Projet de statistiques appliquées
Note d'étape

Sujet 56 :
Sélection des textes applicables à un client
à partir de leur questionnaire

Élèves :
DUSSEAUX Thomas
HUBERT Grégoire
NAOARINE Axel
VIDAL Rémi

Tuteur entreprise :
ABAD Simon

Correspondant ENSAE :
DEPERSIN Jules

Rapport du 15 février 2021

1 Contexte du projet et problématique

Tennaxia propose des prestations de conseil et des logiciels pour servir les démarches RSE de ses clients. L'entreprise les aide ainsi à piloter leur conformité réglementaire en associant une expertise humaine et des outils informatiques.

Ce projet de statistiques appliquées s'insère dans le cadre du logiciel de veille à la conformité réglementaire de Tennaxia. Cette offre permet d'identifier la réglementation qui s'applique au client parmi plus de 6 000 textes de loi. L'identification est aujourd'hui permise par un questionnaire rempli par le client, qui est ensuite analysé par les consultants experts.

Or, ce travail manuel pose deux problèmes majeurs : il est source d'erreurs et prend beaucoup de temps. De plus, le nombre de textes imposés par le législateur croit fortement (multiplié par 2 en 7 ans) et est en constante évolution (10 % sont renouvelés chaque année). Ainsi, Tennaxia souhaite utiliser une solution de Machine Learning pour améliorer la performance de son travail.

L'objectif est donc de **construire un algorithme qui automatise l'identification des textes pour gagner du temps et réduire les erreurs**.

Un critère essentiel est l'absence de faux négatifs, c'est-à-dire n'oublier aucun texte qui devrait s'appliquer au client, quitte à réaliser ensuite un tri manuel par un consultant. Ce critère de performance initial a évolué lors de nos échanges avec notre référent chez Tennaxia : nous nous concentrons dorénavant dans un premier temps sur l'estimation de probabilités que chaque texte s'applique au client donné, ce qui faciliterait déjà considérablement le travail des consultants.

Organisation prospective du travail :

- Comprendre et décortiquer les différentes bases de données : clients, textes, questions, et réponses aux questions.
- Effectuer des statistiques descriptives afin d'approfondir la compréhension des bases, d'identifier des critères de validité des questionnaires (pas tous aussi complets) et différentes corrélations dans les données (relations entre l'application d'un texte et les réponses répondues par exemple).
- Développer les solutions ML sur des données de qualité (clustering sur les questions, arbres de décision...) puis étendre à l'ensemble des données si possible.
- En fonction des résultats, les optimiser : réduire le nombre de faux positifs, travailler sur les "exigences" (un degré plus fin que les textes).

2 Exploration des bases et statistiques descriptives

Une des premiers tâches auxquelles nous avons été confrontées a été de décortiquer les bases de données fournies par Tennaxia. En effet, l'articulation entre les différentes bases était assez complexe. Dix bases nous ont été envoyées, contenant des informations sur les textes appliqués, leur thème, les questions posées, etc... Nous avons concentré notre étude en priorité sur les bases de données relatives aux textes appliqués au client, et sur leurs réponses au questionnaire dans le but d'obtenir un lien entre les réponses et les types de textes appliqués.

En procédant à quelques manipulations (jointures, cleaning...), nous avons pu faire des statistiques descriptives pour mieux comprendre les bases de données et leur articulation. La figure 1 affiche, à gauche, le classement des thèmes les plus représentés parmi les 3515 textes disponibles dans la base de Tennaxia. Il est intéressant de comparer ce résultat au tableau de droite, qui indique le nombre de fois qu'un thème est « appliqué » à un client via un texte.

3515 textes dans la base de Tennaxia :

domain	
1. ENVIRONNEMENT	48.90%
3. SÉCURITÉ	31.38%
6. ENERGIE	8.88%
2. PRODUITS CHIMIQUES	4.38%
4. INSPECTION	3.78%
7. SÛRETÉ	1.82%
5. TRANSPORT DE MARCHANDISES DANGEREUSES ET AUTRES	0.63%
8. SECURITE ALIMENTAIRE	0.23%

329942 textes appliqués sur le total des clients :

domain	
3. SÉCURITÉ	44.50%
1. ENVIRONNEMENT	38.20%
6. ENERGIE	8.02%
4. INSPECTION	3.42%
2. PRODUITS CHIMIQUES	3.05%
5. TRANSPORT DE MARCHANDISES DANGEREUSES ET AUTRES	0.35%
7. SÛRETÉ	0.15%
8. SECURITE ALIMENTAIRE	0.06%

FIGURE 1 – Classement des thèmes les plus présents

Nous avons aussi étudié la répartition du nombre de textes par clients (figure 2) :

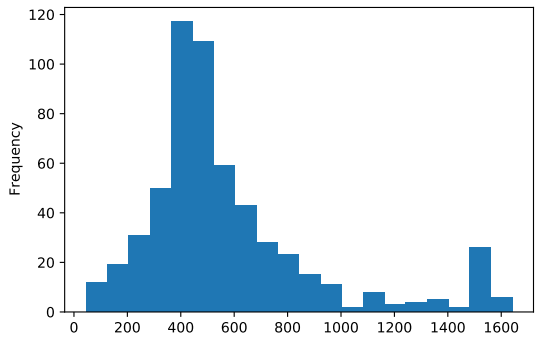

FIGURE 2 – Histogramme du nombre de textes par client

Un client a en moyenne 575 textes, pour une médiane de 483. Bien que la plupart des clients soient plus ou moins normalement centrés autour de cette valeur, on observe que certains clients ont un nombre assez élevé de textes (le max étant de 1641).

Le questionnaire soumis au client comprend 316 questions. Nous nous sommes concentrés dans un premier sur certains types de questions, à savoir les questions à choix multiples et les questions dont la réponse était « oui » ou « non ». Les questions de ce type représentent à elles seules plus de 80% des questions, le reste étant partagé entre les questions ouvertes, et les questions dont la réponse est une valeur réelle (leur analyse aurait demandé une approche différente).
En étudiant la distribution des clients selon le nombre de questions auxquelles ils ont répondu (figure 3), on se rend compte que seulement 5% des clients ont répondu à moins de 100 questions. Nous avons choisi de ne pas intégrer ces clients dans notre analyse, ainsi les questionnaires considérés comme valides sont ceux avec au moins 100 réponses.

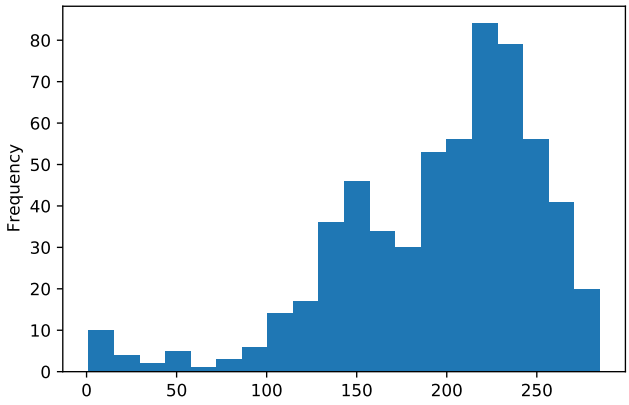

FIGURE 3 – Histogramme du nombre de questions répondues

Une première idée pour savoir dans quelle mesure la réponse à une question influence le fait ou non d'avoir un texte est d'étudier les corrélations entre ces variables. Nous avons créé une matrice de corrélation avec les indicatrices des textes en ligne (i.e une variable qui vaut 1 si le texte en question est appliqué au client et 0 sinon) et les variables indicatrices de réponse en colonne. Il faut cependant noter que pour une même question, plusieurs variables indicatrices de réponse existent (si la question 1 a comme réponse possible «1» et «2», il y aura 3 indicatrices : question1.1, question1.2 et question 1.nan). L'idée sera par la suite de concentrer notre étude sur les textes et questions qui ont un fort coefficient de corrélation.

3 Clustering à partir des réponses aux questions

Une approche retenue pour le clustering consistait à chercher des groupes de clients ayant répondu plus ou moins de la même manière au questionnaire, puis de voir si les clients d'un même cluster ont des textes similaires.

Pour ce faire, il est nécessaire d'utiliser une base de données normalisée. Un travail de cleaning a été réalisé pour créer une base où chaque ligne correspond à un client, et chaque colonne est une indicatrice de la forme "a répondu le choix a à la question i". On intègre le fait de ne pas avoir pas répondu en ajoutant une indicatrice pour la non-réponse :

	id_qu1_0	*id_qu1_1*	*id_qu1_nan*	*id_qu1_0*	*id_qu1_1*	*id_qu1_2*	*id_qu1_nan*	...
id_client_1	1	0	0	0	0	0	1	
id_client_2	0	0	1	0	0	1	0	
id_client_3	0	1	0	1	0	0	0	
...								

Sur l'exemple ci-dessus, le client 2 n'a pas répondu à la question 1 et a répondu le choix 2 (en commençant la numérotation des choix à 0) pour la deuxième question.

Pour former les clusters, on utilise la méthode *AgglomerativeClustering* de sklearn, avec la métrique ward qui minimise la variance entre les clusters. Un dendogramme a été tracé pour voir le nombre optimal de clusters :

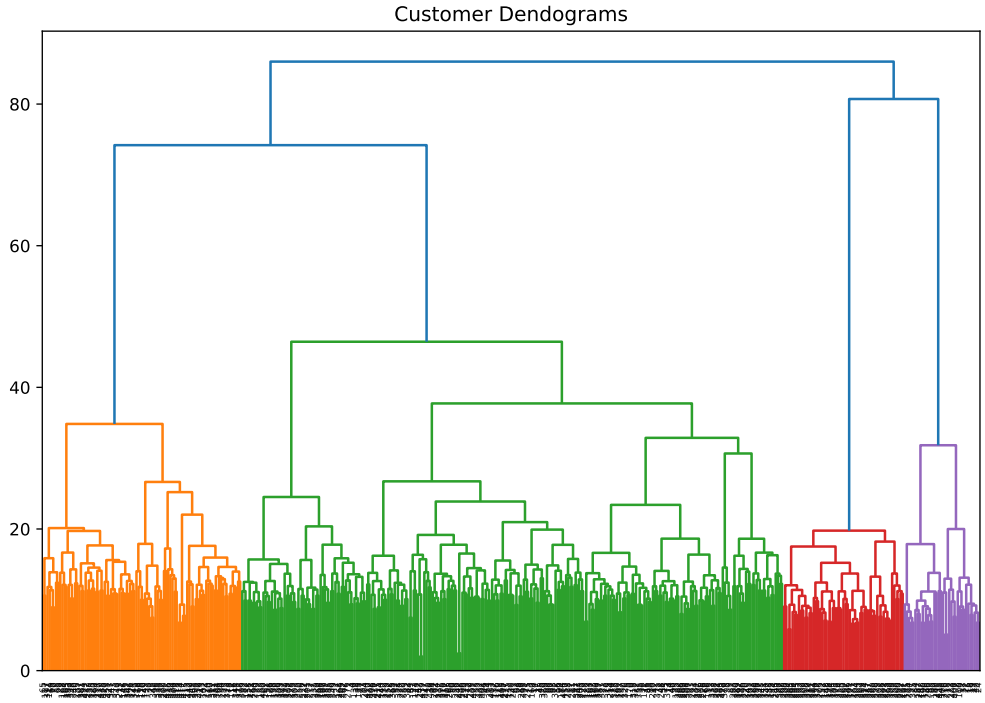

FIGURE 4 – Dendogramme

$n = 4$ clusters semble optimal. Nous avons testé $3 \leq n \leq 8$.

Analyse sur les thèmes des textes En ayant au préalable crée une table qui relie le client au texte appliqué (toujours sous forme d'indicatrice), on essaye de dégager des thèmes récurrents au sein d'un cluster. En réalité, pour $n \geq 4$, seules trois répartitions différentes se dégagent. Voici ci-dessous les répartitions pour $n = 4$:

Numéro du cluster	**0**	**1**	**2**	**3**
Nombre de clients	256	94	57	36
Environnement	39%	37%	32%	73%
Sécurité	47%	44%	54%	6%
Energie	7%	12%	6%	17%
Autres	7%	7%	8%	4%

Les répartitions correspondent bien au dendogramme. On remarque que les clusters 0 et 1 possèdent des répartitions similaires et peuvent être fusionnés. Mis à part dire que les clusters 2 et 3 accordent respectivement plus d'importance à la sécurité et l'environnement, cette analyse par les thèmes des textes semblent assez limitée. Nous avons essayé de comparer ce clustering au clustering par les textes appliqués, mais cela n'a pas été très concluant car sans surprise les trois thèmes prépondérants reviennent, avec des pourcentages plus ou moins similaires.

Nous essayons désormais de chercher certaines réponses comme élément distinctif d'un cluster.

4 Pistes pour la suite du travail

Pour répondre à la problématique, nous allons chercher à assigner les textes applicables à chaque client en se basant sur leurs caractéristiques fournies par leurs réponses au questionnaire. Dans un premier temps, nous ne prenons pas en compte la contrainte d'avoir aucun faux négatif, afin d'avoir une première idée de la pertinence et de l'efficacité de nos modèles. Dans un second temps cette contrainte sera prise en compte via des méthodes de pénalisation.
Nous avons réfléchi à deux approches pour identifier les textes applicables :

1ère approche :

Nous allons chercher à assigner pour les 3515 textes séparément leur probabilité d'applicabilité au client sachant les réponses de celui-ci au questionnaire. Pour cela, deux méthodes nous semblent envisageables, l'arbre de décision et la régression logarithmique. Les deux auront pour Target l'applicabilité du modèle (Y=1 si applicable, 0 sinon) et pour Features l'ensemble des réponses au questionnaire. Une fois ces probabilités trouvées, il nous faudra déterminer un seuil à partir duquel les textes seront ou non proposés au client. On essaiera ensuite d'améliorer cette approche via des méthodes de bagging, afin de réduire les risques d'overfitting de celle-ci.

2ème approche :

En nous basant sur le clustering effectué précédemment sur les réponses au questionnaire, nous allons déterminer n clusters (n sera le nombre de clusters optimal permettant d'obtenir les meilleurs résultats tout en prenant en compte un critère de parcimonie). Chacun de ces n clusters sera ainsi associé un paquet de texte. Ce paquet correspondra à l'union des textes appliqués aux clients qui constituent ce cluster. Chaque nouveau client sera ainsi associé à un des n clusters en fonction de ses réponses et se verra proposer les textes qui constituent le paquet de ce cluster.

Lors du second temps qui cherchera à obtenir 0 faux négatif, une difficulté déjà identifiée sera de traiter les textes qui n'apparaissent que très rarement, par exemple, 219 textes n'apparaissent que chez un unique client.