objective of this pilot study was to generate national classifications that identify the expansion of cities through satellite data and machine learning application. It was expected that use of satellite data and machine learning could contribute to cartographic updates, help incorporate urban growth data into population estimation models, as well as generate new types of statistics for monitoring the evolution of the extension of the cities of Mexico.

The Landsat images were provided by USGS/NASA and INEGI built a Geospatial Data Cube (https://www.opendatacube.org/) in collaboration with Geoscience Australia. The 1km x 1km grid covering the national territory were labelled as urban and rural based on census data. And satellite mage patches corresponding to the 1km-sqaure grid (i.e., 33 pixels, each with 30m resolution) were extracted from the cloud-free national mosaic (calculated using Geomedian algorithm, for more information about Geomedian Landsat: https://www.inegi.org.mx/investigacion/geomediana/). In addition to 6 Landsat channels, 21 additional indices were calculated. Figure below shows the visual representation of one of the images.



Two algorithms were tested, namely, Extremely Randomised Trees (ET) and LeNet Convolutional Neural Network, with the former one being selected given its better performance in validation tests (e.g., 94% for ET and 87% for LeNet) and speed in training and classification (e.g., for training, 67 minutes for ET and 201 minutes for LeNet).

The study has been presented as a proposal in different areas in the organisation as an auxiliary method for planning and generating indicators (e.g., incorporating quarterly predictions in the population estimation models and cartographic updates), and has been well received.

The pilot study proved that machine learning could complement and cover aspects that have reached enough maturity to be automated, especially in the recognition of easily separable categories. Having said that, to monitor the quality and adjust the training sets, internal processes must be developed to ensure a continuous manual validation of the results (e.g., visual interpretation of satellite images by experts).

For the next steps, INEGI will expand the working group within the organisation and start the collaboration with the production areas. They are also going to identify alternative sources of information to improve validation processes, in addition to incorporating manual validation of samples by experts in visual interpretation and fieldwork.

3.3.3. The Value Added from Machine Learning

The pilot studies demonstrated the potential for using machine learning for classifying image data – these models could associate the variables of interest (e.g., building type, land cover type) with the images in the training data set and classify new images with reasonable levels of accuracy. The use of automatic classification frees up human resources to focus on the more complex cases and/or to perform other tasks for which there had previously been insufficient resources to carry out. Automation (or partial automation) of these tasks can allow a large volume of information in the image data to be processed in a reliable and fast way.

In the pilot studies, the organisations also acknowledged that the results achieved provided a foundation for further implementation of machine learning solutions as other applications needs were identified, and that collaboration between methodologists and data scientists has strengthened. This shows the value that machine learning can bring to various existing processes within statistical organisations.

3.3.4. Challenges and Lessons Learned

Machine learning is an iterative and incremental process, and hence, results can continue to improve as experience is gathered in the application of the methods as well as in the specific problem. The organisations considered that the algorithms used are well known in the machine learning field; however, as knowledge was gained from the application of the methods, it was possible to reach customised adjustments that improved the results achieved so far.

The challenges faced in carrying out the pilot studies were diverse. For ABS, it was crucial to have a solid business case to convince their organisation to launch the project, and it was also important to define the scope of the problem to be sufficiently simple to ensure that the goals would be achievable and the value to the organisation demonstrated quickly.

In the case of Statistics Netherlands, one bottleneck was the lack of specialised hardware (e.g., GPU) in its computing centre to avoid having confidential data in open environments while training the Convolutional Neural Networks models. In order to circumvent this problem, their first experiment was carried out with an open data set that allowed to validate the proof of concept while they obtained specialised equipment to work in a secure environment. In Mexico, it was considered that further iterations of

the training process should be made in order to improve the performance of their models.

The lessons learned from pilot studies can be summarised as follows:

- It is important to have a solid business case for the machine learning project, and to narrow down the problem with just enough complexity to demonstrate the value added from the use of machine learning;
- Training of deep learning models often involves the use of specialised hardware, and when a large amount of confidential data is required for the training, this hardware needs to be incorporated into the secure internal computer centres; and
- To be able to carry out classification exercises based on satellite and aerial images, it is essential to have high-quality training sets validated by experts through visual interpretation or fieldwork, as well as complementary data sets from administrative records, surveys or censuses.

4. A Quality Framework for Statistical Algorithms

4.1. Introduction

The aim of national statistical offices (NSOs) is to develop, produce and disseminate high-quality official statistics that can be considered a reliable portrayal of reality. In this context, quality is the degree to which a statistic's set of inherent characteristics fulfills certain requirements [8]. These requirements are typically set out in a quality framework, which is a set of procedures and processes that support quality assurance within an organisation and is meant to cover the statistical outputs, the processes by which they are produced, and the organisational environment within which the processes are conducted. Many widely accepted quality frameworks related to official statistics exist; for example, see the Australian Bureau of Statistics' Data Quality Framework [9], the United Nations' National Quality Assurance Framework [10], Eurostat's European Statistics Code of Practice [11] and Statistics Canada's Quality Assurance Framework [12].

Modern methods such as machine learning (ML) are gaining popularity as tools for official statisticians. In combination with modern hardware and software, these methods allow official statisticians to process new data sources such as text and images, automate existing statistical processes, and potentially make inferences without a sampling design. With this increased interest, quality frameworks may require reassessment to examine if quality implications that arises from new methods are adequately well covered.

In a traditional estimation context, statisticians typically attempt to learn as much as possible about a scientific truth from observed data. As described by [13], the scientific truth can be represented as a surface, and the observed data can be thought of as observations on the surface obscured with noise. Efron calls this the surface plus noise formulation. For example, a simple linear regression uses a formulation $y = \beta_0 + \beta_1 x + \epsilon$, where the surface, or, in this case, the line, is represented as a linear function of a variable x, and the response value, y, is observed with noise ε. Based on a set of observations (or data), the parameters of the line are estimated (e.g., using maximum likelihood or ordinary least squares methods) to obtain the estimated surface.

ML, on the other hand, can be differentiated from the traditional estimation context by its focus on prediction as opposed to estimation. ML algorithms "go directly for high predictive accuracy and [do] not worry about the surface plus noise models" [13]. Rather than searching for a hidden truth about the underlying phenomenon that generated the data or characteristics of the population, ML primarily aims to make predictions about individual cases. Note that this does not mean traditional statistical algorithms cannot be used for prediction. Once the parameters of a regression surface, or line, are estimated (i.e.,$\hat{\beta}_0, \hat{\beta}_1$), they can be used to make a prediction for any given new data point, x (i.e.,$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$). For this reason, some traditional statistical algorithms are commonly found in the ML toolbox, to be used for prediction rather than estimation.

* Reprinted from Yung, W., Tam, S., Buelens, B., Chipman, H., Dumpert, F., Ascari, G., Rocci, F., Burger, J. and Choi, I. (2022) "A quality framework for statistical algorithms", Statistical Journal of the IAOS, vol. 38, no. 1, pp. 291-308, with permission from IOS Press. The publication is available at IOS Press through http://dx.doi.org/10.3233/SJI-210875

With different purposes, it is not surprising that traditional statistical and ML algorithms have different areas of application, where one performs better than the other. For example, city planners who are interested in understanding what factors cause congestion in certain districts may employ statistical methods that have a long history of successfully solving such problems. But companies providing real-time traffic services for commuters are more interested in predicting whether a certain route that a commuter is taking will be congested or not, and this is the area of prediction in which ML specializes. In situations where accurate predictions at the individual level are infeasible, ML methods may also see limited applicability. However, statistical methods can still deliver insight. For example, a statistical model such as a logistic regression allows the assignment of significance to individual predictors when modelling the occurrence of a disease, even if such an ML or classical statistical model cannot accurately predict which individuals will get the disease.

The popularity of ML in social media services, online shopping recommendations and search engine refinement is due to its ability to make predictions for individual cases. In the official statistics field, ML is becoming increasingly popular in areas where such individual prediction tasks are needed. These can be areas where these tasks used to be solved by traditional statistical algorithms (e.g., predicting whether a certain record needs editing) or by manual work (e.g., predicting to which category an open-ended response or satellite imagery pixel should be classified). This popularity may be because machine learning practitioners accept more complex models than traditional statisticians, and this can lead to higher predictive accuracy.

ML is a relatively new tool in the official statistics field. While there is a growing body of work on the methodological aspects of ML, less has been done on the quality considerations needed for the use of ML in Official Statistics or whether existing quality concepts are equally applicable for this new method. Commonly used and accepted quality concepts may require re-evaluation through ML perspectives. For example, the United Nations' National Quality Assurance Framework states, "the accuracy of statistical information reflects the degree to which the information correctly describes the phenomena it was designed to measure, namely, the degree of closeness of estimates to true values" [10]. While this accuracy is often considered as how accurately statistical estimates describe characteristics of the underlying population (e.g., unemployment rate estimate based on the Labour Force Survey), accuracy for ML can also mean how accurate predictions are for individual cases in an intermediate processing task as part of the entire production process. Also, unlike manual classification done by humans, ML methods are scalable but may require initial development and investment. This affects cost effectiveness and timeliness of the end product in a different way than existing methods. The specificity of ML methods may require new quality dimensions (e.g., explainability and reproducibility) that are not considered in existing quality frameworks.

The goal of this document is to propose the Quality Framework for Statistical Algorithms (QF4SA) to provide guidance on the choice of algorithms (including traditional algorithms) for the production process. Throughout this document, we define an algorithm as a process or set of rules to be followed in calculations, derived from an assumed model and a predetermined set of optimization rules, for estimation or prediction. Statistical algorithms are those used within a statistical context. We purposely use the terminology statistical algorithm as it covers both traditional and modern methods typically used by official statisticians. It is impossible to talk about algorithms without thinking of data. However, throughout this document, we do not address data explicitly, but we do recognize that there is an important interplay between algorithms and data. In particular, all quality measures proposed are conditional on the data that are available.

Under the QF4SA, we propose five quality dimensions: accuracy, explainability, reproducibility, timeliness and cost effectiveness. Most of these dimensions are considered in existing quality frameworks for statistical outputs, but, in the QF4SA, they apply specifically to statistical algorithms that typically produce intermediate outputs. For

example, classification and imputation are processes in the production chain whose results are used in subsequent steps. The QF4SA concentrates on these intermediate outputs, as ML algorithms seem to be used, for now, in these contexts. The QF4SA'sdimensions are defined below:

Accuracy

Slightly different definitions of accuracy are given in several internationally accepted frameworks. The definition proposed for the QF4SA is a summary of those given in these existing frameworks: the accuracy of statistical information refers to the degree to which it correctly describes the phenomena it was designed to measure; i.e., it is the closeness of computations or estimates to the exact or true values that the statistics were intended to measure.

Explainability

Explainability is defined as the ability to understand the logic underpinning the algorithm used in prediction or analysis, as well as the resulting outputs. Explainability is greatly assisted by depicting the relationship between the input and output variables and providing the necessary information on the methodology underpinning the algorithm.

Reproducibility

At the basic level, reproducibility is defined as the ability to replicate results using the same data and algorithm originally used. This is known as methods reproducibility. At a higher level, it is defined as the production of corroborating results from new studies using the same experimental methods (results reproducibility), or similar results using different study designs, experimental methods or analytical choices (inferential reproducibility).

Timeliness

For the QF4SA, timeliness is defined as the time involved in producing a result from conceptualization to algorithm building, processing and production. A distinction should be made between timeliness in development and production, with the former generally taking longer than the latter.

Cost effectiveness

Cost effectiveness is defined as the degree to which the results are effective in relation to their cost. It is a form of economic analysis that compares the relative merits of different algorithms. For this purpose, cost effectiveness can be defined as the accuracy (e.g., measured by the mean squared error (MSE) or F1 score) per unit cost. Note that the total cost of doing the work—including fixed costs, such as infrastructure and staff training, and ongoing costs, such as production costs—should be taken into account.

It could be argued that there are other more appropriate definitions for these dimensions, but the purpose of the proposed quality framework is to open a dialogue on what official statisticians should think about when comparing statistical algorithms, be they traditional or modern. In what follows, we elaborate on each of the dimensions and propose aspects of each to consider when comparing algorithms.

4.2. Accuracy

Accuracy has many attributes, and, in practical terms, there is no single aggregate or overall measure of it. Of necessity, these attributes are typically measured or described in terms of the error, or the potential significance of error, introduced through individual sources of error. Accuracy can be said to relate to the concept of measuring the distance between the estimate (output) and the true value in an appropriate way. The closer the estimate is to the true value, the more accurate it is. We note that the deviation may be structural (bias) or random (variance).

The mandate of many NSOs includes developing, producing and disseminating statistics that can be considered a reliable portrayal of reality. To ensure the high quality of these statistics, most NSOs have developed quality frameworks that cover the statistical outputs, the processes by which they are produced and the organisational environment. One of the most important components of every quality framework is accuracy, which is related to how well the data portray reality and has clear implications for how useful and meaningful the data will be for interpretation or further analysis. The concept of accuracy is defined across several frameworks in similar ways; the common fundamental notion is the closeness of the estimate to the true value.

When ML methods are involved, there may be some confusion when discussing accuracy: the term "accuracy" is used for a specific performance indicator in classification and ML (namely the fraction of correctly classified data points). However, in this Chapter, we will present a much wider concept of accuracy and list several indicators to calculate it accordingly, with a special focus when ML methods are used.

The final aim would be to analyse how the Official Statistics Quality Frameworks can deliver a guideline to assess ML, methods and estimates, as well.

4.2.1. Accuracy in Official Statistics

For every framework, qualifying comments are common. For instance, the Australian framework states, "Any factors which could impact on the validity of the information for users should be described in quality statements" [9].The Canadian framework states, "It should be assessed in terms of the major sources of errors that potentially cause inaccuracy. The accuracy of statistical estimates is usually quantified by the evaluation of different sources of error, where the magnitude of an error represents the degree of difference between the estimate and the true value" [12]. These comments relate to the concept of measuring the distance between the estimate and the true value of the target parameter and refer to the closeness between the values provided and the (unknown) true values. This difference is called the error of the estimate, and "error" is thus a technical term to represent the degree of lack of accuracy.

Many measures of accuracy are available, each tailored to the particular estimation method being used and the situation (e.g., the type of data, the type of target parameter). Therefore, measures of accuracy can change according to the process and to the target of the estimator. This target may refer directly to (G1) the data elements (i.e., to the microdata) or (G2) aspects about the distribution of a variable or joint distribution of variables, as in the case of imputation. In addition, a common objective of statistical surveys is to estimate a set of parameters of the target finite population. Therefore, within a quality framework, (G3) the accuracy of the estimates of these parameters is generally also considered a key measure of quality. In all of these cases, the purpose of the measure is to quantify the closeness of the estimate to the true value.

It is important to underline that the existing literature on the performance of an algorithm suggests considering two different aspects when evaluating an estimator (e.g., [14]):

a) In choosing the estimator for the job, one must consider the choice of predictor variables, the estimation of hyperparameters of an algorithm, the exploration of transformations and so on. In this view, when choosing among different estimators, a performance comparison is necessary to choose the most efficient one for the job.
b) After an estimator has been chosen, the estimator's ability to predict the true values of new data must be assessed.

As well, in official statistics it is necessary to add an additional aspect to point (b) above:

c) When the final estimate is released, an estimate of its uncertainty is required.

Therefore, the question naturally arises about which method should be adopted for a particular problem. The answer, of course, depends on what is important for the problem; different estimation methods have different properties, so a choice should be made by matching these to the objective.

4.2.2. Accuracy of Supervised Machine Learning for Classification and Regression

As defined before, accuracy is meant to measure the closeness of an estimate to the true value. This means it depends on the estimation method under study. Therefore, before going into detail on measures of accuracy, we first set the context of how ML algorithms are typically used.

Training, validating and testing principle

To set the context, it is important to describe, in general terms, how the process of estimation and prediction is performed within a supervised ML approach. Suppose that there is a set, S, of labelled data $S: \{(x_1, y_1), \dots, (x_N, y_N)\}$, which belong to two spaces, i.e., $x_i \in \boldsymbol{W}$ and $y_i \in \boldsymbol{Q}$. That is, S is a set of observations of given variables X and Y that take on values over the given spaces. In ML, the existence of a function linking the variables in the two sets is presumed,

$$Y = f(X).$$

An ML algorithm estimates the mapping function f (with $\hat{f}$) from the input to the output. The goal is to approximate the mapping function so well that, when there are new input data (X), it is possible to predict the value of the output variable (Y) for these data. Depending on the nature of the spaces (and thus the variables within), we differentiate the task as follows. If the output space consists of a finite number of elements, then the task is called classification. Otherwise, the task is called regression. In less technical terms, if the output variable is qualitative or categorical, the learning task is called classification; if it is quantitative or numeric, the learning task is called regression.

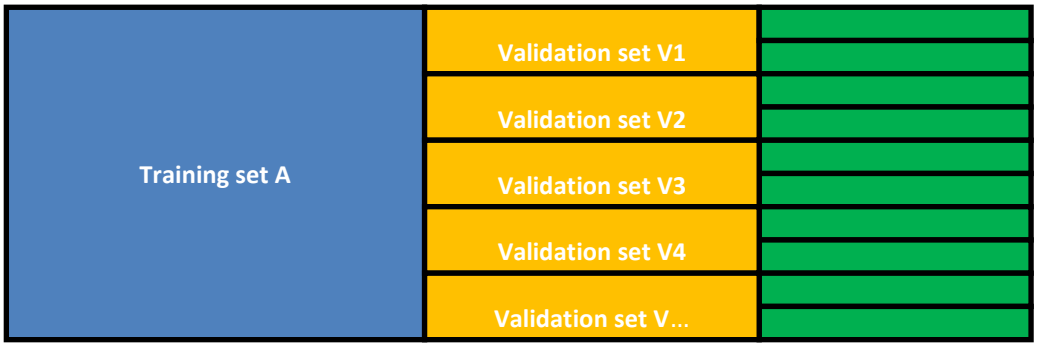
Regardless of whether the task is regression or classification, ML algorithms will attempt to learn the relationship between X and Y based solely on the available data observed in S. As a result, ML algorithms can be much more flexible than traditional modelling methods as they do not tend to presuppose particular relationships between X and Y. Of course, as algorithms become more flexible, the problem of overfitting must always be kept in mind, i.e., the possibility that a learned model fits very well on the observed data (perhaps because it even interpolates the data) but generalizes poorly to as-yet-unseen data. Using pre-specified models with controls to avoid unnecessary complexity (e.g., very high dimensional polynomial terms) can reduce the danger of overfitting. However,

ML simply tries to best estimate the mapping function, so it does not have such a restriction in the form of pre-specified models (in the sense that the set of functions in question is much larger than in traditional modeling). Usually, the class of possible models is much larger and can contain high-order polynomials, which are susceptible to overfitting to the observed data. Regularization, stopping rules and the evaluation of the learned model on test datasets that have not been used during the learning process are schemes to deal with this potential problem, and they help improve the generalizability of estimators or predictors. Therefore, an ML model should be learned in the following way: the set of available data is split randomly into several (ideally independent) subsets, $S \equiv A \cup V \cup T$ (see Figure 4.1). Note that Figure 4.1 is for illustrative purposes and does not suggest an optimal number of validation or testing sets or a ratio between the two.

- The first set, A, is for training the model (blue box).
- The second set (or sets), V (orange boxes), is used to evaluate different (combinations of) hyperparameters of an algorithm (e.g., the k in a k-nearest-neighbour approach, or the cost parameter C in an SVM approach).
- The third set (or sets), T (green boxes), is used to simulate what will happen when we apply the final learned model to new, as-yet-unseen data.

The random attribution of units to A, V and T is important to avoid concept drift, as explained by [13]. The final estimate $\hat{f}$ of the function f is usually obtained on the training set A (in combination with (the orange) validation sets or not) and assessed on the test set or sets using the criteria in Chapter 4.2.3. Having more than one orange and more than one green subset not only allows for point estimates for the accuracy measures, but also enables an estimate of their variance.

Figure 4.1. Training, Validation and Test Sets



The set-up described in Figure 4.1 is the best way to split the set S, but, for various reasons, practitioners may choose other ways. Often, when there is just one validation set (in orange), bootstrapping or cross-validation is used on this single validation set to simulate the ideal situation, in which there are multiple validation sets. At times, because of a lack of data, there is no validation set. In this situation, if optimal values for the parameters have to be found, this can be done via cross-validation or bootstrapping within the training data.

The simplest and most commonly used version is to learn some models based on one training set (perhaps with cross-validation to specify some parameters) and to test them on only one test set (or bootstrap samples created to provide multiple test sets to approximate the situation above). By repartitioning S into A and T multiple times, we have the opportunity to train and test the different algorithms or parameters of the algorithms on multiple datasets, thus showing us their performance in choosing the most efficient algorithm or parameters. For a more detailed exposition of some supervised ML techniques, the reader is referred to [15].

Variance

One common point of criticism of ML concerns the question of how to measure the uncertainty of the outputs. Besides the closeness of computations or estimates to the exact or true values (which can, for example, be expressed by the bias), statisticians also consider the variance of an estimator. This can be used to calculate confidence intervals, or the uncertainty of predictions, which can be used to calculate prediction intervals. In parametric model-based statistics, formulae are usually available for these quantities. The estimated variances of some traditional estimators can be written down in closed formulae; for example, if logistic regression is used, confidence intervals for the parameters and prediction intervals for the predictions themselves are available. As there is currently a lack of mathematical statistical theory for some ML algorithms, results like these cannot be produced at this time for those approaches without making additional assumptions. We note that assumptions are also required in traditional methods. However, in the case of binary classification, [16] have derived estimators of bias and variance for estimates of counts, proportions, differences of counts and growth rates. Their context assumes a binary classifier is used, and the resulting classification is used to produce the estimates mentioned above.

In the context of both ML and traditional statistics, resampling methods such as the jackknife [17], cross-validation [18] and the bootstrap [19] have been developed and can be used to quantify the uncertainty on the three levels, (G1) to (G3), mentioned above. [20] presents an introduction that focuses on the survey sampling context, while studies in the classification and regression context include those of [21] and [22], respectively. Of course, the suitability of using these resampling methods for the algorithm and data at hand has to be demonstrated before they are used. This is emphasized here because there are situations where, for example, the empirical bootstrap does not deliver suitable results (e.g., [23]). However, their examples of bootstrap failures are unlikely to occur in official statistics. Care, however, needs to be exercised to ensure that the dataset to which resampling methods are applied are representative of the population on which valid inference is to be made.

4.2.3. Common Measures for Evaluating Statistical Algorithms or Their Results in Machine Learning

When the focus is on unit wise predictive accuracy (G1)

In the pilot studies undertaken within the ML project of the United Nations Economic Commission for Europe's High-Level Group for the Modernisation of Official Statistics and the literature (e.g., [24], [25] and [14]), the following measures are commonly used to assess the success of ML algorithms:

- in the case of regression, RMSE (absolute or relative), mean error, mean absolute or relative error, R^2 or the standard error of regression
- in the case of classification, predictive accuracy, recall, precision and F1 score per class or on macro levels, G measure, Matthews correlation coefficient, and awareness of the consequences of the different misclassifications (see Chapter 4.3.2 on the Importance of explainability).

The references mentioned above contain many more measures and more discussion about them. A critical point in the case of classification, for instance, is how sensitive are measures to class imbalances (see, e.g., [26]) or whether they need a prespecified threshold in the decision function. In the latter case, areas under curves are used to assess classifiers—for example, the area under the receiver operating curve and the area under the precision recall curve (see [14] for more). Note that when these measures are estimated for a particular task to evaluate how well the learned predictor works, these numbers are valid only for tasks in the same context and based on new data from the

same distribution (or the same data-generating process) as the training and test data used for learning and assessing the predictor. This underlines the importance of having training and test data that are representative of the underlying population. This implies that the accuracy of an ML model must be continuously monitored and underlines the importance of having representative training and test data of the population under consideration.

When the focus is on distributional accuracy (G2)

Distributional accuracy is an important aspect to consider when using statistical algorithms to impute for missing values. In addition to the prediction of the true unknown missing value, relationships between the variables, or distributional accuracy, must be considered. At least in higher dimensions, distributional accuracy cannot be measured easily by only one number. However, in the univariate situation, well-known tests (such as the Kolmogorov-Smirnov test) can check whether two distributions are significantly different from each other. In the multivariate case, interactions of the variables have to be considered. It might be necessary to calculate correlations between the dimensions, but also to calculate extreme values, moments and quantiles separately per dimension and to recombine them in a specified sense. If all this occurs within an imputation step, the number of broken plausibility or edit rules for imputed values (and, if possible, the impact on the downstream task) and the accuracy (ideally also the variance) of the estimation of the target parameters may also be important indicators. When distributional accuracy is measured, the Jensen-Shannon metric appears to be appropriate, as outlined in [27], because of its versatility for handling multivariate distributions with continuous and categorical variables.

When estimating the target parameter (G3)

To quantify the accuracy for the estimate of a (usually continuous) population target parameter, the most common metric used is the MSE.

4.3. Explainability

4.3.1. Description of Explainability

In the QF4SA, explainability is defined as the degree to which a human can understand how a prediction is made from a statistical or an ML algorithm using its input features[46]. Note that this explainability concerns the relationship between input features and the predicted output rather than the "mechanical" understanding of the algorithm. For example, "finding a hyperplane separating data points by the class of output variable Y" is a mechanical understanding of a support vector machine (SVM), while an explanation such as "the higher the value of feature X, the more likely the output Y is classified as a category C" provides an understanding of how the input feature is related[47] to the output. Note that a prediction can be explainable but might not be "interpretable". For example, we may know "how" the output Y behaves depending on the change of the input feature X, but this does not necessarily mean that we know "why" the output Y behaves in such a way, and we reserve the term "interpretability" for this latter type. An ML algorithm is explainable as long as subject matter experts and other users can assess the logic of the way the algorithm makes a decision (see "Importance of explainability" below). Explainability can therefore be considered as a concept between the mechanical understanding and the interpretability.

Predictions from the traditional statistical models are often considered more explainable than those from ML models because they tend to be more explicit in linking the input features to the outputs (e.g., a coefficient from a linear regression model explains the direction and strength of the relationship between a feature and the output). However, explainability is more related to the model complexity than to the model type. For example, a regression model becomes more difficult to explain when more (potentially transformed) features, interactions, non-identity link are added to the model. Also, while a single decision tree is easily explainable, a random forest (an ensemble of decision trees) is less explainable. In both cases, increased model complexity might improve model performance but at the expense of model explainability.

4.3.2. Importance of Explainability

Explainability is important to gain users' trust in ML algorithms, as they are often considered "black-boxes." Understanding how an ML algorithm makes decisions can increase users' trust since they can relate the behaviour of the ML algorithm to their prior knowledge and internal logic. Understanding how algorithms make certain predictions can shed light on hidden patterns within the data that humans cannot easily perceive, which in turn could provide new insights about phenomena for users such as subject matter experts.

Explainability can be also important for model diagnostics for data scientists and statisticians developing the machine learning models. It can help improve the performance of the model and ensure that the model works in a way as expected. While high prediction accuracy may indicate that an ML algorithm performs well, an algorithm can make a correct decision for the wrong reasons. For instance, [28] describes an

[46] In this Chapter, we use the term "feature" to represent input variables. This is synonymous with "explanatory variable", "independent variable" or "regressor" in more traditional contexts

[47] A relationship revealed in any model trained on observational data does not imply causation. For instance, increasing the value of feature *X* through a subsidy or tax benefit may not be a successful policy-making strategy to promote category *Y*

example where an automatic system developed to predict a patient's risk of pneumonia based on X-ray images turned out to have simply learned the type of X-ray machine. The reason was that doctors usually took X-rays with portable X-ray machines for patients in critical condition and in urgent need of diagnosis, whereas patients without serious conditions were sent to a radiology department where their X-ray would be taken with a different type of X-ray machine. If an algorithm is a black-box, the outputs could, at best, be of limited use to the user and, at worst, be misunderstood in the critical decision making. Therefore, by requiring some human intervention, explainability can serve as a safeguard that machines are making correct decisions for the right reasons.

Explainability can play an important role in developing fair, accountable, transparent and ethical artificial intelligence. When decisions made by a machine have a direct and significant impact on the daily lives of people (e.g., medical diagnostics, autonomous driving, fraud detection, social credit), it is important to ensure that such decisions are made fairly and ethically. For example, if an ML model developed for the recruitment of new staff is based on hundreds of features happens to make decisions based mostly on the ethnicity or gender of the candidates, the algorithm is likely to be considered unethical. Therefore, it should be checked before the deployment, regardless of how accurate its prediction is. ML algorithms are often considered neutral and independent as they make decisions solely based on data and free of human bias. However, because of the very fact that they "learn" from data, accidental bias in data can be perpetuated by ML algorithms if careful checks and balances are not in place. While the current focus of exploration is more around the use of ML for intermediate processing that may have limited impact on an individual's life or final statistics, given the increasing awareness that human subjects should be provided with an "explanation of the decision reached [through automated processing]" [29], NSOs, as public agencies, should be aware of these issues with the use of ML.

4.3.3. Making Predictions More Explainable

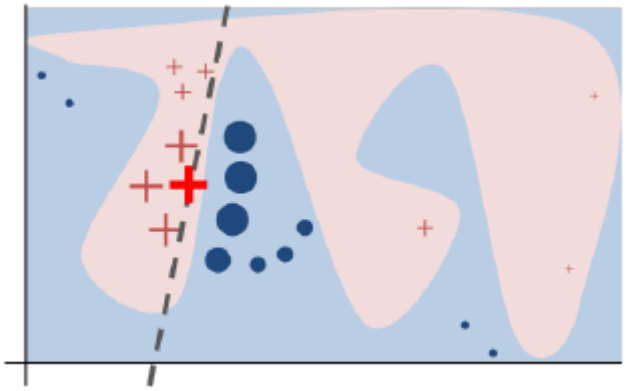
Explainable ML, or explainable artificial intelligence, is a recent but very active field of research. A multitude of methods, each with its own benefits and caveats, have been proposed to make predictions from black-box algorithms more explainable. Note that these methods do not directly make the ML algorithms more explainable. Instead, they make predicted results more explainable, and this sheds lights on the algorithm's behaviour, thus improving understanding of how the algorithm works. The objective of this sub-chapter is not to provide technical or methodological details of those methods but to introduce briefly a few existing methods developed in the ML community. Readers who are interested in further information are encouraged to consult the resources listed in the references (e.g., [30], [31], [32] and [33]).

An important group of explainability methods shows the importance of features, by visual plots, quantitative measures or surrogate models.

- One way to assess feature importance is to plot how the model prediction of an instance changes when the value of one feature is changed. For example, assume there are p features (X_1,...,X_p) and one output variable (Y). For each instance i, changing the value of X_{1i}, while fixing the value of all other features, will create a line of predicted values that shows how the individual prediction changes with the value of feature X_1. Combining all (or a sample of) instances together yields an individual conditional expectation (ICE) plot for feature X_1 [34]. A partial dependence plot (PDP) averages over all instances to show the overall marginal effect of a feature on the model prediction. While ICE plots and PDPs are intuitive and easy to implement, they assume that the feature of interest (plotted on the X-axis) is uncorrelated with other features. This might not be true in real situations.

- Another way to assess the feature importance without having to retrain the model is to measure the increase in prediction error when a feature is permuted, i.e., its values are shuffled to break up the relationship between the feature and the outcome.
- A surrogate model is an explainable model that approximates the relationship between the features and the outcomes predicted by a black-box model. The surrogate model provides an explanation for the prediction by the black-box model. Local Interpretable Model-Agnostic Explanations (LIME) are an implementation of a surrogate model for the purpose of explaining a single prediction [35]. New instances and their black-box predictions are generated around the instance of interest. An explainable model is trained on the generated data, weighted by their distance in feature space to the instance of interest. For example, the figure below shows a complex relationship between the two-dimensional feature space (X-axis and Y-axis) and binary output class (red and blue). An instance of interest is chosen (bold red cross), new instances are drawn from the feature space and their output values are predicted (crosses and points), and an explainable model (dashed line) is fit to the generated data, weighted by their distance from the instance of interest (size of crosses and points).

Figure 4.2. Example of Local Interpretable Model-Agnostic Explanation[48]



- The Shapley value is a measure of the contribution of a single feature value to the prediction of a single instance. It is calculated by comparing the predictions between different values of the feature, averaged over all (or a sample of) possible combinations of values for the other features. The contributions sum to the difference between the individual and average prediction.

Another group of explainability methods find[49] data points in the feature space that are intended to serve as the following:

- Counterfactual example: This is a data point that is as close as possible in the feature space to the instance of interest but with a different predefined outcome. For example, assume that a description of a work-related injury is "I cut my finger while chopping something on a wood board," and the occupation of the person is classified as "a cook." However, if the description had been "I cut my finger while carving something on a wood board," the outcome would have been "a sculptor." The change in feature space between the predicted outcome and the

[48] https://github.com/marcotcr/lime

[49] We focus on describing the data points of interest but omit how to find those data points through optimization of loss functions

counterfactual (e.g., “chopping” for “cook” vs. “carving” for “sculptor”) is a counterfactual explanation.

- Adversarial example: This is a data point when one or more feature values have been slightly perturbed in a way that the right prediction turns into a wrong prediction (e.g., making an image classifier mislabel an image of a stop sign by adding a sticker to it). Although designed to mislead a trained image classifier, adversarial examples can be used to improve model security and robustness, and thus explainability.
- Influential instance: This is a data point in the training set that considerably affects the performance of the algorithm when deleted. For some algorithms, influence functions can approximate an instance’s influence without the need to retrain the model.

Traditional statistical algorithms employ intuitive formulations, which produce results that are often innately explainable. ML algorithms may have higher predictive accuracy than these traditional methods, but, because of their complexity, they are often considered incomprehensible black-boxes. This can hamper the acceptance of ML in statistical organisations. Therefore, as ML becomes more common in the production of official statistics, the QF4SA recommends that if complex algorithms are used in any phase of output production, the official statisticians putting these algorithms in place must not only focus on minimizing the prediction error but also make a strong effort to achieve explainability by adopting some of the methods outlined above.

4.4. Reproducibility

4.4.1. Dimensions of Reproducibility

According to a subcommittee of the U.S. National Science Foundation [36] on replicability in science, "reproducibility refers to the ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator. That is, a second researcher might use the same raw data to build the same analysis files and implement the same statistical analysis in an attempt to yield the same results…. Reproducibility is a minimum necessary condition for a finding to be believable and informative."

It is important to recognize the three dimensions of reproducibility, namely, methods reproducibility, results reproducibility and inferential reproducibility [37].

- Methods reproducibility is defined as the ability to implement, as exactly as possible, the experimental and computational procedures, with the same data and tools, to obtain the same results. This is the same as the minimum necessary condition described in the U.S. National Science Foundation subcommittee recommendation.
- Results reproducibility is defined as the production of corroborating results in an independent study (i.e., with new data) that followed the same experimental methods. This has previously been described as replicability.
- Inferential reproducibility is defined as making knowledge claims of similar strength from a study replication or reanalysis. This is not identical to results reproducibility, because not all investigators will draw the same conclusions from the same results, or they might make different analytical choices that lead to different inferences from the same data.

For the QF4SA, recognizing that it is not feasible for official statisticians to undertake new data collection to corroborate initial findings, it is not proposed to adopt results reproducibility in official statistics.

Consistent with the Fundamental Principles of Official Statistics, methods reproducibility has been invariably embraced by NSOs, and its adoption in the QF4SA when using statistical algorithms to produce official statistics is expected to receive overwhelming support.

For inferential reproducibility, as multiple algorithms can generally be brought to bear on data analysis, there would be multiple ways to reanalyse the data. Official statisticians, when deciding to use a particular algorithm with a decided set of assumptions for analysis, have to be reasonably satisfied that the results from the chosen analysis can be corroborated by analyses using alternative but applicable algorithms and assumptions. This is particularly important for analytical inferences where general assumptions inherent in the algorithms have to be made about the data.

What is the distinction between accuracy and reproducibility? Accuracy is about having large accuracy metrics (e.g., small MSEs for continuous variables or large F1 scores for categorical variables), given a dataset, associated with the algorithm. Inferential reproducibility occurs when the MSE or F1 score of the difference between results obtained from the same dataset—from different choices of study designs, experiments or analytical techniques—is not statistically significant. In other words, inferential reproducibility is an attribute to show whether we can get essentially the same result (within a margin of error, and using algorithms correctly), not whether that result is good.

4.4.2. Importance of Reproducibility for Official Statistics

Reproducibility is a major principle underpinning statistical methods or algorithms used to produce official statistics. For the statistics to be trusted, they need to be reproducible. Publishing information on reproducibility of the statistical methods or algorithms is consistent with the third principle of the Fundamental Principles of Official Statistics, accountability and transparency, adopted by the United Nations Statistical Commission in 2014, stipulates that "To facilitate a correct interpretation of the data, the statistical agencies are to present information according to scientific standards on the sources, methods and procedures of the statistics." [38]. Reproducibility builds and enhances trust in official statistics.

Gleser articulated an underlying rationale for reproducibility in 1996. When commenting on the seminal paper on bootstrap confidence intervals published in Statistical Science [39], Gleser said the "first law of applied statistics" is that "two individuals using the same statistical method on the same data should arrive at the same conclusion." [40].

In the academic world, to ensure this first law of applied statistics is followed, many journals have revised author guidelines to include data and code availability. For example, starting February 11, 2011, the journal Science requires the following:

"*All data necessary to understand, assess, and extend the conclusions of the manuscript must be available to any reader of Science. All computer codes involved in the creation or analysis of data must also be available to any reader of Science. After publication, all reasonable requests for data and materials must be fulfilled. Any restrictions on the availability of data, codes, or materials, including fees and original data obtained from other sources (Materials Transfer Agreements), must be disclosed to the editors upon submission.*"

Trust is the currency of official statistics. While many factors contribute to building trust, an important one, as outlined in the first of the Fundamental Principles of Official Statistics, is impartiality. Impartiality can largely be demonstrated by transparency in the sources, methods and procedures used to compile official statistics. Such transparency allows independent analysts or researchers to assess the integrity of—and, where possible, reproduce and verify—published official statistics.

4.4.3. Demonstrating Reproducibility

Those who develop statistical algorithms to compile official statistics (e.g., methodologists, data scientists and analysts) are encouraged to assess the methods and inferential reproducibility of their algorithms before adoption. Once the reproducibility dimension of the algorithms has been confirmed during the development stage, they can be put into production, but it will be good practice to re-assess both methods and inferential reproducibility – see below – when new training data sets, or new algorithms for testing collaboration, are available.

Methods reproducibility refers to providing enough details about algorithms, assumptions and data so the same procedures could be exactly repeated, in theory or in practice. Documenting methods reproducibility therefore requires, at minimum, sharing analytical datasets (original raw or processed data), relevant metadata, analytical code and related software. For confidentiality reasons, NSOs generally cannot share identifiable raw data for independent analysis. It is therefore proposed that the analyses be replicated in-house and by another individual, who should be at arm's length from the original researcher, to assess reproducibility.

For inferential reproducibility, methodologists should test corroboration of the results from the chosen algorithm with a small set of applicable algorithms and different assumptions. While there are no hard and fast rules to determine what constitutes

corroboration, judgment should be applied when examining the results that are "different" from those of the chosen algorithm and assumption. For example, are the differences statistically significant, i.e., not due to random fluctuations? If they are, can they be explained (e.g., because of an improvement in efficiency), and can the explanation be supported by statistical theory?

Lastly, it is also proposed that the outcomes of methods and inferential reproducibility be documented for longevity and, where possible, published as part of the quality declaration statement normally released together with the official statistical output.

Clearly, reproducibility of statistical algorithms is fundamental for upholding trust in official statistical outputs. While three types of reproducibility are recognized in this Chapter, we propose NSOs adopt methods and inferential reproducibility to support their choice of statistical algorithms in producing outputs.

4.5. Timeliness

4.5.1. Timeliness for Statistical Algorithms

The quality guidelines and frameworks of many NSOs (Statistics Canada, the Australian Bureau of Statistics, the UK Office for National Statistics, and the Organisation for Economic Co-operation and Development) define timeliness as the length of time between the reference period and the availability of information. The QF4SA is advocating for development and processing time to be considered as well as the normal timeliness measures. More broadly, the concept of timeliness should be expanded to cover the period of time between a decision to fill an identified need for data has been made and the release of the information to meet that need. With the increased use of large datasets, the speed at which ML algorithms can be trained and run could lead to significant improvement in timeliness. This is particularly true for processes that are typically done manually, such as coding. Coding applications can be developed quickly using ML, particularly if past manually coded data can be used as training data. In addition to being fairly quick to set up, once developed, ML algorithms can process vast amounts of data in a short time. Compared with manual processes, ML algorithms could lead to significant savings in processing time which makes the release of the final output more timely.

4.5.2. Importance of Timeliness

Official statistics are useful only when they are relevant, and this means that they need to be available in a timely fashion. Indicators of an economic downturn are not relevant if they are available only six months after the downturn has occurred. Many quality frameworks define timeliness as the length of time between the reference period and the availability of information (for example, see [12]). However, for the QF4SA, we consider two additional dimensions of timeliness:

- the length of time it takes to develop or put in place a process
- the amount of time it takes to process data.

These two dimensions deserve consideration, as we feel that ML can offer some advantages over commonly used methods that can lead to improvements in the commonly used definition of timeliness.

4.5.3. Aspects to Consider

Clearly, measuring the time required in production to develop, set in place and use something is straightforward. In this Chapter, we list some aspects that need to be considered during the evaluation.

Data cleansing

It is highly likely that all potential methods will require that similar data cleansing be performed. However, if certain methods require specialized preparation of input data, for some reason, then this should be recorded.

Informatics infrastructure

If the method requires an informatics infrastructure that is not currently available, then the time required to set up such an environment should be considered. The time required to put it in place should not be underestimated.

Preparation of training data

Supervised ML algorithms require high-quality training data and, depending on the method, a large quantity of data can be required. Existing data should be considered for training data, if appropriate. Note that some traditional approaches also need auxiliary data, which can be time-consuming to obtain. For instance, non-ML coding algorithms typically need a data dictionary that is complete, accurate and up-to-date, but that can be very time-consuming to create.

Evaluation of data quality

Many well-established methods have processes for evaluating data quality. For instance, a well-developed theory exists for variance because of imputation. However, new approaches may not have well-defined processes to estimate quality indicators and may rely on resampling-type algorithms (e.g., cross-validation or bootstrap) to evaluate quality. Depending on the algorithm, these resampling methods could take significant time to compute.

Scalability of the approach

As data sources continue to grow in size, the time required to process large datasets should be considered. Manual processes are not a viable choice when the number of records to process becomes large, so ML algorithms may be preferable.

Model re-training

As ML models depend heavily on the training data, and hence the dataset, the model can quickly become outdated as patterns or concept schemes in the data start to change. Therefore, ML models should be continuously monitored and re-trained when needed. The validation of the model often requires the combined effort of data scientists and subject matter experts, which can make it time consuming and potentially costly [41].

4.6. Cost Effectiveness

4.6.1. Cost Effectiveness for Statistical Algorithms

Cost effectiveness can be defined as the degree to which results are effective in relation to the costs of obtaining them. Results in statistics are mainly measured in terms of accuracy; therefore, it is natural to link cost effectiveness to the accuracy dimension and try to measure it from this perspective. In this Chapter, we will define cost effectiveness as the accuracy (measured by the MSE for continuous data and F1 score or similar metrics for categorical data) per unit cost.

This is an operative definition that makes comparisons between different methodologies possible. In the case of ML, an organisation may compare the accuracy of an ML algorithm with the accuracy of a traditional method for the same statistical process, expressing both approaches in terms of their unit costs. The assessment of accuracy in ML is usually based on the consideration of a loss function; in traditional methods, uncertainty is expressed by the variance of an estimator, but resampling methods may be used as well. The same comparison could be made, of course, between two or more ML algorithms if the objective were to choose the most cost-effective one, all other aspects considered. However, some practical issues may need to be considered with this method, especially related to which costs should be included in the analysis.

Whenever a new method is introduced in a production process, an organisation will have to face some initial expenses to implement it. Such costs may be broadly defined as fixed costs, as they usually represent costs that must be paid to launch the infrastructure for the new method. ML, which can rely heavily on the underlying information technology (IT) infrastructure, may pose some challenges in this regard. In fact, fixed costs for ML mainly include the IT-related costs for acquiring new software and hardware and the costs of training the organisation's staff. These are different from the other category of costs that can be identified—ongoing costs—which derive from regular efforts to keep the whole system running and up to date. The following table lists the possible costs of an ML project. It may be useful to note that traditional methods also present fixed costs. However, NSOs have been investing in these over many years, so additional fixed expenditures are not usually required for them.

Table 4.1. Potential Additional Fixed and Ongoing Costs for Machine Learning Adoption

Cost component	Type	Purpose
Information technology (IT) infrastructure	Fixed	Acquiring necessary hardware and software
Cloud storage	Ongoing	Acquiring necessary cloud storage space
IT maintenance	Ongoing	Maintaining IT infrastructure
Initial staff training	Fixed	Training current staff on ML; may include hiring new staff
Ongoing staff training	Ongoing	Keeping staff up to date with new ML developments
Data acquisition	Fixed/ongoing	Acquiring and processing new data sources
Quality assurance	Ongoing	Conducting quality assurance and control

The details of these components will be explained later in this Chapter. For now, it should be noted that ML methods, by themselves, are not necessarily more expensive than traditional methods. In some cases, as they generally rely on less theoretical assumptions than classical statistics, they could be even simpler to implement and could be applied to traditional datasets without much difficulty. In such cases, where big data are not included, ML methods may present few additional costs.

It is true, however, that some specific machine learning applications are more prone to introducing significant costs than others. Typically, this is the case of deep learning neural networks algorithms that require large numbers of parameter weights for training, sometimes in the order of millions. This, of course, has an impact on hardware infrastructure and computational resources. Indeed, in literature the memory costs of the deep learning model's parameters have been used to represent hardware complexity and the minimization of the used memory bits can be incorporated in the model's cost function [42]. Of course, the direct translation from hardware requirements to monetary costs may not be straightforward and requires further analysis.

This approach stems from the fact that the term "costs" in ML context may assume a diverse set of meanings. In ML procedures it may refer to the monetary costs of implementation, to the energy and resources used for the required operations, to computational time, the memory requirements and so on. In this discussion we will focus mainly on the economic and the computational aspects, as the concepts related to timeliness have already been covered in the previous Chapter, although some mentions to temporal aspects of ML will be discussed.

The elements shown in Table 4.1 can be considered a starting point for comparing ML and traditional methods; such comparison can be made by (a) analysing whether the running costs for ML methods are cheaper than those of traditional methods or (b) computing the number of years to recoup the investment needed for the extra elements outlined in the table. Finally, we note that fixed costs should not be associated with a single instance of implementing an ML algorithm (unless the NSO intends to implement a single instance of ML). Fixed costs should be spread over the number of ML algorithms under consideration and future possible applications. At some point, the costs of new ML instances will be nil.

4.6.2. Advantages of Cost Effectiveness

The last decade has seen an explosion in data production, because of improvements in computer processing speed and innovations in communication networks. Official statistics have therefore been forced to compete with an increasing pool of data producers, while often being limited by tight budget constraints. Statistical offices are facing a challenge in meeting the required high-quality standards of official statistics with the resources that are made available to them. Cost effectiveness has guided many statistical institutes in recent years: the European Statistics Code of Practice, for example, dedicates its principle 10 to cost effectiveness, stating that resources should be used effectively. Current statistical processes may be revised to achieve the same or better levels of accuracy using sources or methodologies that would allow the organisations to save some costs; new data sources may be explored to save costs in data collection procedures. Indeed, cost effectiveness is one of the reasons behind the shift by NSOs from survey-centred data production to processes involving administrative and alternative sources of data. The introduction of ML can be seen as a further step in this evolution.

4.6.3. Organisational Considerations

ML in official statistics is still a field under investigation, although it has shown promising results. However, every organisation is different in terms of available budget and statistical production, so the convenience of introducing ML into current production has to be looked at on a case-by-case basis. If an organisation is new to ML algorithms and to big data sources in general, it would probably need to implement a suitable infrastructure from the start. Therefore, it will have to take into account the start-up costs and evaluate them against its budget, the cost of the current production, and the expected accuracy and timeliness improvements. Fixed costs may represent the main challenge in this case and may take a toll on the organisation's budget, but they also have to be compared with the future savings that ML would grant. As a result, fixed costs could actually be considered an investment that would allow greater savings in the future. Such savings may depend on the characteristics of the statistical production itself, as some processes may be more suitable for a migration toward ML than others. A given organisation may be involved in many projects that can easily—and beneficially—adopt an ML approach, while another organisation may have too few such projects, in which case the initial investment would be harder to justify. As [43] has shown, the decision of deploying a machine learning system from the prototype stage to the production stage can be regarded as an investment decision like any others, in which the cash flow at different points of time becomes a key factor to be considered. According to this approach, the gain in cash flow should be at least a fraction (given by the ratio of the rate of return over the number of decisions per year) of the cost of deployment. Under an alternative perspective, which is more apt to the official statistics context, the deployment should be carried out when the compounded savings per decision are greater than the cost of deployment, especially when the new system emulates the old decision process that it has replaced within the organisation.

4.6.4. The Potential Costs of Machine Learning

As can be seen in Table 4.1, IT-related and staff-related costs are a big part of the costs linked to the adoption of ML. To illustrate these, two of the main advantages of ML methods—scalability and automation—are introduced.

Scalability implies that a procedure can be applied with no or few modifications to a larger scale—for example, to a bigger data source with a greater set of units or features. As noted earlier, ML methods per se do not necessarily require any additional effort in terms of computation or resources. However, when used in conjunction with big data, they can quickly become computationally intensive. ML algorithms are often based on iterative methods and, of course, the better the hardware, the faster such iterations will be. An organisation's existing infrastructure may require some adjustments (e.g., central processing units, graphics processing units, storage space) before it can be used for computationally intensive operations or large datasets. Furthermore, IT costs should also include the resources needed for cloud storage and computation in the cloud, which are usually ongoing costs. In conclusion, when planning to introduce ML in a statistical process, an organisation could require an IT infrastructure that is optimized for a level above its current needs to accommodate potentially more intensive processing or bigger data sources.

Automation, on the other hand, enables an organisation to save on human resources. As listed in the table, the cost of training staff should be included in the initial costs of introducing ML methods, as the staff of statistical institutes is usually trained in classical statistics and may need appropriate training to use ML. This cost has to be sustained whether the application of ML is planned for small datasets or large datasets. However, the staff's underlying domain knowledge and statistical preparation should ensure that such training is not too extensive; consequently, the transition training costs may not be

high. But, as the field of ML is subject to rapid innovation and its application in official statistics is still new, the need for continuous learning cannot be neglected. For this reason, staff training is also an ongoing cost.

Once the fixed and ongoing costs of training are considered, automation should make it possible to save in terms of staff needed to execute operations. This should enable organisations to free up human resources for employment in other sectors of the statistical production cycle. In turn, the staff employed on ML procedures could then focus on aspects important for official statistics, such as explainability and the methods and inferential reproducibility of results.

Lastly, the adoption of ML algorithms opens new possibilities for data collection and data sources. From an IT point of view, acquiring big data sources presents the challenges that were illustrated before: expansion of storage space, both local and in the cloud, improvement of hardware, and so on. Additionally, acquisition costs must also be factored in, as big data sources are often held by private companies. Such costs may be either fixed or ongoing, depending on the agreements with data providers. In such cases, of course, it is advisable for an organisation to try to obtain a test dataset to assess its usefulness for the current production before committing to an agreement. It is also worth reiterating that some big data sources can be freely accessed, for example, through web scraping or open data portals. In such cases the access to the data is free but the subsequent activities may not be: the operations aimed at the storage, transformation and processing of the data – that is, the ones that fall under the category of ETL (Extract, Transform and Load) actions - can be considered as "hidden" costs because the effort for their completion could go undetected or underestimated, especially if the organisation is new to ML adoption. Such procedures often require the intervention of a domain-level expert and may not be fully automatable.

The structure of the data themselves is a factor to be taken into account. ML can be used either in a supervised or in an unsupervised context, depending on the need of labelling of the data, and these definitions can be regarded as a spectrum crossing through different algorithms that can be fed with semi-labelled data. Therefore, if the data require additional labelling under a supervised or semi-supervised approach, time and resources spent in annotating the data should be factored in. On the other hand, unsupervised procedures (including procedures, like neural networks, that are typically used under a supervised approach but can also be employed for the analysis of unlabeled or partially labeled data) may not need additional labelling but the information required as input data could exceed the immediate availability by the organisation. The time spent in the search of such large sets, which are required for the parametrization and the tuning of the models, along with the potential costs of their acquisition, would be difficult to ignore, even if labelling is not required.

From the elements described above, some tests can be formulated to include the various aspects of cost effectiveness into the assessment of accuracy. First of all, the accuracy per unit cost metric described in Chapter 4.6.1 could be regarded as a cost-effectiveness test, useful for investigating the costs linked to an accuracy improvement deriving from the adoption of a new method. For this purpose, this test should include only the variable costs in its assessment, especially if used to compare an ML method with a traditional one, for which fixed costs have probably already been paid in previous years.

Another possible test focuses on the return on investment, which is useful to assess the fixed costs and the time needed to recoup the initial investment in ML. Two or more ML algorithms can be compared over a specific period of time (e.g., five years) to assess which offers more savings and whether such savings are enough to compensate for their introduction in the production process.

An ML algorithm should be chosen only if both tests produce satisfactory results, that is, if the algorithm is cost effective and the cumulative savings it guarantees are bigger than the net present value of the investment in ML.

The same ML and IT infrastructure can be—and usually is—shared between multiple ML procedures. This should happen as NSOs become more confident in ML methodologies and increase their adoption of ML. In this case, when the metrics are computed to evaluate the costs and savings of an ML implementation, fixed costs should be apportioned between the relevant algorithms.

4.6.5. Conclusions

The previous illustration of the potential costs of implementing ML should shed some light on the metric that was introduced at the beginning of the Chapter, accuracy per unit cost. When this measure is computed, it can be convenient to differentiate between specific elements of the potential expenses, depending on the needs and the current state of the statistical organisation. In other words, the accuracy per unit cost metric does not have a given single use, as it has to be considered in the context of each organisation. For example, decomposing it into different cost components is useful to better assess potential savings and accuracy improvements against future ongoing costs. This would also help estimate the time needed to recoup the initial investment.

Lastly, if ML allows an organisation to improve the accuracy of its estimates while saving some resources, the question of where best to redirect these resources should be investigated. Of course, this is another case-by-case question, and a general answer is impossible. In the context of official statistics, it is important to highlight that the experimental nature of the processes and the novelty of some of the techniques may call for additional quality measures and controls. Since the mission of official statistics programs is to produce transparent, accurate and accessible data, it may be worth spending some of the additional resources to maintain regular quality assurance and quality control operations for the processes involving ML. This would ensure greater transparency for data users and give data producers deeper insight into the technical aspects of ML.

4.7. Summary and Recommendations

National statistical offices (NSOs) around the world are modernising, and many are looking at modern statistical algorithms as a significant part of their modernisation journey. Modern statistical algorithms have plenty to offer in terms of increased efficiency; potentially higher quality; and the ability to process new data sources, such as satellite images. The challenge comes from deciding when modern algorithms should be used to replace or complement existing algorithms. Many modern algorithms were developed in a prediction context and are designed to minimize prediction error. However, most algorithms currently used in official statistics were developed to produce inferentially correct outputs. Comparing methods developed under these two paradigms is not easy.

The proposed Quality Framework for Statistical Algorithms (QF4SA) is a first attempt to lay down some groundwork to guide official statisticians in comparing algorithms (be they traditional or modern) in producing official statistics. The QF4SA's five dimensions are applicable to traditional and modern algorithms and provide food for thought to official statisticians when choosing between different algorithms. Based on the QF4SA, the following recommendations are proposed for NSOs considering the use of machine learning in the production of official statistics:

1) It is recommended that all five dimensions of the QF4SA be considered when deciding on the choice of an algorithm, particularly when choosing between traditional and machine learning (ML) algorithms.
2) Ideally, NSOs should estimate the expected prediction error using methods such as cross-validation or other appropriate resampling methods. These methods are valid only if the training sets are generated from the data in the same way the data are generated from the population. This underlines the importance of properly constructed training sets. For instance, training data about only women should not be used to train a model to predict the income of men. Therefore, it is recommended that NSOs calculate the expected prediction error, as well as the prediction error, to protect against potentially poor-quality training data. In addition, it is recommended that high-quality training data (data that are representative of the population in question) be created when applying supervised ML algorithms. NSOs may want to leverage their sampling expertise to research the creation of high-quality training data.
3) Given that explainability is a major barrier to wide acceptance of ML algorithms, it is recommended that NSOs explore and use the methods outlined in the Explainability Chapter to help users understand the relationship between input and output variables. Data users' understanding can help eliminate some of the black-box concerns associated with ML. This will contribute to increased acceptance of and trust in ML algorithms.
4) Given the role that reproducibility plays in gaining the trust of data users, the QF4SA recommends that, as a minimum, NSOs take action to implement methods reproducibility. Inferential reproducibility should be carried out as well, when possible and desirable, limited to only the replication of the analysis using different but applicable algorithms and assumptions. We note that for inferential reproducibility, the results of a chosen method need only be corroborated by alternative algorithms or assumptions. They do not need to be the same. When the alternative algorithms or assumptions do not corroborate the original results, NSOs should ensure they understand why and determine whether the chosen method is warranted.
5) Timeliness is covered by most, if not all, existing quality frameworks. However, the timeliness dimension commonly used is defined as the time between the end of the reference period and the availability of the information sought. For certain processes leading to the production of statistical outputs, it is recognized that modern algorithms could lead to significantly shorter development and processing

times, in comparison with traditional algorithms. Examples of these processes include industry and occupational coding and image processing. Therefore, the QF4SA recommends that development and processing time be added to the commonly used concept of timeliness.

6) A motivating factor of NSOs' modernisation is cost effectiveness. By considering alternative data sources, NSOs want to reduce collection costs and respondent burden. For some alternative data sources (e.g., satellite images), modern algorithms are the only available way to process them. When evaluating the cost of potential algorithms, NSOs must consider fixed costs, as well as ongoing costs. Examples of fixed costs include establishing information technology (IT) infrastructure and retraining employees to work with the new infrastructure. We note that fixed costs can be amortized over time or across projects. Examples of ongoing costs include IT maintenance, cloud storage for the data, the cost of acquiring the data and processing time. Processing time in particular could be significantly reduced under certain circumstances by using modern methods. Given these costs, the QF4SA recommends that NSOs consider two aspects in particular when considering cost effectiveness: cheaper operating costs and time to recoup fixed costs.

5. Journey from Machine Learning Experiment to Production

5.1. Introduction

The pilot studies in Chapter 3 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*"[50]. 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[51]. 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[52]. 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 5.1). Also, while the "experiment environment" often 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.

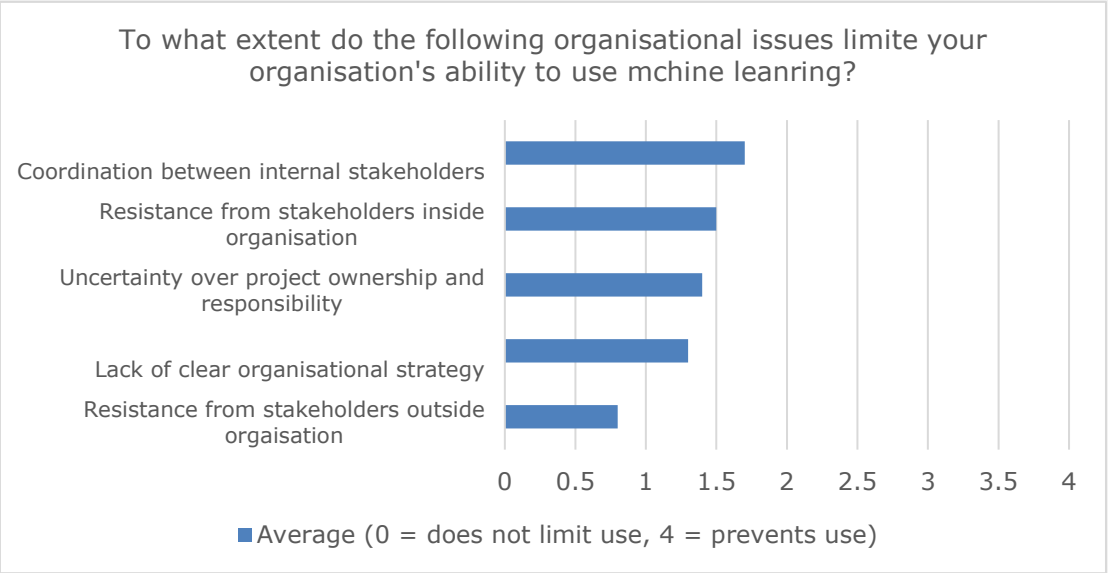
In this Chapter, typical steps that statistical organisations would take from the machine learning experiment to its deployment in production are described with some of technical and organisational issues and constraints often experienced in each stage. Note that, while the steps are in the logical order, they do not need to be followed in the sequential

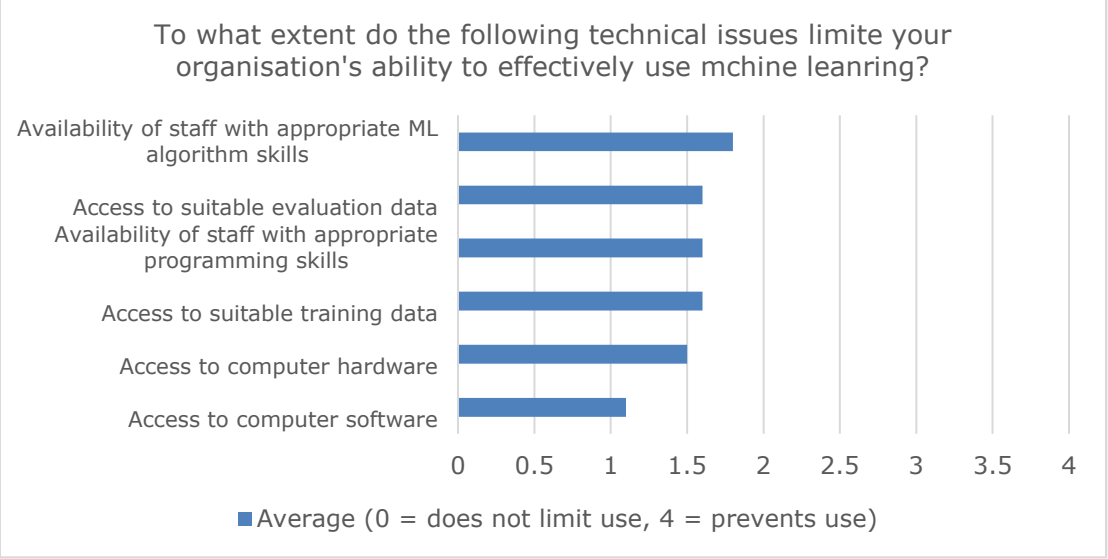
[50] https://venturebeat.com/2019/07/19/why-do-87-of-data-science-projects-never-make-it-into-production/
[51] https://algorithmia.com/state-of-ml
[52] Fundamental Principles of Official Statistics https://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx

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 5.1. 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[53].





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

5.2. Journey from Machine Learning Experiment to Production

5.2.1. 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 machine learning solution in production. However, regardless of how it started in the beginning, a machine learning 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; see Chapter 4) 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., machine learning 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 machine learning can realistically accomplish [41].

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 Chapter 5.2.6 Deploy the Model).

Machine learning is a relatively new area of work in the statistical organisation. Some organisations are equipped with a centralised 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.

5.2.2. 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 machine learning 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[54] 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

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

available. The proof of concept (Chapter 5.2.3) or business cases development (Chapter 5.2.4) stages could then be used to show that the method in question works well for the particular problem using real data.

5.2.3. 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[55].

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, machine learning methods need labelled/annotated data set to train and test the model. Given that machine learning 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 Chapter 3 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*"[56]. 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

[55] See Chapter 2.2.2 for more details on accuracy measures
[56] https://statswiki.unece.org/display/ML/WP1+-+Theme+1+Coding+and+Classification+Report

textual data, this is where the original form is converted into a numerical form (e.g., vectorization of text data[57]);
- **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[58]; and
- **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.

5.2.4. 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 machine learning (e.g., if manual coding can be replaced by rule-based coding, why machine learning?);
- **Business value addition**: description of how machine learning 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 machine learning 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 Chapter 5.2.6), 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;

[57] See Chapter 3.1.2 for more details on the text data preprocessing
[58] https://scikit-learn.org/stable/modules/cross_validation.html

- **Project plan**: identification of tasks and steps to follow from the development of the machine learning 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 machine learning-based prediction can also be included;
- **Operational business process**: description of the process steps and flow to be followed when the machine learning 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, machine learning 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 5.2), privacy and security; and
- **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 5.2. The Ethics of Machine Learning

Prepared 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[59], 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,

[59] 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.

5.2.5. 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); and
- **IT Environment**: the production environment might be different from the experiment environment (where PoC model is built and tested). The software used for machine learning 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 machine learning codes in a software language that is supported by the production system. Note that some machine learning 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 machine learning algorithms may not be readily available in some software languages, hence affect the choice of the machine learning algorithms to try in this model development stage.

The machine learning 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 Chapter 5.2.6); and
- For re-usability: some of features and model components can be re-used for the development of other machine learning models in the organisation.

It is also important to have workflow around the machine learning 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 machine learning 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 machine learning 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 finalised as the model development stage may take a long time.

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

5.2.6. Deploy the Model

What is the model deployment?

The machine learning model is a tool designed to address a business problem identified in the stage 1. To provide its business value, therefore, the machine learning 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 machine learning 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 machine learning 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 5.3). 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. machine learning 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 machine learning 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 machine learning models, it may be useful to consider training sessions as a part of the model handover process.

Monitoring plan after deployment

Machine learning 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 5.3. Designing and Deploying a Machine Learning Solution for Official Statistics: The IMF Experience

Prepared by International Monetary Fund (IMF) Statistics Department

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[60].

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 against the use of ML solutions is that they are black boxes and very difficult

[60] 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

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; and

- **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; and
- Retrain the model using inputs adjusted by the subject matter experts.

5.3. 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 machine learning 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 Chapter described the six stages toward the operationalisation of machine learning 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 machine learning 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;
- Machine learning 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; and
- 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.

6. Key Messages and Conclusion

Based on the knowledge and experience gained during the HLG-MOS Machine Learning Project, the advancement of machine learning for the production of official statistics can be summarised in two words: acceptance and facilitation.

Most of the burden of making machine learning *accepted* within a statistical organisation lies upon those who develop and operationalise machine learning methods, while most of the burden for *facilitating* its development and operationalisation lies upon the organisation. Most importantly, both the acceptance and facilitation require the support of all employees. The next two Chapters elaborate on the key aspects for the acceptance and facilitation of machine learning within statistical organisations.

6.1. Key Aspects for Acceptance of Machine Learning

Alignment with Business Needs

Ultimately, machine learning solutions must be accepted by the people responsible for producing data (usually subject matter experts) and, more importantly, those who use the data. Like any approach or technology, machine learning is one possible means to an end, and as such, it should not be considered or adopted simply for what it is, but for what it can do to better address the business needs (e.g., increased relevance, detail, timeliness, accuracy, cost efficiency). The pilot studies in Chapter 3 generally focused on improving timeliness and accuracy for three statistical processes. Applications of machine learning to address other business needs in other processes abound, and some such examples are listed in "Other applications of Machine Learning for some examples" on the UNECE statistics wiki.[61]

Guidance from a Quality Framework

Machine learning solutions must contribute to results of as good as, or better-quality than previously used approaches for fulfilling business needs. In order to do this, one needs to define what "quality" means. Definitions of quality are provided by many widely accepted quality frameworks that have been developed by national and international statistical organisations.

The Quality Framework for Statistical Algorithms (QF4SA) presented in Chapter 4 provides a supplement to these frameworks, and focuses on aspects that are more prominent to the acceptance of machine learning solutions. QF4SA provides guidance on the choice of algorithms (including traditional algorithms) for statistical production processes. It deliberately uses the terminology "statistical algorithm" since this term covers both traditional and modern methods that are typically used by official statisticians to strengthen the mutual comprehension between proponents of each of these types of methods. There is no set formula to ascertain when results from machine learning solutions are good enough or better than alternatives, and as with most quality frameworks, QF4SA proposes a number dimensions of quality that must be considered jointly. One may choose to place more emphasis on one or two of these dimensions, but none should be ignored.

[61] https://statswiki.unece.org/display/ML/Other+applications+of+Machine+Learning

Demonstration of Added Value

Most of the pilot studies described in Chapter 3 emphasised the importance of demonstrating added value. For classification and coding (Chapter 3.1), the studies demonstrated that machine learning can deliver better quality results than a strictly manual method.

The common challenge faced by these pilot studies was a lack of a statistically sound baseline against which to compare the machine learning results. Consequently, many studies start with the goal of replicating the results of an existing method (e.g., producing the same product classes as manual classification), and focusing on added value in terms of timeliness (and indirectly in cost).

There are issues with this goal. First, the accuracy of the existing (or competing) method is often either not known or not supported by a sound assessment method. Second, machine learning may not be able to replicate existing methods. The pilot studies described in Chapter 3 had accuracy between 40% to 85% of the results from an existing (manual or other automated) operation.

The goal of machine learning should not be limited to replicating another approach, unless it can do so much more quickly and at a significantly lower cost, but rather to improve the approach by combining the respective strengths of each.

For example, in the context of a coding and classification process, this could mean:

- Using machine learning predictions to automatically assign a class on the predictions known to be very accurate, for example, over 98% confidence;
- Using the less confident but good enough predictions to aid coders; and,
- Ignoring the machine learning predictions with low level of confidence, and instead relying on human coders to classify these cases (usually for rare classes).

Variants of this strategy can be also used in production, as shown in several pilot studies (Workplace injury and illness[62], Industry and occupation[63] and Standard Industrial Classification[64])[65]. On editing and imputation (Chapter 3.2), the studies showed results ranging from having no added value (a simple imputation method did better than other options) to being promising. There are no indications that machine learning methods cannot work. They may require less programming and be quicker to implement than current methods.

On the downside, creating and maintaining good training data for machine learning algorithms is a challenge. Additionally, explaining what machine learning solutions produce and how they produce it, even if it is quicker or more accurate, can be difficult, making it hard for stakeholders to accept them. More studies and foundational developments[66] are needed to guide the use of machine learning in this area and to determine the characteristics of a favourable context within which to apply it.

---

[62] https://statswiki.unece.org/download/attachments/285216428/ML_WP1_CC_USA-BLS.pdf?version=2&modificationDate=1605171512748&api=v2

[63] https://statswiki.unece.org/download/attachments/285216428/ML_WP1_CC_Canada.pdf?version=1&modificationDate=1605171571083&api=v2

[64] https://statswiki.unece.org/download/attachments/285216428/ML_WP1_CC_Norway.pdf?version=1&modificationDate=1605171509316&api=v2

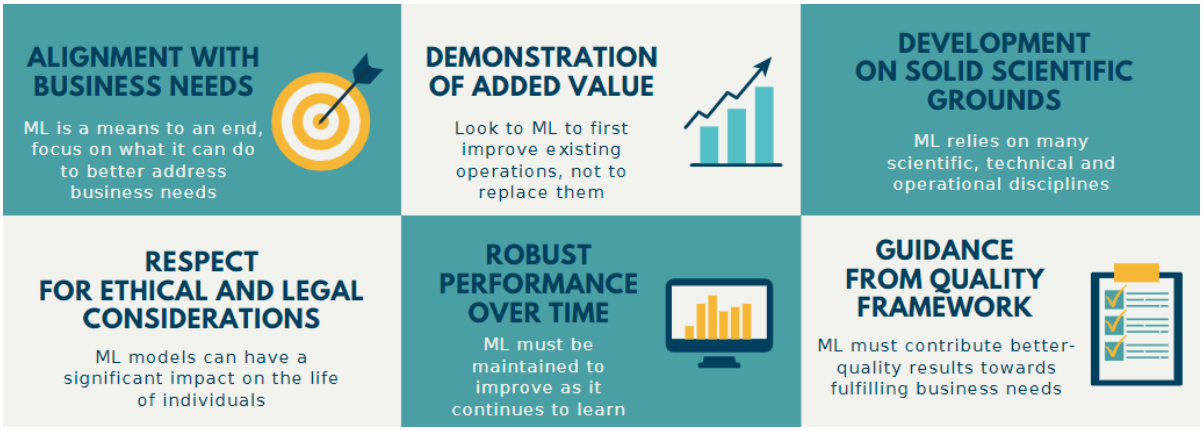
[65] An experiment on this strategy using ML code and data shared was also conducted (https://statswiki.unece.org/display/ML/A+user%27s+experiences+with+the+ML+code+and+data+shared)

[66] For example, Hints and Ideas for Data Cleaning (https://statswiki.unece.org/download/attachments/285216428/ML_WP1_EI_Italy_Rocci.pdf?version=1&modificationDate=1605171577247&api=v2)

From the beginning of the Machine Learning Project, machine learning was presumed important for exploiting large volumes of data in an efficient manner, and this was confirmed in the pilot studies on the analysis of imagery data (satellite and aerial images) in Chapter 3.3. As access to large volumes of such data increases, one of the challenges is to provide users with information on the complex processes needed to correctly and efficiently exploit them, including when machine learning is to be called upon. The Generic Pipeline developed during the Project aimed to provide some of this information[67].

Going forward, statistical organisations are encouraged to continue advancing their current machine learning developments towards their operationalisation, and to do so while they continue to collaborate and share with others. Their development could be broadened to other areas of interest to those organisations[68], particularly for business needs that are labour intensive, stable over time and offer considerable data to train the models.

Figure 6.1. Keys to Accepting Machine Learning



Robust Performance over Time

The pilot studies in Chapter 3 focused on assessing the added value of different machine learning algorithms and identifying the best model (algorithm and parameters), based on the available data. As stated before, there are still many challenges in bringing a demonstrated machine learning solution into production. It is just as important that, the machine learning solution not only continues to perform as well, but for its performance to increase as it "learns" and adapts to the new data. During the pilot studies, the questions of when to update or refresh the machine learning models, how frequently and how to proceed were raised.

The central element in putting in place and maintaining an efficient machine learning solution is the data used for training, not only at the start when determining the initial model and its parameters, but throughout the use of the machine learning solution. Another key element is the data used for evaluation, that is needed to assess not only how the machine learning model performs, but the entire operation, which usually includes some clerical operations. This data must be independent from the training data. These data are essential, but they usually come at a significant cost, and must respect

67 https://statswiki.unece.org/download/attachments/285216428/ML_WP1_Imagery_UNECE.pdf?version=1&modificationDate=1605171593842&api=v2

68 See Other applications of Machine Learning for some examples (https://statswiki.unece.org/display/ML/Other+applications+of+Machine+Learning)

certain characteristics (e.g., collection of ground truth data, texts classified by subject matter experts).

Respect for Ethical and Legal Considerations

The ethical issues that arise from use of machine learning are important considerations when using this technology within statistical organisations, as stressed by Statistics Netherlands' "Fair Algorithms in Context": "*machine learning has become more powerful over the past decade, sparking an expansion of new applications. Some of these applications fall within the social domain, in which models based on data profiles can have a significant impact on the life of individuals. To prevent unwanted discrimination in these models, different methods have been proposed within the field of algorithmic fairness*"[69].

Going forward, these are important issues that should be addressed as future work in collaboration with other working groups looking at this issue in a broader context (e.g., statistical laws, policy, data governance)[70]. In these developments, it will be important to distinguish the issues about the data sources from the methods used to exploit them. It will also be important to focus on issues specific to official statistics, rather than the consequences of using machine learning algorithms to make decisions influencing individuals (e.g., acceptance for loan applications, medical diagnosis), as often raised in discussions of ethics in the context of other application areas of machine learning.

Development on Solid Scientific Grounds from Many Disciplines

National and international statistical organisations are responsible for producing relevant and trusted information based on sound methods and processes. When machine learning methods are developed and implemented on the same basic principles, they can go a long way towards dealing with the above-mentioned issues and encouraging their acceptance.

The scientific grounds needed to underpin statistical production processes encompasses knowledge and skills from many disciplines: subject matter domains, statistics, informatics, methodology, data science and operations. Compared to the use of traditional methodologies, implementing machine learning requires specialists within these disciplines to work even more closely together from the start (fleshing out an idea and connecting it with a business need) to the end (operationalisation).

This is particularly the case for subject matter domain, where machine learning is not just another solution that has to work for subject matter business needs, but also a solution that particularly needs subject matter knowledge to work (e.g., to create "gold standard" data sets). While the idea to use machine learning can come from a single individual (as in the case of some of the pilot studies), the development of the idea needs to involve other disciplines, notably subject matter specialists, to correctly and efficiently advance. One may count more on one or two particular disciplines, but none should be left out.

Having experts from many disciplines (e.g., data science, subject matter, IT) allows learning and sharing different perspectives and issues to consider in developing, assessing and advancing the machine learning solutions. Statistical organisations will however continue to face challenges to acquire, develop and organise the varied expertise needed to effectively and efficiently use machine learning to address their business needs. The acquisition and development of expertise (e.g., training) was the most pressing need expressed in a poll conducted during the Machine Learning Project

[69] https://www.cbs.nl/en-gb/background/2020/17/fair-algorithms-in-context
[70] See Box 5.2

webinar[71]. This issue is further discussed in the next Chapter on facilitating machine learning solutions.

[71] https://statswiki.unece.org/display/ML/HLG-MOS+ML+Project+webinar

6.2. Key Aspects for Facilitation of Machine Learning Solutions

Combination of Multi-disciplinary Skills

It is important to combine knowledge and expertise from many disciplines for the production of official statistics. This is still the case, to an even greater extent, with the proliferation of alternative data sources (big, medium or small), the demand from users wanting to exploit them, and the technologies enabling their use.

While many of these skills are present within the field of data science (a relatively new discipline), the breadth and depth of skills needed in each of these disciplines cannot be found in a single individual or a small group of them. Therefore, one of the main challenges facing statistical organisations to advance the use of machine learning in the organisations is bringing together the required skills. This can be broken down into four sub-challenges:

- Identification;
- Acquisition;
- Development; and
- Organisation.

Some of these were addressed by the Project (e.g., the Machine Learning Project Work Package 3 Report[72]). Several concrete actions, many of which are very recent, by NSOs to facilitate and expand the use of machine learning were discovered[73]. These initiatives include setting up separate divisions or organisations dedicated to data science, laboratories and internal or external forums to exchange experience and knowledge on data science and machine learning. The leaders who manage these entities may connect and interact with others on an informal basis.

Going forward, it is recommended that senior management should create a formal network to share challenges, practices, experiences and results. This network should focus on managerial issues such as corporate strategies, alignment with needs, culture change and communication. The future machine learning network and group are encouraged also connect with other groups working on different issues (e.g., organisational frameworks for collaboration, change management, building competencies, culture, communication) to exchange knowledge and insights to advance machine learning in the organisations.

Computing Infrastructure

Some machine learning models, in particular the state-of-art deep learning models that are often used for image data analysis, are computationally intensive and might require a specialised hardware (e.g., GPU, TPU). While one can utilise cloud computing services to avoid the trouble of acquiring the hardware, the use of cloud might raise security issues, for example, when transferring data or interacting with existing in-house components of the production workflow. The computing infrastructure is an important aspect to facilitating the machine learning in the organisation that should be considered in light of broader corporate-level strategies.

[72] https://statswiki.unece.org/display/ML/WP3+-+Integration
[73] See Initiatives to accelerate the integration of machine learning solutions (https://statswiki.unece.org/display/ML/Initiatives+to+accelerate+the+integration+of+machine+learning+solutions)

Research and Development

The first key aspect for the acceptance of machine learning solutions, as mentioned above, is to align them with business needs. There are different approaches to this issue that were observed in the Machine Learning Project. Some emphasised the importance of starting with a business need, moving to Research and Development (R&D), producing a prototype and then involving other functional areas such as IT. Others emphasised the importance of building machine learning experience first, through R&D, which in turn allows one to identify suitable business problems which might be solved by machine learning.

Going forward, whichever path taken to advance the use of machine learning, it should be driven, if not by a specific business need (e.g., from a single statistical program), then at least by a clear corporate-wide strategy to continuously increase its relevance, by giving access to more information of better quality in a timelier manner and potentially at a lower cost.

Figure 6.2. Keys to Facilitating Machine Learning

Sharing and Collaboration

Within the Machine Learning Project, members shared working documents, methodological and technical references, links to learning resources, and presentations at meetings. Sharing of data and machine learning code greatly facilitated and accelerated learning and experimentation by others. Many of these documents and other resources were packaged and released on the UNECE Statistics Wiki[74] to allow anyone in the official statistics community to benefit from the knowledge and materials accumulated by the Project team over two years.

Going forward, sharing and collaboration will not only be beneficial to the advancement of machine learning, but also to quickly avoid its application in areas or contexts where it does not add much value. It will be important to continue collaboration through virtual platforms, networks as well as face-to-face meetings.

Senior Management Support

The Machine Learning Project would not have existed and been successful without the engagement of many people from numerous organisations, and the support from Chief Statisticians through the HLG-MOS. Going forward, it will be essential to continue to

[74] https://statswiki.unece.org/display/ML

count on their support in order to pursue research and development within their respective organisations and with others in collaborative initiatives. In return, these initiatives must be accountable to the priorities of statistical organisations.

Engagement from All Employees

New technologies, such as machine learning or artificial intelligence, have a significant impact on the culture of an organisation. Machine learning changes what an organisation and each individual employee can do and how to do it. All of the pilot studies described in Chapter 3 used supervised machine learning methods that need essential input, notably from subject matter experts and clerical staff, who are likely to be the most impacted by the resulting changes to business processes.

The studies conducted indicate that these machine learning methods cannot totally replace the work of staff and should not be perceived as such, but rather a means of introducing partial automation or strengthening automated processes, in order to achieve better results at the same or lower costs, and to allow time for staff to focus on high value work.

As machine learning solutions demonstrate their added value in more business processes, employees in all functions and at all levels of the organisation must be encouraged to consider machine learning as a potential solution to their business needs. They should also have access to experts or a centre of expertise on machine learning, to quickly determine if their proposed use case for machine learning should be considered for further investigation.

6.3. Conclusion - Is Machine Learning a Buzz, a Must or a Bust?

In its 2018 position paper, the HLG-MOS Blue-Sky Thinking Network wrote that: "*although ML seems promising there is only limited experience with concrete applications in the CES statistical community, and some issues relating to e.g., quality and transparency of results obtained from ML still have to be solved*". At that time, one could have shortened this statement to the following question in the context of producing official statistics: "Is machine learning a buzz, a must or a bust?"

Two years later, the work of the Machine Learning Project leads it to conclude that: machine learning is not just a buzz; that it is a must where it can add value, and it should not be used where it does not (i.e., avoid it becoming a bust); but that its use still has challenges in being accepted and facilitated.

Machine learning is a must where it has proven to contribute to producing data that is more relevant, with better quality, in a faster or more cost-efficient manner, without any significant reduction to any of these dimensions. The Project showed that machine learning is advantageous in processes that are labour intensive, repetitive and stable, such as in classification and coding. Machine learning can play an important role in many applications involving large volumes of data. It is more challenging to use machine learning for processes that had a higher degree of subjectivity such as in editing and imputation.

With all new and evolving technologies comes a certain degree of resistance from different parties. Some will challenge them with scientific arguments. Others will simply resist them like most changes. The former will be convinced as long as the machine learning solutions are developed on solid scientific grounds from the different disciplines that they need, and the development of those machine learning solutions is guided by a quality framework and ethical considerations. The latter can be dealt with through clear and strong senior management support. Sharing and collaboration within and between statistical organisations are also essential to advance the use of machine learning based on lessons learned on where it adds value, where it shows promise and where it offers less value.

Reference

[1] Cao L. (2017). Data science: a comprehensive overview. ACM Computing Surveys, 50(3), 1–42.

[2] Toth, C., & Jóźków, G. (2016). Remote sensing platforms and sensors: A survey. ISPRS Journal of Photogrammetry and Remote Sensing, 22-36.

[3] Curzi, G., Modenini, D., & Tortora, P. (2020). Large Constellations of Small Satellites: A Survey of Near Future Challenges and Missions. Aerospace, 7, 133. doi:10.3390/aerospace7090133

[4] Safyan, M. (2020). Handbook of Small Satellites, Technology, Design, Manufacture, Applications, Economics and Regulation. 1057-1073. doi:10.1007/978-3-030-36308-664.

[5] Holloway, J., & Mengersen, K. (2018). Statistical Machine Learning Methods and Remote Sensing for Sustainable Development Goals: A Review. Remote Sensing, 10, 1365. doi:10.3390/rs10091365.

[6] Ferreira, B., Iten, M., & Silva, R. G. (2020). Monitoring sustainable development by means of earth observation data and machine learning: a review. Environmental Sciences Europe, 32, 120. doi:10.1186/s12302-020-00397-4.

[7] Youssef, R., Aniss, M., & Jamal, C. (2020). Machine Learning and Deep Learning in Remote Sensing and Urban Application: A Systematic Review and Meta-Analysis. Proceedings of the 4th Edition of International Conference on Geo-IT and Water Resources 2020, Geo-IT and Water Resources 2020. New York, NY, USA: Association for Computing Machinery. doi:10.1145/3399205.3399224.

[8] Eurostat (2014). Handbook on Methodology of Modern Business Statistics, CROS Portal, https://ec.europa.eu/eurostat/cros/content/handbook-methodology-modern-business-statistics_en.

[9] Australian Bureau of Statistics (2009). The ABS Data Quality Framework, https://www.abs.gov.au/websitedbs/D3310114.nsf//home/Quality:+The+ABS+Data+Quality+Framework.

[10] United Nations (2019). National Quality Assurance Frameworks Manual for Official Statistics, https://unstats.un.org/unsd/methodology/dataquality/.

[11] Eurostat (2017). European Statistics Code of Practice, https://ec.europa.eu/eurostat/web/quality/european-statistics-code-of-practice.

[12] Statistics Canada (2017). Quality Assurance Framework, 3rd edition, https://www150.statcan.gc.ca/n1/pub/12-539-x/12-539-x2019001-eng.htm.

[13] Efron, B. (2020). Prediction, estimation, and attribution. Journal of the American Statistical Association, 115(530), 636–655, doi: 10.1080/01621459.2020.1762613.

[14] Hand, D.J. (2012). Assessing the performance of classification methods. International Statistical Review, 80(3), 400–414, doi: 10.1111/j.1751-5823.2012.00183.x.

[15] Hu, l., Chen. J., Vaughan, J., Aramideh, S., Yang, H., Wang, K., Sudjiant, A., and Nair, V.N. (2021). Supervised Machine Learning Techniques: An Overview with Applications to Banking, International Statistical Review, https://doi.org/10.1111/insr.12448.

[16] Scholtus, S., and van Delden, A. (2020). On the Accuracy of Estimators Based on a Binary Classifier, discussion paper, CBS, The Hague/Heerlen.

[17] Quenouille, M.H. (1956). Notes on bias in estimation. Biometrika, 43, 353–360.

[18] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society B, 36, 111–147, doi: 10.1111/j.2517-6161.1974.tb00994.x.

[19] Efron, B. (1979). Bootstrap methods: Another look at the jackknife. The Annals of Statistics, 7(1), 1–26.

[20] Wolter, K.M. (2007). Introduction to Variance Estimation.2nd edition. Springer.

[21] Kim, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. Computational Statistics and Data Analysis, 53, 3735–3745, doi: 10.1016/j.csda.2009.04.009.

[22] Borra, S., and Di Ciaccio, A. (2010). Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods. Computational Statistics and Data Analysis, 54, 2976–2989, doi: 10.1016/j.csda.2010.03.004.

[23] Bickel, P.J., and Freedman, D.A. (1981). Some asymptotic theory for the bootstrap. The Annals of Statistics, 9(6), 1196–1217.

[24] Japkowicz, N., and Shah, M. (2011). Evaluating Learning Algorithms. Cambridge University Press.

[25] Pepe, M.S. (2003). The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford University Press.

[26] Luque, A., Carrasco, A., Martín, A., and de las Heras, A. (2019). The impact of class imbalance in classification performance metrics based on the binary confusion matrix. Pattern Recognition, 91, 216–231, doi: 10.1016/j.patcog.2019.02.023.

[27] Prasath, V.B.S., Alfeilat, H.A.A., Hassanat, A.B.A., Lasassmeh, O., Tarawneh, A.S., Alhasanat, M.B., and Salman, H.S.E. (2019). Effects of distance measure choice on K-nearest neighbour classifier performance: A review. Big Data,7(4), 221–248, doi: 10.1089/big.2018.0175.

[28] Szabo, L. (2019). Artificial intelligence is rushing into patient care—and could raise risks. Scientific American, December.

[29] European Union (2016). Regulation 2016/679: General Data Protection Regulation, Recital on Profiling, https://gdpr-info.eu/recitals/no-71/.

[30] Arrieta, B.A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115, doi:10.1016/j.inffus.2019.12.012.

[31] Vilone, G., and Longo, L. (2020). Explainable artificial intelligence: A systematic review. arXiv:2006.00093.

[32] Molnar, C. (2019). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. https://christophm.github.io/interpretable-ml-book/.

[33] Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J.M.F., and Eckersley, P. (2020). Explainable machine learning in deployment. arXiv:1909.06342.

[34] Goldstein, A., Kapelner, A., Bleich, J., and Pitkin, E. (2014). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. arXiv:1309.6392.

[35] Ribeiro, M.T., Singh, S., and Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144, doi:10.1145/2939672.2939778.

[36] Stodden, V., Seiler, J., and Ma, Z. (2018). An empirical analysis of journal policy effectiveness for computational reproducibility. PNAS,115(11), 2584–2589, doi: 10.1073/pnas.1708290115.

[37] Goodman, S., Fanelli, D., and Ioannidis, J. (2016).What does research reproducibility mean?Science Translational Medicine, 8(341), 341ps12, doi: 10.1126/scitranslmed.aaf5027.

[38] United Nations (2014). Fundamental Principles of National Official Statistics, https://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx.

[39] DiCiccio, T., and Efron, B. (1996). Bootstrap confidence intervals. Statistical Science, 11(3), 189–228.

[40] Gleser, L.J (1996). Comment on 'Bootstrap confidence intervals' by DiCiccio and Efron, Statistical Science, 11(3), 219-221.

[41] Baier, L., Fabian, J. and Stefan, S. (2019). Challenges in the deployment and operation of machine learning in practice. Presented at the 27th European Conference on Information Systems.

[42] Doshi, R., Hung, K., Liang, L. and Chiu, K. (2016). Deep learning neural networks optimization using hardware cost penalty, 2016 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1954-1957, doi: 10.1109/ISCAS.2016.7538957.

[43] Domingos, P. (1998). How to get a free lunch: a simple cost model for machine learning applications. AAAI Technical Report WS-98-16.

Machine Learning for Official Statistics

Machine Learning holds a great potential for statistical organisations. It can make the production of statistics more efficient by automating certain processes or assisting humans to carry out the processes. It also allows statistical organisations to use new types of data such as social media data and imagery.

Many national and international statistical organisations are exploring how machine learning can be used to increase the relevance and the quality of official statistics in an environment of growing demands for trusted information, rapidly developing and accessible technologies, and numerous competitors. While the specific business environments may vary depending on the country, these statistical organisations face similar types of challenges which can benefit from sharing knowledge, experiences and collaborating on developing common solutions within the broad official statistical community.

This publication presents the practical applications of machine learning in three working areas within statistical organisations and discusses their value added, challenges and lessons learned. It also includes a quality framework that could help guiding the choice of methods, challenges that arise when integrating machine learning into statistical production, and key steps for moving machine learning from the experimental stage to the production stage and concludes with key messages on advancing the use of machine learning for the production of official statistics.

This publication is based on the results from two international initiatives: the UNECE High-Level Group on Modernisation of Official Statistics (HLG-MOS) Machine Learning Project (2019-2020) and the United Kingdom's Office for National Statistics (ONS) – UNECE Machine Learning Group 2021, and approved by the HLG-MOS.