Machine Learning for Edit and Imputation procedures in Official Statistics

A brief discussion on the occasion of the virtual Workshop on ML in E&I under the HLG-MOS, on 15 October 2020.

Mark van der Loo

Statistics Netherlands

mpj.vanderloo@cbs.nl

On 15 October a working group under the HLG-MOS blue-sky thinking network presented the final report on a set of projects where Machine Learning (ML) techniques were tested in a series of proof-of-concept projects (Dumpert 2020). During the project, various forms of ML techniques were put to use to solve 'classical' data editing problems. For example, for improved error localization or for creating better models for imputation. The report, and the wiki covers an impressive amount of work and very interesting worked examples that no doubt will serve as an important reference to the E&I community.

The results may very briefly be summarized as: ML shows promise, but there is no free lunch. Projects showed that issues that are typically related to projects involving ML are also encountered here: the need for extensive training and validation of models and the loss of interpretability. On the positive side there are shorter processing times and higher consistency with data editing rules.

Taking a step back, we see that the projects have focused on an area where the official statistics industry has a decades-long history of automating and improving E&I procedures. This includes areas like error localization, imputation based on statistical models, outlier detection, and selective editing. The main themes in these previous efforts have been twofold. The first is separating domain knowledge from the technique of data processing. For example, by defining data editing rules, using deductive imputation, or user-defined data correction rules. The second theme has been an ever further refining and improving of imputation models, data correction methods, quality measurement, and selective editing. The results of these efforts have yielded a body of literature and software tools that have been used in production for years – a testament to the success of this research program.

All these efforts have one important property in common: they are almost completely based on structured information that is available in surveys or administrative data. Indeed, one may wonder after four or five decades of research, how much extra information we might expect to gain from these sources by applying refined or newer models.

On the other hand, there is one area that has largely been untouched by the Official Statistics industry. This is the part of E&I where domain experts manually update the data. Research has focused on selecting records to be treated manually, but not on supporting E&I staff in making adjustments. The main difference with automated procedures is that domain experts typically use 'fuzzy' forms of information. Examples include written reports, financial statements, articles in trade magazines and (financial) news outlets, company websites and so on.

Fuzzy information is where traditional methods cannot be used, but where ML-based methods or artificial intelligence (AI) may offer a way forward. For example, at Statistics Netherlands the analyses of statutory documents were significantly sped up by a combination of optical character recognition (an application of pre-existing ML models) and text mining techniques. Similarly, identification of entities where (local) governments have a substantial interest (financial or otherwise) were identified manually but are now mostly identified automatically applying a combination of text mining techniques and ML to written (scanned) budget statements as input.

Conclusion

The largest opportunities for ML and AI are not in replacing or improving already automated E&I procedures. After decades of research these all but completely utilize the information available to them. The E&I community will probably benefit most from projects where structured information is extracted or derived from text or image sources in order to support or replace manual work.

References

Dumpert, F (2020) Theme Report of the Editing & Imputation Group [pdf]. UNECE/HLG-MOS.