Work Package #2
Quality

Gabriele Ascari, Joep Burger, Hugh Chipman, InKyung Choi, Florian Dumpert, Fabiana Rocci, Siu-Ming Tam and Wesley Yung

Delivering insight through data for a better Canada

Outline

- Introduction
- Quality Framework for Statistical Algorithms
  - Explainability
  - Accuracy
  - Reproducibility
  - Timeliness
  - Cost effectiveness
- Summary

Introduction

- National Statistical Offices (NSOs) produce many important statistical outputs
  - They have a responsibility to ensure that the highest quality outputs are produced
- Quality frameworks exist to support quality assurance in
  - Statistical outputs
  - Processes producing these outputs
  - Organizational environment producing the outputs
- Many quality frameworks exist, such as those from
  - Australian Bureau of Statistics , Eurostat, the Office for National Statistics and Statistics Canada to name a few
- These frameworks are typically geared towards statistical outputs

Introduction

- With increased interest in machine learning methods, existing quality frameworks need to be expanded
  - Machine learning methods, for now, are typically used for intermediate outputs in the production process
- We are proposing a Quality Framework for Statistical Algorithms (QF4SA)
  - Guide official statisticians on the choice of algorithms to be used in the production process
  - We define algorithm as
    - A process or set of rules to be followed in calculation, derived from an assumed model and a predetermined set of optimisation rules, for estimation or prediction.
  - Statistical algorithms cover both *traditional* and *modern* statistical methods

Introduction

- The QF4SA proposes five quality dimensions
  - Explainability
  - Accuracy
  - Reproducibility
  - Timeliness
  - Cost Effectiveness
- As with most quality frameworks, the five dimensions have to be considered jointly
  - One may chose to place more emphasis on particular dimensions but none should be ignored

QF4SA - Explainability

- This dimension will be presented by Joep Burger shortly

Accuracy

- In the context of ML, we distinguish accuracy of a classification algorithm and the wider concept of accuracy presented here
- Accuracy of an algorithm refers to the degree to which it correctly describes the phenomena it was designed to measure
  - It is the closeness of computations or estimates to the exact or true values that the statistics were intended to measure
- A simple example of a measure of accuracy is the prediction error $(y - \hat{y})$ where $\hat{y}$ is a prediction of y based on features x obtained from a model trained on a training set A
- The prediction error for a model is estimated by the average prediction error over all units in the test set
  - Note that this prediction error is conditional on the training set A

Accuracy

- From a purely prediction point of view, minimizing the prediction error is the goal
- Drawing analogies from survey sampling, NSOs may prefer to minimize the expected prediction error
  - The predication error averaged over all possible training sets
  - Protection from an unfortunate training set – representative training data is essential
- Fortunately, the expected prediction error can be estimated using resampling methods such as cross validation
- The QF4SA is suggesting that when considering accuracy, resampling methods like cross validation be used to estimate expected prediction error
  - These methods are valid when training data is ‘representative’ of population
  - Underlines importance of having good quality training data

Reproducibility

- The ability of a researcher to duplicate the results of a prior study using the same materials as were used by the original investigator
- The QF4SA recognizes three types of reproducibility
  - **Methods reproducibility** occurs when the same results are obtained using the same methods and data
  - **Results reproducibility** occurs when *corroborating* results are obtained using the same methods but in a new study
  - **Inferential reproducibility** occurs when knowledge claims of similar strength are obtained 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.

Reproducibility

- Methods reproducibility is already embraced by most if not all NSOs
- Results reproducibility may be difficult for NSOs as new data collections are not usually possible
- Inferential reproducibility is possible in NSOs during the stage of selecting an algorithm
  - The official statistician should be reasonably satisfied that the results from the chosen analysis can be **corroborated** using similar algorithms and assumptions
- Therefore, the QF4SA recommends
  - At a minimum, methods reproducibility is implemented
  - Where possible and desirable, inferential reproducibility, limited to only the replication of the analysis using different but applicable algorithms and assumptions, should be carried out as well

Timeliness

- The time involved in producing a result from conceptualization, algorithm building, processing and production
  - Distinction should be made between timeliness in development and production. The former generally takes longer than the latter.
- Most quality frameworks define timeliness as the the length of time between the reference period and the availability of information
- Machine learning can lead to substantial gains in timelines
  - Development of processes
  - Time required to process data
  - Scalability for use with Big Data

Timeliness

- In particular, processes requiring a lot of manual work
  - Coding is a prime example
  - Development of a ML coding process can be significantly quicker than traditional methods and can process significantly more data in a shorter time period
  - Examples include coding for scanner data and image processing
- QF4SA is recommending that development and processing time be considered as well as the commonly used concept of timeliness

Cost Effectiveness

- The degree to which results are effective in relation to the costs that have been expended to obtain them
- Costs can be either fixed or ongoing

Cost component	Type	Purpose
IT infrastructure	*fixed*	Acquisition of necessary hardware and software
Cloud storage	*ongoing*	Acquisition of necessary on cloud storage space
IT maintenance	*ongoing*	Maintenance of IT infrastructure
Initial training of staff	*fixed*	Training of current staff on ML; may include hiring of new staff
Staff training	*ongoing*	Keeping staff up to date regarding new ML developments
Data acquisition	*fixed / ongoing*	Acquisition and processing of new data sources
Quality assurance	*ongoing*	Quality assurance and control

Cost Effectiveness

- We note the following
  - Traditional methods also have (or had) fixed costs
  - Most costs apply to both traditional and modern methods
- ML may have advantages due to scalability and automation
  - However, initial costs may be significant but when amortized over many years, could be reasonable
- The QF4SA is recommending that the following two aspects of cost be evaluated when considering algorithms
  - Cheaper operating costs
  - Time to recoup fixed costs

Summary

- Evaluating quality for statistical algorithms is multi-dimensional
- The proposed QF4SA presents five dimensions to help guide official statisticians when comparing statistical algorithms (ML and non-ML)
- The QF4SA is not a replacement for existing quality frameworks but is a supplement to them
- More details are available in the report