Amélioration des explications contrefactuelles pour les recommandations à l'aide de SHAP

Jinfeng Zhong*, Elsa Negre*

*Paris-Dauphine University, PSL Research University,
CNRS UMR 7243, LAMSADE, 75016 Paris France
jinfeng.zhong@dauphine.eu
elsa.negre@lamsade.dauphine.fr

Résumé. Les explications dans les systèmes de recommandation aident les utilisateurs à mieux comprendre pourquoi de telles recommandations sont générées. Expliquer la recommandation est crucial pour renforcer la confiance et la satisfaction des utilisateurs. Étant donné que les systèmes de recommandation deviennent de plus en plus impénétrables, expliquer directement les recommandations devient parfois impossible. Les méthodes d'explication post-hoc qui n'élucident pas les mécanismes internes des systèmes de recommandation sont des approches populaires. Les méthodes d'explication post-hoc telles que SHAP génèrent des explications en construisant des modèles de substitution plus simples pour se rapprocher des modèles originaux. Cependant, l'application directe de telles méthodes pose plusieurs soucis : (1) Il se peut que les explications post-hoc ne soient pas fidèles aux modèles de recommandation originaux puisque les mécanismes internes ne sont pas élucidés; (2) Les résultats retournés par des méthodes telles que SHAP ne sont pas triviales à comprendre pour la plupart des utilisateurs, car des connaissances mathématiques sont nécessaires. Dans ce travail, nous présentons une méthode d'explication des recommandations à l'aide de SHAP qui peut générer des explications contrefactuelles facilement compréhensibles avec une forte fidélité au modèle original.

1 Introduction

L'article résumé ici, intitulé "Shap-enhanced counterfactual explanations for recommendations" (Zhong et Negre, 2022a), a été accepté et présenté à la conférence internationale *37th ACM/SIGAPP Symposium on Applied Computing (SAC '22)*. Dans Zhong et Negre (2022a), nous nous intéressons à l'explicabilité dans les systèmes de recommandation (SRs). Plus spécifiquement, nous étudions des explications contrefactuelles à l'aide de SHAP (Lundberg et Lee, 2017). Les SRs sont de plus en plus déployés pour aider les utilisateurs à trouver des articles qui les intéressent. Par conséquent, les SRs sont devenus des outils d'aide à la décision importants et omniprésents. Cependant, les techniques actuelles des SRs deviennent de plus en plus impénétrables en raison de la complexité des modèles et de la grande dimensionnalité des données avec lesquelles les modèles fonctionnent. Des modèles "boîte noire" tels que les