LeGes

Michael Debétaz

La data science au service de l'évaluation

Cet article offre un aperçu des sujets et des réflexions traités dans la publication « Big data, machine learning et évaluation : la data science au service des administrations », récipiendaire du prix annuel SEVAL récompensant les travaux scientifiques contribuant au développement de théories et de pratiques en évaluation. L'article s'articule autour d'extraits choisis et présente deux cas d'analyse automatisée de textes libres : la gestion des questions soumises par les utilisateurs d'un service et l'analyse des commentaires libres des participants à un congrès SEVAL.

Catégories d'articles : Reflets de la pratique

Proposition de citation : Michael Debétaz, La data science au service de l'évaluation, in : LeGes 33 (2022) 1

ISSN 2624-5477, leges.weblaw.ch, Weblaw AG, info@weblaw.ch, T +41 31 380 57 77

EDITIONS WEBLAW

Table des matières

1. Introduction sur la publication originale : « Big data, machine learning et évaluation »

[1] La publication originale est une adaptation de mon travail de MAS en évaluation réalisé à l'Université de Berne.[1] Il reprend des méthodes et des concepts utilisés dans ma collaboration avec le *Service Gestion des Données et de l'Information* de *l'Office cantonal des systèmes d'information et du numérique (OCSIN)* de l'État de Genève.

[2] Désireux de promouvoir des pratiques modernes dans le monde de l'évaluation, la forme et le contenu de mon travail ont été adaptés à un format pensée pour le web : le contenu est ainsi interactif, à l'affichage adaptatif et facilement réutilisable pour ceux qui désirent exploiter les morceaux de code utilisés pour les analyses (en langage de programmation Python, actuellement populaire en data science). La publication partage ainsi les algorithmes de résolution des problèmes auxquels ont fait face les administrations.

[3] Néanmoins, en raison de la confidentialité des données originales qui n'étaient accessibles que pour mon équipe (par exemple, les communications des usagers des services publics), les algorithmes présentés dans ce travail utilisent des données similaires publiques ou fictives. Grâce aux outils libres et gratuits permettant de reproduire ces codes, les évaluateurs sont invités à se familiariser de façon pratique à la data science et aux perspectives qu'elle peut leur offrir dans leurs pratiques professionnelles.

2. L'avènement de la data science grâce à la révolution numérique

[4] Nous vivons une époque inédite dans laquelle nous évoluons connectés, de façons multiples et instantanées, à des biens et des services qui étaient encore inaccessibles un siècle auparavant. L'avènement du big data et la démocratisation des puissances de calculs permettent de recourir désormais couramment à des modèles de machine learning dans des domaines comme l'imagerie, la santé, le marketing, ou encore l'aide à la conduite automobile.

[5] La data science offre ainsi également des opportunités nouvelles pour les professionnels de l'évaluation. Par exemple, à partir d'un système de données informatisé, il a déjà été possible dans le secteur médical d'automatiser l'analyse, les recommandations et même la prise de décision. Bien entendu, il n'est actuellement pas raisonnable de penser que ces nouvelles méthodes

[1] Michael Debétaz, Big data, machine learning et évaluation, disponible sous : https ://greval.ch/big-data-machine-learning-et-evaluation-1-introduction/.

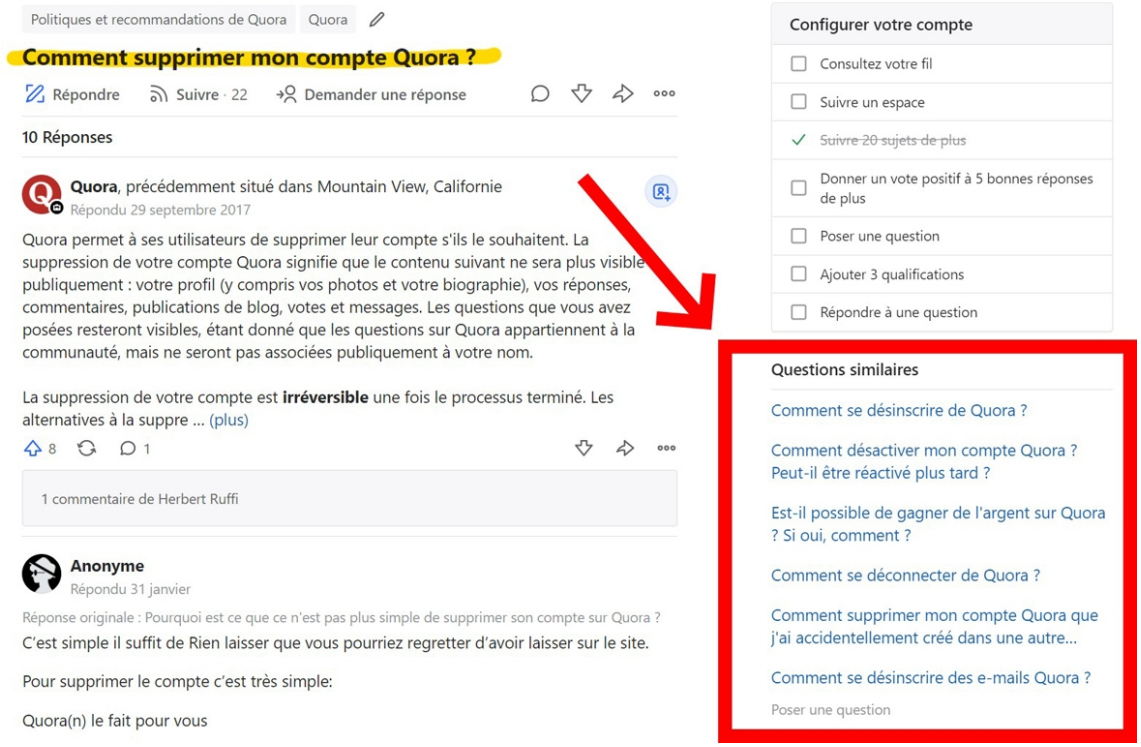
sauraient remplacer l'entier du processus évaluatif, en raison du travail humain incompressible requis en évaluation.

3. Exemple d'application : la gestion des tickets d'incident par l'administration

[6] En délivrant des services informatisés (e-démarches, poursuites, etc.), l'administration assume également la gestion des problèmes des usagers. Comme pour beaucoup de supports techniques, les problèmes se ressemblent et les équipes passent un temps important à répéter les mêmes instructions. Une façon de résoudre ce problème est de créer une foire aux questions (FAQ) à disposition des usagers. Néanmoins, ces derniers préfèrent le plus souvent ouvrir un ticket, imposant un temps d'attente ainsi qu'une charge de travail pour l'équipe de support.

[7] Une autre solution, plus ambitieuse, consiste à automatiser l'analyse de tickets pour trouver des contenus similaires et rediriger immédiatement l'utilisateur vers une réponse appropriée. Une telle application contribue grandement à l'expérience des usagers, tout en déchargeant l'administration qui reçoit une centaine de milliers de tickets par année. À défaut de pouvoir partager les données de l'administration, nous utiliserons les données publiées par Quora, un site web appartenant à une société privée qui met en relation des utilisateurs et leurs questions avec des experts du domaine concerné. Le problème de Quora est en effet similaire à celui de la gestion des tickets d'incident : rediriger efficacement l'usager et sa question vers la réponse adéquate, économisant ainsi la création d'un ticket doublon.

3.1. Exemple de suggestions de questions similaires sur quora.com



[8] Quora a ainsi publié un jeu de données de 404'209 paires de questions posées par les internautes dans des domaines divers et labellisées comme des doublons (36,92 %) ou des questions différentes (63,08 %). Les paires de questions rassemblent 537'361 textes distincts.

3.2. Exemple de quelques lignes et colonnes du jeu de données

	id	qid1	qid2	question1	question2	is_duplicate
0	0	1	2	What is the step by step guide to invest in sh...	What is the step by step guide to invest in sh...	0
1	1	3	4	What is the story of Kohinoor (Koh-i-Noor) Dia...	What would happen if the Indian government sto...	0
2	2	5	6	How can I increase the speed of my internet co...	How can Internet speed be increased by hacking...	0
3	3	7	8	Why am I mentally very lonely? How can I solve...	Find the remainder when [math]23^{24}[/math] i...	0
4	4	9	10	Which one dissolve in water quikly sugar, salt...	Which fish would survive in salt water?	0
...	...	...	...	...	...	...

Exemple d'une paire de doublons :	**Exemple d'une paire de questions différentes :**
• Do you believe there is life after death ? • Is it true that there is life after death ?	• What is the step by step guide to invest in share market in india ? • What is the step by step guide to invest in share market ?

[9] Appliquer des calculs mathématiques sur du texte demande tout d'abord de le transformer en valeurs numériques (en vecteurs plus précisément). Cette étape est réalisée mot par mot, au moyen d'un ou plusieurs modèles entraînés sur des millions de textes. Une question devient ainsi un ensemble de vecteurs-mots à 300 dimensions, dont la moyenne donne un vecteur-question également à 300 dimensions.

3.3. Quelques valeurs d'un vecteur à 300 dimensions représentant une question

```
[-4.01725955e-02  2.29274958e-01  4.54928875e-02  2.78494023e-02
  2.38825336e-01 -8.93444493e-02 -2.42939349e-02 -3.92701365e-02
  6.05317988e-02  2.15563869e+00 -5.42740464e-01 -2.80755777e-02
  5.60646690e-02  2.89211012e-02 -3.68169397e-02 -6.68809861e-02
 -4.13626023e-02  1.45753133e+00 -7.53424019e-02  2.51612859e-03
  6.11104108e-02 -8.85446072e-02 -1.69211999e-02 -7.62557704e-03
  7.98301864e-03  1.11381322e-01  8.01155269e-02 -1.35535384e-02
 -1.46775529e-01  6.18697777e-02 -1.43805463e-02  2.09816638e-02
  5.32343984e-02  9.07924026e-02  1.83719605e-01 -2.46358868e-02
  1.01320043e-01  5.29148243e-02  2.11667344e-02 -2.85674930e-01
 -4.51723300e-02 -6.28473097e-03  5.12836725e-02 -1.81825459e-01
 -5.29006757e-02  6.63057491e-02 -3.17010060e-02 -8.20088461e-02
 -4.50024009e-02 -9.95443612e-02  2.73730047e-02  8.01710505e-03
 -1.43034324e-01 -6.11312762e-02  1.35051116e-01 -1.37194812e-01
  3.52874026e-02 -8.69693905e-02  8.63466598e-03 -7.61130750e-02
```

[10] Pour préserver la simplicité du cas, les paires de questions sont rudimentairement jointes pour donner un nouveau vecteur de 600 dimensions. Ainsi, notre jeu de données consiste désormais en une foule de vecteurs catégorisés 0 (différentes) ou 1 (doublons). Le jeu est enfin prêt à entraîner un modèle de machine learning.

[11] Le modèle[2] répète enfin une centaine d'itérations de calculs pour ajuster de lui-même une prédiction efficace de catégorie pour chaque paire de questions proposée. Les prédictions ensuite sont comparées aux catégories effectives afin de fournir une métrique d'évaluation de sa justesse.

[12] L'entraînement résulte sur un modèle qui prédit correctement 85 % des paires différentes et 70 % des questions doublons. La plus faible performance sur les questions dupliquées est utile pour choisir la façon dont les recommandations sont soumises aux internautes. Par exemple, dans le cas de Quora, afficher plusieurs questions rapidement lisibles diminue la gravité d'un faux positif pour l'expérience utilisateur, comme illustré dans la capture d'écran ci-dessus.

[13] Similairement au cas de Quora, l'administration publique a également la possibilité de labelliser les messages qu'elle reçoit comme des questions doublons ou non, pour entraîner le modèle prédictif de son choix et ainsi offrir des suggestions pertinentes aux usagers qui cherchent une solution à leur problème. Si le raisonnement général est le même pour Quora comme pour les administrations, les détails de la solution ne seront jamais identiques : typiquement, des courriels demandent de nettoyer les signatures et autres textes non pertinents.

2 Le modèle utilisé ici est le classificateur XGBoost (Extreme Gradient Boosting), un algorithme de machine learning basé sur un ensemble d'arbres de décision et dont la performance a fait la renommée dans beaucoup d'applications (Chen/Guestrin 2016). L'apprentissage est paramétrable de plusieurs manières et cette étape requiert typiquement plusieurs itérations qui ne seront pas abordées ici par souci de concision.

4. Un exemple pour les évaluateurs : l'analyse qualitative automatisée

[14] Si le cas des administrations est une bonne illustration des possibilités offertes par la data science, certains évaluateurs peuvent peiner à se retrouver dans les problèmes présentés dans le travail. Dans ce souci, cette partie aborde une autre application des technologies prédictives, cette fois sur des données fréquemment présentes dans le cadre d'évaluation : les retours des participants du congrès SEVAL/GREVAL 2020. En effet, via un formulaire en ligne, les participants avaient la possibilité de laisser un commentaire libre sur chaque partie de l'événement (les conférences du matin, les ateliers de l'après-midi, puis le congrès en général).

[15] Humainement, l'analyse de textes libres est une tâche chronophage et pratiquement irréalisable au-delà d'une certaine échelle. Il est pourtant aujourd'hui possible d'utiliser des modèles de machine learning pour déterminer, par exemple, le sentiment général d'un texte. Parallèlement, la data science permet également de produire des représentations tirant pleinement parti des fonctionnalités du web, ajoutant notamment de l'interactivité aux graphiques.

[16] L'analyse des commentaires repose ici sur les algorithmes pré-entraînés[3] fournis par Google Cloud. Ceux-ci allègent considérablement le travail et s'adaptent automatiquement à la langue des textes. L'analyse des sentiments retourne, pour chaque texte, deux valeurs :

- **le score :** l'émotion globale du texte, allant de – 1 (négative) à +1 (positive).
- **la magnitude :** l'intensité émotionnelle du texte (de 0 à +∞), souvent proportionnelle à la longueur du document. Cette métrique permet notamment de distinguer les textes au sentiment neutre (un score proche de zéro et une magnitude faible) et les textes au sentiment mixte (un score proche de zéro et une magnitude élevée).

5. Exemple d'interprétation des scores et magnitudes

Sentiment	Exemples de valeurs
Clairement positif*	`"score"` : 0.8, `"magnitude"` : 3.0
Clairement négatif*	`"score"` : -0.6, `"magnitude"` : 4.0
Neutre	`"score"` : 0.1, `"magnitude"` : 0.0
Mixte	`"score"` : 0.0, `"magnitude"` : 4.0

[17] Pour éviter d'analyser des commentaires longs aux sentiments mixtes, ceux-ci sont réduits en phrases. Les résultats de l'analyse de sentiment peuvent ensuite être représentés graphiquement sur un axe à deux dimensions, comme présenté ci-dessous. Notons que, comme pour toute opération destinée à un grand nombre de données, la préparation des données n'est pas toujours parfaite et certaines coupures de commentaire ne sont pas forcément pertinentes. Aussi, la magni-

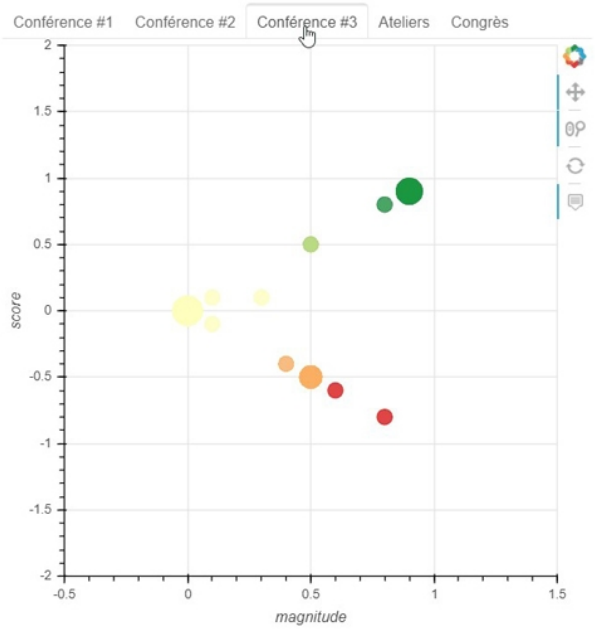
[3] Un modèle pré-entraîné est un modèle de machine learning dont les paramètres ont déjà pu être ajustés par un entraînement préalable. Ce modèle peut être utilisé, comme dans notre exemple, pour obtenir un résultat sans devoir passer par l'entraînement, ou bien de construire un nouveau modèle performant malgré un petit jeu de données (par exemple, un algorithme de détection de tumeurs cancéreuses peut reposer sur un modèle de reconnaissance de chats).

tude étant une agrégation des composantes du texte (à longueur égale, un texte avec un sentiment plus positif aura une magnitude plus élevée), la distribution en « » est un résultat attendu.

Notes sur l'utilisation du graphique interactif

- Cliquez sur les onglets pour naviguer entre les parties du congrès.
- Utilisez la molette de la souris pour zoomer/dézoomer.
- Cliquez sur le bouton sur la droite du graphique pour réinitialiser l'affichage.
- Si vous rencontrez des problèmes pour afficher le graphique, essayez de le charger dans une nouvelle page.

https://greval.ch/wp-content/uploads/2020/11/sentiments-1.html



[18] L'application rejoint le cas du traitement des questions des internautes, dans le sens où elle utilise également des modèles de machine learning pour transformer des textes en valeurs numériques. Cependant, l'analyse des retours sur le congrès SEVAL ne cherche pas à comparer les sens des phrases, mais à retourner un indicateur de sentiment représentable graphiquement. De plus, développer un graphique interactif permet également d'organiser un grand nombre d'informations sans alourdir ce qui se présente à l'écran.

[19] Une telle représentation est particulièrement utile pour un décideur confronté à un grand nombre de commentaires et qui désire un aperçu rapide des avis, ou bien qui cherche une solution économique pour ne retenir que les plus positifs ou négatifs. Notez également que, malgré un travail initial important, le code utilisé pour produire ce graphique demeure réutilisable à l'infini si appliqué à des inputs de même forme : la plus-value réside par conséquent dans les économies d'échelle réalisables lors d'évaluations menées cycliquement.

6. Conclusion

[20] Les tentatives d'intégration de nouveaux modèles statistiques et de leurs applications informatiques dans le pilotage de projets, de programmes ou de politiques publiques ouvrent des voies prometteuses pour l'évaluation. La croissance exponentielle des données ainsi que la diffusion gratuite d'outils de développement accélèrent la recherche et les innovations dans le domaine. La relation que développeront les évaluateurs avec les nouvelles technologies de l'information et de la communication demeure incertaine et le futur séparera les pionniers des suiveurs dans le développement de nouveaux standards de qualité.

Michael Debétaz, Consultant en Data Science, Développement web et Évaluation, Datafame, e-mail : michael@datafame.ch.

Quelques références bibliographiques

Chen, Tianqi / Guestrin, Carlos (2016) : *XGBoost : A Scalable Tree Boosting System.* Arxiv, pp. 785–794.

Google (2022) : *Natural Language API Basics : Sentiment analysis.* https ://cloud.google.com/natural-language/docs/basics#sentiment_analysis (accès le 17.02.2022).

Mikolov, Tomas / Chen, Kai / Chen, Greg / Dean, Jeffrey (2013) : *Efficient Estimation of Word Representations in Vector Space.* Arxiv, pp. 1–12.

Quora (2017) : *Quora Question Pairs : Can you identify question pairs that have the same intent ?* https://www.kaggle.com/c/quora-question-pairs/overview (accès le 27.09.2020).

Schwab, Klaus (2017) : *La Quatrième Révolution industrielle.* Malakoff : Duno.

Taylor, Nick Paul (2018) : *FDA approves diabetic retinopathy-detecting AI algorithm.* https ://www.fiercebiotech.com/medtech/fda-approves-diabetic-retinopathy-detecting-ai-algorithm (accès le 15.09.2020).