Investigating the Intelligibility of Plural Counterfactual Examples for Non-Expert Users: an Explanation User Interface Proposition and User Study

Clara Bove, Marie-Jeanne Lesot, Charles Albert Tijus, Marcin Detyniecki

Investigating the Intelligibility of Plural Counterfactual Examples for Non-Expert Users: an Explanation User Interface Proposition and User Study

Clara Bove
clara.bove@lip6.fr
clara.bove@axa.com
Sorbonne Université, CNRS, LIP6
F-75005 Paris, France
AXA, Research lab
Paris, France

Marie-Jeanne Lesot
marie-jeanne.lesot@lip6.fr
Sorbonne Université, CNRS, LIP6
F-75005 Paris, France

Charles Tijus
tijus@lutin-userlab.fr
Laboratoire des Usages en Technologies d'Information Numériques, University Paris 8
Paris, France

Marcin Detyniecki
marcin.detyniecki@axa.com
AXA, Research lab
Paris, France

ABSTRACT

Plural counterfactual examples have been proposed to explain the prediction of a classifier by offering a user several instances of minimal modifications that may be performed to change the prediction. Yet, such explanations may provide too much information, generating potential confusion for the end-users with no specific knowledge, neither on the machine learning, nor on the application domains. In this paper, we investigate the design of explanation user interfaces for plural counterfactual examples offering comparative analysis features to mitigate this potential confusion and improve the intelligibility of such explanations for non-expert users. We propose an implementation of such an enhanced explanation user interface, illustrating it in a financial scenario related to a loan application. We then present the results of a lab user study conducted with 112 participants to evaluate the effectiveness of having plural examples and of offering comparative analysis principles, both on the objective understanding and satisfaction of such explanations. The results demonstrate the effectiveness of the plural condition, both on objective understanding and satisfaction scores, as compared to having a single counterfactual example. Beside the statistical analysis, we perform a thematic analysis of the participants' responses to the open-response questions, that also shows encouraging results for the comparative analysis features on the objective understanding.

CCS CONCEPTS

• **Human-centered computing** → **User studies**.

KEYWORDS

explainable AI, XAI, human-centered AI methods, user studies, interface design, XUI

ACM Reference Format:
Clara Bove, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki. 2023. Investigating the Intelligibility of Plural Counterfactual Examples for Non-Expert Users: an Explanation User Interface Proposition and User Study. In *28th International Conference on Intelligent User Interfaces (IUI '23), March 27–31, 2023, Sydney, NSW, Australia.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3581641.3584082

IUI '23, March 27–31, 2023, Sydney, NSW, Australia

ACM ISBN 979-8-4007-0106-1/23/03…$15.00
https://doi.org/10.1145/3581641.3584082

1 INTRODUCTION

The research field of XAI (eXplainable Artificial Intelligence) has proposed numerous interpretability methods to extract various type of information that act as explanations about the behavior of Machine Learning (ML) models and the outcome they predict (see e.g. [1, 33] for recent surveys). Examples of such methods include feature importance vectors [28, 37], rules [38] and counterfactual explanations [34, 44] to name a few. However, this step of explanation identification must be completed by a step regarding their presentation and display to the end-users [40]. In the human-computer interaction research field, the concept of *eXplanation User Interface* (XUI) has been introduced [6] to that aim, "as the sum of outputs of an XAI system that the user can directly interact with". This challenge is real for all kinds of ML explanations, it has been especially addressed for the case of local feature importance vectors [2, 5, 11, 46].

In this paper, we study counterfactual explanations, that have been argued to be a highly relevant form of explanation based on arguments from cognitive sciences regarding their resemblance to human explanations [3, 31, 44, 45, 49]. Among others, they possess contrastive properties, i.e. they are formulated as answers to *Why not?* questions [31]. Such explanations are particularly useful to users who are trying to understand why they did not get a desired outcome (e.g. using a canonical example, if a ML model predicts their loan application is denied) [36]. From a computational point

of view, in the reference case of binary classification, counterfactual examples are basically defined as data instances that are (i) predicted to be in the other class, and (ii) as similar as possible to the user instance. They allow to underline the minimal changes that should be performed to be predicted in the desired class. Numerous methods and variants have been proposed to implement these principles, see e.g. [13, 43] for recent surveys, listing more than 50 approaches. This paper focuses on the case of plural counterfactual examples, i.e. when counterfactual explanations contain several examples. Indeed, it has been proposed to build so-called diverse counterfactual examples [10, 21, 34], claimed to constitute more relevant and appropriate explanations. This paper only takes into account the fact that they provide several examples instead of a single one and does not study the extent to which they differ one from another. Thus, we favor the word "plural" instead of "diverse". From an XUI point of view, the interfaces that have been proposed to display explanations in the form of counterfactual examples mostly focus on the case of single counterfactual example. Moreover, as detailed in Section 2, they are overall more adapted to an expert audience: they most often remain difficult to exploit for users with no knowledge, neither in machine learning nor in the application domains for which the explanations are provided.

Recent works underline the issue that most of these explainability methods have not been tested with real users, and that there is a lack of empirical research in understanding the users' needs for counterfactual explanations in their usage [19, 39, 43]. This also applies to the case of explanations in the form of plural counterfactual examples, for instance investigating experimentally whether too many examples may create confusions [4, 20].

In this work, we investigate the intelligibility of explanations expressed in the form of plural counterfactual examples that are presented to users. We consider the case of non-expert users, as it has been demonstrated that they are generally the ones struggling most with the understanding of complex systems [2, 5]. The paper contributions are, first, a process for designing and evaluating an XUI for such explanations: we investigate (i) if plural counterfactual examples are indeed better than having a single one, and (ii) if we can mitigate the users' confusion with a comparative analysis enhancement when there is a high number of examples. To that aim, we propose an implementation of such enhanced explanations in an XUI for a financial scenario related to a loan application. As a second contribution, the paper provides the first experiment, to the best of our knowledge, evaluating the intelligibility of plural counterfactual examples for the non-expert users. We propose to consider two components for this quality, distinguishing between subjective satisfaction and objective understanding that have been shown to be crucial criteria [2, 5, 46]. The evaluations of the effectiveness of the propositions are performed through a moderated study in lab with 112 participants. To analyze the results, we propose to combine quantitative and qualitative approaches, exploiting both the numerical collected data and the textual answers provided to open-response questions. The results of the statistical analysis demonstrate the effectiveness of the plural condition, both on objective understanding and satisfaction scores, as compared to having a single counterfactual example. The qualitative analysis based on a thematic analysis of the textual answers shows that the proposed comparative analysis features are promising approaches to improve the intelligibility of such explanations, even if the participants partially report they are not satisfied by the counterfactual explanations, as they perceive them as incomplete and too complex.

The paper is structured as follows: Section 2 describes related works on counterfactual explanations and introduces our research questions and hypotheses. Section 3 presents the enhancements we propose for the explanation intelligibility in the form of plural counterfactual examples. Section 4 presents the implementation we propose of such enhanced explanations for a financial scenario The material we use for the evaluation of the proposals in a moderated lab study is described in Section 5. Section 6 presents and describes the results of the quantitative and qualitative analyses of this study. Finally, we discuss limitations and future works in Section 7 and conclude in Section 8.

2 RELATED WORKS AND RESEARCH QUESTIONS

Among the large variety of explanation forms proposed in the XAI domain (see e.g. [1, 33] for recent surveys), counterfactual explanations possess the key property to be contrastive [6, 31, 49]: they allow to answer to questions such as "Why Q rather than P ?". It is argued that they are much more causally informative than factual explanations [45, 47], which is another crucial property of explanations. Yet, it can also be argued that they can be misleading or deceptive [22]. User studies thus appear to be necessary to assess the relevance of such explanation methods, but most methods lack user studies [6, 19]. In this section, we briefly review existing methods for generating counterfactual explanations, and focus on the specific case of diverse counterfactual explanations. Then, we analyze recent works in XAI for presenting such explanations to end-users and evaluate their intelligibility.

2.1 Counterfactual explanations in XAI

Given a user-defined instance and the prediction associated to it by a block-box machine learning model, a counterfactual example proposes to explain this prediction by identifying minimal changes that can be applied to the instance so as to get another prediction. These changes can be interpreted as causes that explain why the model did not predict the desired class in the first place. Unlike other XAI methods that generate factual explanations (e.g. SHAP for the weight of features on the model's decision), the contrastive and selective aspects of these explanations give them a causal dimension [45, 47].

Numerous methods have been proposed to generate such counterfactual explanations (see e.g. [13, 43] for recent surveys), for instance varying the definition of change minimality between the initial instance and the generated counterfactual example. They are defined by minimizing the distance, for which various definitions can be considered, such as the Euclidean distance, the L1 norm [44] or combinations with other norms so as to improve the change sparsity, in terms of the number of modified features [23–25]. In the latter case, the notion of minimal changes combines the amount of changes for each feature as well as the number of features involved, looking for a tradeoff: it is considered as relevant to change more one feature if it makes it possible to achieve the desired class with changing less features. Other variants integrate constraints in the

formulation of this distance, for instance so as to generate only plausible [27] or feasible [35] counterfactual examples, to take into account causal reasoning [18, 29] or users' knowledge [17, 23].

Recent works on counterfactual explanations propose approaches to generate plural counterfactual examples, and claim that having several examples can help users to better interpret them [10, 21, 34]. Indeed, a single counterfactual can be considered misleading as it may suggest changes that are not feasible or plausible for example [9]. Similarly to the methods cited above, in particular for combining the constraints of proximity (using the L2 norm) and the sparsity (L1 or L0 norm), these methods propose different approaches to generate a set of multiple counterfactual instances, most often emphasizing the need for diversity among them. The latter can be understood in different ways: diversity in terms of optimized metrics [9] or in the feature space [14, 34]. This paper only takes into account the fact that they provide multiple examples instead of a single one, not studying the extent to which they differ one from another. Thus we favor the word "plural" instead of "diverse".

As mentioned in the introduction, these methods are most often proposed from a computational point of view, but lack empirical research in understanding users' needs of counterfactual explanations in their usage [19, 39, 43]. There is no or little empirical evidence to prove the relevance of one approach as compared to another, and in particular to establish empirically the claim that plural counterfactual examples are helpful to users.

2.2 Counterfactual explanations in XUI

Recent works propose XUIs for counterfactual explanations [4, 11, 12, 42, 46, 48] as well as evaluation methods to measure their effectiveness on users' understanding [2, 5, 30].

Several forms have been proposed to present counterfactual explanations, such as textual [46], visual [11, 12, 48] or vocal [42]. Regarding the textual approach, the user study shows that it increases users' objective understanding and satisfaction [46]. Regarding the visual presentation, *AdViCE* [12] is an XUI with visual and interactive counterfactual explanations that enables the comparison of decisions on user-defined data subsets. Although, this XUI has not been user tested, recent work has demonstrated the effectiveness of enabling comparison on the users' understanding for example-based explanations [4]. Finally, Glass-Box [42] is a voice-enabled device that provides class-contrastive counterfactual explanations when questioned by users for the understanding of automated decisions.

Similarly to XAI methods, most of these XUIs have not been tested with users. Evaluating the effectiveness of XUIs remains a challenging task [5, 6, 26, 31, 32], for which numerous methods and quality criteria have been proposed (see e.g. the survey proposed by Hoffman et al. in [15]). A consensus has recently been reached, according to which this assessment needs to take into account two distinct components, evaluating both objective understanding and subjective satisfaction [2, 5, 46]. Also, it has been demonstrated that performing both quantitative and qualitative analyses helps assess users perceptions of the quality of the explanations [30].

2.3 Research questions

This paper aims at studying first the presentation of counterfactual explanations with plural examples in an XUI for non-expert users. Second, we study the enhancement of these explanations with comparative analysis for non-expert users. More precisely, the aim is to examine how effective they are to improve the explanation quality for users with no expertise, neither in the ML nor in the involved application domains. As discussed in more details in Section 5.1, we consider two components for this explanation quality, distinguishing between objective understanding, which assesses the extent to which users actually understand the explanation, and subjective satisfaction, which assesses the extent to which users appreciate the interface.

More precisely, the study is driven by the following research questions and hypotheses:

- **RQ1** : How effective are plural examples for improving understanding and satisfaction of counterfactual explanations for non-expert users?
  - H.1.1 : Plural examples improves understanding of counterfactual explanations, as compared to one example.
  - H.1.2 : Plural examples improves satisfaction of counterfactual explanations, as compared to one example.
  - H.1.3 : Comparative analysis on plural examples improves understanding of counterfactual explanations, as compared to one example only.
  - H.1.4 : Comparative analysis on plural examples improves satisfaction of counterfactual explanations, as compared to one example only.
- **RQ2** : How effective is comparative analysis for improving understanding and satisfaction of plural counterfactual explanations for non-expert users?
  - H.2.1 : Comparative analysis improves understanding of plural counterfactual explanations
  - H.2.2 : Comparative analysis improves satisfaction of plural counterfactual explanations

In order to answer these questions, we develop an interface offering plural enhanced counterfactual explanations for the case of non-expert users, as described in the next sections.

3 DESIGN PRINCIPLES

This section presents the XAI principles we propose for offering features to allow comparative analysis of counterfactual explanations with plural examples. Their description, purpose and level are discussed in turn below and summarized in Table 1. We propose an implementation of these principles in a financial usage scenario described in Section 4.

3.1 Card-based design

We apply a card-based design for the display of the explanations, as illustrated in Figure 1: each counterfactual example is represented on its own card. Compared to the displays of counterfactual explanations reminded in Section 2.2, this design choice allows us to associate more content and interactions with the explanations provided by the machine learning tools described in Section 2.1. Thus, we first consider counterexamples individually and adapt the length of the card to the amount of content to display. A card

contains two parts with different pieces of information related to the feature.

The top part displays the labels of the data descriptive features whose values are modified in the considered counterexample, together with these new values that allow to reach the class opposite to the predicted one. We believe it is important for labels to be user-friendly so we propose to name them with non-technical labels: we use the names known from the user (see Section 4.1). Also, we propose to group the features into contextual categories [2] so users can identify quickly what category of information is impacted by the suggested change: e.g. for a change on the age, the feature is displayed under the category "Personal Information".

The bottom part of the card is dedicated to additional information on the counterfactual example. In particular, we highlight the level of feasibility of the example with integrated expert knowledge. As presented in Section 2, counterfactual examples are generated based on different criteria that do not always take into consideration how feasible the proposed changes are in the context of use. Thus, we allocate a significant area on the card to display this information to help users identify visually the feasibility of each example. We present in details how this additional information is obtained in Section 3.2.2.

The cards are presented in a 3-column grid. When generating plural counterfactual examples, it can be difficult to present a rich set in one screen. Here, we use the grid so that the users can scroll on the page to discover the different counterfactual examples. To navigate between the cards, we add a search bar using key words that automatically filters the set of cards accordingly. We also add a "sort by" button above the grid: users can sort the counterfactual examples by increasing or decreasing number of modified data feature. In addition, as we propose to add three different levels of feasibility of the examples (see above and in Section 3.2.2), users can also sort the card by increasing or decreasing level of feasibility.

3.2 Comparative analysis of counterfactual explanations with plural examples

To mitigate potential confusion in explanations with plural counterfactual examples, we propose to offer the non-expert users features that should make it easier for them to compare and analyze this rich set. We present in turn below the purpose of comparative analysis and the different features we propose as XAI principles to enhance these explanations with plural counterfactual examples.

3.2.1 Comparative analysis' purpose. As discussed in the previous section, having plural counterfactual examples may bring confusion to the users. Indeed, in the process of explanation assimilation, the users may need to compare and analyze various information and we argue that they need guidance and complementary information due to their lack of knowledge in both machine learning and the applied domain. We propose two XAI principles to do so: features for highlighting singularities of each example, presented in Section 3.2.2, and features for guiding the non-expert users to compare the diversity in a rich set of counterfactual examples, presented in Section 3.2.3. They respectively apply at two levels we propose to distinguish: the first one corresponds to each counterfactual example represented in a card, individually; the second considers all counterfactual examples globally.

3.2.2 Highlighting examples' singularities. The first principle aims at visualizing and assessing singularities in order to help the users differentiating one counterfactual example from another.

As discussed in Section 2, when interacting with plural counterfactual examples, users do not know how to interpret the proposed changes. As to help users better interpret and assess a counterfactual, it is important to be more precise regarding the meaning of the provided explanation. Also, the users may need extra information on the value of each counterfactual. Thus, we propose to highlight these two elements on the example card. On the top of the card, we highlight the non-zero differences with initial values (as opposed to information retrieved by plural counterfactual methods such as *DiCE* [34] which displays the changed values as the counterfactual explanations). At the bottom, we add new information derived from expert knowledge. For example, we add a level of feasibility of the suggested variations on the example. For each data descriptive feature, three levels of feasibility are distinguished for the suggested variations: they can be either feasible, moderately feasible or hardly feasible, depending on the context of use. For counterfactual examples with more than one data descriptive feature variation, we adopt a pessimistic approach and display the lowest level of feasibility between all the involved modified data features. This level of feasibility should provide non-expert users with an additional element for the good assessment of the counterfactual. Moreover, this information should help them to compare the rich set of counterfactual examples.

We propose that these highlighted singularities features are accessible on each card of counterfactual examples, so that the non-expert users can better interpret and assess them.

3.2.3 Guided comparison. As previously presented, we should also provide more guidance to the non-expert users on how to analyze a set of various example-based explanations. We propose a **guided comparison** XAI principle, that aims at underlining what makes the difference from one example to another.

The non-expert users could get lost when exploring various examples, not knowing where to start and how to navigate in between. Thus, they should have more guidance towards the directions they should follow when exploring and analyzing a set of counterfactual examples. We propose to offer users filtering buttons for the display of the plural counterfactual examples. These options aim at underlining the differences between examples, and should match the users needs when comparing and analyzing them. In this context, as the suggested examples are not necessarily diverse, we add a filtering button to display only the cards with the most diverse ones. We describe the implementation of this feature in Section 4.2.2.

We propose that these guided comparison features are accessible above the card grid, so that users can filter the cards to better navigate between them.

4 APPLICATION: IMPLEMENTATION OF THE DESIGN PRINCIPLES IN A FINANCIAL SCENARIO

This section presents the application of the XAI principles we propose, as described in Section 3, into an finance-related interface. We

Principle	Description	Purpose	Level
Highlight singularities	Enhance the counterfactual examples by highlighting two complementary information: the non-zero differences with initial values and the added value of the example as compared to others.	Help the users for an accurate interpretation of each counterfactual example	At the example level
Guided comparison	Offer the users pre-defined filtering options for the display of the plural counterfactual examples. These options should match the users needs when comparing and analyzing a set of examples.	Ease the analysis and comparison of plural counterfactual examples towards the predicted output	At the explanation level

Table 1: Design principles to improve the intelligibility of counterfactual explanations with plural examples for non-expert users. We describe and define the purpose of each principle we propose for adding comparative analysis features. We also define the level of the ML explanations where the described principle is more valid: "explanation level" refers to principles that apply to the overall ML explanations for one prediction; "example level" refers to the principles that apply to each counterfactual example.

describe the usage scenario in Section 4.1 and the design process for implementing the principles we propose in this XUI in Section 4.2.

4.1 Usage scenario

We apply the principles we propose in a solvency evaluation interface. In this considered scenario for the usage of the interface we propose, a user connects to a platform and starts applying for a loan by providing several pieces of information such as the the desired loan settings (loan amount, duration, installment rate), bank information (bank account and savings values, current and/or loan history), personal information (age, number of dependents, phone number), professional situation (current job occupation and duration, foreigner worker status), as well as current lodging situation. This information is usually required by financial organizations to evaluate the solvency of the applicant according to each individual risk for the payment and reimbursement of the loan. We additionally consider that the names used in this form define non-technical labels the user understands, as he/she fills them: they thus constitute the labels used in the explanation interface.

An ML model uses this information to estimate the solvency of this user. The aim of the XAI interface is to present the estimated solvency to the user, together with explanations to help him/her understand how the provided information impact the evaluation.

4.2 XUI Interface

The implementations of the explanations in the form of plural counterfactual examples, as well as the XAI principles we propose to enhance such explanations with comparative analysis features, as presented in Section 3, are illustrated in Figures 1 and 2. On the left, this interface presents the solvency predicted for the considered loan application whose characteristics are supposed to have been inputted to the system in a preliminary step. We provide the user with transparency on the ML system's scope and basic operations above the explanations, as it has been demonstrated that it can help the users understand how the model works and how to read the following explanations [2]. We describe in the following paragraphs the design of these explanations with the implemented principles.

4.2.1 Implementing plurality. We implement multiple counterfactual examples based on the objective to provide users with maximum diversity in the built explanation. We use the card-based design approach presented in Section 3 to display this set: each card present a counterfactual example, that can suggest variations on one or more data descriptive features.

The search bar and the "sort by" are offered to the user above the grid, as illustrated in Figure 1, to allow users to search for specific information and sort the cards. We implement the filtering buttons on top of the grid as well, as presented below.

4.2.2 Implementing comparative analysis. This section presents the proposed implementation of the two principles we propose to ease comparison and analysis of the counterfactual explanations with plural examples, that we present in turn bellow.

Highlight singularities. Each example-associated card contains two complementary pieces of information: highlighted information about the counterfactual change on the top and a feasibility score on the bottom.

For all features modified by the counterfactual example, we display the initial value, striking it through, and we highlight, in bold and green color, the new, counterfactual, value. We also add the legend "Good solvency" next to the counterfactual value, which is the opposite class. We do so to highlight that this change (or the combination of these changes for counterfactual examples with plural changes) would have made the model to predict the opposite class. These highlights should allow users to understand the value of the change and its effect on the predicted outcome.

In addition, we pair the counterfactual example with a feasibility score regarding the suggested changes at the bottom of the card. We provide this information by injecting expert knowledge into the model, as described in Section 3.2.2. In the user interface, we add a color code for the three levels of feasibility (green color for feasible, orange color for moderate and red for difficult) in order to ease the visual screening of the examples for non-expert users.

Guided comparison. We design filter buttons above the list of the feature-associated cards (see Figure 2), allowing users to change the ordering and/or the filtering of the cards to better compare them. The first button corresponds to the generic display of the cards

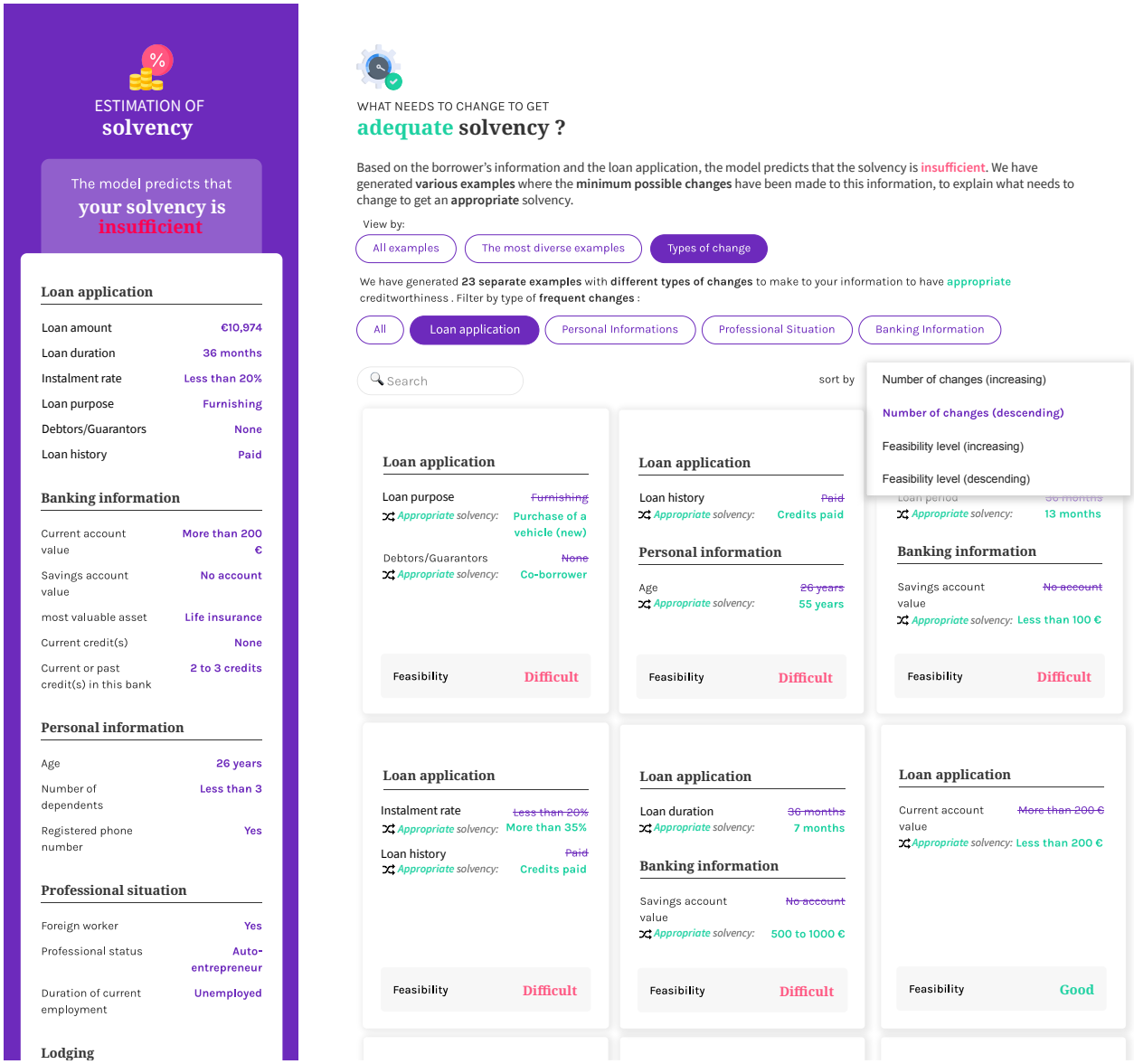

Figure 1: Implementation of plural counterfactual examples and comparative analysis principles in a fictitious finance-related scenario. (Left part of the interface) estimated solvency for the considered loan application, (right) provided explanations: grid presentation of the cards associated with each counterfactual example. The highlighted singularities are implemented as enhanced changes and added expert knowledge on feasibility scores. The guided comparison is implemented with contextual filtering and sorting options on top of the grid. *Note: The interface has been translated from the original language used for the evaluation*

as generated by the explainer. We propose two additional buttons with sorting/filtering options to guide comparison.

First, we add a button that filters the cards to display only the most diverse ones with the closest proximity to referent values (e.g. when there are plural cards that suggest changes on a similar feature, it will only display the one that is the closest to the instance value), so that users can have a synthetic overview of all the closest and diverse counterfactual examples in real value.

Second, we add a button that offer an option to filter by the frequency of changes by data feature categories: there can be several counterfactual examples that suggest changes on the same data descriptive feature, which leads to define frequently modified features and further on to frequently modified feature categories. We propose to add dynamically a button for each frequent category of change suggestions, in order to filter and only display the counterfactual examples offering such changes. We propose to display

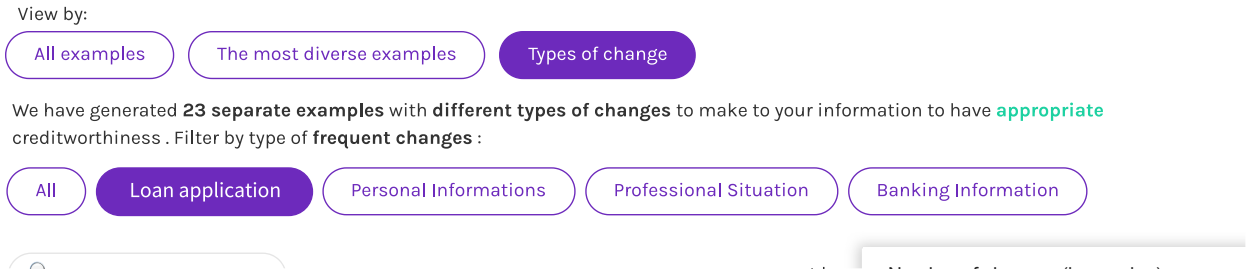

Figure 2: Application of the guided comparison principle: if the user selects the "Types of change" button, an additional row of button appears and offers different filtering options of the cards display, according to frequency of similar changes (in this context, the most frequent types of changes are for the "loan application" settings, and the least frequent are for the "Banking information"). *Note: The interface has been translated from the original language used for the evaluation*

these buttons in a frequency decreasing order (i.e. from the category with most counterfactual examples to the one with the least), so as to analyze in which category of data descriptive features there are the most suggested variations.

5 EXPERIMENT

To answer the studied research questions and to evaluate the effectiveness of the XAI principles we propose, we describe in turn below the prototype we build and the method we used to conduct the monitored study at the INSEAD-Sorbonne University Behavioural Lab. We use this prototype to test our hypothesis towards the effectiveness of the XAI principles we propose on two dimensions of user's understanding, as described in Section 5.2.

5.1 Prototype

This section presents the interactive prototype we develop as the basis for the evaluation. We use this prototype to build the different versions of the interface for the user evaluation, as described in Section 5.3.2. We discuss in turn below the data set we use to train a ML model for the estimation of the solvency for a prospective loan customer and the method we use to extract diverse counterfactual explanations.

We develop an interactive prototype for a solvency estimation service, as described in Section 4.1. We use the German Credit dataset [16] which is a public dataset downloaded from the UCI Machine Learning Repository. It contains the description of 1000 loan applicants on 20 data descriptive features and their labels as having a good or bad solvency. We use this data set to train a ML model to compute a predicted solvency for each user, namely a Random Forest trained with sklearn tool[1]. On the estimated solvency we get for one instance, we use the *DiCE* method[2] [34] to generate diverse counterfactual examples to explain the given output. To obtain diversity in this set, *DiCE* requires the number of desired examples, the weight on distance and sparsity, as well as the definition of user knowledge such as the list of features that can be modified and their associated range of accepted variation. For the instance we select from the training set and the chosen DiCE configuration, that for instance excludes modifying "foreigner worker status" and "phone number", DiCe generates 23 counterfactual examples that suggest changes on at most two data descriptive descriptive features.

[1]https://scikit-learn.org/stable/index.html
[2]we follow the authors' implementation guidelines as documented on https://github.com/interpretml/DiCE

5.2 Hypothesis testing

We use the three versions of this interactive prototype (see Section 5.3.2) to answer the studied research questions on the effectiveness of the principles we propose towards the objective understanding and satisfaction of non-expert users. We expect that (RQ1) the plural condition as compared to one counterfactual example, and (RQ2) the comparative analysis features for plural counterfactual examples, increase both the objective understanding and user's satisfaction, as presented in Section 2.3. More formally, we consider null hypotheses of the form "the considered condition or enhancement provides no significant improvement of the considered metric". To answer RQ1, we compare the two scores and answers (for objective understanding and satisfaction) for each of the two enhanced interfaces (interface B with plural examples, and interface C with comparative analysis on plural examples) as compared to the baseline interface (interface A with a single counterfactual example). To answer RQ2, we compare again the two scores and answers between the two enhanced interfaces (B and C).

5.3 Method

We describe in turn the participant recruitment, the evaluation material, the study procedure and the method to analyze the collected results. The method has been approved by the INSEAD Institutional Review Board (IRB). We pre-tested it with 2 participants to validate the understanding of the XAI interfaces and questionnaires presented in this section, and to adjust the vocabulary used in the questions.

5.3.1 Participant recruitment. We recruited 112 participants from a large open network of volunteers at the INSEAD-Sorbonne University Behavioural Lab (in Paris, France), filtered to meet the requirements of our experiments, i.e. participants with little to no basic knowledge in AI nor in finance. Participants were aged from 19 to 39 (in average 25.5 ± 5.3), 73 were women and 39 were men, and there were various demographics (e.g. job position, level of

study, previous experience in loan application). To ensure the participants were non-experts in both AI and finance, we asked them to self-report their literacy for both topics on a 5-point Likert scale. We excluded the data of 1 participant who reported literacy scores between 4 to 5 at the end of the experiment, despite the initial filtering. After checking the data collected, we also excluded 2 participants who answered all open-response questions with in total less than five words. The results analyzed in the next sections thus rely on the evaluation collected from 109 participants, randomly and evenly distributed across the three versions of the interfaces we propose (see Section 5.3.2). The participants were distributed in independent groups in a between-subjects setting, allowing us to compare (RQ1) the impact of the plural condition, and (RQ2) the impact of comparative analysis features, on the objective understanding scores and the satisfaction rates. All participants received a 6-euro compensation at the end of the experiment.

5.3.2 Material. We present in the following paragraph the material that we use for the user study. We describe in turn below the three tested interfaces, the questionnaires for objective understanding and satisfaction evaluation and the additional collected data.

Tested interfaces. In this monitored experiment, we use three versions of our interface corresponding to all three conditions required for the hypothesis testing. We do so in order to be able to evaluate the impact of having plural examples as compared to having a single, as well as the impact of comparative analysis when having plural examples as compared to having plural examples only. More precisely, the different versions are designed as follows:

- Interface A is the baseline interface. It simply displays one counterfactual example with the card-based design described in Section 3.1.
- Interface B is the interface with plural examples. It adds to interface A plural counterfactual examples, as described in Section 4. Yet, none of the design principles we propose are applied in this version.
- Interface C is the interface offering the features of comparative analysis on plural examples. It adds to interface B the two principles of comparative analysis, as described in Section 3. Figures 1 and 2 present screenshots of this version.

Objective understanding questionnaire. Similarly to [2, 5], we design a questionnaire with 14 statement questions (see questionnaire in B), for which users can either answer "I agree", "I disagree" or "I don't know" . We propose three types of questions to capture different components of user understanding when evaluating the intelligibility of XAI interfaces:

- *(i) Explanations' nature* questions measure the extent to which users understand what type of explanations is provided by a counterfactual example. In our experiment, we measure participants understanding that the provided information is a counterfactual example. *e.g. "The interface provided examples that suggest changes on the initial values that would have made the model predict a different solvency."*
- *(ii) Explanations' effects* questions measure the ability of users to understand how to interpret the explanation towards the predicted outcome. In our experiment, we measure participants understanding of the value of the changes compared to the initial values. *e.g. "The model would have predicted a good solvency if the loan duration was reduced by 10 months."*
- *(iii) Explanations' specificity* questions measure the users' understanding of one complex component specific to the explanation provided. In our experiment, we measure participants understanding of the diversity in the generated counterfactual examples and how they compare them. Thus, these questions apply only to participants using interfaces with plural explanations. *e.g. "It is easier to reduce the loan amount than to change job position."*

For each question, an expected answer is predefined. We consider a participant provides a correct answer if his/her answer is identical to the expected one.

Self-reported satisfaction questionnaire. We adapt the self reporting questionnaire from the Explanation Satisfaction Scale [15] in order to assess users' satisfaction (see questionnaire in C). It gives the participant satisfaction statements in the form of *"The explanations provided by the interface are..."*, followed by one of the eight satisfaction dimensions (respectively "*understandable*", "*satisfying*", "*sufficiently detailed*", "*complete*", "*intuitive*", "*useful*", "*accurate*", "*trustworthy*"). Participants are required to answer on a 6-point Likert scale, from "Strongly disagree" (1) to "Strongly agree" (6), as it has been shown that 6-point response scales are a reasonable format for psychological studies [41].

open-response questions. Also, we ask participant two open response questions to qualitatively measure the intelligibility of the provided explanations. For the objective understanding, we ask participants *what examples would they select to explain the predicted outcome*. For the satisfaction, we ask participants *if they are satisfied with the provided explanations*. We perform a thematic analysis on answers for both questions [8].

Demographics. In addition to the previous items which are related to our research questions, a demographic questionnaire includes two questions regarding the participant literacy in artificial intelligence/machine learning and finance, again using 6-point Likert scales, from "Not familiar at all" to "Strongly familiar", to ensure that participants are indeed non-expert users.

Finally, we collect basic demographic information such as age, gender, education level and current occupation. We also ask participants their experiences with loan applications. Participants can also share their insights and comments on the study in open-response questions.

5.3.3 Study procedure. We conduct the user study in a lab setting at INSEAD-Sorbonne University Behavioural Lab, as it has been shown that the presence of a moderator increases participants' focus [7]. It also allows them to ask questions throughout the evaluation to make sure they understand the instructions.

After giving written consent and prior to the experiment, participants are introduced to the following experimental scenario, translated and resumed from original language : "26-years old freelance graphic designer, Swann will be moving to a new place to work and live in Bordeaux, France. Swann is applying for a loan to the bank in order to fully furnish and equip this new apartment. Swann has previous experiences with loans (for studies first and

travel then) and is confident that it will be accepted. Yet, Swann's solvency is estimated as being not acceptable on the XAI platform used to submit the loan application and some explanations are provided".

This scenario allows us to present the same information and explanations to all participants, which makes the comparison and the statistical analysis significantly easier than if participants inputted their own information into the ML system.

Then, each participant is randomly assigned to one version of the interface for the evaluation. While interacting with the interface, they take the objective understanding questionnaire, answer the subjective satisfaction questionnaire, and then answer the open-response questions and demographics, as described in Section 5.3.2.

5.3.4 Data analysis. As the preprocessings of the collected data show that it is normally distributed, we use one-way ANOVA to analyze (RQ1) the impact of the plural condition, and (RQ2) the impact of comparative analysis features, to test our hypotheses as presented in the previous Section. Table 2 displays the results for the scores and rates obtained in the experiment.

To answer the first research question towards the effectiveness of having plural examples (See Section 2.3), we use the seven questions from the objective understanding questionnaire presented in Section 5.3.2 that are relevant for this comparison, both on the nature and the effects of the explanations. As we compare the intelligibility of the explanations between participants having one counterfactual example (interface A) and participants having plural examples (interfaces B and C), we need to ask questions all participants can answer with the provided information. Thus, we focused on one counterfactual example that is provided on all interfaces. The first score of objective understanding score can vary from 0 to 7 corresponding to the number of correct answers for the 7 related questions of the questionnaire.

To answer the second research question towards the effectiveness of comparative analysis features (see Section 2.3), we use all the questions from the objective understanding questionnaire presented in Section 5.3.2. We compare the intelligibility of the explanations between participants having plural counterfactual examples (interface B) and participants using comparative analysis on plural counterfactual examples (interface C). This second score of objective understanding can vary from 0 to 14 corresponding to the number of correct answers for the 14 questions of the objective understanding questionnaire. Finally, the user's satisfaction is reported from 1 to 6 corresponding to the average score over the eight satisfaction's dimensions presented in Section 5.3.2.

We use one-way ANOVAs to compare the difference between the independant groups. The significance level is defined as $\alpha = .05$. We use the Tukey post-hoc test to get adjusted p-values for multiple pairwise comparisons. Table 2 shows descriptive statistics for each group and their statistical significant differences.

6 RESULTS

We use the results presented in Table 2 and Figure 3, to answer the two research questions we consider regarding the plural condition in Section 6.1 and the comparative analysis in Section 6.2.

RQ1	**Interface A**	**Interface B**	**Interface C**
Objective understanding	4.16 (±1.5)	5.14 (±1.3)	5.0 (±1.1)
ANOVA *as compared to A*	-	+.98 (p=.003)	+.84 (p=.01)
Satisfaction	2.1 (±1)	2.5 (±1.1)	2.9 (±0.7)
ANOVA *as compared to A*	-	-	+.8 (p=.0009)

RQ2	-	**Interface B**	**Interface C**
Objective understanding	-	9.78 (±1.4)	9.66 (±1.4)
Satisfaction	-	2.5 (±1.1)	2.9 (±0.7)

Table 2: Descriptive analysis of the results for the two objective understanding scores and the satisfaction rates, as well as the results of one-way ANOVAs. For RQ1: mean (standard deviation) of the scores for objective understanding from 0 to 7, and rates for satisfaction from 0 to 5. We compare the scores and rates obtained for group B and group C to the ones for group A. For RQ2: mean (standard deviation) of the scores the objective understanding from 0 to 14, and rates for satisfaction from 0 to 5. We compare the scores and rates obtained between groups B and C.

6.1 RQ1: How effective are plural examples for improving understanding and satisfaction of counterfactual explanations for non-expert users?

We measure the significant effectiveness of having plural counterfactual examples on users' objective understanding scores and satisfaction rates. We use 7 questions of the objective understanding questionnaire, as described in Sections 5.3.2 and 5.3.4. This score can vary from 0 to 7. The analysis of Table 2 leads to two main observations commented in turn below. First, having plural examples improves significantly objective understanding. Second, it also improves users satisfaction but there is only a significant difference when there are comparative analysis features.

Having plural examples improves significantly objective understanding. Table 2 shows that interface B (plural counterfactual examples) have the highest improvement in objective understanding with an average score of 5.14 correct answers out of 7, i.e. .98 point more than interface A (one counterfactual example only). The one-way ANOVA shows that this difference is significantly higher ($F_{1,72}$=9.18; p=.003). The Tukey post-hoc test also reveal significant pairwise differences between interfaces A and B (p=.005). In addition, we observe that participants interacting with interface C (plural examples paired with comparative analysis) obtain also higher scores for objective understanding with an average score of 5 out of 7, i.e. which is .84 point higher than for interface A. This difference is statistically significant ($F_{1,70}$=6.76; p=.01) and the Tukey post-hoc test also reveals significant pairwise differences between interfaces A and C (p=.03).

Thus, we reject the null hypotheses as the scores for interfaces with plural counterfactual explanations are greater than the claimed value and conclude that **having plural examples in counterfactual explanations (significantly improves objective understanding of non-expert users,** both with (H1.3) and without (H1.1) comparative analysis features.

Having plural examples improves satisfaction. We observe that participants interacting with interface B give higher satisfaction rates regarding the provided explanations with an average rate of 2.5 out of 5, which is .4 point higher than participants interacting with interface A. Yet, this difference is not statistically significant. Participants interacting with interface C (plural examples paired with comparative analysis) also give higher satisfaction rates with an average rate of 2.9 out of 5, i.e. which is .8 point higher than for interface A. This difference is statistically significant ($F_{1,70}$=11.82; p=.006) and the Tukey post-hoc test also reveals significant pairwise differences between interfaces A and C (p=.005).

Based on these observations, we fail to reject the null hypothesis and are not able to demonstrate the positive effect of having plural examples only on counterfactual explanations on users satisfaction (H1.2). Yet, we reject the null hypothesis as the average rate for the interface with comparative analysis features is higher than claimed value and conclude that **having plural examples when paired with comparative analysis significantly improves satisfaction of non-expert users** (H1.4).

6.2 RQ2: How effective is comparative analysis for improving understanding and satisfaction of plural counterfactual explanations for non-expert users?

We measure the significant effectiveness of comparative analysis of plural counterfactual examples, as compared to plural counterfactual examples only, on users' objective understanding scores and satisfaction rates. We use here all 14 questions in the objective understanding questionnaire described in Section 5.1, and the same satisfaction rates as for RQ1. The analysis of Table 2 and Figure 3 leads to two main observations commented in turn below. First, having plural examples does not improve the objective understanding of plural counterfactual explanations. Second, it improves satisfaction but this difference is not statistically significant.

Comparative analysis on plural counterfactual explanations does not improve objective understanding. Table 2 shows that participants using comparative analysis features (interface C) have slightly lower scores of objective understanding, with an average score of 9.66 out of 14, i.e. which is .12 point lower than for participants without these features (interface B). This difference is not statistically significant. Yet, when analyzing Figure 3, we observe that the minimum score for interface C is 2 point higher than for interface B.

Thus, we fail to reject the null hypothesis and are not able to demonstrate the impact of comparative analysis for plural counterfactual explanations on users' objective understanding (H2.1).

Comparative analysis on plural counterfactual explanations improves satisfaction but the difference is not significant. When comparing the average rates for satisfaction among participants interacting with plural counterfactual explanations, we can see on Table 2 that those who are using comparative analysis features (interface C) rates their satisfaction higher, with an average rate of 2.9 out of 5, i.e. which is .4 point higher than the average rate of participants interacting with interface B. Yet, this difference is not statistically significant.

Again, we fail to reject the null hypothesis and are not able to demonstrate the positive effect of comparative analysis for plural counterfactual explanation on users satisfaction (H2.2).

6.3 Thematic analysis

In combination with the statistical analysis done on the participants' scores and rates, we also analyze their answers for the two open-response questions presented in 5.3.2. We conduct a thematic analysis with an iterative coding process [8]: for each question separately, we analyze in an iterative process the answers without knowing the version of the interface that they were associated with, and identify codes. Then, we analyze the codes by versions of the interface and define themes for both objective understanding and satisfaction.

6.3.1 Objective understanding. For the objective understanding open-response question, we identify 11 codes, and define 4 themes discussed in turn below.

Interpretation of counterfactual examples. We identify codes related to the level of understanding of the counterfactual examples presented to participants. For each version of the interface, the same number of participants (between 12 and 13 in each group) understand that with the minimum suggested changes, the predicted outcome would have been different. Similarly, when having plural examples (interfaces B and C), the same number of participants (respectively 11 and 10) partially understand counterfactual examples. Most of these participants do not refer to the change values when suggesting modifications to the input values to get the loan accepted. Indeed, they all suggest features to change but only one participant in interface C provides the new value as suggested on the examples (C2 says "to lower the loan duration to 26 months, to lower the bank account value to 200 euros and to change the investment rate for 20% to 25%" as suggested on different examples). Finally, we observe 2 participants interacting with interface A that do not understand that the example suggested can be used to explain the predicted outcome, as well as for 1 participant interacting with interface C.

Personal beliefs. We also identify codes related to participants' personal beliefs. Regarding interface A, most participants (21 out of 35) propose alternative explanations based on their own beliefs to explain the predicted class. In addition to the suggested change on the one example provided, participants propose additional changes based on the input data they have (e.g. A6 says that in addition to have a shorter loan duration, the applicant should "open a saving account, find a stable position and find a warrant for the loan"). For interfaces B and C, there are less participants who suggest personal beliefs' based explanations for the predicted class (15 participants for interface B, and 11 for interface C).

Feasibility of the examples. For participants interacting with plural counterfactual examples (interfaces B and C), we identify codes related to the assessment of the feasibility for each suggested example. When disposing of comparative analysis (interface C), 10 participants are capable of selecting the most feasible examples to explain the predicted class. For interface B, only 4 participants are able to do so.

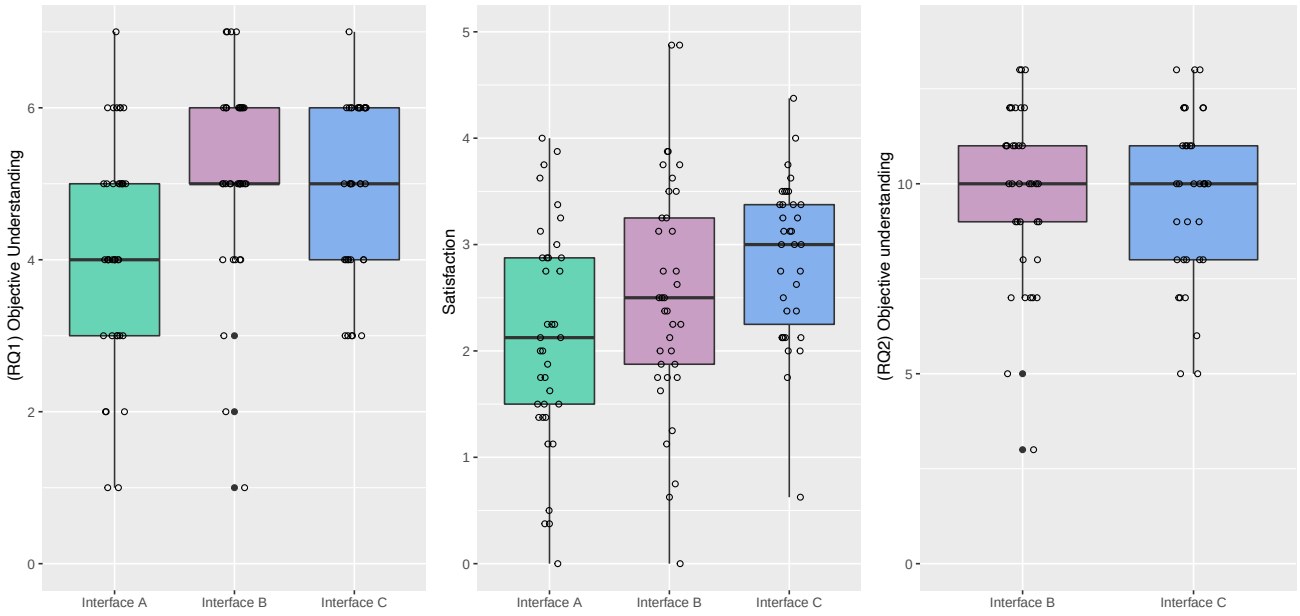

Figure 3: Measuring the intelligibility of the different versions of the interface: overview of (left) the objective understanding scores for evaluating the effect of the plural condition, (middle) the satisfaction rates for evaluating the effect of the plural condition and comparative analysis features, and (right) the objective understanding scores for evaluating the effect of the comparative analysis features.

Association of different examples. Finally, for participants interacting with plural counterfactual examples (again, interfaces B and C), we identify codes related to the ability of the participants to differentiate among the counterfactual examples. They understand that the examples can be used to explain the reject of the loan, yet most of them believe that the suggested changes from different examples can be associated (8 participants with interface B; 7 participants with interface C). For example, participant B1 believes that the best changes that would have made the model to accept the loan application are "to lower the amount of the loan, to find a new position and to wait 10 years", which are three changes on three different examples in the provided set of counterfactuals.

Review. Overall, having plural examples seems to increase the intelligibility of counterfactual explanations and to reduce the inference with personal beliefs. Yet, it also can increase the risk of believing that the proposed changes can be associated. Adding comparative analysis features on counterfactual explanations with plural examples may reduce this risk, and allows users to better assess the feasibility of each suggested change. Thus, these observations lead us to believe that both having plural examples and comparative analysis are promising features to increase objective understanding of counterfactual explanations.

6.3.2 Satisfaction. Similarly for satisfaction, we identify 20 codes and define 4 themes discussed in turn below.

Dissatisfaction. First, we identify codes related to participants expressed of satisfaction. More specifically, we identify three levels of satisfaction. The first level and most observed is expressed dissatisfaction of the provided explanations. Among the 65 participants who report they were unsatisfied, 27 of them are interacting with interface A (one counterfactual), 19 with interface B (plural examples) and 19 with interface C (comparative analysis on plural examples). Participants report that the reasons why they are not satisfied are either because there is no explanation according to them (A20 says "I am not satisfied because there are no explanations provided") or because the information provided is incomplete (A31 says "it misses more justifications, explanations and contextual information"). Some participants also blame the complexity of the explanations (e.g. C22 did not know "how to interpret" the examples, despite them being "very clear and detailed").

The second level expresses partial satisfaction. In total, 20 participants report they were partially satisfied with the explanations (7 for interface A, 7 for interface B and 6 for interface C). Most participants appreciate the clarity of the provided interfaces, but still believe that the explanations are too complex (B28 suggests to introduce better how to interpret the examples "so that it could be easier to understand why this value should change" for the model to accept the loan application).

Finally, the last level expresses satisfaction. Among the 24 participants who reported they are satisfied with the explanations, only 3 of them are interacting with interface A, 11 with interface B and 10 with interface C. In particular, participants report they like to get actionable changes (e.g. B10 says "We immediately understand which values we can change so that we can get the loan application to be accepted"). Overall, there are more participants satisfied with the explanations in interfaces B and C as compared to interface A.

Missing explanations. We identify codes related to missing content in the presented explanations. For participants interacting with interface A, 11 of them feel that there is no explanations provided (A20 says "there are no explanations provided"). For interfaces B and C, the number of participants who share similar opinion is lower (2 for B, and 2 for C).

Expressed needs. Also, we identify codes related to needs explicitly expressed. Whether participants are satisfied or not with the explanations and no matter which interface they are interacting with, 36 participants say they need further information or details about the provided information. For example, participant C16 says that "there should be more detailed information for some examples [...] that are counterintuitive". In addition, 12 participants say they need more contextualization of the provided explanations. Participant A32 says that "this information could be further explained and detailed so that the borrower understands what aspect of his/her application is problematic". Finally, participants interacting with interface C express some additional needs, such as the need to have human contact (1 participant), more transparency over the model (1 participant), and other explanations such as the weight of features on the predicted solvency (1 participant).

Perceived complexity. We identify codes related to the complexity of the provided information. The expressed complexity is different from one version of the interface to another. For interface A, 9 participants say that the provided example is either "difficult to understand", "not intelligible" or "not feasible" according to them. For interface B, 17 participants report that the explanations are complex. Most reported complexity regards the organization of the provided examples. For example, participant B28 says that "the 23 examples are a bit scattered all over the interface and could have been grouped by category (bank information, duration of loan...)". These insights are particularly valuable to us as we aim at addressing this issue with the comparative analysis in interface C. Others report that the examples are also difficult to understand, and that all examples are not always feasible. For interface C, also 17 participants report that the explanation are complex. More specifically, 7 participants report that the explanations are difficult to understand completely because of lack of knowledge in the applied domain. Participant C12 says that "without previous knowledge in loans, it is difficult to understand the reasons why some examples are proposed." Also, 2 participants say that it is difficult to understand how to interpret the explanations because of the plurality of examples (for example, participant C23 says that "the accumulation of cards make it difficult to use the platform"). Moreover, 2 participants say that the explanations are counter-intuitive: participant C16 also says that some "suggested changes" are "unclear and counter intuitive". Finally, other participants also question the feasibility of the suggested examples and that they are not well organized on the interface.

Review. Overall, these observations lead us to believe that participants are mostly unsatisfied with the provided explanations. They are even more unsatisfied when there is only one example suggested. Reasons are multiple: the provided information does not act as explanations, the needs for more details and justification about the suggested examples, the needs for contextualized information, the difficulty to interpret them. Also, we believe that participants are more inclined to believe one counterfactual example is not an explanation. Yet, the complexity seems to be increased when there are plural examples. Participants suggest that the explanations would need to be better organized, and that they would need to be guided on how to analyze and interpret them. These insights encourage us to believe that comparative analysis features are promising tools to improve the intelligibility of such explanations.

7 LIMITATIONS AND FUTURE WORKS

While this paper demonstrate that having plural examples and offering comparative analysis feature can improve the intelligibility of counterfactual explanations, there are some limitations to our work which are important to mention.

We acknowledge that the DiCE method [34] is not adapted to the Random Forest model we trained with sklearn tool, resulting in presenting the participants with surprising counterfactual explanations. For instance, according to one of the provided example, a huge change in the loan amount would be needed to yield a positive outcome which is quite unrealistic. This might explain why some participants are unsatisfied with the explanations and find it difficult to interpret them. We believe there is room for improvement on that point and consider as future works conducting additional user experiments applying the same protocol with other configuration of the explainer as well as other explainer models.

Moreover, we use 23 counterfactual examples based on the objective to provide users with maximum diversity in the built explanation. We believe that another experiment with fewer counterexamples would be also an important topic to address.

Future works will also aim at investigating new modalities to evaluate the objective understanding and the satisfaction, in particular extending the conducted study with a new method for analyzing the collected results. We believe a qualitative evaluation might help to have move comprehensive view of what the users understand or not about the provided explanations, as well as their points of satisfaction or disappointment when using the XUI. Other directions for refining the conducted study will focus on other possible effects of interest. The latter for instance include a possible correlation between objective understanding and subjective satisfaction, the scores per type of questions we ask to the participants for the evaluation of the objective understanding, or the users demographics. Conducting more detailed analyses to evaluate the effect of the collected demographic information will also make it possible to obtain more detailed insights about the effectiveness of explanations provided to non-expert users, tackling one of the major current challenges of the XAI community.

8 CONCLUSION

In this work, we investigate the intelligibility of explanations expressed in the form of plural counterfactual examples that are presented to the non-expert users. The paper contributions are, first, a process for designing and evaluating an XUI for such explanations. We investigate (i) if plural counterfactual examples are indeed better than having a single one, and (ii) if we can mitigate the users' confusion through a comparative analysis enhancement

when there is a high number of examples. We propose an implementation of such enhanced explanations in an XUI for a financial scenario related to a loan application. We perform quantitative and qualitative evaluations of the collected data. In the quantitative analysis, the results show that having plural examples does improve significantly the objective understanding of counterfactual explanations, as compared to having one example only. It does also improve the satisfaction, but this difference is significant only for the interface offering comparative analysis features. On the contrary, the comparative analysis features do not appear to improve significantly neither the objective understanding nor the subjective satisfaction of plural counterfactual explanations. Yet, the qualitative analysis of the collected open-response answers shows that they may reduce the inferences with personal beliefs and help the users to better assess the feasibility of the suggested changes. These observations lead us to believe that the comparative analysis features are promising tools to improve the intelligibility of counterfactual explanations for the non-expert users. These results are of course dependent on the quality of the explanations generated by the machine learning explainer model in the first place, prior to the question of the presentation.

REFERENCES

[1] Alejandro Barredo Arrieta, Natalia Díaz Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.

[2] Clara Bove, Jonathan Aigrain, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki. 2022. Contextualization and Exploration of Local Feature Importance Explanations to Improve Understanding and Satisfaction of Non-Expert Users. In *27th Int. Conf. on Intelligent User Interfaces*. Association for Computing Machinery, New York, NY, United States, 807–819.

[3] Ruth MJ Byrne and Alessandra Tasso. 2019. Counterfactual reasoning: Inferences from hypothetical conditionals. In *Proc. of the Sixteenth Annual Conf. of the Cognitive Science Society*. Routledge, New York, NY, United States, 124–130.

[4] Carrie J. Cai, Jonas Jongejan, and Jess Holbrook. 2019. The Effects of Example-Based Explanations in a Machine Learning Interface. In *Proc. of the 24th Int. Conf. on Intelligent User Interfaces, IUI'19*. Association for Computing Machinery, New York, NY, United States, 258–262.

[5] Hao Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 2019. Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In *Proc. of the Int. Conf. on Human Factors in Computing Systems, CHI'19*. Association for Computing Machinery, New York, NY, United States, 1–12.

[6] Michael Chromik and Andreas Butz. 2021. Human-XAI interaction: a review and design principles for explanation user interfaces. In *Proc. of the 18th TC13 Int. Conf. on Human-Computer Interaction, INTERACT 2021*. Springer, New York, NY, United States, 619–640.

[7] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI. In *26th Int. Conf. on Intelligent User Interfaces* (College Station, TX, United States) *(IUI '21)*. Association for Computing Machinery, New York, NY, United States, 307–317. https://doi.org/10.1145/3397481.3450644

[8] Victoria Clarke, Virginia Braun, and Nikki Hayfield. 2015. Thematic analysis. *Qualitative psychology: A practical guide to research methods* 222, 2015 (2015), 248.

[9] Susanne Dandl, Christoph Molnar, Martin Binder, and Bernd Bischl. 2020. Multi-objective counterfactual explanations. In *Int. Conf. on Parallel Problem Solving from Nature*. Springer-Verlag, Berlin, Heidelberg, 448–469.

[10] Michael D. Ekstrand, F. Maxwell Harper, Martijn C. Willemsen, and Joseph A. Konstan. 2014. User Perception of Differences in Recommender Algorithms. In *Proc. of the 8th ACM Conf. on Recommender Systems* (Foster City, Silicon Valley, California, United States) *(RecSys '14)*. Association for Computing Machinery, New York, NY, United States, 161–168. https://doi.org/10.1145/2645710.2645737

[11] Oscar Gomez, Steffen Holter, Jun Yuan, and Enrico Bertini. 2020. ViCE: visual counterfactual explanations for machine learning models. In *Proc. of the 25th Int. Conf. on Intelligent User Interfaces, IUI'20*. Association for Computing Machinery, New York, NY, United States, 531–535.

[12] Oscar Gomez, Steffen Holter, Jun Yuan, and Enrico Bertini. 2021. AdViCE: Aggregated Visual Counterfactual Explanations for Machine Learning Model Validation. In *IEEE Visualization Conf., VIS 2022*. IEEE, New York, NY, United States, 31–35.

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

[14] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. 2018. Local rule-based explanations of black box decision systems. arXiv:1805.10820

[15] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. arXiv:1812.04608

[16] Hans Hofmann. 1994. Statlog (German Credit Data). UCI Machine Learning Repository.

[17] Adulam Jeyasothy, Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, and Marcin Detyniecki. 2022. Integrating Prior Knowledge in Post-hoc Explanations. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, Davide Ciucci, Inés Couso, Jesús Medina, Dominik Ślęzak, Davide Petturiti, Bernadette Bouchon-Meunier, and Ronald R. Yager (Eds.). Springer Int. Publishing, Cham, 707–719.

[18] Amir-Hossein Karimi, Julius von Kügelgen, Bernhard Schölkopf, and Isabel Valera. 2022. Towards Causal Algorithmic Recourse. In *Int. Conf. on Machine Learning, ICML 2022 - Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*. Springer, Cham, 139–166.

[19] Mark T. Keane, Eoin M. Kenny, Eoin Delaney, and Barry Smyth. 2021. If Only We Had Better Counterfactual Explanations: Five Key Deficits to Rectify in the Evaluation of Counterfactual XAI Techniques. In *Proc.of the Thirtieth Int. Joint Conf. on Artificial Intelligence, IJCAI-21*, Zhi-Hua Zhou (Ed.). International Joint Conferences on Artificial Intelligence, 4466–4474. https://doi.org/10.24963/ijcai.2021/609 Survey Track.

[20] René F. Kizilcec. 2016. How Much Information? Effects of Transparency on Trust in an Algorithmic Interface. In *Proc. of the 2016 CHI Conf. on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, United States, 2390–2395. https://doi.org/10.1145/2858036.2858402

[21] Matev Kunaver and Toma Porl. 2017. Diversity in Recommender Systems A Survey. *Know.-Based Syst.* 123, C (may 2017), 154–162. https://doi.org/10.1016/j.knosys.2017.02.009

[22] Himabindu Lakkaraju and Osbert Bastani. 2020. "How Do I Fool You?": Manipulating User Trust via Misleading Black Box Explanations. In *Proc. of the AAAI/ACM Conf. on AI, Ethics, and Society* (New York, NY, USA) *(AIES '20)*. Association for Computing Machinery, New York, NY, USA, 79–85. https://doi.org/10.1145/3375627.3375833

[23] Michael T. Lash, Qihang Lin, Nick Street, Jennifer G. Robinson, and Jeffrey Ohlmann. 2017. Generalized Inverse Classification. In *Proc.of the 2017 SIAM Int. Conf. on Data Mining, SDM2017*. SIAM, Philadelphia, PA, United States, 162–170. https://doi.org/10.1137/1.9781611974973.19

[24] Thibault Laugel, Marie Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. 2019. The dangers of post-hoc interpretability: Unjustified counterfactual explanations. In *Proc. of the Int. Joint Conf. on Artificial Intelligence, IJCAI'19*. International Joint Conferences on Artificial Intelligence, 2801–2807.

[25] Thai Le, Suhang Wang, and Dongwon Lee. 2020. GRACE: Generating Concise and Informative Contrastive Sample to Explain Neural Network Model's Prediction. In *Proc. of the 26th ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining*. Association for Computing Machinery, New York, NY, United States, 238–248.

[26] Zachary C. Lipton. 2016. The Mythos of Model Interpretability. In *Proc. of the Int. Conf. on Machine Learning, ICML'16 - Workshop on Human Interpretability in Machine Learning*. Association for Computing Machinery, New York, NY, United States, 36–43.

[27] Arnaud Van Looveren and Janis Klaise. 2021. Interpretable counterfactual explanations guided by prototypes. In *Joint European Conf. on Machine Learning and Knowledge Discovery in Databases*. Springer, 650–665.

[28] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proc. of the Int. Conf of Advances in Neural Information Processing Systems, NeurIPS'17*. Curran Associates Inc., Red Hook, NY, United States, 4765–4774.

[29] Divyat Mahajan, Chenhao Tan, and Amit Sharma. 2019. Preserving causal constraints in counterfactual explanations for machine learning classifiers. arXiv:1912.03277

[30] Christian Meske and Enrico Bunde. 2022. Design principles for user interfaces in AI-Based decision support systems: The case of explainable hate speech detection. *Information Systems Frontiers* (2022), 1–31.

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

[32] Sina Mohseni, Niloofar Zarei, and Eric D Ragan. 2018. A survey of evaluation methods and measures for interpretable machine learning. arXiv:1811.11839

[33] Christoph Molnar. 20120. *Interpretable Machine Learning - A Guide for Making Black Box Models Explainable.*

[34] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proc.of the 2020 Conf. on fairness, accountability, and transparency, FAccT 2020*. Association for Computing Machinery, New York, NY, United States, 607–617.

[35] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. 2020. FACE: feasible and actionable counterfactual explanations. In *Proc.of the AAAI/ACM Conf. on AI, Ethics, and Society*. Association for Computing Machinery, New York, NY, United States, 344–350.

[36] Yanou Ramon, Tom Vermeire, Olivier Toubia, David Martens, and Theodoros Evgeniou. 2021. Understanding Consumer Preferences for Explanations Generated by XAI Algorithms. arXiv:2107.02624

[37] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should i trust you?" Explaining the predictions of any classifier. In *Proc. of the 22nd ACM SIGKDD Int. Conf. on knowledge discovery and data mining, KDD 16*. Association for Computing Machinery, New York, NY, USA, 1135–1144.

[38] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Proc. of the AAAI Conf. on artificial intelligence*, Vol. 32. AAAI Press, Palo Alto, CA, United States, Article 187, 9 pages.

[39] Ruoxi Shang, K. J. Kevin Feng, and Chirag Shah. 2022. Why Am I Not Seeing It? Understanding Users' Needs for Counterfactual Explanations in Everyday Recommendations. In *Proc.of the 2022 Conf. on fairness, accountability, and transparency, FAccT 2022* (Seoul, Republic of Korea). Association for Computing Machinery, New York, NY, United States, 1330–1340. https://doi.org/10.1145/3531146.3533189

[40] Ben Shneiderman. 2020. Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 10, 4 (2020), 1–31.

[41] Leonard J Simms, Kerry Zelazny, Trevor F Williams, and Lee Bernstein. 2019. Does the number of response options matter? Psychometric perspectives using personality questionnaire data. *Psychological assessment* 31, 4 (2019), 557.

[42] Kacper Sokol and Peter A Flach. 2018. Glass-Box: Explaining AI Decisions With Counterfactual Statements Through Conversation With a Voice-enabled Virtual Assistant.. In *Proc.of the 27th Int. Joint Conf. on Artificial Intelligence, IJCAI-18*. International Joint Conferences on Artificial Intelligence, 5868–5870.

[43] Sahil Verma, John Dickerson, and Keegan Hines. 2020. Counterfactual explanations for machine learning: A review. https://doi.org/10.48550/ARXIV.2010.10596

[44] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. 2017. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law* 7, 2 (2017), 76–99.

[45] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proc. of the Int. Conf. on Human Factors in Computing Systems, CHI'19*. Association for Computing Machinery, New York, NY, United States, 1–15.

[46] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th Int. Conf. on Intelligent User Interfaces* (College Station, TX, United States) *(IUI '21)*. Association for Computing Machinery, New York, NY, United States, 318–328. https://doi.org/10.1145/3397481.3450650

[47] Greta Warren, Mark T Keane, and Ruth MJ Byrne. 2022. Features of Explainability: How users understand counterfactual and causal explanations for categorical and continuous features in XAI. https://doi.org/10.48550/ARXIV.2204.10152

[48] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. 2019. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 56–65.

[49] Wencan Zhang and Brian Y Lim. 2022. Towards Relatable Explainable AI with the Perceptual Process. In *Proc.of the 2022 CHI Conf. on Human Factors in Computing Systems* (New Orleans, LA, United States) *(CHI '22)*. Association for Computing Machinery, New York, NY, United States, Article 181, 24 pages. https://doi.org/10.1145/3491102.3501826

A TESTED INTERFACES

We present the two other versions of the interface we use for the user study, as presented in 5.3.2 *(Note: The interfaces have been translated from the original language used for the evaluation).*

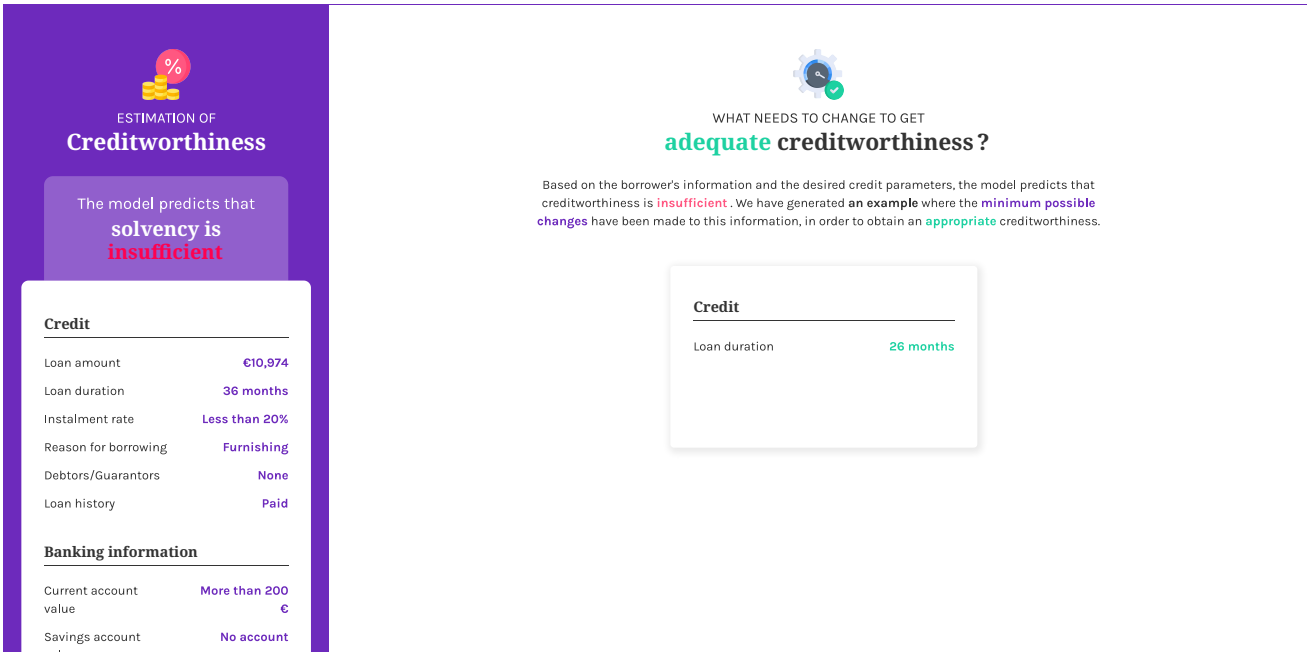

Figure 4: Interface A is the baseline version with a single counterfactual example.

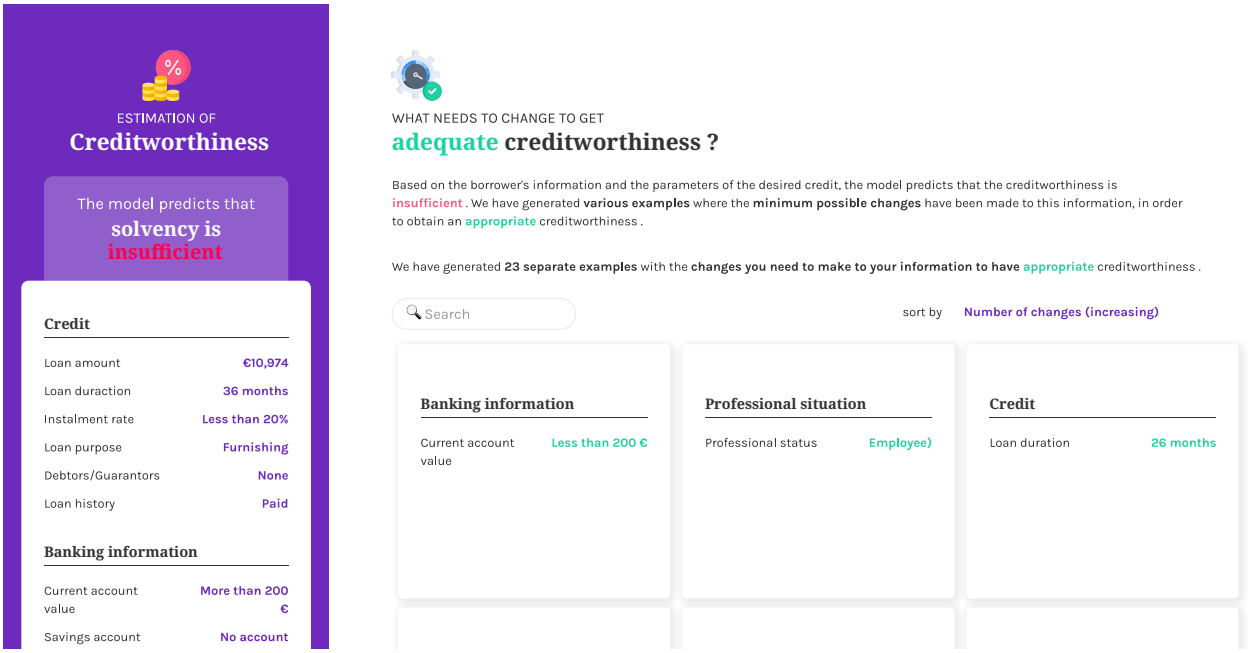

Figure 5: Interface B is the XUI we propose for implementing counterfactual explanations with plural examples

B OBJECTIVE UNDERSTANDING QUESTIONNAIRE

This section lists the questions asked to the user-lab experiment participants to evaluate the proposed interface, translated from the original language. In all cases, except for the open-response question, the participant must choose between three answers:

- I agree with Swann
- I disagree with Swann
- I don't know

B.1 Explanations' nature questions: counterfactual examples

- Swann thinks that the information displayed indicates what variations can be made on his/her information, to be predicted as having an adequate solvency.
- Swann thinks that the proposed changes are always on the parameters of his/her credit application (amount, duration, loan rate, and so on).
- Swann thinks that all the provided information needs to be changed, in addition to the proposed changes, to be predicted as having an adequate solvency.
- Swann thinks this system proposes changes to be predicted as having an adequate solvency.

B.2 Explanations' effects questions: if...then...

- Swann thinks that if the loan term was 20 months instead of 36 months, his/her solvency would have been predicted as adequate.
- Swann thinks that for the solvency to be predicted as adequate, the loan term could be reduced by 10 months.
- Swann thinks that with a loan rate of 22%, the solvency would be predicted as adequate.
- Swann believes that his/her solvency would have been predicted as adequate if there had been a co-borrower.
- Swann thinks that the solvency would have been predicted as adequate if he/she was not a foreigner.

B.3 Explanations' specificity question: plurality

- Swann believes that the only way to be predicted as having an adequate solvency would be to reduce the loan duration to 26 months.
- Swann thinks that the least common changes suggested concern the employment status.
- Swann thinks that among all the proposed changes, some are more feasible than others.
- Swann thinks that he/she would have had to be at least 54 years old in order to be predicted as having an adequate solvency.
- Swann believes that for all the examples provided, only one or two pieces of information would need to be changed as indicated in order to be predicted as having an adequate solvency

B.4 Open-response question

What strategy/changes would you recommend to Swann to make his/her solvency to be predicted as adequate?

C SATISFACTION QUESTIONNAIRE

This section presents the self-reporting questionnaire we propose, adapted from the Explanation Satisfaction Scale [15] in order to assess users' satisfaction (translated from the original language).

C.1 Explanation Satisfaction Scale adapted

- In your opinion, the explanations for obtaining appropriate creditworthiness are understandable
- In your opinion, the explanations for obtaining appropriate credit are satisfying
- In your opinion, the explanations for obtaining appropriate credit are sufficiently detailed
- In your opinion, the explanations for obtaining appropriate credit are complete
- In your opinion, the proposed explanations indicate how they should be interpreted to fully understand how to obtain appropriate credit
- In your opinion, the explanations for obtaining appropriate credit are useful in helping you make an informed decision
- In your opinion, the explanations for obtaining appropriate credit are accurate
- n your opinion, the explanations for obtaining appropriate credit are trustworthy

C.2 Open-response question

How satisfied are you with the explanations the interface provides to achieve appropriate creditworthiness?

Received 10 October 2022; revised 23 January 2023; accepted 3 February 2023