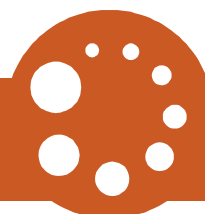




The Data Innovation Project "NOGAuto"

Management summary

EXPERIMENTAL STATISTICS



Neuchâtel, 2021

Published by: Federal Statistical Office (FSO)
Information: cindia.ducsfez@bfs.admin.ch, tel. +41 58 463 63 37
Editors: Lorenz Helbling (lorenz.helbling@bfs.admin.ch)
Topic: 00 Statistical Basis and Overviews
Original text: English
Translation: Section URD

Layout concept: Section DIAM
Downloads: www.statistics.ch
Copyright: FSO, Neuchâtel 2021
Reproduction with mention of source authorised
(except for commercial purposes)

The Data Innovation Project "NOGAuto"

Management summary

This document describes how the “NOGAuto” machine learning project was conducted during the ten months of internship at the Swiss Federal Statistical Office (FSO) by an university intern as part of its Master work in Science of Business Analytics at the Geneva School of Economics and Management (GSEM) of the University of Geneva.

In 2017, the FSO launched its data innovation strategy to stay ahead of technological developments. The “NOGAuto” project, which has been submitted by the Business Register Data section (URD) and was one of the main five projects, has been retained in line with FSO's data innovation strategy with the goal to augment and/or complement the existing basic official statistical production at the office.

According to the essential element of statistical production of the General Classification of Economic Activities (NOGA) at the FSO, the “NOGAuto” project has been realised to automate the coding of the business economic activity. This prediction system aims to automate the assignment or, more precisely, the prediction of economic activity codes to enterprises, based initially on the information already available within the FSO (e.g., survey data, Commercial Register descriptions, explanatory notes for classifications) and using "supervised machine learning" methods in the interest of discharging enterprises and to support coding.

Through the “NOGAuto” project, all the FSO statistical users could benefit from:

- a “tool” which automatically verifies the NOGA codes of enterprises;
- a standardisation of the codification;
- a minimisation of the interpretation factor;
- an optimisation of the time spent on codification;
- a quality improvement of the NOGA coding and consequently of all enterprise statistics.

In the first part of the document, the Master thesis highlights the importance of applying a well-defined process, where the product goes through three phases (Prototype, Minimal Viable Product and Production) and a proper understanding of the business context. In a second part, the paper describes the techniques of data mining (e.g. text cleaning, etc.) and of feature engineering (e.g. word embedding, etc.) that have been used for preparing the data for the modeling part, where machine learning methods (e.g. gradient boosting machine) were applied. Finally, in the “lessons learnt” chapter, some of the highlights of the project will be discussed such as its presentation at the European New Techniques & Technologies for Statistics (NTTS) 2021 conference as well as what can be improved for the future of the project.