SEO Prédictif : interprétation d'un modèle de Machine Learning pour mieux se positionner dans les moteurs de recherche

Garaud, Damien, damien.garaud@cogniteev.com, Cogniteev, Mérignac (Orateur)

Terrasi, Vincent, vincent.terrasi@cogniteev.com, Cogniteev, Mérignac

Mondon, Élodie, elodie.mondon@cogniteev.com, Cogniteev, Mérignac

Thématique : Thème 4 "Analyse des données : Prévisions, statistiques, corrélations, Machine Learning,..."

Résumé : L'interprétation des modèles de Machine Learning est stratégique pour aider nos clients à prendre les bonnes décisions. Nous avons identifié des méthodes pour garantir la qualité des données, et surtout contrôler les performances du modèle et s'assurer que les résultats restent exploitables en innovant dans la conception de vues fonctionnelles pour rendre actionnable les résultats.

Mots clés : SEO, Machine Learning, Classification, NLP, Moteur de recherche, Facteurs de ranking, Valeurs de Shapley

1. Introduction

La présence dans les résultats de recherche, i.e. le référencement naturel, est devenue un enjeu majeur pour la visibilité des entreprises sur Internet. De plus, les algorithmes de ranking des moteurs de recherche (Google, Bing, Yandex, Baidu, …) prennent en compte de plus en plus de critères afin d'améliorer la pertinence de leurs résultats ainsi que l'expérience utilisateur·rice.

Au sein de tous ces critères, l'application Oncrawl développée par Cogniteev (un des leaders du marché) permet de récupérer et analyser les données de site Web afin d'améliorer leur SEO (Search Engine Optimization). Une des problématiques, c'est la quantité de données et de métriques SEO à analyser. Autrement dit, quels leviers une entreprise peut-elle actionner efficacement pour améliorer le ranking de certaines pages Web de leur site ?

Pour répondre à ces questions, nous proposons une approche reposant sur des algorithmes de Machine Learning ainsi que leur interprétation afin d'expliquer et comprendre les résultats de moteur de recherche et ainsi mieux guider les décisions business liées au SEO.

2. Méthodologie

La première étape de notre démarche concerne la récolte de données. Nous distinguons deux types de données : (1) les données de l'application Oncrawl et (2) les données tierces comme par exemple les liens entrants (backlinks), la position des pages sur des centaines de milliers de mot-clés… Notre *pipeline* de données nous permet rapidement de récupérer et traiter des données de plusieurs sites + mot-clés afin de réunir une soixantaine de *features*. Côté backend, nous utilisons des flows de traitement via la plate-forme Dataiku. Par ailleurs, dans le but d'obtenir une *feature* qui mesure la pertinence entre un mot-clé et le contenu d'une page Web, nous n'utilisons pas les méthodes classiques de Recherche d'Information telles que BM25 ou TF-IDF. Nous avons au contraire une approche plus "sémantique" en utilisant les dernières techniques de NLP, e.g. les *transformers*, afin de donner un score de pertinence entre le mot-clé et le contenu des pages.

Dans un second temps, nous faisons tourner un algorithme de classification en utilisant le modèle XGBoost dont l'objectif est de prédire si une paire (mot-clé, url) est bien ou mal positionnée sur les résultats de recherche. À titre d'exemple, une bonne position serait de se trouver dans les 10 premiers résultats de Google. Nous évaluons le modèle sur des données de test afin de s'assurer de ses performances (accuracy, f1-score, ROC AUC, …). Nous avons par ailleurs comparé les performances d'autres modèles de classification tels que la régression logistique ou *Random Forest*, XGBoost reste à ce jour le modèle qui donne les meilleurs résultats.

Nous utilisons ensuite SHAP (pour SHapley Additive exPlanations) afin d'interpréter et d'expliquer les prédictions du modèle entraîné. Par site, par groupe d'URLs, nous sommes capable de déterminer ce qui explique le plus une bonne ou une mauvaise position dans les résultats des moteurs de recherche. À l'aide de ces résultats, nous proposons des vues de type dashboard afin de donner aux métiers une vision claire des résultats. Ces derniers ont pu étudier et trouver des critères d'amélioration de leur référencement naturel.

3. Originalité / perspective

- L'apport des derniers modèles de NLP tels que les *transformers* pour scorer sémantiquement un mot-clé et un document (issu d'une page Web)
- L'utilisation des prédictions interprétables afin de comprendre les critères qui favorisent ou non le ranking des résultats de moteur de recherche
- Des résultats centrés sur le métier pour qu'ils soient exploitables facilement

Indiquez aussi les prochaines étapes de vos travaux.

Pour la suite, nous voyons plusieurs aspects :

- Monter à l'échelle et automatisation
- Ajouter des *features* dont le but est de réunir le plus de critères de ranking possibles
- Rendre la *User Interface* la plus claire et pédagogique possible pour des personnes non initiées à l'interprétation des modèles de Machine Learning
- Challenger notre score de pertinence sémantique avec d'autres modèles

Références

- Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In 22nd SIGKDD Conference on Knowledge Discovery and Data Mining, 2016.
- Scott M. Lundberg, Su-In Lee. A Unified Approach to Interpreting Model Predictions In *NIPS*, 2017.
- J.Guo. A Deep Relevance Matching Model for Ad-hoc Retrieval. 2016
- Leonid Boytsov and Zico Kolter. Information Retrieval: Interpretability, Effectiveness, and Efficiency Benefits. *ArXiv 2102.06815*, 2021.
- Lundberg, Scott M and Lee, Su-In. A Unified Approach to Interpreting Model Predictions. In *NIPS*, 2017.