UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS
UNECE Machine Learning for Official Statistics Workshop 2023
5-7 June 2023, Geneve Switzerland
04 April 2023

Lessons learned when applying Machine Learning in Official Statistics: Why it helps to be a survey statistician and a data scientist!

Piet Daas, Marco Puts (Statistics Netherlands)
pjh.daas@cbs.nl

Abstract

At Statistics Netherlands the use of Machine Learning (ML) to extract information from texts and images has been studied since 2016. During this period, many so-called Big Data based statistics have been developed that are either in production (online platform economy detection), very close to production (innovative companies, cybercrime, land use identification, skills extraction from job ads, and other Natural Language Processing based applications) or have ended in the experimental phase (quite a lot). As a result of these studies, we learned important lessons on the quality issues that arise when applying ML in an official statistics production environment. The most important ones are: i) begin with a thorough preliminary investigation ii) create an as good as possible (preferably representative) training and test set, iii) examine the effect of various metrics during the model's training phase, iv) prefer 'transparent' ML-algorithms, v) perform extensive manual checks, vi) focus on both the internal and external validity of the model developed, vii) include the statistics production department when results start to look promising, and viii) anticipate answering all kinds of questions raised by traditional statisticians (non-data scientists).

In the presentation and paper, these issues will be discussed in the context of the work we have conducted regarding the development of an ML-model focused on the identification of online platform companies by using website texts. Online platforms are defined by the OECD as "a digital service that facilitates interactions between two or more distinct but interdependent sets of users (whether firms or individuals) who interact through the service via the Internet." The model was trained on a set of positive and negative examples provided by experts and was used to identify the subpopulation of all (potential) online platform companies in the Business Register of Statistics Netherlands. The companies identified were subsequently checked by experts after which they received the Dutch Platform Economy survey. The first two questions in the latter survey focused on a) checking if the correct website was found for the particular company and b) checking the findings of the model regarding the correct identification of the platform status of the company. This enabled a proper external validation of the model developed. It also gave us an indication of the number of online platform companies in the Netherlands and revealed a remarkable conflict between some of the Statistics Netherlands experts' opinions and those of the companies themselves.