Too good to be true? Machine learning in the editing process

Solveig Bjørkholt, Eva Charlotte Berner (Statistics Norway)

solveig.bjorkholt@ssb.no

Abstract

One of the most common methods of gathering data in statistics agencies is through surveys. However, since surveys are prone measurement error, the organizations also work to assure the quality of these data. Processes of quality assurance involve a validation process that is usually manual, which renders the process time consuming, less transparent and dependent on individuals. At the same time, learning from previous data and other data sources to predict the likelihood of a correct survey answer is a task that is well suited for machine learning. In this paper, I use the survey from Statistics Norway on enterprises' research and development efforts to test how machine learning algorithms can improve efficiency, transparency and transferability in the validation process. Showing that machine learning has promising results, I also discuss how these methods, despite their efficiency, might reinforce already existing bias in the validation process, and make suggestions on how to amend this.