Machine Learning for Coding and Classification procedures in Official Statistics

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

Jo Edwards

Australian Bureau of Statistics

jo.edwards@abs.gov.au

Great to see such a wide variety of different techniques used. The report shows the breadth and maturity of applying a wide variety of ML techniques in this space, plus how it fits with other techniques – such as text processing. It also shows where this can go, with (US Bureau of Labor Stats) going down the NN path.

Of course, the breadth of options, in of itself, provides a challenge. Particularly to those starting out.

Questions such as: What technique to choose, where to start (is it targeting a particular use case or more generic implementation), can we just borrow yours…. (and yes, it appears in some situations you can as a starting point!)

I think this theme report and all the individual reports, provides a great map or a rough sketch of the different options and allows the new person good information to help with knowing where to start

I was interested in reading about the pre-processing text analysis techniques and whether they make a difference or not. A lot of these are currently turned off in the Intelligent Coder in the ABS, but we suspect they may make an important difference in some situations and probably not in others.

I liked the chapter on the Quality Measures, something that we have spent a fair bit of time on recently, in particular how an area can use these to determine if they are achieving what they want from the intelligent coder.

I want to spend a bit of time on implementation and benefits.

I want to start with an anecdote

Anecdote: when we started on creating a coding tool using ML methods, 10 years ago, reduced cost was the main benefit targeted, with the main classifications (think SIC, SOC) and we focused on a replacing a tool – our index-based coder – and the auto-coding part of it. We had a good method, but we stumbled on trying to articulate the quality measures and in comparing the results between the two methods. We made many mistakes.

An observation around benefits, having been involved with defining, trying to reap benefits, and building systems to meet benefits over the last 5 years is that barriers to implementation are not often considered and when it comes to getting the green light for implementation, it can ‘outweigh’ or ‘outcost’ the benefit trying to be realised.

Barriers include complicated systems in place and changing one component is either very difficult or impossible (and often only have the resource to change that component), understanding the new tool and new measures and the change required to adopt the new tool and new measures (so change in culture, change in process, change in thinking – and often the areas try to 'convert' rather than 'adopt'), time series of estimates – depending on the collection, this can be a barrier in terms of 'when' to make the change, or even that some of the advantages of the new tool aren't utilised in the beginning.

Intelligent coder is now beginning to be adopted but the main benefits for those adopting is from areas where they currently code by hand (or not at all) and use subject matter specific classifications (Time Use survey, CPI).

- Hidden benefits – e.g., maintenance of index-based coders often hidden from users of the index-based coder (at least in ABS)

My other observation that has been touched upon, indirectly, is creating a ML tool versus creating a new business process that includes a ML tool, and also Quality Assurances processes, manual coding, and updating of the 'knowledge file' ready for the next cycle.

I want to leave with some questions – questions that we are asking ourselves:

- Acknowledging that there is still a 'human' part – what does this look like (e.g., human QA processes or manual coding) and how does this affect adoption in different areas within an organisation
- What should we be communicating a new tool or a new business process? How to communicate what it does to a non-technical audience?
- When do we review the method we have chosen? Have you built it such that the engine can be switched out?