


Leveraging Causal Relations to Provide Counterfactual Explanations and Feasible Recommendations to End Users

Riccardo Crupi¹, Beatriz San Miguel González², Alessandro Castelnovo¹ and Daniele Regoli¹ ^a

¹*Intesa Sanpaolo S.p.A., Turin, Italy*

²*Fujitsu Research of Europe, Madrid, Spain*

Keywords: Explainable Artificial Intelligence, Counterfactual Explanations, Causality, Recourse.

Abstract: Over the last years, there has been a growing debate on the ethical issues of Artificial Intelligence (AI). Explainable Artificial Intelligence (XAI) has appeared as a key element to enhance trust of AI systems from both technological and human-understandable perspectives. In this sense, counterfactual explanations are becoming a *de facto* solution for end users to assist them in acting to achieve a desired outcome. In this paper, we present a new method called Counterfactual Explanations as Interventions in Latent Space (CEILS) to generate explanations focused on the production of feasible user actions. The main features of CEILS are: it takes into account the underlying causal relations by design, and can be set on top of an arbitrary counterfactual explanation generator. We demonstrate how CEILS succeeds through its evaluation on a real dataset of the financial domain.

1 INTRODUCTION


Note-worthy governmental initiatives, such as the General Data Protection Regulation (GDPR) (The European Union, 2016) and Ethics guidelines for trustworthy Artificial Intelligence (AI) (High-Level Expert Group on AI, 2019) in Europe, and the Defence Advanced Research Projects Agency’s Explainable Artificial Intelligence (XAI) program of the United States (Gunning and Aha, 2019), endeavour to promote the creation of trustworthy AI systems based on human oversight, prevention of harm, transparency, interpretability, accountability, etc. In this sense, the XAI field has appeared as a crucial set of technologies to improve and ensure trustworthiness of AI systems.

XAI addresses different purposes sought by the stakeholders of AI systems (Arrieta et al., 2020). XAI can provide reasons and justifications for the whole logic of an AI model or a specific outcome, considering both technical and non-technical forms. For example, AI developers can take advantage of explanations to verify that AI outcomes are not erroneous, biased or insecure; to ensure the efficiency and functionality; and to get new insights to improve the system. Moreover, non-technical profiles (i.e. regula-

tors, domain experts, executives and end users) can receive explanations to assess and certify regulatory compliance, gain business knowledge and get insights to their specific situation in a human-understandable way.

In this paper, we present CEILS¹: Counterfactual Explanations as Interventions in Latent Space, a new method to generate counterfactual explanations and recommendations. Counterfactual explanations (Wachter et al., 2017) are a set of statements to communicate to end users what should change in their features in order to receive a desired result. These are gaining large acceptance in technical, legal, and business contexts (Barocas et al., 2020). Their advantages include: help end users whose life is impacted by automatic decisions to interact with AI systems, do not disclose technical details of the models, thus protect trade secrets and commercial interests and is appropriate to legal frameworks.

While there are significant efforts to generate counterfactual explanations (Verma et al., 2020; Stepin et al., 2021), they generally fall short of generating feasible actions that end users should carry out in practice. CEILS, on the other hand, is designed to leverage the underlying causal relations to generate

^a  <https://orcid.org/0000-0003-2711-8343>

¹CEILS code is publicly available at the repository <https://github.com/FLE-ISP/CEILS>

feasible recommendations to end users on how to act to reach a desired outcome. Moreover, CEILS does that by building on top of an arbitrary counterfactual generator, thus avoiding the need of dealing with an *ad hoc* optimization problem.

It is important to point out that counterfactual explanations and counterfactuals in the causal inference field are separated concepts. Generally speaking, counterfactual explanations do not account for any causal relationship among data, and these are not to be confused with counterfactuals in the causal inference setting (i.e. instances answering to questions of the form “*what would have happened if...?*”) (Pearl et al., 2000). However, CEILS, by explicitly considering the underlying causal structure, provides a bridge between these two otherwise unrelated concepts (see Section 3.1.4 for more details).

The main contributions of this work are:

- to provide a novel and straightforward way to account for causality in generating counterfactual explanations of AI models,
- to provide counterfactual explanations along with (*causally*) *feasible actions* to reach them.

The structure of this paper is the following. First, Section 2 summarizes prior work for the generation of counterfactual explanations. Then, Section 3 details the CEILS proposal and Section 4 describes its evaluation in a real dataset of the financial domain. Finally, Section 5 concludes the paper and outlines future work.

2 RELATED WORK

Most of the prior approaches for the generation of counterfactual explanations relies on establishing an optimization problem, usually to find the nearest counterfactual with respect to the input to be explained. This has been understood in different ways: minimizing the distance among the explanation and the original instance (*proximity*) (Wachter et al., 2017) or the training data (*data manifold closeness*) (Joshi et al., 2019), keeping a low number of feature changes (*sparsity*) (White and Garcez, 2019), adhering to observed correlations or generating a set of *diverse* explanations (Mothilal et al., 2020).

Moreover, other works incorporate *causality* to produce explanations more grounded with reality (Mahajan et al., 2019). In general, current solutions provide a common understanding of a decision. However they fall short of assisting end users with feasible recommendations to act and achieve a desired outcome (Karimi et al., 2020). A set of pro-

posals addresses this last issue through the definition of a recourse problem (Ustun et al., 2019; Karimi et al., 2020). Our work is focused on the production of counterfactual explanations with recommendations and actions more realistic and feasible for end users.

3 PROPOSED METHOD

Figure 1 depicts a general overview of the CEILS context and its main building blocks. Like other counterfactual explanation methods, CEILS offers to end user a set of statements to indicate what should be changed in order to achieve a desired outcome. However, CEILS produces not only explanations, but also feasible actions.

By way of example, consider a simplified scenario of an AI model used for loan approvals that takes into account age, income and credit score of an end user to decide if a loan would be accepted or rejected. When a person is denied a loan, counterfactual explanations indicate what input data (features) should be altered to access to the loan. For instances, the person may receive as counterfactual explanations: “*reduce your age*”, “*increase your income and credit score*”, “*improve your credit score value*”, etc. As can be deduced from the examples, these explanations contain unfeasible and impractical actions (i.e. for end users, it is impossible to reduce their age, or usually they have no control over their credit scores values). With our proposal, feasible actions are delivered to end users due to the fact that causal relations are taken into account by design, e.g. credit score is in general influenced by age and income and age can only increase. Therefore, using CEILS, a person could receive as a recommended action “*increase your income*”, which would impact the credit score and guarantee the loan approval.

The CEILS methodology consists of three main steps:

1. creation of a model in a latent space (i.e. space of unobserved variables) through a Structural Causal Model – comprised by a Causal Graph and a set of Structural Equations;
2. generation of counterfactual explanations with an arbitrary generator using the aforementioned model; and
3. translation of counterfactual explanations to the original feature space.

It is important to note that our method requires as inputs the historical data used to build the AI model that will be explained and the causal graph that describes the causal relations among the features of the

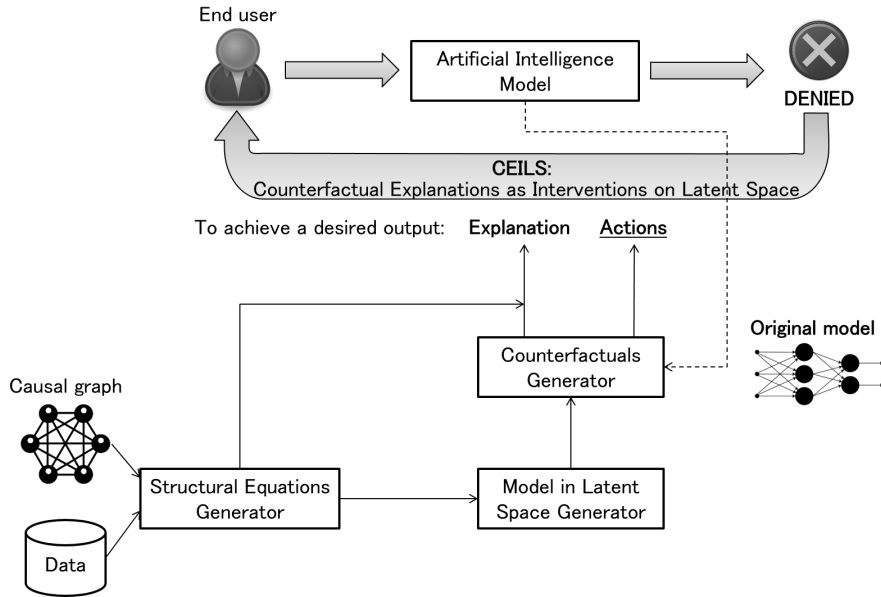


Figure 1: CEILS context and building blocks. The context in which CEILS operates is similar to other explainability methods: a user interacting with an AI model receiving a negative outcome together with an explanation of the outcome (provided by CEILS). Unlike other methods, CEILS provides not only explanations but also realistic actions. The main building blocks of CEILS are shown: Data, Casual Graph, Structural Equations Generator, Model in Latent Space Generator and Counterfactuals Generator.

dataset. Moreover, CEILS needs to have access to the original model to generate the explanations.

On the bottom part of Figure 1 the main CEILS building blocks are shown. Next, we detail each block of the proposal.

3.1 CEILS: Building Blocks

3.1.1 Causal Graph

CEILS requires to access to a predefined causal graph that encodes the causal relations among the features of the dataset. Modeling this causal knowledge is complex and challenging since this requires understanding relations beyond statistical dependencies. Different causal discovery algorithms have been proposed to identify causal relationships from observational data through automatic methods (Kalainathan et al., 2020; Kalisch et al., 2012). In general, it is important that domain experts validate the relations detected by the causal discovery routine, or include new ones when deemed necessary.

Causal relationships among features are modelled via a Directed Acyclic Graph (DAG) $G = (V, E)$, with the set V of vertices (or nodes) and the set E of directed edges (or links). Nodes of the graph G are composed by:

- $X = (X_1, \dots, X_d)$: endogenous variables, representing the actual (observed) variables used as predictors in the model;

- $U = (U_1, \dots, U_d)$: exogenous variables, representing (unobserved) factors not accounted for by the features X ; and
- Y represents the dependent variable to be predicted by means of X .

In Figure 2 an example of a causal graph is included, pointing out the relations among the different kind of nodes (X, U, Y).

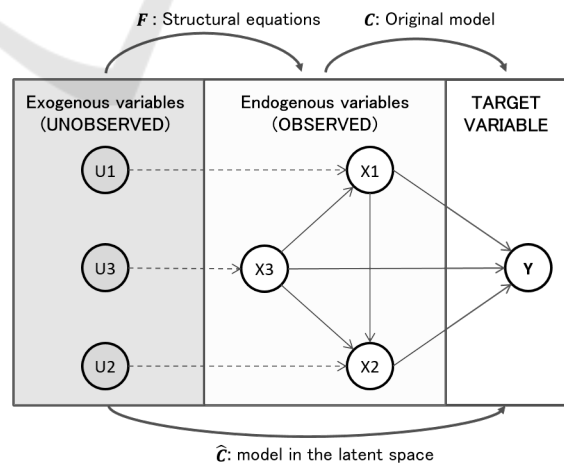


Figure 2: Example of a causal graph, indicating the different kind of nodes (U, X , and Y), and the functions associated with the CEILS methodology (F, C and \hat{C}).

3.1.2 Structural Equations

The Structural Causal Model encodes a more detailed description of causal relationships. This is defined by a triplet (X, U, F) where F represent a set of functions called Structural Equations $F : \mathcal{U} \rightarrow X$, mapping the exogenous (unobserved) variables living in the domain \mathcal{U} to the endogenous (observed) ones $X \ni X = F(U)$.

The first step of the CEILS methodology is to infer the Structural Equation from observations (X) , given the DAG. It does so by assuming an Additive Noise Model (see, e.g., (Peters et al., 2014)):

$$X_j = f_j(\mathbf{pa}(X_j)) + U_j, \quad \forall j = 1 \dots d. \quad (1)$$

where $\mathbf{pa}(X_j)$ denotes the parent nodes of X_j with respect to the DAG. Thus, in the example of Figure 2 the variable X_2 , whose parents are the nodes X_1 and X_3 , would be modeled by the equation: $X_2 = f_2(X_1, X_3) + U_2$.

A regressor model \mathcal{M}_j is trained to estimate each function f_j , predicting a variable X_j from its parent nodes. A node without parents (e.g. X_3 in Figure 2) is referred to as a root node, while residuals are an estimate of the latent variables: namely, to calculate the value of the latent variable U given X one would compute:

$$\begin{cases} \hat{U}_j = X_j, & \text{for all roots } j, \\ \hat{U}_j = X_j - \mathcal{M}_j(\mathbf{pa}(X_j)), & \text{for all non-roots } j. \end{cases} \quad (2)$$

Therefore, for the example of the Figure 2, the estimated value of the latent variable U_3 would be equal to X_3 since this has no parents, while for the variable X_2 — with parents X_1 and X_3 — the estimation of the latent variable would be $\hat{U}_2 = X_2 - \mathcal{M}_2(X_1, X_3)$.

Structural Equations for root nodes r simply reduce to $F_r(U) = U_r$. Once all models \mathcal{M}_j are learned from a training dataset, it is possible to recursively compute the actual function F connecting U to X — namely $X = F(U)$ — following the causal flow in the DAG (i.e from the root nodes down to the leaves):

$$\begin{cases} F_j(U) = U_j, & \text{for all roots } j, \\ F_j(U) = \mathcal{M}_j(\{F_v(U)\}_{v \in \mathbf{pa}(X_j)}) + U_j, & j \text{ non-root.} \end{cases} \quad (3)$$

Thus, starting from root nodes — for which the relation is trivial — one can recursively construct the full relation $F : U \mapsto X$. In the example of Figure 2 equation (3) would read:

$$X = F(U) = \begin{pmatrix} U_3 \\ \mathcal{M}_1(U_3) + U_1 \\ \mathcal{M}_2(\mathcal{M}_1(U_3) + U_1, U_3) + U_2 \end{pmatrix}. \quad (4)$$

3.1.3 Model in the Latent Space

Once the relations $X = F(U)$ are available, it is possible to build the model:

$$\hat{C}(U) = C \circ F(U), \quad (5)$$

where C is the original model that predicts the target variable Y given X , of which we need to provide counterfactual explanations. For example, this could be an arbitrary machine learning classifier. Therefore, \hat{C} has domain in the (latent) space of variables U and leverages the Structural Causal Model by means of F in order to replicate C in estimating Y .

Figure 2 shows the relations among the functions F , C and \hat{C} in a toy example: the original model C and the Structural Equations F , are composed to get the model in the latent space \hat{C} .

3.1.4 Counterfactuals Generator

Given the latent-space model \hat{C} as by equation (5), an arbitrary counterfactual explanations generator can be employed. As we mentioned in Section 2, most approaches to generate counterfactual explanations rely on an optimization process based on different metrics (i.e. proximity, sparsity, etc.).

More precisely, given an instance x for which we want to find a counterfactual explanation, CEILS prescribes the following three steps:

1. compute the latent variables u corresponding to x by means of residuals of models $\{\mathcal{M}_j\}_{j=1}^d$ (eq. (2));
2. run a counterfactual generator for the model \hat{C} relative to the point u , thus obtaining $u^{\text{cf}} = u + \delta$;
3. compute $x^{\text{cf}} = F(u + \delta)$.

Notice that these three points correspond precisely to the well-known steps of counterfactual computation, e.g. as described in chapter 4 of (Pearl et al., 2016), namely:

1. *abduction*, i.e. update the unobserved variables to account for the observed data $X = x$;
2. *action* — the crucial step — that consists in intervening on variables to change the observed values;
3. *prediction*, i.e. use the new variables to actually compute the counterfactual instance.

Namely, the CEILS approach — similarly to (Karimi et al., 2020) — provides a natural bridge between the two separate concepts of XAI counterfactual explanations and causal counterfactuals in the sense of (Pearl et al., 2016).

It is important to point out that the main advantage of CEILS is that it provides not only the counterfactual explanation x^{cf} , but also the *action* δ needed to reach it. In particular, in standard counterfactual explanation methods the action is seen as a shift *in feature space*, but this kind of actions represent, in general, *unfeasible* recommendations, since they completely neglect the fact that a change in a feature has impacts on others. Our method, instead, recommending actions as shift in the *latent space*, takes into account the underlying causal flow. Incidentally, notice that the latent space action δ is an example of *soft intervention* (Eberhardt and Scheines, 2007) in that it is performed on top of other changes in the variable due to changes in its parents.

4 EVALUATION

We demonstrate the advantages of CEILS on a real dataset of the financial domain. The evaluation of the results is based on a set of known metrics and new ones that we propose to capture the particularities of CEILS.

4.1 Dataset and Causal Graph

We use a proprietary dataset of past loan applications (Castelnovo et al., 2020). This dataset comprises 220,304 applications and 8 features (namely gender, age, citizenship, monthly income, bank seniority, requested amount, number of installments and rating) to determine whether the loan application is accepted or rejected. The features are related according to the causal graph (Figure 3) that we obtain using different causal discovery algorithms. In particular, we employed the Python Causal Discovery Tool-Box (Kalainathan and Goudet, 2019), including different graph modelling algorithms on observational data (i.e. SAM, PC) and the NOTEARS algorithm (Zheng et al., 2018) included in the Python library CausalNex. Additionally, a manual revision has been performed by a group of domain experts to validate each causal relation detected or to include new ones.

Among the features, rating indicates the creditworthiness of an applicant, i.e. an estimation of the probability that a customer will repay loans in a timely manner. Usually, end users cannot directly intervene on this feature, being it a complicated function of other variables. Thus, we set up our experiment constraining rating in such a way that no direct intervention is made on it. Namely, rating is considered a feature non-actionable but that can vary due to changes in other variable. Gender and citizenship, on

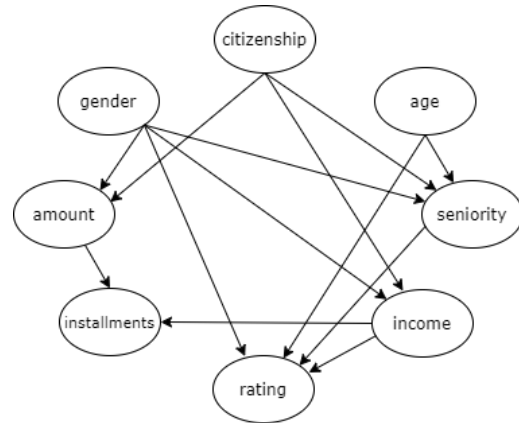


Figure 3: Causal graph related to the proprietary loan applications dataset.

the other hand, are constrained to be immutable features (they cannot change in any way), meanwhile age and bank seniority can only increase.

4.2 Experiments Setup

First, we build the original AI model, which would be explained, using the 8 features of the dataset to predict the target variable that indicates if a loan application will be accepted or rejected. This model is established through a feed-forward neural network with 2 hidden layers with ReLU activation functions.

On the other hand, the Structural Causal Model relies on the DAG represented in Figure 3. Additionally, Structural Equations are calculated to map unobserved variables to the 8 observed variables using feed-forward neural networks with 2 hidden layers. Similarly, the model in the latent space is based on a feed-forward neural network with 2 hidden layers with ReLU activation functions.

The popular open source Machine Learning library, TensorFlow, is used for setting up and training the neural networks.

We rely on counterfactual explanations guided by prototypes proposed by (Van Looveren and Klaise, 2019) included in the open source library Alibi (Klaise et al., 2019) as our baseline generator of counterfactuals explanations. In particular, this is used as counterfactuals generator for the CEILS method (check Figure 1 and Section 3.1.4 for more details) and on the other hand, to generate counterfactuals explanations that will be compared with the explanations generated by CEILS.

Therefore, we compute the corresponding counterfactual explanations first using the baseline approach (Van Looveren and Klaise, 2019) and then overlaying our proposed CEILS method. 1,000 ran-

Table 1: Comparison results of the baseline approach and the CEILS method based on 8 metrics.

	baseline	CEILS
validity	22%	82%
continuous proximity	289.57 ± 830.79	43.23 ± 109.46
categorical proximity	0.0 ± 0.0	0.09 ± 0.15
sparsity	2.86 ± 0.95	2.83 ± 1.17
sparsity action	-	2.28 ± 1.04
distance	2.16 ± 1.1	1.72 ± 0.87
cost	2.51 ± 1.24	1.35 ± 0.81
feasibility	0.064	1.0

dom instances of the dataset are used as factual observations to be explained.

As mentioned above, gender and citizenship are treated as immutable features while rating can vary only as a consequence of changes in other variables and cannot be directly acted upon. To implement this, in CEILS we simply keep the corresponding latent variables fixed (gender and citizenship are root nodes, thus fixing their latent counterpart is equivalent to constrain them to be immutable), while in the baseline approach we directly fix the variables in the X space. This means that, for the baseline approach, rating is effectively immutable, while in CEILS it is non-actionable but mutable.

The two sets of explanations are evaluated based on state-of-the-art metrics, such as validity, proximity and sparsity (refer to (Mothilal et al., 2020) for more details). Moreover, we define three new metrics (i.e. distance, cost, and feasibility) to capture the particularities of our approach. Next, we explain the meaning of each metrics with the report of the results.

4.3 Results

Table 1 summarizes the results obtained for both methods (baseline and CEILS) with respect to the different metrics: validity, proximity (for continuous and categorical features), sparsity (including the action provided by CEILS), distance, cost and feasibility.

First, **validity** refers to the fraction of generated explanations that are valid counterfactuals, i.e. that are given a different outcome y with respect to the factual instance. Here, we can observe the first big difference among the two approaches: only 22% of instances have an associated counterfactual explanation with the baseline approach, while 82% of explanations are found with the CEILS method.

Intuitively the aforementioned difference in the validity is explained by the rating feature, which is crucial to estimate the granting of loans and we configure it to be non directly actionable, since end users cannot control it. This explains why the baseline

method falls short in providing valid counterfactual explanations: this approach has no way of changing the rating, thus for it is either impossible to find actual counterfactual explanations or they are too far to be considered valid. On the other hand, CEILS is much more efficient in providing valid explanations, since it can indirectly act on rating by changing variables that causally impact it (i.e. seniority or income). Moreover, CEILS provides to end users the action to be performed on these variables that are needed to change the rating appropriately.

To clarify the differences among the methods, consider the example included in Table 2, where an application has been rejected and the counterfactual explanation indicates how to act to have it approved ($decision = 0 \rightarrow decision = 1$). As expected, immutable features (i.e. gender and citizenship) do not vary their values, while age and bank seniority show equal or higher values with respect to the original instance. However, the baseline method produces a counterfactual explanation with values far away from the factual profile (i.e. increase the income to 3643.3K and almost double the amount requested with less number of installments). On the other hand, if we focus on the action recommended by CEILS, this only suggests to increase the bank seniority and the requested amount². Increasing the bank seniority (action seniority = 5.4) results in a better rating (ACEILS rating = -0.375), thus, ultimately, in loan approval. Evidently, an increase in the seniority is impossible without a corresponding increase in age: actually, we have treated bank seniority as an actionable feature, but it would have been more appropriate to consider it as mutable only as a consequence of age changes, since seniority cannot be controlled independently of age. Nevertheless, we have decided to keep seniority actionable to focus our discussion on rating and not to limit too much the baseline method (for which it would have been impossible to change seniority as well as rating).

Regarding the rest of metrics in Table 1, it is important to note that they are computed over the explanations common to both methods (22% according to the validity), i.e. on observations to which both CEILS and the baseline were able to provide a valid explanation. In particular, results included in the table correspond to the mean and standard deviation of the metrics over all valid explanations.

²Both methods apparently provide the counter-intuitive suggestion of increasing the requested amount: this is due to the fact that the baseline method searches for explanations as close as possible to the data distribution. In other words, a too small requested amount is not plausible with respect to the other suggested features.

Table 2: Example of an instance, counterfactual explanations obtained by the baseline approach and CEILS method, differences among the original instance and the counterfactual explanations for both methods, and action provided by CEILS method.

variable	instance	counterfactuals		Δ baseline	Δ CEILS	CEILS
		baseline	CEILS			action
decision	0	1	1			
gender	1	1	1	0	0	0
age	30	30	30	0	0	0
citizenship	1	1	1	0	0	0
income	56.7K	3700K	56.7K	3643.3K	0	0
seniority	0	8.1	5.4	8.1	5.4	5.4
amount	210K	409K	320K	199K	110K	110K
installments	48	33.4	53.1	-14.6	5.1	0
rating	10	10	9.625	0	-0.375	0

Thus, **proximity** refers to the distance between the original instance and the counterfactual explanation. We distinguish among proximity for continuous and categorical features. For continuous features, we measure proximity as feature-wise L_1 distance rescaled by the Median Absolute Deviation from the median (MAD). For categorical features, we consider a distance of 1 if there is values mismatch. The results show better values of proximity taking into account continuous features for the CEILS method. This means that the explanations obtained are closer to the original input of the end user, which is a desirable propriety. Regarding the proximity for categorical features, there is no major difference among the methods.

Moreover, we measure the **sparsity** of the explanations obtained as the number of features that need to change with respect to the original input³. In the feature space the difference among the methods is not remarkable, but notice that in terms of recommended actions (“sparsity action” in Table 1) CEILS is able to slightly improve this metric as well.⁴

Closely related to the proximity metric, the **distance** measure the L_1 distance between counterfactual explanation and the original instance⁵. Again, the CEILS method obtains better values (1.72) with respect to the baseline method (2.16).

Finally, we propose to measure cost and feasibility: these 2 metrics are designed to compare CEILS actions with the actions that the baseline approach would have recommended considering the Structural

³For continuous feature a tolerance threshold is considered similarly to (Mothilal et al., 2020).

⁴Notice that for non-causal counterfactual explanations actions actually boil down to the simple difference in feature space, thus sparsity and sparsity action coincide. Another way to compute actions for non-causal explanations could be what we call *ex-post* actions, mentioned below in the text.

⁵Notice that this metric is computed over standardized feature values.

Causal Model *ex post*, i.e. the actions — given the SCM — that one should perform in order to reach the baseline explanations.

On the one hand, **cost** is defined as the L_1 norm of the action that has to be done in order to reach a counterfactual explanation⁶. In the evaluation, a lower value of cost is obtained with the CEILS method since the causal influence reduces the effort in order to reach an explanation.

On the other hand, the **feasibility** metric shows the percentage of actions that are compatible with the feasibility constraints over features, e.g. if the variable bank seniority can only increase, the action over bank seniority must be a positive value for the corresponding explanation to be feasible. We observe that the CEILS method perfectly preserves the underlying causality as expected (a value of 1.0), while the reconstructed actions of the baseline spoil almost completely the actionability of the features (0.064) (e.g. to keep the rating fixed while changing income, the baseline should actually recommend a non-null action on rating, which is unfeasible).

5 CONCLUSIONS AND FUTURE WORK

In this paper, we present the CEILS methodology for generating counterfactual explanations focused on providing feasible actions to end users who want to achieve a desired outcome. The main novelty, which is reached via the Structural Causal Model and by building a model in the latent space, lies in the possibility of embedding it in any existing counterfactual generator, effectively producing realistic recommendations.

Indeed, because of variable interdependence, static counterfactual explanations generators (in the sense of non-causal) fail, in general, to provide feasible actions, recommending a set of interventions that may be either impossible to perform or sub-optimal in reaching a desired result, as outlined in Section 4 (check the example included in Table 2). On the other hand, CEILS leverages the underlying causal relationships among variables to provide recommendations that are feasible and also to reduce the effort in terms of actions to reach valid counterfactual profiles (e.g. acting on seniority to influence rating in the example of Table 2).

Our first practical evaluation employs a real dataset of the financial domain and is based on well-

⁶Notice that this metric is computed over standardized feature values.

known metrics and others that capture the particularities of CEILS that confirm the efficiency of our method in generating feasible actions with respect to its baseline counterfactual generator.

Despite the growing research on the field of counterfactual explanations, there are a lot of open questions and challenges yet to be tackled (Verma et al., 2020). In particular, we are interested in relaxing the assumption of having a complete and reliable causal graph and work with incomplete causal relations. Moreover, as for future work, we consider to extend our evaluation by employing other counterfactual generators as baselines to analyze how it could contribute to the overall results, and also to compare the explanations produced by other causal methods with the ones found with CEILS, and possibly involving end users to obtain feedback that will guide towards better explanations. The preprint (Crupi et al., 2021) includes detailed evaluations and experiments on additional datasets.

REFERENCES

- Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115.
- Barocas, S., Selbst, A. D., and Raghavan, M. (2020). The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 80–89.
- Castelnovo, A., Crupi, R., Del Gamba, G., Greco, G., Naseer, A., Regoli, D., and Gonzalez, B. S. M. (2020). Befair: Addressing fairness in the banking sector. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 3652–3661. IEEE.
- Crupi, R., Castelnovo, A., Regoli, D., and San Miguel González, B. (2021). Counterfactual explanations as interventions in latent space. *arXiv preprint arXiv:2106.07754*.
- Eberhardt, F. and Scheines, R. (2007). Interventions and causal inference. *Philosophy of science*, 74(5):981–995.
- Gunning, D. and Aha, D. (2019). DARPA’s explainable artificial intelligence (XAI) program. *AI Magazine*, 40(2):44–58.
- High-Level Expert Group on AI (2019). Ethics guidelines for trustworthy AI. <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>.
- Joshi, S., Koyejo, O., Vijitbenjaronk, W., Kim, B., and Ghosh, J. (2019). Towards realistic individual recourse and actionable explanations in black-box decision making systems. *arXiv preprint arXiv:1907.09615*.
- Kalainathan, D. and Goulet, O. (2019). Causal discovery toolbox: Uncover causal relationships in python. *arXiv preprint arXiv:1903.02278*.
- Kalainathan, D., Goulet, O., and Dutta, R. (2020). Causal Discovery Toolbox: Uncovering causal relationships in Python. *Journal of Machine Learning Research*, 21(37):1–5.
- Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H., and Bühlmann, P. (2012). Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, 47:1–26.
- Karimi, A.-H., Schölkopf, B., and Valera, I. (2020). Algorithmic recourse: from counterfactual explanations to interventions. *arXiv preprint arXiv:2002.06278*.
- Klaise, J., Van Looveren, A., Vacanti, G., and Coca, A. (2019). Alibi: Algorithms for monitoring and explaining machine learning models. <https://github.com/SeldonIO/alibi>.
- Mahajan, D., Tan, C., and Sharma, A. (2019). Preserving causal constraints in counterfactual explanations for machine learning classifiers. *arXiv preprint arXiv:1912.03277*.
- Mothilal, R. K., Sharma, A., and Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617.
- Pearl, J. et al. (2000). Models, reasoning and inference. Cambridge, UK: Cambridge University Press, 19.
- Pearl, J., Glymour, M., and Jewell, N. P. (2016). *Causal inference in statistics: A primer*. John Wiley & Sons.
- Peters, J., Mooij, J. M., Janzing, D., and Schölkopf, B. (2014). Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(58).
- Stepin, I., Alonso, J. M., Catala, A., and Pereira-Fariña, M. (2021). A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, 9:11974–12001.
- The European Union (2016). EU General Data Protection Regulation (GDPR): Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union. <http://data.europa.eu/eli/reg/2016/679/2016-05-04>.
- Ustun, B., Spangher, A., and Liu, Y. (2019). Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 10–19.
- Van Looveren, A. and Klaise, J. (2019). Interpretable counterfactual explanations guided by prototypes. *arXiv preprint arXiv:1907.02584*.
- Verma, S., Dickerson, J., and Hines, K. (2020). Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596*.

- Wachter, S., Mittelstadt, B., and Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841.
- White, A. and Garcez, A. d. A. (2019). Measurable counterfactual local explanations for any classifier. *arXiv preprint arXiv:1908.03020*.
- Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. (2018). Dags with no tears: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, pages 9472–9483.

