UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

UNECE Machine Learning for Official Statistics Workshop 2023

04 April 2023

5-7 June 2023, Geneva Switzerland

Creating a modern business index: Machine learning record linkage at scale

Isabela Breton, Dana Seman-Bobulska, Joanne Sheppard, Angela Collin (Office for National Statistics, United Kingdom of Great Britain and Northern Ireland)

isabela.breton@ons.gov.uk

Abstract

Situation: The Office for National Statistics has a successful product called the Inter Departmental Business Register (IDBR). This product is a statistical business register consisting of various Administrative and statistical data sources ie HMRC, ONS surveys. The product is very successful; however, it is based on legacy technology, clerically intensive, and thus limited in its data sources.

Task: Our task is to replace the IDBR with a new Business Index: a more efficient approach (utilising a much smaller production team) on a modern platform and extending the number of data sources to go beyond the existing administrative sources. Business Index is not the complete product replacement of IDBR but nevertheless a core component. Discussions of additional components of the IDBR and combining indexes are out of scope.

Product: We have used a hybrid of probabilistic record linkage and defined rules – based on subject matter expertise. Our pipeline updates the Business Index daily for births, deaths and amends. Making it an up to date (near real time) resource for researchers and policy makers - with 99.9% accuracy against truth data. The final product will be used as a stand-alone resource or in combination with other indices.

Our work demonstrates efficiency and accuracy of using a hybrid approach. The hybrid consists of:

- Scalable machine learning: Splink: a machine learning implementation of Fellegi-Sunter approach
- Rules based approach based on exploratory data analysis and subject matter experts (whom have extensive knowledge from working with the data for many years)

Development approach : What we have learnt during the production of the index will be of use for others working on record linkage. Core recommendations:

- Well developed existing approach: Splink produces weight match scores and is based on a well understood matching process.
- Dedicated subject matter (SME) expertise: is high value for identifying exceptions, edge cases and general understanding. This is facilitated by (code) notebooks and a daily log.
- Daily log: we deployed this early in our development. This enables rapid feedback from SMEs and has enabled quick product evolution.

- Utilising the existing Indexes: (in our case the IDBR) as a truth data. This sped up our development: by utilising developed / existing matching knowledge.
- Early output deliveries for QA: We regularly deliver outputs for customers’ and SMEs inspection: Speeds up improvements and corrections.