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

Classifying companies in France using machine learning

Thomas Faria, Tom Seimandi (National Institute of Statistics and Economic Studies (INSEE), France)
thomas.faria@insee.fr

Abstract

The French company registry, SIRENE, lists all companies in France and assigns them a unique identifier, the Siren number, for use by public institutions. As part of the registration process, companies must provide a description of their economic activity, which is then classified into an industry using the French classification of activities (NACE Rev. 2) by a sophisticated codification tool named Sicore. The engine works by detecting specific patterns in the description of the company's economic activity, which then guides the process down a decision tree-like structure. As a result, the final NACE code is determined based on the path taken through the expert rules. However, this classification method has limitations, such as a low automatic classification rate of 66% and a difficult maintenance process.

To address these limitations, this work presents an experiment that applies machine learning methods to classify economic activities based on the descriptions provided by companies in SIRENE. More precisely, we test methods based on embeddings of the activity descriptions, which are expected to perform well on such classification tasks, ranging from a simple linear classifier on top of a word embedding layer to more complex transformer models. Models were trained using a dataset of more than 8 million rows of data, which was both labeled by Sicore and manually.

The results of this experiment showed that machine learning classifiers are suitable candidates to replace Sicore, significantly reducing the need for manual classification and offering easier maintenance through regular re-training. As a result of these promising results, a machine learning model has been deployed in production for SIRENE.

We propose to present our work through a 15-20 min presentation that focuses on the performance and deployment of our ML classifier in a production environment. We will highlight the results of our experiment and discuss the challenges faced during deployment, including optimization, versioning, and maintenance, which are crucial aspects of MLOps. The successful transition of our machine learning model to production had a substantial impact, enhancing the productivity of INSEE agents and reducing their workload significantly.