Rapport de stage de Master

Déploiement et intégration d'un projet d'intelligence artificielle : de la phase exploratoire à la mise en échelle

Auteur :
Paul Géhin

Référent :
Benjamin Girault
Tuteurs :
Yves-Laurent Bénichou
Geoffroy Wyckaert

8 septembre 2021

Résumé

Tous les ans, environ 8 millions de bulletins individuels sont recueillis lors du recensement. Parmi ceux-ci, 1.7 millions sont sélectionnés pour un traitement statistique poussé. En particulier dans ce rapport, on veut connaître leur lieu de travail et leur profession. Actuellement, cela se fait en deux parties : batch et reprise manuelle. La partie batch permet de coder environ 44% des lieux de travail et 88% des professions automatiquement. La partie reprise manuelle occupe 70 équivalents temps-plein (ETP) pendant 6 mois. Finalement, il reste environ 10% des lieux de travail non codés chaque année par manque d'informations.

Ce rapport étudie la faisabilité d'une solution de *machine learning* pour remplacer la partie batch et proposer de meilleures recommandations à la partie reprise manuelle. La solution retenue est l'association d'un réseau de neurones siamois (les bulletins et les lieux/professions rentrent par la même entrée) et d'un transformeur (pour la gestion du langage).

Sur des données de tests du recensement 2020, une augmentation de 5 points du codage automatique du lieu de travail est observée pour un rappel équivalent.

Une évaluation mobilisant 15 ETP a eu lieu de mi-mai à mi-juillet. Celle-ci donne des premiers résultats encourageants. Les retours des utilisateurs sont également globalement positifs. Ce projet devrait donc se concrétiser en production dans les prochaines années.

Table des matières

Chapitre 1

Milieu pro

1.1 Institut National de la Statistique et des Études Économiques

L'Institut National de la Statistique et des Études Économiques (Insee) est une administration publique qui dépend à la fois du ministère des Finances et des Comptes publics et du ministère de l'Économie, du Redressement productif et du Numérique. Il a été créé par la loi de Finances du 27 avril 1946.

L'Insee emploie plus de 5000 personnes réparties sur l'ensemble du territoire national : il est composé d'une direction générale située à Metz et Paris, et de directions régionales.

L'Insee collecte, produit, analyse et diffuse des informations sur l'économie et la société françaises, à l'échelle nationale mais aussi régionale.

Ces informations intéressent les pouvoirs publics, les administrations, les entreprises, les chercheurs, les médias, les enseignants, les étudiants et les particuliers. Elles leur permettent d'enrichir leurs connaissances, d'effectuer des études, de faire des prévisions et de prendre des décisions. Le principal objectif de l'Insee est d'éclairer le débat économique et social.

1.2 Service National de Développement Informatique de Paris

Le Service National de Développement Informatique de Paris (SNDIP), dans lequel je travaille actuellement, assure au profit des maîtrises d'ouvrage de l'Insee principalement deux types de mission : le développement de nouvelles applications et la maintenance d'applications en production au sein de l'Institut.

La maintenance des applications (corrections, adaptations, évolutions) est pilotée par un responsable informatique d'application (RIA). La maintenance recouvre les tâches de fonctionnement (réception d'un projet qui se termine, corrections, adaptation aux changements, assistance aux utilisateurs, gestion et suivi) et des tâches d'investissement (prévention des risques, évolutions, nouveaux développements, formation des utilisateurs, formation aux nouvelles techniques, études).

Au sein du SNDIP, les applications et projets sont regroupés en domaine, qui couvrent un champ large de clientèle. En règle générale, un secteur est en relation avec plusieurs maîtrises d'ouvrage de la direction générale (DG). Cependant, les secteurs ne sont pas calqués exactement

sur les directions de la DG.

J'ai effectué mon stage au sein du SNDIP, plus précisément au domaine "Recensement de la population".

1.3 Le recensement de la population

Pourquoi un recensement ?

Le recensement de la population (RP) est une opération annuelle qui vise à dénombrer la population par communes, mais aussi à déterminer les caractéristiques de cette population, à décrire les conditions de logement et les déplacements domicile-travail.

Ces informations servent à l'État pour définir les politiques publiques nationales, mais aussi aux communes. En effet, la contribution de l'État au budget des communes, le nombre d'élus ou encore le nombre de pharmacies, dépendent du nombre d'habitants. Plus de 300 lois communales dépendent du chiffre de la population légale. Mais le recensement permet aussi à une commune d'analyser sa population pour décider des équipements collectifs et des rénovations (crèches, gymnases...). Le recensement est utile aussi pour les entreprises privées et les associations, qui peuvent étudier leurs clientèles et leur main-d'œuvre.

Chapitre 2

Service de prédiction

2.1 Contexte

Pour rappel, tous les ans, environ 8 millions de bulletins individuels (BI) sont recueillis lors du recensement. Parmi ceux-ci, 1.7 millions sont sélectionnés pour un traitement statistique poussé.

On s'intéresse ici au traitement consistant à encoder les variables liées à l'emploi du répondant.

On dispose de champs textuels, ouverts (non guidés), courts (nom et adresse de l'employeur, activité principale de l'employeur, profession du répondant, etc.) et sujets à une très forte variabilité et contenant souvent des réponses peu fiables (par exemple le nom de la franchise au lieu du franchisé).

L'objectif est de coder :

- Le numéro Siret [1] de l'établissement employant le répondant
- Le code NAF [2] de l'établissement (ou APET [3])
- Le code PCS [4] du répondant

2.1.1 Existant

L'application P7 permet de codifier et redresser les données du recensements (individus-logements-liens, création du ménage et de la famille). Elle est composé d'une partie batch pour le codage et redressement automatique, d'une IHM Recap [5] utilisée annuellement d'avril à octobre, ainsi que d'une IHM Recap Qualité utilisée tous les 3-4 ans pour mesurer la qualité des codages automatiques et manuels.

1. Système d'identification du répertoire des établissements
2. nomenclature d'activités française
3. Activité principale exercée pour l'établissement
4. Professions et catégories socioprofessionnelles
5. RECodification de l'Activité et de la Profession

P7 Batch

Le module batch est composée de :

- de la création de la base MCA [6] à partir des fichiers du répertoire Sirus [7]
- du codage
- du redressement
- d'une partie qualité et arbitrage

Le codage automatique de l'activité se base sur la recherche de l'établissement employeur. En effet, les déclarations d'activité sont peu fiables. Ainsi, on privilégiera la recherche de l'établissement employeur par le biais d'une mise en concordance automatique (MCA) qui associe aux informations du bulletin individuel (nom de l'établissement, lieu de travail notamment) l'établissement le plus concordant parmi ceux trouvés dans une extraction du répertoire Sirus. En plus de connaître précisément l'activité professionnelle de la personne, les éléments recueillis concernant l'établissement permettent de coder plus facilement la profession.

En cas de doute ou de codage plus incertain (moins bonne évaluation de la pertinence du codage), on s'appuie alors sur la déclaration de l'activité de la personne.

Ainsi, on distingue deux modes de processus automatique de codage : par la MCA uniquement ou par la MCA, appuyée de la codification automatique de l'activité par Sicore [8] (MCA + Sicore activité).

Chaque année environ 44% des Siret sont codés automatiquement.

Le codage automatique de la profession est réalisé à partir de Sicore. Le libellé de la profession principale donnée par l'individu est traduit en une modalité de la nomenclature PCS en s'appuyant sur d'autres informations issues du questionnaire ou associées à l'établissement employeur.

Chaque année environ 88% des professions sont codés automatiquement.

P7 Recap

En cas d'échec de codification automatique, en DR, des équipes de codeurs affectés au recensement reprennent le codage à l'aide de l'application web Recap. Cette activité occupe environ 70 ETP pendant 6 mois.

De la même façon que pour le codage automatique, on cherchera d'abord à retrouver l'établissement employeur. À défaut, on essayera de retrouver l'activité, selon une nomenclature (NAF ou APET) la plus détaillée possible de 5 (ou à défaut 2) caractères. Ainsi, on distingue le codage manuel de l'activité par identification de l'établissement employeur ou le codage direct de l'activité. Pour le codage de la profession, la codification recherchée est au niveau de la PCS uniquement.

Cette campagne permet de coder manuellement environ 45% des Siret.

Environ 10% de Siret ne sont pas codés tous les ans car impossibles à traiter (champs manquants ou mal renseignés)

6. Mise en Cohérence Automatique pour les codes activités
7. Système d'Immatriculation au Répertoire des Unités Statistiques
8. Système de codification automatique selon un apprentissage des pays, communes, activités et professions permettant de passer d'un libellé, e.g. pompier, à un code, e.g. 533A dans la PCS2003

2.1.2 Premières tentatives

Un premier hackaton a mis en avant l'intérêt d'introduire des outils tels qu'ElasticSearch pour rechercher des similarités textuelles entre les bulletins individuels et les entreprises.

Dans la continuité du hackaton, un projet a démontré cet intérêt, ses résultats montrant qu'il était possible de faire remonter le bon résultat dans le top 5 dans environ 60% des cas (40 à 75% selon la précision du géocodage).

Ces premières tentatives ont permis de démontrer la faisabilité de l'amélioration de la solution actuelle. Les principales problématiques ont été d'améliorer le top 1 et d'avoir un algorithme de décision du codage automatique.

2.1.3 Enjeux

1. Augmenter le taux de codage automatique
2. Détecter les BI non codables avant la reprise manuelle par Recap
3. Améliorer le process de reprise manuelle en améliorant la pertinence des suggestions

L'objectif est donc d'avoir moins de BI à reprendre manuellement et de passer moins de temps à reprendre chaque BI.

2.2 Présentation

2.2.1 Modèle de détection des non codables

L'objectif est de de créer un modèle qui prédise si un bulletin est codable ou non avant de transmettre à l'application Recap.

Malgré que chaque année environ 270k bulletins sont traités pour "rien", l'objectif principal est d'obtenir le plus d'information possible de l'enquête du recensement. On va donc privilégier la précision du modèle à son rappel.

Cette détection est un problème de classification avec deux classes très déséquilibrées. En effet, seulement un peu moins de 2% du dataset de test est non codable.

Approches testées

Plusieurs approches de machine learning ont été testées dont certaines relevant du deep learning.

Le modèle mlpTransformer a été retenu en fonction de la précision des prédictions.

	Vrai négatif	Faux positif	Faux négatif	Vrai positif	**Rappel**	**Précision**
Régression Logistique	771862	2036	47260	8675	0.155091	0.809915
NaiveBaye	722081	51716	29352	26684	**0.476194**	0.340357
Xgboost	768584	5314	43286	12649	0.226137	0.704170
mlpTransformer	771704	2194	45617	10318	0.184464	**0.824648**

TABLE 2.1 – Tableau récapitulatif des solutions

Modèle mlpTransformer - en théorie

Réseaux de Neurones

Les réseaux de neurones font partis du machine learning et sont au coeur du deep learning. Comme leur nom l'indique, ils sont inspirés du fonctionnement des neurones du cerveau humain. Ils imitent ainsi la manière dont les neurones propage le signal de proche en proche.

Une réseau de neurones artificiel est ainsi composé de couches de neurones artifiels ou perceptrons. Il contient ainsi une couche d'entrée, une ou plusieurs couches cachées et une couche de sortie [9]. Pour un perceptron donné k qui reçoit n signaux de la couche précédentes $x = (x_1, \ldots, x_n)$, des poids associées $w_k = (w_{k,1}, \ldots, w_{k,n})$ et une fonction φ d'activation. On a alors : $y_k = \varphi\left(\sum_{i=1}^{n} w_{k,i} x_i\right)$
Généralement, la fonction φ est une fonction de seuil avec une valeur de seuil t_k. Ce qui donne :

$$y_k = \mathbb{1}_{\mathbb{R}^+}\left(t_k - w_k \cdot x^T\right)$$

Afin d'entrainer le modèle, il faut évaluer sa précision. Pour se faire, on peut introduire une fonction de coût. Elle est habituellement référée comme la *mean squarred error* (MSE).

Soit :

- $\forall i \in [\![1, m]\!], \hat{y_i}$ la prédiction du réseau
- $\forall i \in [\![1, m]\!], y_i$ la valeur réelle
- m la taille de l'échantillon

$$MSE = \frac{1}{2m} \sum_{i=1}^{m} \left(\hat{y_i} - y_i\right)^2$$

L'objectif est donc de minimiser cette fonction de coût en ajustant les poids et seuils. L'algorithme se renforce par descente de gradient. La MSE descend donc avec chaque exemple d'entraînement jusqu'à converger vers un minimum local.

Multilayer Perceptron Neural Network

Cette catégorie de réseaux se compose de plusieurs couches de perceptrons, généralement inter-connectées selon le principe de la propagation directe (feedforward). Chaque neurone d'une couche a des connexions dirigées vers les neurones de la couche suivante.

Ces réseaux utilisent une variété de techniques d'apprentissage, la plus populaire étant la rétropropagation du gradient (backpropagation). Dans ce cas, les valeurs de sortie sont comparées à la réponse correcte pour calculer la valeur d'une fonction d'erreur prédéfinie. Grâce

9. illustration en annexe A

à diverses techniques, l'erreur est ensuite renvoyée dans le réseau. En utilisant cette information, l'algorithme ajuste les poids de chaque connexion afin de réduire la valeur de la fonction de coût d'une petite quantité. Après avoir répété ce processus pendant un nombre suffisamment important de cycles d'apprentissage, le réseau converge généralement vers un minimum local de la MSE. Dans ce cas, on peut dire que le réseau a appris une certaine fonction cible. Pour ajuster correctement les poids, on applique une méthode générale d'optimisation non linéaire appelée descente de gradient. Pour cela, le réseau calcule la dérivée de la fonction d'erreur par rapport aux poids du réseau et modifie les poids de manière à ce que l'erreur diminue (en descendant sur la surface de la fonction de coût). Pour cette raison, la rétropropagation du gradient ne peut être appliquée que sur des réseaux ayant des fonctions d'activation différentiables.

Transformer

Un transformer est un modèle d'apprentissage profond qui adopte le mécanisme de l'attention, en pondérant de manière différentielle l'importance de chaque partie des données d'entrée. Il est principalement utilisé dans le domaine du NLP [10] et de la CV [11].

Les transformers sont conçus pour traiter des données d'entrée séquentielles pour des tâches telles que la traduction et le résumé de texte.

Cependant, les transformers ne traitent pas nécessairement les données dans l'ordre. Au contraire, le mécanisme d'attention fournit un contexte pour toute position dans la séquence d'entrée. Par exemple, si les données d'entrée sont une phrase, le transformer n'a pas besoin de traiter le début de la phrase avant la fin. Au contraire, il identifie le contexte qui confère un sens à chaque mot de la phrase. Cette caractéristique permet une plus grande parallélisation que les réseaux de neurones et réduit donc les temps d'apprentissage.

Les transformers sont le modèle de choix pour les problèmes de NLP. La parallélisation supplémentaire de l'apprentissage permet de s'entraîner sur des ensembles de données plus importants que ce qui était possible auparavant. Cela a conduit au développement de systèmes pré-entraînés tels que BERT [12] et GPT [13] qui ont été entraînés avec de grands ensembles de données linguistiques et peuvent être affinés pour des tâches spécifiques.

Le transformer adopte une architecture encodeur-décodeur. L'encodeur est constitué de couches d'encodage qui traitent l'entrée de manière itérative, une couche après l'autre, tandis que le décodeur est constitué de couches de décodage qui font la même chose à la sortie de l'encodeur.

La fonction de chaque couche d'encodage est de générer des encodages qui contiennent des informations sur les parties des entrées qui sont pertinentes les unes par rapport aux autres. Elle transmet ses encodages à la couche d'encodage suivante comme entrées. Chaque couche de décodage fait l'inverse, en prenant tous les codages et en utilisant les informations contextuelles qu'ils contiennent pour générer une séquence de sortie. Pour ce faire, chaque couche d'encodage et de décodage fait appel à un mécanisme d'attention.

Pour chaque entrée, l'attention pondère la pertinence de toutes les autres entrées et en tire parti pour produire la sortie. Chaque couche de décodeur possède un mécanisme d'attention supplémentaire qui puise des informations dans les sorties des décodeurs précédents, avant que la couche de décodeur ne puise des informations dans les encodages.

Les couches d'encodage et de décodage disposent toutes deux d'un réseau neuronal feedforward pour le traitement supplémentaire des sorties, et contiennent des connexions résiduelles

10. Natural Language Processing (traitement automatique des langues)
11. computer vision
12. Bidirectional Encoder Representations from Transformers
13. Generative Pre-trained Transformer

et des étapes de normalisation des couches.

Résultat

Dans le jeu de test d'environ 270k BI non codables : on écarte environ 60k BI dont 49800 effectivement non codables.

Pour cela,

- On utilise une représentation ngrams
- L'utilisation d'embeddings fasttext améliore les résultats
- Pas d'amélioration avec des techniques de rééquilibrage de classes [14]
- On utilise un seul bloc d'encodeur Transformer

2.2.2 Modèle de codage des nomenclatures

Tous les inputs sont :

- Textuels
- Extrêmement divers car non guidés
- Complémentaires - la profession donnée peut dépendre du contexte, les nomenclatures étant très précises

Les nomenclatures sont constituées d'arbres dont les enfants sont des sous éléments de leurs parents. Pour chaque nœud, on connait sa place dans la hiérarchie de la nomenclature et possède une description textuelle associée.

La nomenclature NAF possède environ 2000 nœuds dont 35 feuilles.

La nomenclature PCS, quant à elle, possède environ 575 nœuds dont 85 feuilles.

L'idée est d'essayer de rapprocher les champs déclaratifs des descriptions des codes nomenclatures.

Choix technologique

On décide d'approcher par le paradigme des problèmes de distance learning simplifiés. Il n'y aura donc pas nécessairement besoin de filtrer à l'exécution.

Lors du training, on apprend au réseau à projeter une entrée vers un vecteur à k dimensions. Pour se faire, à chaque batch de N paires d'inputs qui se correspondent, le réseau apprend à "rapprocher" N paires de vecteurs qui se correspondent, et à éloigner $N(N-1)$ paires de vecteurs qui ne se correspondent pas. Ainsi, on peut calculer la distance cosine entre deux vecteurs prédits pour chaque élément de la paire à comparer.

Lors du runtime, on calcule le vecteur correspondant à l'entrée à coder puis la distance à l'ensemble des vecteurs connus. On peut alors choisir le code correspondant au vecteur le plus proche.

Une illustration se trouve en annexe B

14. SMOTE (Synthetic Minority Oversampling Technique)

Modèle

Le modèle retenu est un modèle Transformer.

On commence par séparer l'information du BI en diverses couches d'embeddings qui s'additionnent

- Embeddings du vocabulaire (par exemple Fasttext)
- Embeddings du champ → tous les tokens du champ reçoivent le même
- Embeddings de la position dans le champs → tous les 1er tokens de chaque champ reçoivent le même, etc

Puis on aligne les champs entre les deux types d'input. En sortie, la projection est normalisée.

Plus l'influence de la représentation choisie est riche (et le vocabulaire large), meilleurs sont les résultats. Les mots sont donc préférables aux trigrammes. Eux-même préférables aux bigrammes. On a également remarqués que l'influence des embeddings pré-entrainés (Fasttext) est non négligeable. De plus, la taille de ces embeddings doit bien être choisie. Effectivement, si elle est trop grande, l'information a tendance à trop se disperser. Et si elle est trop petite, le réseau n'arrive pas à apprendre.

Enfin, la taille de la sortie joue sur les performances. Si trop petite, le modèle ne trouve pas assez d'expressivité par rapport à la distance. Et si elle est au contraire trop grande, il y a rapidement plus de progression de performance malgré un temps de calcul et des ressources plus élevés.

Finalement, on se retrouve avec un petit modèle.

Beaucoup d'information se retrouve donc dans les embeddings.

- 1 bloc encodeur
- 3 têtes
- 1 seule couche de sortie

	MSE (training)	Top k pre optim			Top k post optim		
		Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
NAF	0.1186	37,4%	68,7%	76,7%	55,2%	72,1%	78,5%
PCS	0.1126	53,4%	77,9%	85,2%	54,0%	78,7%	85,6%

TABLE 2.2 – Taux de prédiction correcte par nomenclature

Ces résultats montrent que les post process sont pertinents pour la NAF. Principalement, leur effet est de réorganiser les similarités les plus hautes pour que le bon code remonte en première position.

En revanche, ils ne changent quasiment rien à la PCS.

2.2.3 Modèle de codage Siret

On utilise le même modèle siamois, avec une distance plus simple **1 si l'entreprise et le BI matchent, 0 sinon**.

Une illustration se situe en annexe C

Filtrage par

- Géocodage des entreprises et des BI via des API Insee (BANO, POI)
- Requêtes ElasticSearch basées sur
  - Similarité textuelle, proximité géographique pour les entreprises
  - Similarité sur l'ensemble des champs proximité géographique pour les BI déjà codés. On en prend ensuite le Siret associé.

Une fois codés (via batch ou reprise manuelle), les BI sont ajoutés dans ElasticSearch et peuvent apparaître dans les requêtes futures.

Performances : Le siret cherché est présent dans au moins un résultat de requête dans 98,5% des cas.

Meta-modèle

Pour améliorer les résultats, on combine plusieurs scores en un score final

$$score\_final = sim\_siret + \alpha \times sim\_naf + \beta \times score\_elasticsearch$$

- sim_siret : similarité entre la projection du candidat Siret et le BI
- sim_naf : score du modèle de similarité NAF si le code APET du candidat est dans le top 10 des prédictions Siret (sinon 0)
- $score\_elasticsearch$: nombre de fois où le candidat apparait dans les résultats du filtrage d'ElasticSearch
- α et β sont des coefficients optimisés par une procédure d'exploration de l'espace des valeurs (optuna [15]) sur une partie du dataset de test du training.

	MSE	Top k similarité Siret			Top k meta modèle		
	(training)	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
Siret + filtr. orig.	0.144	44%	74%	85%	69%	81%	85%
Siret + filtr. optim.	0.144	45%	75%	86%	75%	84%	86%

TABLE 2.3 – Taux de prédiction correcte par filtrage ElasticSearch

On voit ici le clair apport des scores complémentaires dans l'amélioration des résultats.

On note aussi que le top 10 ne change pas le méta modèle permet de réordonner les candidats pour optimiser le top 1.

Optimisation du filtrage

- Les performances du filtrage ont peu changé (ie. le recall du bon Siret)
- En revanche le score associé a beaucoup évolué (basé sur le nombre de fois où le bon Siret est remonté), ce qui a permis d'optimiser sa prise en compte dans le méta-modèle.

15. optuna est un logiciel d'optimisation automatique des hyperparamètres, particulièrement conçu pour l'apprentissage supervisé

Décision de codage automatique

Un classificateur est entraîné sur les top k afin de décider du codage automatique.

Pour se faire, on utilise une variable "écart" entre les scores des top 1 et top 2. Puis, on apprend à classer sur les similarités top 1 et top 2 et sur cette variable supplémentaire. Une fois le modèle entraîné, on choisit un seuil de décision permettant la précision / rappel voulus. On ne code que les prédictions dont le score est supérieur au seuil.

Le modèle utilisé est XGBoost [16] (interprétable et de performance élevée).

	Codage automatique		
Précision	Total codé	Taux non codé par erreur	Taux codé par erreur
0.8	94%	0.6%	18.5%
0.9	77%	6.9%	7.9%
0.95	64.5%	15%	3.5%

Table 2.4 – Taux de codage automatique par précision

16. Le eXtreme Gradient Boosting est un algorithme glouton qui peut rapidement s'adapter à un ensemble de données d'apprentissage. Des méthodes de régularisation sont donc utilisées pour améliorer les performances de l'algorithme en réduisant le surapprentissage.

Chapitre 3

Opération codification employeur

3.1 Introduction

3.1.1 Contexte

Afin d'évaluer les performances de l'expérimentation, il est nécessaire de faire un test de la nouvelle application et de l'algorithme de codification, via une labellisation d'un échantillon de bulletins individuels du recensement. Ce projet s'inscrit dans une perspective de gain d'efficience pour le recensement.

Ces travaux ne nécessitent pas de formation particulière s'ils sont effectués par des agents des divisions recensement habitués à travailler sur l'application Recap. Seulement un guide d'utilisation de la nouvelle application et une présentation des consignes de labellisation ont été suffisants pour lancer l'opération.

L'évaluation a durée du 19 mai au 15 juillet.

3.1.2 Objectifs

1. Avoir un retour d'expérience sur l'interface développée
2. Étudier la pertinence des échos proposés dans l'application
3. Évaluer les gains liés à la méthode d'apprentissage statistique sur les futures codifications

Objectifs stratégiques

Cette nouvelle application doit permettre une amélioration de la précision des statistiques domicile-travail.

De plus, il est envisageable que ces nouvelles données du recensement puissent enrichir le répertoire Sirene dans une demarche de "crowdsourcing".

Objectifs opérationnels

De nombreuses démonstrations d'améliorations sont attendues de cette opération. Les principales attendues sont :

- l'augmentation de la part des bulletins traités automatiquement.

- l'augmentation de la productivité des gestionnaires pour la tâche de reprise de l'établissement employeur
- l'augmentation du nombre d'échos pertinents pour les bulletins à reprendre manuellement
- l'amélioration de l'exactitude de la reprise par les gestionnaires

3.1.3 Fonctionnalités

L'opération a pour objectif de tester en conditions réelles une partie de la solution apportée.

En particulier, on s'intéresse particulièrement à la partie codification du Siret. Pour se faire, on utilise le nouveau service de prédiction ainsi qu'une nouvelle interface de reprise (modernisation de Recap).

Service de prédiction

Le service de prédiction calcule trois points :

1. le taux de confiance pour la classification en codable / non codable.
2. le top 10 des Siret
3. le taux de confiance en la classification automatique du Siret (top 1)

Interface de reprise

L'interface de reprise doit reprendre les fonctionnalités principales de Recap. Elle sera à compléter par la suite s'il est décidé d'une mise en production. Elle doit également permettre de nouvelles fonctionnalités :

- L'interface permet de mettre en pause la reprise d'un BI.
- L'interface inclut un outil de suivi des BI traités.
- L'interface doit permettre de fournir des recommandations de Siret en fonction des champs remplis manuellement.

3.2 Méthode

3.2.1 Données

Les données sont issues de l'enquête de recensement 2020.

Elles concernent les départements 92, 63, 14 et 57.

Dans un premier temps, la MOA a formulée le souhait d'inclure des données qui avaient de fortes chances d'être non codable dans ce jeu. Les champs de raison sociale, d'activité et d'adresse notamment ne sont pas remplis. Cependant, l'algorithme d'attribution des BI à reprendre a fait ressortir ces bulletins. On a donc eu dans les premières semaines beaucoup de BI jugé non codable. Par la suite, nous avons exclus ceux-ci.

3.2.2 Déroulement

La recette de l'application et un premier test avec des utilisateurs de la DR Ile-de-France ont été effectués sur un échantillon de 1k de BI. Un entretien qualitatif avec un gestionnaire et le chef de la section "Enquêtes de recensement" de la Direction Régionale d'Ile-de-France a permis de récolter les premières impressions. De même, un entretien avec la responsable et deux agents de la Division Qualité des Traitements du SeRN [1] a permis de cibler les premières améliorations de l'IHM.

L'opération s'est déroulée avec des gestionnaires de Recap dans 8 directions pilotes durant deux mois. Des données sur les résultats choisis par les gestionnaires ont été récoltées par l'application (lean time, numéro de la proposition choisie). De plus, un questionnaire en ligne a été transmis aux gestionnaires participant au test en fin d'opération.

3.2.3 Mesures / Indicateurs

Ergonomie et facilité de prise en main de l'interface

- Appréciation qualitative sur la facilité de prise en main et l'ergonomie de la nouvelle interface

Impact de l'utilisation de l'outil sur le processus métier

- Appréciation qualitative sur la pertinence des suggestions fournies par l'outil
- Évolution du numéro d'ordre de la proposition sélectionnée au cours de l'expérimentation
- Temps de traitement manuel d'un bulletin pendant l'expérimentation
- Appréciation qualitative sur l'impact de l'utilisation de l'outil sur la tâche de reprise manuelle des bulletins (facilité, rapidité, etc.)

Impacts de l'évolution induite sur les parties prenantes

- Taux de bulletins correctement repris par les gestionnaires
- Appréciation qualitative sur l'impact de l'utilisation de l'outil sur le travail des gestionnaires
- Appréciation qualitative sur l'impact de la mise en place de l'outil sur l'organisation du travail dans la direction
- Estimation quantitative des gains de productivité permis par l'outil

3.3 Résultats préliminaires

3.3.1 Résultats techniques

Augmentation du taux de bulletins codés automatiquement

Avec un seuil de codification automatique à 85% du Siret et du non codable. Nous avons une précision similaire à l'algorithme MCA/Sicore mais avec un taux de codification automatique supérieur (de 45% à 51%).

1. Service Recensement National de la Population

Recommandations aux gestionnaires afin d'accélérer le codage manuel

- Les gestionnaires ont traités presque 57k BI dans le temps estimé pour en traiter 50k
- Choix du gestionnaire parmi les proposition Siret

Top 1	Top 2	Top 5	Top 10
54%	62%	69%	73%

3.3.2 Appréciation qualitative

Ergonomie et prise en main de l'outil

L'application ressemble très fortement à celle utilisée actuellement, avec une prise en main facile ne nécessitant pas une formation particulière. Cependant des fonctionnalités non reprises seraient pourtant utiles.

Quelques bugs techniques ont été identifiés. Certains restent encore à corriger.

Pertinence des échos

Les échos manquent encore de pertinence, notamment sur des bulletins facilement codables, avec des propositions parfois incohérentes (sans rapport avec la raison sociale, la commune, ou même l'activité).

Principales fonctionnalités supplémentaires et améliorations souhaitées

Les gestionnaires aimeraient pouvoir choisir le mode de recherche à partir de l'adresse déclarée dans le questionnaire (pour le cas où la raison sociale inscrite est différente de la dénomination commerciale couramment utilisée).

La possibilité de pouvoir ajouter une base de formation pour accompagner les nouveaux arrivants à la prise en main de l'outil serait appréciée.

Les agents en charge du pilotage aimeraient pouvoir agir sur les bulletins d'un lot affecté à une Direction Régionale et de réattribuer les droits.

L'onglet de suivi des bulletins doit être amélioré avec plus d'informations et de détails.

3.3.3 Risques identifiés

Une présence de bulletins incodables dans l'échantillon test pourraient avoir un impact négatif sur la perception sur les performances de l'outil par les utilisateurs, ne comprenant pas pourquoi ces bulletins n'ont pas été écartés par l'algorithme.

On note également un risque de sorties incohérentes pour les bulletins codés automatiquement (le niveau de risque dépendra du seuil retenu).

3.4 Discussion

Succès

L'association des utilisateurs de la Direction Régionale Ile-de-France au projet a été appréciée. Ils ont eu l'occasion de partager leurs difficultés rencontrées au quotidien et besoins d'amélioration sur l'application.

La phase de tests au sein des Directions Régionales a bien été anticipée, avec la possibilité pour les chefs de division d'inscrire les travaux dans les plans de charge des agents.

La sollicitation du SeRN pour présenter la phase de tests aux utilisateurs pilotes a permis une présentation plus proche des réalités et attentes métiers.

Difficultés rencontrées

Communiquer sur le projet et ses bénéfices potentiels au sein de l'organisation et auprès des utilisateurs s'est révélé compliqué. Deux raisons principales ont été identifiées. Premièrement, il est complexe de vulgariser les éléments très techniques du projet sur les aspects d'IA. Il semble qu'un décalage entre la conception idéalisée des utilisateurs sur l'IA et la réalité des performances du POC [2] se soit créé. Enfin, un manque de sensibilisation et d'expérience des parties prenantes sur les projets d'innovation et le caractère de POC des solutions développées s'est fait ressentir.

Le temps très court dédié aux tests de recette de l'application, en raison de retards pris dans les développements techniques, n'a pas permis de régler l'ensemble des problèmes identifiés en amont de l'expérimentation au sein des Directions Régionales.

La répartition des rôles et tâches des membres de l'équipe projet (MOA et équipe innovation) a manquée de clarté.

Apprentissages

L'utilisation de l'IA ne peut pas se passer de la définition de règles métiers "en dur" dans le cadre des développements. Dans le cas contraire, un modèle performant est très difficile à trouver.

Il est nécessaire de sensibiliser au sein de l'Insee sur l'innovation et le concept de POC. Cela permet une meilleure adhésion au projet et une plus grande compréhension des limites actuelles de la solution développée.

On a noté l'importance d'intégrer et de prendre en compte dès les premières phases du projet la dimension et les enjeux métiers, en lien avec la MOA, pour faciliter ensuite l'adhésion à l'outil.

Le nouvel outil impliquera obligatoirement une évolution du métier du gestionnaire. Une phase d'accompagnement sera donc nécessaire.

2. Proof Of Concept

Remerciements

Je tiens à remercier toutes les personnes qui ont permis la bonne réalisation de ce projet.

Tout d'abord, l'équipe que j'ai rejoins sur ce projet : Yves-Laurent Bénichou et Jeremy L'Hour. Merci de m'avoir accompagné et les sessions de travail ensemble, bien que ponctuelles, furent un plaisir.

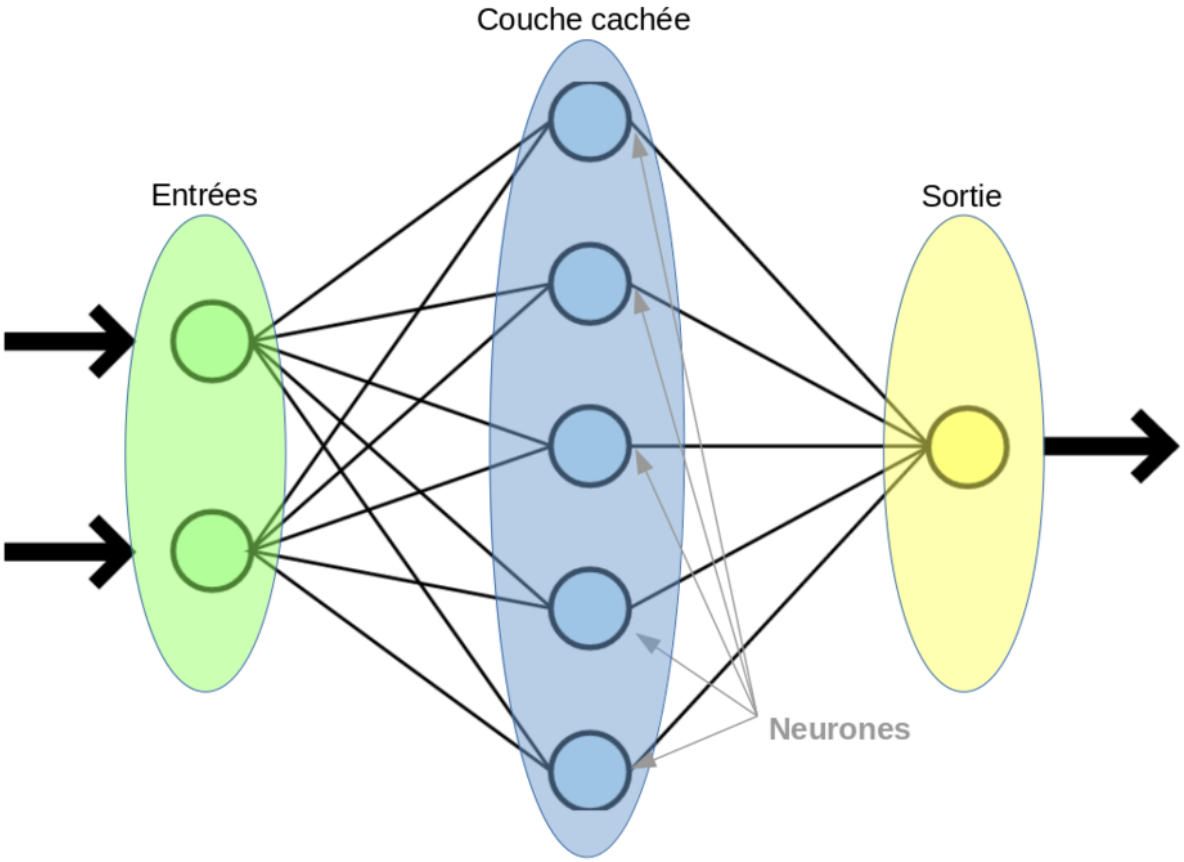
Je remercie aussi l'équipe informatique du recensement de la population et notamment Audrey Schleininger et Geoffroy Wyckaert pour leur soutien et pour avoir permis que j'ai du temps à consacrer au stage (notamment dans les dernières semaines).

Enfin, je remercie Laurence Blanc-Garin et Ludovic Vincent pour leur intérêt porté à mon master et pour m'avoir fait réaliser l'approche des différentes deadlines.

Annexes

Annexe A

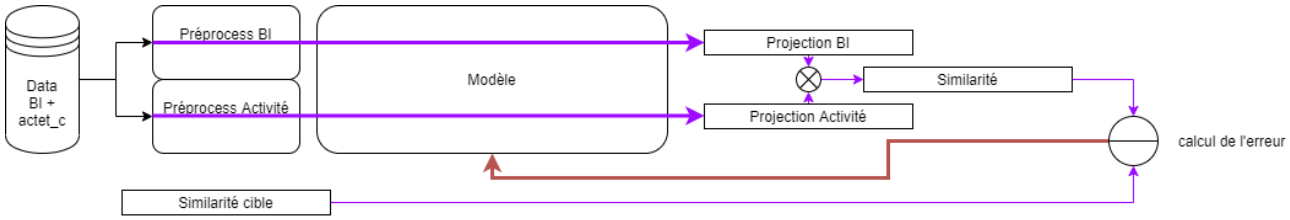
Réseau de neurone



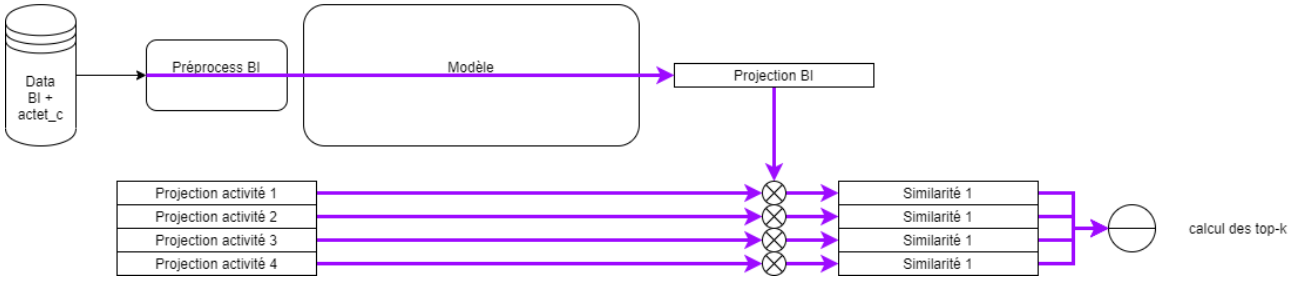
Annexe B

Modèle de codage des nomenclatures

Training



Runtime

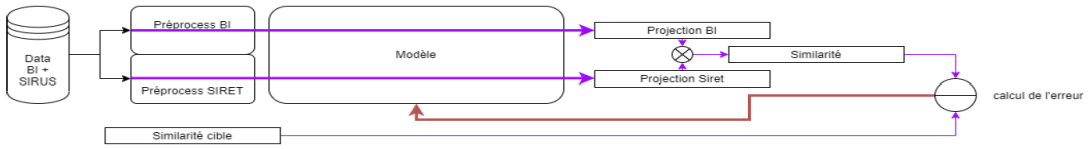


Les projections des nœuds de la nomenclature sont pré-calculées.

Annexe C

Modèle de codage Siret

Training



Runtime