Synthesis Report

Deployment and integration of an artificial intelligence project: from exploratory to scale-up

Author:
Paul Géhin

Referent:
Benjamin Girault
Tutors:
Yves-Laurent Bénichou
Geoffroy Wyckaert

September 8, 2021

Each year, approximately 8 million individual ballots are collected during the census. Of these, 1.7 million are selected for advanced statistical processing. We particularly seek to know their workplace and occupation. The fields are textual, unguided, short, subject to high variability and often contain unreliable answers.

Actually, 44% of workplaces and 88% of occupations are automatically encoded. If automatic coding fails, teams of coders assigned to the census take over the coding using a web application. This activity occupies about 70 FTEs for 6 months. At the end of the year, 10% of workplaces remain uncoded due to insufficient data.

The automatic coding of the activity is based on the search for the workplace. Indeed, the activity declarations are not very reliable.

In case of doubt or more uncertain coding (less good assessment of the relevance of the coding), the person's activity report is used.

A first hackaton highlighted the value of introducing tools such as ElasticSearch to search for textual similarities between individual newsletters and companies.

Following the hackaton, a project demonstrated this interest, its results showing that that it was possible to move the correct result into the top 5 in about 60% of cases (40 to 75% depending on the accuracy of the geocoding).

These first attempts have demonstrated the feasibility of improving the current solution. The main issues were to improve the top 1 and to have an automatic coding decision for automatic coding decisions.

This report aims to propose an Artificial Intelligence (AI) solution to detect non-codable data and find a better solution to automatically encode workplaces. Several machine learning solutions were tested. And a solution using Siamese neural networks and transformer headings was selected.

Although every year about 270,000 ballots are processed for "nothing", the main objective is to get as much information as possible. We will therefore privilege the precision of the model to its recall. This detection is a classification problem with two very unbalanced classes. Indeed, only a little less than 2% of the test dataset is non-codable.

The focus here is on the processing that consists of encoding the variables related to the respondent's workplace. We use a Siamese model, with a simple distance : 1 if the company and the ballot match, 0 otherwise. We also use a filtering using Elasticsearch based on geocoding and proximity of the fields. At last, we compute - using another mlpTransformer model - the occupation. To improve the results, we combine these several scores into a final score :

$$final\_score = distance + \alpha \times occupation\_score + \beta \times elasticsearch\_score$$

An optimization of ElasticSearch queries increased the performance of the model. Without this optimization, the marginal gain of these queries for the metamodel was almost zero.

A classifier - XGBoost - was trained in order to decide whether to code automatically or not the top 1. With the same precision as the actual coding model, this model can encoded automatically approximately 51% of workplaces.

Using data from the 2020 census, an evaluation was conducted with eight regional branches of the Insee. The goal was to test recommendations of the new model in real conditions. This evaluation started in May and ended in July.

The rate of automatically coded ballots has increased.

With a confidence level of 85% for the coding of the Siret and the non-codable. We have a accuracy similar to the MCA/Sicore algorithm but with a higher automatic coding rate (from 45% to 51%).

The new application provides better recommendations to managers to speed up manual coding.

Managers processed almost 57k in the estimated time to process 50k.

The manager chose the first proposal in 54% of the cases and one of the proposals in 73% of the cases.

We do not have all returns of those involved but the preliminary results are encouraging.

The involvement of users from the Ile-de-France Regional Directorate in the project was appreciated.

They had the opportunity to share the difficulties they encountered on a daily basis and their needs needs of the application.

The test phase in the Regional Directorates was well anticipated, with the possibility for division heads to include the work in their agents' work plans.

Soliciting the SeRN to present the test phase to pilot users enabled a presentation closer to the realities and expectations of the business.

We however encountered some difficulties.

Communicating the project and its potential benefits within the organization and to users proved to be complicated. Two main reasons were identified. First, it is complex to popularize the very technical elements of the project on the AI aspects. It seems that a gap between the users' idealized conception of AI and the reality of the POC's performance was created. Finally, a lack of awareness and experience of the stakeholders on innovation projects and the POC character of the developed solutions was felt.

The very short time dedicated to the acceptance tests of the application, due to delays in technical developments, did not allow for the resolution of all the problems identified before the experimentation in the Regional Directorates. The distribution of the roles and tasks of the project team members (MOA and innovation team) was not clear.

But it has allowed us to learn from it.

It is necessary to raise awareness within INSEE about innovation and the POC concept. This allows for a better adhesion to the project and a greater understanding of the current limits of the solution developed.

We noted the importance of integrating and taking into account the business dimension and challenges The importance of integrating and taking into account the business dimension and issues from the very first phases of the project, in conjunction with the project owner, to facilitate subsequent adoption of the tool.

The new tool will necessarily involve changes in the manager's job. A transition phase will therefore be necessary.