# Feasibility in counterfactual explanations

Kleopatra Markou
klmark@di.uoa.gr
National and Kapodistrian University of Athens
Athens, Greece

Dimitrios Tomaras, Vana Kalogeraki
{tomaras,vana}@aueb.gr
Athens University of Economics and Business
Athens, Greece

Dimitrios Gunopulos
dg@di.uoa.gr
National and Kapodistrian University of Athens
Athens, Greece

## ABSTRACT

The imminent need to interpret the output of a Machine Learning model with counterfactual (CF) explanations – via small perturbations to the input – has gained significant research interest. Although a variety of alternatives from CF examples is important, being feasible at the same time does not necessarily apply in their entirety. This work uses different datasets to examine through the preservation of the causal relations of their attributes, whether CF examples can be created, be feasible and actually useful to the end-user in real-world cases. In our evaluation we used four commonly used datasets and we managed to generate feasible CF examples that satisfy all possible predefined causal constraints and confirmed the importance of causal relations between the attributes in a dataset.

**ACM Reference Format:**
Kleopatra Markou, Dimitrios Tomaras, Vana Kalogeraki, and Dimitrios Gunopulos. 2024. Feasibility in counterfactual explanations. In . ACM, New York, NY, USA, 1 page.

## 1 INTRODUCTION

Training machine learning models in distributed systems can affect their performance and add bias to the model output. Even simple transformations can impact models learnt in a distributed setting and this phenomenon is exacerbated in such systems. Thus, researchers often ask themselves whether the output of a Machine Learning (ML) model can be interpretable and applicable to the real world. Counterfactual explanations [2] are often used as a type of explanations as they are consistent with the ML model and can be interpretable. As CF explanations we consider a representation of the small perturbations of an input feature, that can lead to a change in the prediction of the model. For instance: "what an individual should change in order to be granted with a loan that now cannot get?". All possible what-if scenarios form different CF explanations[2]. Are all these scenarios applicable to the real world? Feasibility can determine this answer, since the scenarios stem from real-world applications. A way to quantify feasibility is through a causal model, granted from the relations between the attributes of a dataset. A CF explanation is feasible if the changes satisfy the constraints entailed by the causal model [2]. Different datasets have been used to examine the important role of causal relations between their attributes, by creating feasible CF examples.

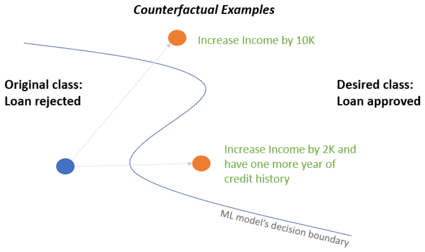

**Figure 1: Counterfactual example**

## 2 RESEARCH PROBLEM

Suppose that we design constraints that capture the feasibility of a CF, using a causal model or even basic domain knowledge, for a certain dataset and its attributes. For example, if an individual wants to take a loan, a CF example decreasing the "age", will be considered infeasible since it violates the natural causal constraint that age can only increase with time. As a first step in our approach, a black box model (with two linear layers) is used to classify the input data into two classes. This satisfies the definition of a CF example of always having the input and the desired class. The result will be later used in the validity loss function as a pretrained model. As a final step, a Variational Autoencoder (VAE) [1] will generate feasible CF examples. The desired result will be CF examples that satisfy all the given causal constraints as well as a feasibility score. The highest the feasibility score, the better the training of our model in the distributed system. To conclude, our experimental evaluation has shown that causal relations of the attributes play a significant role to the performance, as the more complex the causal relations between the attributes the more complex the satisfaction of the constraints. This validates our intuition for more realistic results using binary constraints rather than using unary constraint models.

## ACKNOWLEDGMENTS

This research has been financed by the European Union through the H2020 AutoFair project (No. 101070568).

## REFERENCES

[1] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[2] Sandra Wachter and et al. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* 31 (2017), 841.

*Conference'17, July 2017, Washington, DC, USA*

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM