FRI Data Science Project Competition 2024
Final report

Catterfactuals a CATegorical Counterfactual Visualisation

Cristian Bassotto, Camile Lendering, and Nikolay Kormushev

Abstract

Counterfactual explanations in AI involve altering input data to demonstrate how changes can affect model outcomes, providing insights into decision-making processes. While many studies focus on generating these explanations, few address their visualization. This paper introduces a model-agnostic UI for counterfactuals, emphasizing simplicity and user-friendliness. Guided by psychological principles and validated through user feedback, our UI aims to enhance the accessibility and interpretability of counterfactual explanations.

Keywords
XAI, UX, Counterfactuals

Advisors: prof. dr. Erik Štrumbelj

Introduction

Counterfactuals are established as a key approach in explainable AI (XAI). They offer insights into model decisions by showing the feature changes needed for different outcomes. While numerous studies focus on generating counterfactuals [1], few address their visualization or user experience. Most existing visualization methods are either overly complex or underdeveloped and lack user validation. This research addresses these gaps by developing a user-friendly, model-agnostic counterfactual visualization designed for non-experts.

Related works

Existing counterfactual visualization methods vary in complexity. SDA-Vis [2] features a well-developed interface and includes a user study, but it is domain-specific and complex, catering more to expert users. AdViCE [3] and DECE [4] also visualize counterfactuals but are not suitable for non-experts.

Simpler UIs like VICE [5] and its improvement [6] handle numeric and categorical features respectively. However, they do not scale well with increased number of features, making the interface less interpretable. Additionally, [6] requires pre-generated counterfactuals in JSON format, needing expert intervention.

All mentioned UIs overlook the psychological aspects of explanation. Warren et al. [7] suggest that users interpret categorical values more easily than continuous numbers, advocating for binning continuous values.

Our research identifies a lack of suitable UIs that make counterfactuals accessible to non-expert users. To address this, we developed a model-agnostic UI requiring minimal expert setup. We aim to simplify counterfactual generation and interpretation, using configurable binning on tabular data and validating our visualization through a user study.

Methods

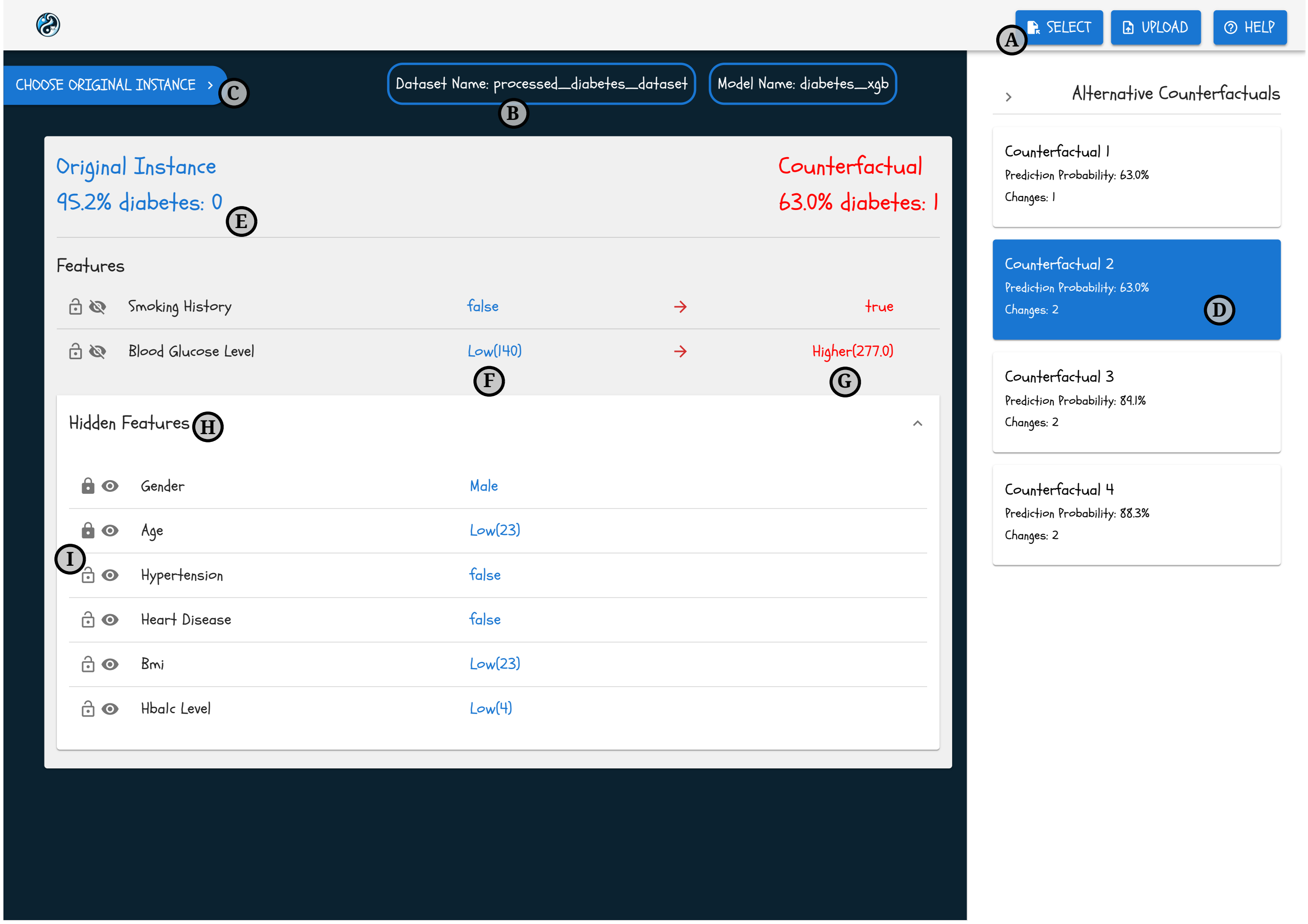
In this section we outline the methods for developing and evaluating our user-centered, model-agnostic counterfactual explanation visualisation system for tabular data.

Catterfactual UI

To see our UI and a basic description of the user flow refer to Figure 1. Here I will outline some details on the visible sections.

UI Sections

- **(A)**: Menu to upload or select dataset. Help displays the same image as in Figure 1 with explanations for the different parts of the UI to assist the user.
- **(B)**: Displays the current model and dataset in use. They can be swapped from the select button or uploaded from the upload button.
- **(C)**: This button opens the menu where a query can be input and sent to generate a counterfactual.
- **(D)**: A menu showing the generated counterfactual where the number of changes is seen and a counterfactual can be chosen for display.



Dataset Name: processed_diabetes_dataset

Model Name: diabetes_xgb

Original Instance
95.2% diabetes: 0

Counterfactual
63.0% diabetes: 1

Features

Smoking History	false	→	true
Blood Glucose Level	Low(140)	→	Higher(277.0)

Hidden Features

Gender	Male
Age	Low(23)
Hypertension	false
Heart Disease	false
Bmi	Low(23)
Hbalc Level	Low(4)

Alternative Counterfactuals

Counterfactual 1
Prediction Probability: 63.0%
Changes: 1

Counterfactual 2
Prediction Probability: 63.0%
Changes: 2

Counterfactual 3
Prediction Probability: 89.1%
Changes: 2

Counterfactual 4
Prediction Probability: 88.3%
Changes: 2

Figure 1. Catterfactual UI offers options to generate multiple counterfactuals. First we selected a model and a dataset (A) which we can see in (B). We are currently using a diabetes dataset with an xgboost model. After we have input a query (C) and a list of counterfactuals is displayed (D). A binned version of the original (F) and counterfactual instance (G). We see that we need to start smoking and that our BLood glucose level is low and that it needs to become 277 to get diabetes. Higher shows that 277 is greater than 140. The probabilities of being the respective class show that the original instance was 95.2% not diabetic but the counterfactual is 63% diabetic (D). A hiding option is available for the features (H) and a locking features displaying which features were not allowed to change during generation like age and gender (G)

- **(E)**: An indicator of the probability of the original instance having income class 0 and the probability of the counterfactual having class 1. Here the values represent True and False but that is not always true. In case like this the values are encoded before providing the dataset and we can not invert that encoding without input from the user or a decoder which he can provide. More work on this will be done in future.
- **(F)**: A binned value of the original instance. We bin the value based on in which tertile of the dataset it is or in other words we want to show if the value is **Low**, **Medium** or **High** for that feature
- **(G)**: Used local binning to demonstrate if the new value is **Higher** or **Lower** than the previous one
- **(H)**: A menu of hidden features so the user can focus only on the ones he is interested in
- **(I)**: A locking feature showing which features were allowed to change during generation. Useful since you can not change your age or race for example.

Technological stack

Our UI is developed using React, with a Python and Django backend to implement a REST API for counterfactual generation. We use DiCE [8] for its features and ease of use. We designed our application to support any generation model, ensuring extensibility.

Model-agnostic Counterfactual Generation

Our back-end handles counterfactual generation requests in a model-agnostic manner, allowing users to upload their own trained models and datasets for generating counterfactual instances. The only requirement is that the model is differentiable, though many non-differentiable models (e.g., Decision Trees, Rule-based systems) are inherently more interpretable.

Model-Agnostic Binning

To make the counterfactual explanations more interpretable, we implement a model-agnostic binning method which transforms continuous feature values into discrete bins, aligning with the findings of Warren et al. [9], which show improved user understanding of counterfactual explanations with categorical features. For the counterfactuals, we use the local transformation method, CAT-CFLocal, as described in [7], where continuous feature values in the counterfactual instance are re-labeled as ”higher” or ”lower” relative to the values in the original instance. For the query instance, we apply global binning [7], categorizing continuous values as Low, Medium, or High. These categories represent the tertile in which the feature falls.

User Study

While there exist many numerical evaluation metrics for counterfactual explanations, such as sparsity and proximity [1], these measures are not suitable for evaluating discretized counterfactual instances. Hence, a user study, conducted by the Zurich User Study group, was used to evaluate the effectiveness of our proposed visualization method.

Participants

The participants for this study come from diverse academic backgrounds, including Computer and Information Science, Electronic Engineering, and Chemistry.

User Interfaces Evaluated

- **Numerical UI**: Displayed exact numerical values for features and counterfactuals.
- **Binning UI**: Used bins (high, medium, low) based on tertiles to represent feature values without exact numbers.
- **Combined Numerical and Binning UI**: Showed both the numerical value and the corresponding categorical bin.

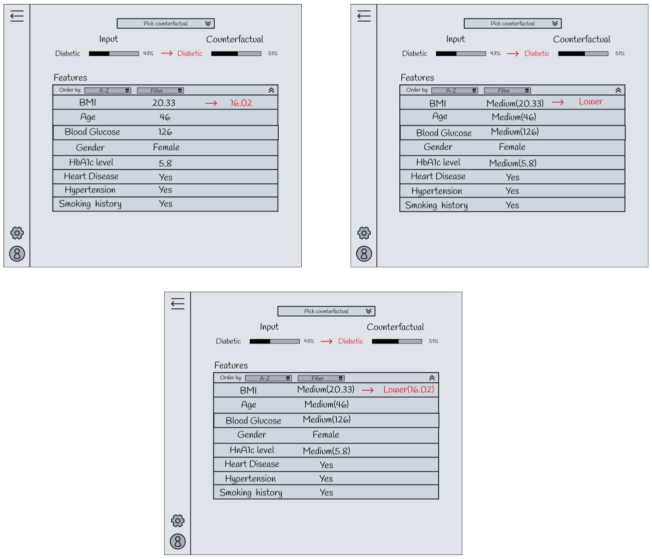
These UIs were evaluated using a diabetes dataset that included features such as HbA1c, blood glucose, gender, age, and BMI. Figure 2 shows the UIs tested in the user study, with the Numerical UI on the top left, the Binning UI on the top right, and the Combined UI on the bottom.

Procedure

Through structured interviews, participants rated each UI on clarity and effectiveness in communicating counterfactual explanations, providing feedback on their understanding and preferences.

Results of the User Study

The study evaluated three UIs for presenting counterfactual explanations: numerical values, binning, and a combination of both. Participants from diverse academic backgrounds provided feedback on their experiences and preferences.



BMI	20.33	→	16.02
Age	46		
Blood Glucose	126		
Gender	Female		
HbA1c level	5.8		
Heart Disease	Yes		
Hypertension	Yes		
Smoking history	Yes		

BMI	Medium(20.33)	→	Lower
Age	Medium(46)		
Blood Glucose	Medium(126)		
Gender	Female		
HbA1c level	Medium(5.8)		
Heart Disease	Yes		
Hypertension	Yes		
Smoking history	Yes		

BMI	Medium(20.33)	→	Lower(16.02)
Age	Medium(46)		
Blood Glucose	Medium(126)		
Gender	Female		
HnA1c level	Medium(5.8)		
Heart Disease	Yes		
Hypertension	Yes		
Smoking history	Yes		

Figure 2. UI’s tested in the user study. **Numerical** top left, **Binning** top right and **Combined** bottom

Overall, participants appreciated the simplicity and clean design of the UIs, especially the use of separate colors to indicate changes in the counterfactuals. Many identified potential model flaws through the counterfactuals. However, there were requests for clearer explanations and more context, particularly for the global bins, as well as suggestions to reorder features based on importance.

The Combined Numerical and Binning UI was the highest rated. Participants found it useful for understanding feature importance, appreciating both absolute values and relative scales. They suggested adding a legend for the bins and incorporating domain knowledge for better clarity.

While participants liked having exact numerical values in the Numerical UI, some found it challenging to understand the counterfactuals without additional explanation.

The Binning UI was the least favored. Participants found it less informative and intuitive, struggling with the meaning of high, medium, and low without reference points. The lack of precise numerical data led to mistrust in the generated counterfactuals.

Discussion

The results of the user study were not as favorable as anticipated. On the positive side the users appreciated our UI. However, many found it challenging to interpret elements like the binning. This highlights that striving for simplicity can sometimes introduce ambiguity when information is reduced to the bare minimum. It became clear that more context was needed to improve their understanding. To address this, we added the previously mentioned help menu with more information on the UI and the binning process. Additionally, we added a home page explaining counterfactuals with an example scenario.

Despite these issues, we discovered that the combined view was the most liked by the study participants. This suggests that providing numerical information alongside the binning, as discussed in [7], is beneficial. Further research can be conducted to evaluate the classification accuracy of other samples, as this study only estimated subjective satisfaction and understanding.

Another noteworthy finding is that counterfactuals effectively highlight the flaws in AI models. Users were able to identify where the model fell short and gain a better understanding of how they work. We believe this is crucial when deciding whether to trust a model's decision.

After implementing the improvements, we interviewed three additional users. These interviews indicated an improved understanding of the UI and the binning process, validating our enhancements.

Future improvements

- Add information tooltips for the features with information given by an expert user to further give users domain knowledge and ease their understanding. Here also optimal data ranges can be added as per the suggestions.
- Incorporate an option to pass an encoder and or decoder to the backend so categorical features and labels can be fed into the model in an easy to interpret format and returned in the same way.
- Implement more generator models to see which provide the most understandable counterfactuals for non-expert users.
- Support setting of feature bounds to not generate impossible values making counterfactuals more interpretable.
- Add an option to display counterfactuals already present in the dataset to account for plausability.
- Display multiple counterfactuals side by side in the main menu for comparison.
- Implement customisable binning ranges (not just tertiles) by an administrator/expert when uploading the model.
- Support expressing preference for which features should change first.
- Include multi-label classification and regression counterfactuals.

Conclusion

To conclude, we believe counterfactuals can help people understand why and how AI models make decisions. While binning can be useful, it can also be confusing without prior explanations. Providing clear context and integrating numerical values with bins can enhance interpretability in AI models. We believe our UI showed great potential although it is a work in progress.

References

[1] Mark T. Keane and Barry Smyth. Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable ai (xai), 2020.

[2] Germain Garcia-Zanabria, Daniel A Gutierrez-Pachas, Guillermo Camara-Chavez, Jorge Poco, and Erick Gomez-Nieto. Sda-vis: A visualization system for student dropout analysis based on counterfactual exploration. *Applied Sciences*, 12(12):5785, 2022.

[3] Oscar Gomez, Steffen Holter, Jun Yuan, and Enrico Bertini. Advice: Aggregated visual counterfactual explanations for machine learning model validation. In *2021 IEEE Visualization Conference (VIS)*, pages 31–35. IEEE, 2021.

[4] Furui Cheng, Yao Ming, and Huamin Qu. Dece: Decision explorer with counterfactual explanations for machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1438–1447, 2020.

[5] Oscar Gomez, Steffen Holter, Jun Yuan, and Enrico Bertini. Vice: Visual counterfactual explanations for machine learning models. In *Proceedings of the 25th international conference on intelligent user interfaces*, pages 531–535, 2020.

[6] Victor Guyomard, Françoise Fessant, Thomas Guyet, Tassadit Bouadi, and Alexandre Termier. Interactive visualization of counterfactual explanations for tabular data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 330–334. Springer, 2023.

[7] Greta Warren, Barry Smyth, and Mark T Keane. "better" counterfactuals, ones people can understand: psychologically-plausible case-based counterfactuals using categorical features for explainable ai (xai). In *International conference on case-based reasoning*, pages 63–78. Springer, 2022.

[8] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 607–617, 2020.

[9] Greta Warren, Ruth M. J. Byrne, and Mark T. Keane. Categorical and continuous features in counterfactual explanations of ai systems. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, page 171–187, New York, NY, USA, 2023. Association for Computing Machinery.