 ISBN: 9789073592421

IPS Paper

Business sector classification and beyond using machine learning

Author: Mr Alejandro Morales

Submission ID: 1141

Reference Number: 1141

Presentation File

abstracts/ottawa-2023_2efefe3aafb45a921e16f65adb6af0f8.pdf

Brief Description

The objective of this presentation is to summarize the automation works done using Data Science and Machine Learning to identify the economic activity developed by entities.

Abstract

The main goal of the work done is to use Data Science and Machine Learning to identify the economic activity developed by entities. This must include business classification of economic activities (NACE), as well as determining whether if their sector is financial or not.
The first step has been to identify the economic activity Holding Company / Head Office or not Holding Company / not Head Office. To this effect, diverse data sources of accounting information obtained from the Central Balance Sheet Data Office of Banco de España have been integrated and several supervised machine learning models have been selected to compete. For this, it has been essential to have supervised information, due to the fact that these models learn from previously labelled data. Once done, it has been chosen the best performing model and interpretation of features have been performed, to have a better understanding of the model, from a business perspective. After other iterations for its improvement, a model with a reduced number of variables and with no overfitting has been chosen.
Secondly, similar accounting data have been used to assign Financial or non-Financial institutional sector. In order to do so, other different variables are important to classify these companies, and the classification problem is statistically more difficult. However, with a good election of the model and a new method to select the features, good results have been reached.
Next steps include using Balance of Payments data to determine Financial or non-Financial sector. The objective of this work is to provide an early classification of entities, as this data is available sooner than balance sheet data. However, the accounting information is much scarcer and therefore contextual data (Entity name, address, etc.) will needed, and possibly NLP methods and transformers will be required.