Visualiser des explications contrefactuelles pour des données tabulaires

Victor Guyomard[*,**], François Wallyn[****], Françoise Fessant[*], Thomas Guyet [***]
Tassadit Bouadi[**], Alexandre Termier[**]

[*] Orange, Lannion, France
[**] Univ Rennes, Inria, CNRS, IRISA, Rennes, France
[***] Inria – Lyon, Villeurbanne, France
[****] ENSAI – Rennes, France

Résumé. Dans cet article, nous présentons un outil de visualisation interactif destiné à la visualisation d'explications contrefactuelles. Une explication contrefactuelle se présente sous la forme d'une version modifiée de l'exemple à expliquer qui répond à la question : que faudrait-il changer pour obtenir une prédiction différente ? Ces explications visent à fournir aux utilisateurs des informations personnalisées et exploitables qui leur permettent de comprendre, et éventuellement contester ou améliorer les décisions automatisées. Les résultats sont affichés dans une interface où les explications contrefactuelles sont mises en évidence. Des méthodes interactives sont également fournies pour que les utilisateurs puissent explorer différentes solutions. Le fonctionnement de l'outil est illustré sur un cas d'usage de rétention client. L'outil est compatible avec n'importe quel générateur d'explications contrefactuelles et modèle de décision.

1 Introduction

L'apprentissage automatique est désormais massivement utilisé pour automatiser la prise de décision dans de nombreux domaines, et en particulier dans des domaines qui impactent notre vie quotidienne tels que la santé, le crédit ou encore la justice. Les modèles utilisés sont généralement complexes et opaques. C'est le phénomène de la « boite noire ». L'IA explicable (ou XAI) vise à limiter ce problème en fournissant un ensemble de méthodes pour qu'un utilisateur humain comprenne les facteurs qui ont motivé la décision d'un modèle. L'enjeu de l'explicabilité devient crucial que ce soit pour l'acceptation de l'IA ou le respect des réglementations existantes [1] et à venir [2]. Par exemple, si une personne se voit refuser son crédit, à la suite d'une décision algorithmique, la banque doit être en mesure de lui expliquer les raisons de cette décision. Dans un tel contexte, il pourrait être intéressant de fournir une explication sur ce que cette personne devrait changer pour influencer la décision du modèle.

Les explications contrefactuelles sont un type d'explication permettant d'expliquer la décision du modèle de prédiction à l'aide d'un exemple, proche de l'exemple à expliquer, qui

1. https://gdpr-info.eu/art-22-gdpr/
2. https://artificialintelligenceact.eu/

montre comment celui-ci devrait changer pour que sa prédiction change. L'explication fournit ainsi un retour utile à l'utilisateur qui va pouvoir identifier les différences entre son dossier et un autre, et donc également les actions à mener pour espérer faire changer la décision à l'avenir.

Un enjeu important réside dans la présentation d'un exemple contrefactuel à un utilisateur. Au delà de la génération de cet exemple contrefactuel, il est nécessaire que sa présentation soit effectivement comprise pour que l'utilisateur sache exploiter cette information.

Nous proposons dans cette démonstration un outil de visualisation d'explications contrefactuelles pour faciliter le dialogue avec un utilisateur. L'outil est destiné à des utilisateurs non spécialistes des algorithmes d'apprentissage machine. Ce peut être un expert métier ou un utilisateur final impacté par les décisions du modèle de prédiction. Par le biais de l'outil, l'utilisateur accède aux explications et peut interagir avec le système de décision. L'outil est également indépendant de l'algorithme utilisé pour générer les explications contrefactuelles. Cependant, pour illustrer ses différentes fonctionnalités, nous nous sommes appuyés sur VCNet (Guyomard et al., 2022) un modèle adapté au traitement de données tabulaires mixtes, capable de fournir simultanément prédiction et explication contrefactuelle.

2 Contexte et travaux connexes

De nombreux travaux récents traitent de l'explicabilité des modèles de décision basés sur l'apprentissage automatique. On renvoie à Molnar (2022) pour une revue des méthodes, des enjeux et des défis du domaine. On peut chercher à expliquer globalement le modèle de décision ou s'intéresser plus spécifiquement à expliquer une décision prise pour un individu en particulier.

La plupart des travaux existants sur les outils de visualisation pour l'explicabilité s'intéressent à la première catégorie, c'est-à-dire l'explication globale de modèles. Ainsi, What-If (Wexler et al., 2020) est une interface interactive permettant de visualiser les données, les décisions du modèle et d'explorer différents scénarios en modifiant les caractéristiques des variables. RuleMatrix (Ming et al., 2019) propose la visualisation interactive d'explications à base de règles. ExplainExplore (Collaris et van Wijk, 2020) quant à lui combine exploration globale et locale avec des approches basées sur l'importance des variables.

Notre focus est sur l'explication de décisions individuelles (locales) à l'aide d'exemples contrefactuels pour des données tabulaires. Miller (2019) pense qu'un tel mode d'explication est facilement appréhendé par des utilisateurs non-experts. L'explication contrefactuelle consiste à proposer un changement minimal des valeurs des caractéristiques qui permet à la prédiction de l'instance de changer pour un résultat différent. Cela peut se formaliser comme trouver une perturbation de l'exemple de sorte à changer la décision. Par exemple, trouver la plus petite perturbation des caractéristiques qui changerait la prédiction d'une demande de prêt de *rejetée* à *approuvée*. Ce nouvel exemple est appelé exemple contrefactuel ou bien contrefactuel, et le changement associé explication contrefactuelle.

Il y a encore peu de travaux dédiés à la visualisation des explications individuelles de type contrefactuelles. Gomez et al. (2020) ont proposé ViCE, un outil qui permet de générer les explications contrefactuelles et de les visualiser dans le cadre de la classification d'octroi de crédit. ViCE ne traite que les variables numériques. Une extension dédiée à l'explication globale a été proposée récemment Gomez et al. (2021). Avec DECE, Cheng et al. (2021) auto-

risent l'analyse exploratoire des décisions du modèle au niveau des instances mais également au niveau d'un groupe d'instances. Enfin, récemment, Garcia-Zanabria et al. (2022) ont proposé SDA-Vis un outil de visualisation d'explications contrefactuelles dans un contexte d'aide à l'analyse du décrochage scolaire.

Bove et al. (2022) ont réalisé une étude utilisateur pour identifier les informations visuelles que ceux-ci estimaient être les plus intéressantes à recevoir dans un contexte d'explications de décisions automatisées. Leur étude portait sur les explications individuelles par importance de variables (obtenues à l'aide de SHAP (Lundberg et Lee, 2017)). Le cas d'usage évalué concernait la prédiction du prix d'une prime d'assurance par apprentissage supervisé. Les auteurs ont montré que les utilisateurs de l'étude accordaient une forte importance à la mise en contexte et à l'interactivité de l'outil de visualisation. La mise en contexte correspond principalement à une description des variables qui sont utilisées pour la prédiction tandis que l'interaction laisse de la liberté à l'utilisateur pour explorer plus en détail chaque explication.

Nous nous sommes appuyés sur ces différents travaux pour spécifier les différentes fonctionnalités de notre outil interactif de visualisation d'explications contrefactuelles.

3 Description de l'outil

L'objectif principal de l'outil proposé est de fournir une représentation visuelle intuitive des explications contrefactuelles fournies par un algorithme d'explicabilité (ici l'algorithme VCNet). Plus précisément notre objectif est de montrer, pour une instance donnée, 1) quelles caractéristiques doivent être modifiées pour que la décision du modèle change, 2) quelle doit être l'amplitude du changement et 3) de permettre l'exploration de solutions alternatives.

3.1 Génération des explications contrefactuelles

La plupart des méthodes d'explications contrefactuelles sont basées sur la perturbation de l'instance originale grâce à l'optimisation d'une fonction de coût Wachter et al. (2018). Selon les propriétés souhaitées pour l'explication on rajoute des contraintes dans le processus d'optimisation sous la forme de termes supplémentaires dans la fonction de coût. Par exemple, on peut souhaiter un contrefactuel le plus proche possible de l'exemple à expliquer, avec le moins de variables perturbées possible, actionnable (où seules certaines variables peuvent être perturbées) ou encore réaliste. Pour une revue récente de ces approches, on peut se reporter à Guidotti (2022).

L'algorithme de génération d'explications contrefactuelles que nous avons utilisé dans le cadre de cet article est décrit dans Guyomard et al. (2022). L'originalité du modèle (VCNet) est qu'il apprend simultanément à prédire et à générer une explication associée à la prédiction. Un des intérêts de l'approche est qu'elle assure un meilleur alignement entre la prédiction et l'explication, et ainsi la génération de contrefactuels valides (au sens qu'ils ont bien une classe différente de la classe de l'exemple). Un autre intérêt réside dans le temps de génération des contrefactuels. Contrairement aux approches post-hoc l'explication est ici générée de façon immédiate une fois le modèle entraîné. De plus, VCNet est un modèle à base de réseaux de neurones de type autoencodeur conditionnel variationnel permettant la génération de contrefactuels réalistes.

3.2 Description de l'interface

La figure 1 illustre la présentation d'une explication, pour une instance donnée, pour un cas d'usage de désabonnement client (appelé *chrun*). Le problème de décision auquel on s'intéresse est un problème de classification à deux classes (*churn*/*non churn*). Une instance est décrite par 20 variables. On trouve différentes informations sur la partie haute de l'interface concernant l'exemple et sa prédiction. La partie centrale de l'interface est dédiée aux informations relatives aux valeurs des variables : la valeur actuelle pour l'exemple et la valeur proposée pour le contrefactuel. La partie basse de l'interface est dédiée à la traduction sous forme textuelle de l'explication. Un code couleur permet l'identification de chacune des classes (ici orange pour un *churner*, et vert pour un *non churner*). Plus précisément :

- En ① on trouve les informations concernant la classe prédite par le modèle de décision pour l'individu à expliquer (ici le client est étiqueté comme *churner*) ainsi que la probabilité avec laquelle le client a été prédit dans la classe (69%).
- En ② on trouve les informations concernant la classe prédite pour le contrefactuel correspondant et la probabilité associée (prédiction de *churn* à 21% i.e. *non churn* à 79%). On observe que la classe du contrefactuel est, comme attendu, bien différente de celle de l'individu à expliquer.
- Le camembert en ③ présente un résumé des changements entre l'exemple et son contrefactuel. Il indique la proportion de variables modifiées. Le camembert est interactif. En cliquant sur celui ci on peut naviguer entre les variables modifiées par le contrefactuel et celles qui sont restées inchangées.
- Les exemples à analyser sont sélectionnés individuellement grâce au menu déroulant ④ à l'aide de leur identifiant.
- La partie centrale du graphique s'intéresse aux variables de l'exemple qui ont fait l'objet d'une modification afin d'obtenir le contrefactuel. Dans l'exemple présenté, 7 variables ont été modifiées, chacune étant identifiée par son label ⑤, ⑥. Une flèche associée à chaque variable indique le sens et l'amplitude du changement dans le cas d'une variable numérique ⑤ ou la nouvelle modalité dans le cas d'une variable catégorielle ⑥. Le code couleur associé aux variables correspond à celui de la classe du contrefactuel (vert pour un *non churner* ici).
- Les différents changements de variables sont résumés sous forme textuelle ⑧.
- Une information supplémentaire concernant l'erreur de classification de l'exemple par le modèle de décision est fournie (si elle est disponible) sous la forme d'un code graphique particulier ⑨. On raye le rond ① correspondant à l'exemple quand il a été mal classé par le modèle de décision.

L'outil dispose de plusieurs autre fonctionnalités accessibles par navigation à partir de la page principale de l'interface :

- L'utilisateur peut interagir avec l'interface et demander à sélectionner un autre contrefactuel ⑦. Il est alors redirigé vers une page illustrée dans la figure 2. Plusieurs contrefactuels lui sont proposés et il peut choisir un contrefactuel selon les critères qu'il souhaite privilégier : parcimonie (le moins de changement de variables possibles) ou performance de prédiction (le score le plus faible prédit pour la classe de l'exemple). Le contrefactuel proposé par défaut est le contrefactuel qui nécessite le moins de changements. On rappelle en partie haute de l'interface, les informations liées à l'exemple en cours d'analyse.

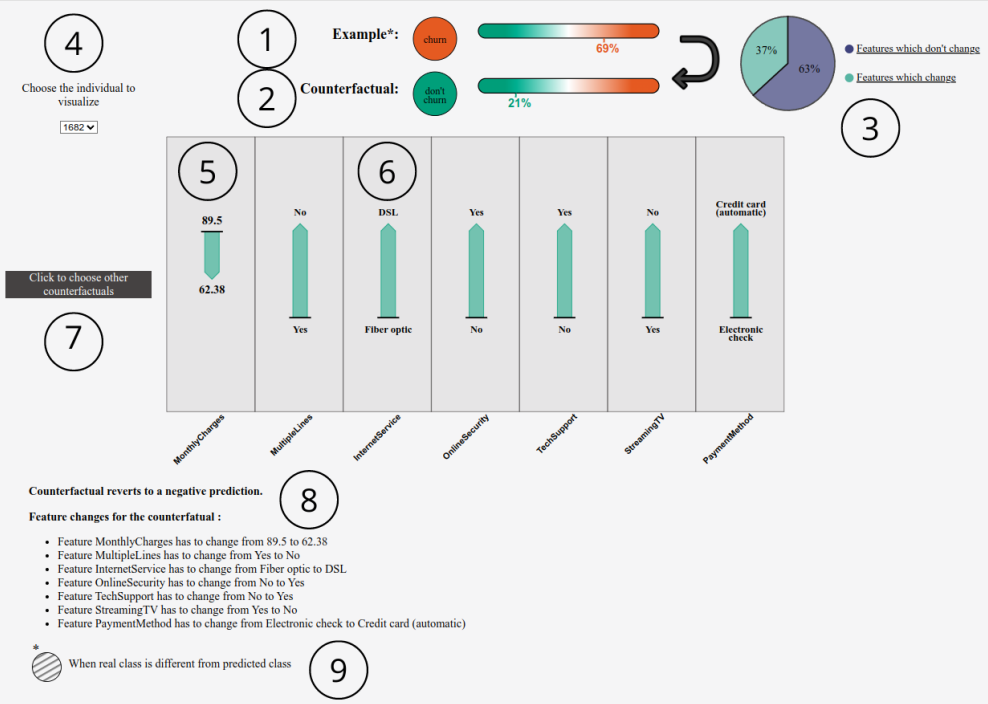

FIG. 1 – *Interface de présentation d'un exemple à expliquer et d'un contrefactuel associé.*

— Une page d'accueil qui rappelle les caractéristiques principales de l'algorithme de génération de contrefactuels, et donne une description des données analysées (caractéristiques et sémantique des variables).

3.3 Implémentation

Pour la réalisation de notre outil de visualisation, nous avons utilisé une application Flask, qui est un micro-framework de développement web en Python permettant de présenter les données et d'afficher les pages web. Les visualisations et interactions sont créées grâce à JavaScript et d3js. Nous utilisons également HTML et CSS pour créer les pages web. L'interface est compatible avec n'importe quel modèle de prédiction, ainsi qu'avec n'importe quel générateur d'explication contrefactuelle. Les données nécessaires à la visualisation sont fournies via un fichier JSON. Ce fichier doit contenir :

— les noms des variables,
— une matrice variables/instances, contenant les instances à expliquer et une autre contenant les contrefactuels,
— les probabilités de prédiction du modèle ainsi que les classes prédites, pour les instances à expliquer ainsi que pour les contrefactuels.

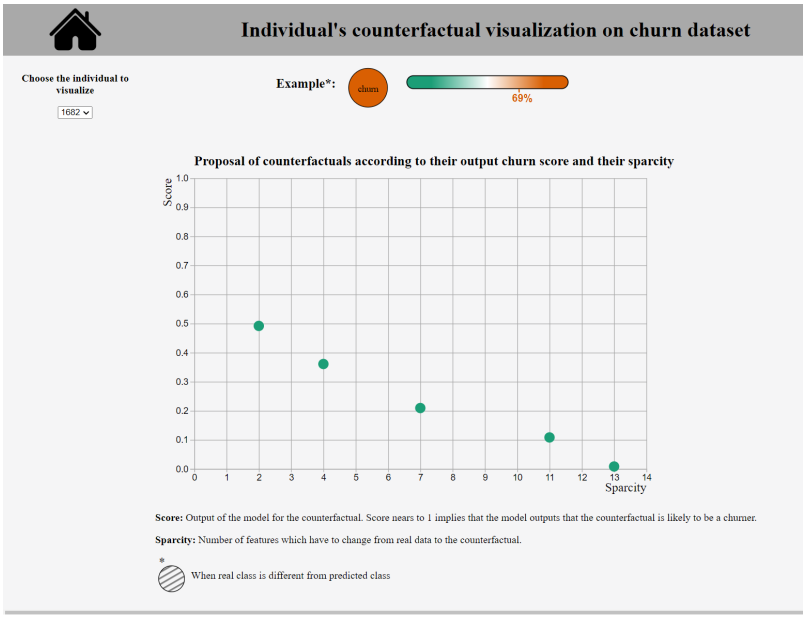

FIG. 2 – *Interface de sélection d'un contrefactuel alternatif selon les axes à privilégier (parcimonie/score de classification).*

4 Étude d'un cas d'usage

Nous illustrons l'outil sur un exemple de prédiction de la résiliation de clients d'un opérateur télécom. Le jeu de données utilisé, Telco Customer Churn [3], comprend 7 043 clients décrits par vingt variables (informations personnelles, services souscrits, type de contrat), dont la résiliation (*oui*/*non*). Il s'agit donc d'un problème de classification binaire sur des données tabulaires. Les données ont été découpé en 60% des exemples pour l'apprentissage du modèle, 20% pour la validation, et 20% pour le test.

On discute ici l'analyse de l'exemple présenté Figure 1. L'exemple correspond à un individu (*Id 1682*) qui a été étiqueté par le modèle de décision comme *churner* avec une probabilité de 69%. Le contrefactuel proposé pour l'exemple fait changer la classe de l'exemple de *churner* à *non churner* avec une probabilité de *non churn* de 79% (probabilité de *churn* à 21%). 7 variables de l'exemple initial ont été modifiées pour obtenir le contrefactuel (37% des variables). En termes d'analyse métier, les modification consistent en un changement de mode d'accès internet (passer de la fibre à l'ADSL), la suppression de certains services (streamingTV), la diminution de la facture mensuelle de 89.5$ à 62.4$, la souscription à un service de protection et de support technique. La figure 2 indique que 4 contrefactuels alternatifs sont disponibles. Un premier contrefactuel, qui propose la modification de 2 variables (diminution de la facture mensuelle de 89.5$ à 77.25$ et modification de la méthode de paiement), ramène la probabi-

3. https://www.kaggle.com/datasets/blastchar/telco-customer-churn

lité de *churn* de 69% à 49%, un second qui propose la modification de 4 variables ramène la probabilité de *churn* à 36% (suppression de plusieurs services et modification de la méthode de paiement), etc. L'expert métier a ainsi la possibilité de choisir le critère qui lui parait le meilleur entre parcimonie et score de classification.

Une autre utilisation possible de notre outil, dans un contexte métier, consisterait à observer de manière préventive les clients *non churners* et leurs contrefactuels. L'objectif étant de détecter une évolution dans les variables qui indiqueraient que le client est sur le point de *churner*.

5 Conclusion et évolutions futures

Cet article a présenté un outil de visualisation d'explications contrefactuelles. Pour chaque instance à analyser, on présente, via une interface graphique les variables descriptives qui ont été modifiées (ainsi que comment elles on été modifiées) pour obtenir le contrefactuel et faire changer la décision du modèle. L'utilisateur a la possibilité d'interagir avec l'interface pour explorer des contrefactuels alternatifs. Cet outil est dédié pour l'instant à des utilisateurs de type expert métier ou utilisateur destinataire de la décision. Il est compatible avec tout type de modèle de décision et générateur d'explications contrefactuelles. L'outil a été illustré sur un cas d'usage de rétention client.

Le travail présenté ici est une première étape dans l'objectif de doter les utilisateurs d'outils de visualisations simples et intuitifs pour l'explicabilité des modèles d'intelligence artificielle. L'outil pourrait à terme s'enrichir de différentes fonctionnalités. Par exemple, pour l'instant les interactions avec l'utilisateur sont limitées au choix d'un contrefactuel dans un ensemble possible selon des critères de parcimonie ou de performance de classification. L'utilisateur pourrait également être intéressé par une sélection des variables qui composent le contrefactuel. Un autre axe d'amélioration concerne la formalisation textuelle de l'explication qui est pour l'instant très limitée. Un travail sur l'ergonomie de l'interface serait également d'intérêt, ainsi qu'une étude utilisateur.

Références

Bove, C., J. Aigrain, M.-J. Lesot, C. Tijus, et M. Detyniecki (2022). Contextualization and exploration of local feature importance explanations to improve understanding and satisfaction of non-expert users. In *Proceedings of the 27th International Conference on Intelligent User Interfaces (IUI)*, pp. 807–819. Association for Computing Machinery.

Cheng, F., Y. Ming, et H. Qu (2021). DECE : Decision explorer with counterfactual explanations for machine learning models. *Transactions on Visualization and Computer Graphics 27*(2), 1438–1447.

Collaris, D. et J. J. van Wijk (2020). ExplainExplore : Visual exploration of machine learning explanations. In *Proceedings of the Pacific Visualization Symposium (PacificVis)*, pp. 26–35. IEEE.

Garcia-Zanabria, G., D. A. Gutierrez-Pachas, G. Camara-Chavez, J. Poco, et E. Gomez-Nieto (2022). SDA-Vis : A visualization system for student dropout analysis based on counterfactual exploration. *Applied Sciences 12*(12), 5785.

Gomez, O., S. Holter, J. Yuan, et E. Bertini (2020). ViCE : Visual counterfactual explanations for machine learning models. In *Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI)*, pp. 531—535. Association for Computing Machinery.

Gomez, O., S. Holter, J. Yuan, et E. Bertini (2021). AdViCE : Aggregated visual counterfactual explanations for machine learning model validation. In *Proceedings of the Visualization Conference (VIS)*, pp. 31–35. IEEE Computer Society.

Guidotti, R. (2022). Counterfactual explanations and how to find them : literature review and benchmarking. *Data Mining and Knowledge Discovery*, 1–55.

Guyomard, V., F. Fessant, et T. Guyet (2022). VCNet : A self-explaining model for realistic counterfactual generation. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD)*, pp. 10.

Lundberg, S. M. et S.-I. Lee (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, et R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Volume 30. Curran Associates, Inc.

Miller, T. (2019). Explanation in artificial intelligence : Insights from the social sciences. *Artificial Intelligence 267*, 1–38.

Ming, Y., H. Qu, et E. Bertini (2019). RuleMatrix : Visualizing and understanding classifiers with rules. *Transactions on Visualization and Computer Graphics 25*(1), 342–352.

Molnar, C. (2022). *Interpretable Machine Learning* (2 ed.).

Wachter, S., B. Mittelstadt, et C. Russell (2018). Counterfactual explanations without opening the black box : Automated decisions and the GDPR. *Harvard journal of law & technology 31*, 841–887.

Wexler, J., M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, et J. Wilson (2020). The What-If Tool : Interactive probing of machine learning models. *Transactions on Visualization and Computer Graphics 26*(1), 56–65.

Summary

In this paper we present an interactive visual analytics tool that exibits counterfactual explanations to evaluate model decisions. Each sample is assessed to identify the set of changes needed to flip the model's output. These explanations aim to provide end-users with personalized actionable insights with which to understand automated decisions. The functionality of the tool is demonstrated by its application to a customer retention dataset.