Proceedings 64th ISI World Statistics Congress - Ottawa, Canada ISBN: 9789073592421

CPS Paper

Machine learning for coding occupations in the Census: lessons from experiment to production

Author: Mr Lucas Malherbe

Coauthors: Lucas Malherbe, Elise Coudin, Tom Seimandi, Théo Leroy

Submission ID: 941

Reference Number: 941

Presentation File

abstracts/ottawa-2023_64103c8d2603ea93a213c9e77782f727.pdf

Brief Description

This
paper presents the approach undertaken by INSEE to select and implement classification
of the occupational variables of the annual census survey in the new national
occupational classification (PCS 2020).

The coding process will
use a combination of automatic approaches (list auto-completion and
supervised ML prediction models) and manual coding.

An ad hoc annotation
campaign conducted in 2021 provides a first set of training and testing
of the algorithms.

A two-layer neural
network algorithm (fastText embeddings of words and n-grams and classifier)
allows to achieve overall accuracy goals fixed as conditions for going into
production.

Abstract

Occupational classifications are useful tools used by statisticians, economists, sociologists to provide descriptors both accounting for similarities in job tasks and contents and similarities in economic and institutional contexts. To provide realistic social or economic analyses, occupational classification dictionaries have to be regularly updated. In 2020, a new dictionary of the French occupation classification (PCS 2020) was disseminated, accompanied with an autocompletion tool, which links perfectly a list of 5,000 jobs to their classification category. Only responses not in this list remain to be coded. INSEE has chosen not to adapt its rule-based automatic coding system set to code within the previous dictionary (PCS 2003) to the new dictionary. INSEE rather has chosen to experiment the use of machine learning techniques to perform this type of classification task for which they are expected to perform well. In 2021, a large campaign of manual labelling was conducted: around 100,000 Census job answers were labelled in PCS 2020, each twice, by two different manual coders, and a third arbitrage when required, with the aim of ensuring the quality of the training/test sets on which the algorithms would be trained/tested. A two-layer neural network algorithm (FastText embeddings of n-grams and classifier) was finally selected. The experiment suggests that the combination of the two automatic coding modes (list and supervised learning on non-lists) allows to reach or even exceed the accuracy rates of the previous system at the finest level for the current occupation, but not for the previous occupation (retired and unemployed) which has more paper slips. The combination with a part sent to manual work allows to gain some points of accuracy.
Based on these results, the integration of predicting and training tools into the Census production chain is investigated in 2022, with the aim of having the 2024 Census campaign coded in PCS 2020. This covers evaluating costs and gains of the integration of (part of) the modules developed during the experiment. This covers defining the new organization of the Census production relative to occupational coding, defining different roles and strategy to evaluate and control the quality of coding by the algorithms. This covers also keeping as optimal target another, much more ambitious, challenge which is the construction of a completely mutualized tool to code in PCS 2020 data from different sources and for different actors.