Interactive Visualization of Counterfactual Explanations for Tabular Data

Victor Guyomard, Françoise Fessant, Thomas Guyet, Tassadit Bouadi, Alexandre Termier

Interactive visualization of counterfactual explanations for tabular data

Victor Guyomard[1,2], Françoise Fessant[1], Thomas Guyet[3], Tassadit Bouadi[2], and Alexandre Termier[2]

[1] Orange Innovation, Lannion, France
victor.guyomard@orange.com
[2] Univ Rennes, Inria, CNRS, IRISA, Rennes, France
[3] Inria, AIstroSight, Lyon, France
[4] ENSAI, Rennes, France

Abstract. In this paper we present an interactive visualization tool that exhibits counterfactual explanations to explain model decisions. Each individual sample is assessed to identify the set of changes needed to flip the output of the model. These explanations aim to provide end-users with personalized actionable insights with which to understand automated decisions. An interactive method is also provided so that users can explore various solutions. The functionality of the tool is demonstrated by its application to a customer retention dataset. The tool is compatible with any counterfactual explanation generator and decision model for tabular data.

Keywords: Counterfactual explanation · Interactive visualisation tool.

1 Motivation

A counterfactual explanation is a modified version of an example to be explained that answers the question: what would have to change to get a different prediction? These explanations are intended to provide users with personalised and actionable information that allows them to understand, and possibly challenge or improve, automated decisions [5]. Beyond the generation of this counterfactual explanation, it is necessary that its presentation be understood so that the user knows how to exploit this information. There is still little work dedicated to the visualization of individual explanations of the counterfactual type. Gomez et al. [4] proposed ViCE, a tool that allows the generation of counterfactual explanations and visualise them as part of the credit granting classification. ViCE deals only with numerical variables. SDA-Vis [3] is another example used in a context of helping analysis of school drop-out. Bove et al. [2] were able to identify through a user study that the most interesting visual information for them were contextualisation, with a description of the variables that are used for prediction, and the interactivity of the visualisation tool which gives the user freedom to explore an explanation. Their study focused on individual explanations by feature importance, in a context of car insurance. We have built on this work to specify the functionalities of our tool in the context of counterfactual explanations.

2 Demonstrator

The tool we propose is intended for users who are not specialists in machine learning algorithms. It can be a business expert or an end user impacted by the decisions of a model. Through the tool, the user has access to explanations and can interact with the decision system. The main objective of the tool is to provide an intuitive visual representation of the counterfactual explanations provided by any algorithm. More precisely, our objective is to show, for a given instance, 1) which features must be modified for the model decision to change, 2) what the magnitude of the change must be and 3) to allow the exploration of alternative solutions. [5]

2.1 Interface description

Figure 1 displays a counterfactual explanation for a binary classification problem of customer churn (more details in use case study section below).

Various information can be found on the upper part of the interface concerning the example and its prediction. ① gives the predicted class for the individual to explain (labelled churner by the scoring model, with a probability of 69%). ② gives the predicted class for the proposed counterfactual (labelled non churner as with a probability of churn of 21%). A colour code allows the identification of each class (here orange for a churner, and green for a non churner). As expected the class of the counterfactual is different from those of the observed individual. The pie chart ③ shows the proportion of variables in the individual that have been modified to generate the counterfactual. By clicking on it, one can navigate between the modified variables and those that have remained unchanged. A drop-down menu ④ allows you to select the individuals.

The central part of the interface is dedicated to the modified variables between the individual and the counterfactual. Here 7 variables were changed. The direction of the change is specified by an arrow with its magnitude in the case of a numerical variable ⑤. In the case of a categorical variable, the change is indicated by an upward arrow pointing to the new modality ⑥.

Additionally, the variable changes are summarized in textual form in the lower part of the interface ⑧. The text also precises whether the individual had been misclassified by the decision model (if the information is available) by a circle with a hatched pattern ⑨.

By clicking on ⑦, the user accesses another screen (see Figure 2), where he/she can select another counterfactual depending on whether he/she wants to focus on sparsity (as few modified variables as possible) or prediction performance (the lowest predicted score for the counterfactual for the individual class). The counterfactual that requires the least number of modified variables is proposed by default. Finally, a home page (not shown here) gives a description of the analyzed data (characteristics and semantic of variables).

[5] `https://drive.google.com/file/d/1yog5J1QVq2zQ9WK4P3ujg4Zxn_NZScJB/view?usp=share_link`

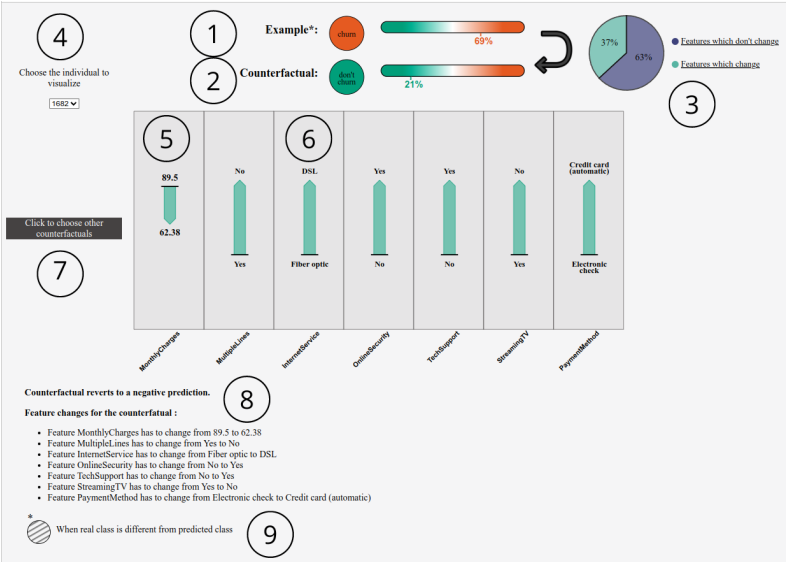

Fig. 1. Interface for presenting an example to be explained and an associated counterfactual.

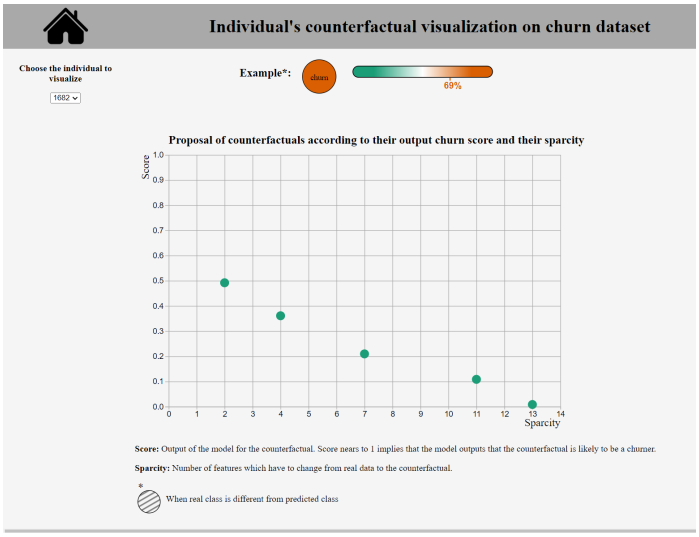

Fig. 2. Interface for alternative counterfactual selection according to the two axes sparsity/classification score.

2.2 Implementation

The implementation is based on Flask, which is a micro-framework for web development in Python to present data and display web pages. Visualizations and interactions are created using JavaScript and d3js. Flask applications can easily be embedded in website or even in Jupyter Notebooks. In this demo, we use HTML and CSS to create the web pages. We can interface with any prediction model, and any counterfactual explanation generator. The data needed for the visualization is provided via a JSON file which must include the variables names and 2 variable/instance matrices, one with the instances to explain and another with the counterfactuals. The file must also contain some classification results: the prediction probabilities of the model and the predicted classes both for the instances to explain and their counterfactuals.[6]

3 Use case study

We illustrate the tool on the Telco Customer Churn dataset [1] which contains 7,043 instances described by 20 input variables. The goal is to predict the churn of a telecom operator's customers (with 2 classes: *churn* vs *no churn*). For our experiments we used VCNet, an architecture that is able to generate at the same time the decision and a counterfactual explanation and is well adapted for processing mixed tabular data [6]. We discuss the analysis of the example shown in Figure 1. It corresponds to an individual (Id 1682) labelled by the decision model as a churner with a probability of 69%. The displayed counterfactual changes the class of the example from churn to no churn with a 79% probability of no churn (21% probability of churn). The counterfactual was obtained by the modification of 7 variables from the initial example (37% of the input variables). The reader who wants more details about the changes can look at the details in Figure 1. Figure 2 shows that other counterfactuals with a good compromise on the performance and sparsity objectives are available. A first counterfactual that proposes the modification of 2 variables (decrease of the monthly bill from 89.5$ to 77.25$ and modification of the payment method) reduces the probability of churn from 69% to 49%. The other counterfactuals can be discussed in the same way. The business expert is thus able to choose the criterion that seems the best between sparsity and classification score.

4 Further developments

The tool presented will evolve to include new features. For now, interactions with the user are limited to the choice of a counterfactual in a possible set according to criteria of sparsity or classification performance. The user could also be interested in selecting the variables that make up the counterfactual. Another area for improvement concerns the textual formalization of the explanation, which is currently very limited. Work on the ergonomics of the interface would also be of interest, as would a user study.

[6] `https://github.com/fwallyn/counterfactualViz`

References

1. https://www.kaggle.com/datasets/blastchar/telco-customer-churn
2. Bove, C., Aigrain, J., Lesot, M.J., Tijus, C., Detyniecki, M.: 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 (2022)
3. Garcia-Zanabria, G., Gutierrez-Pachas, D.A., Camara-Chavez, G., Poco, J., Gomez-Nieto, E.: SDA-Vis: A visualization system for student dropout analysis based on counterfactual exploration. Applied Sciences **12**(12), 5785 (2022)
4. Gomez, O., Holter, S., Yuan, J., Bertini, E.: 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 (2020)
5. Guidotti, R.: Counterfactual explanations and how to find them: literature review and benchmarking. Data Mining and Knowledge Discovery pp. 1–55 (2022)
6. Guyomard, V., Fessant, F., Guyet, T.: 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). p. 10 (2022)