

Feature-based Learning for Diverse and Privacy-Preserving Counterfactual Explanations

Vy Vo
Monash University, Australia

He Zhao, Edwin V. Bonilla
CSIRO's Data61, Australia

Trung Le, Van Nguyen
Monash University, Australia

Gholamreza Haffari, Dinh Phung
Monash University, Australia

ABSTRACT

Interpretable machine learning seeks to understand the reasoning process of complex black-box systems that are long notorious for lack of explainability. One flourishing approach is through counterfactual explanations, which provide suggestions on what a user can do to alter an outcome. Not only must a counterfactual example counter the original prediction from the black-box classifier but it should also satisfy various constraints for practical applications. Diversity is one of the critical constraints that however remains less discussed. While diverse counterfactuals are ideal, it is computationally challenging to simultaneously address some other constraints. Furthermore, there is a growing privacy concern over the released counterfactual data. To this end, we propose a feature-based learning framework that effectively handles the counterfactual constraints and contributes itself to the limited pool of private explanation models. We demonstrate the flexibility and effectiveness of our method in generating diverse counterfactuals of actionability and plausibility. Our counterfactual engine is more efficient than counterparts of the same capacity while yielding the lowest re-identification risks.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Security and privacy** → **Privacy protections**.

KEYWORDS

Explainable AI, Algorithmic Recourse, Privacy

ACM Reference Format:

Vy Vo, Trung Le, Van Nguyen, He Zhao, Edwin V. Bonilla, and Gholamreza Haffari, Dinh Phung. 2023. Feature-based Learning for Diverse and Privacy-Preserving Counterfactual Explanations. In *Proceedings of 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23)*, August 6–10, 2023, Long Beach, CA, USA (KDD '23). ACM, New York, NY, USA, 17 pages. <https://doi.org/XXXXXXX.XXXXXXX>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

KDD '23, August 6–10, 2023, Long Beach, CA, USA

© 2023 Association for Computing Machinery.

ACM ISBN 978-1-4503-XXXX-X/23/08...\$15.00

<https://doi.org/XXXXXXX.XXXXXXX>

1 INTRODUCTION

The eminence of deep neural networks in recent years has proliferated the use of machine learning in various real-world applications. Such models provide remarkable predictive performance yet often at a cost of transparency and interpretability. This has sparked controversy over whether to rely on algorithmic predictions for high-stakes decision making such as graduate admission [1, 50], job recruitment [3], credit assessment [26] or criminal justice [15]. Progress in interpretable machine learning offers interesting solutions to explaining the underlying behavior of black-box models. One useful approach is through counterfactual examples¹, which sheds light on what modifications to be made to an individual's profile that can counter an unfavorable decision outcome from a black-box classifier. Such explanations explore what-if scenarios that suggest possible recourses for future improvement. Counterfactual explainability indeed has important social implications at both personal and organizational levels. For instance, feedback like 'getting 1 more referral' or 'being fluent in at least 2 languages' would help unsuccessful candidates better prepare for future job applications. By advocating for transparency in decision making, organizations can improve their attractiveness to top talents while inspecting possible prejudice implicitly introduced in historical data and consequentially embedded in the classifiers producing biased decisions.

The ultimate goal of this line of research is to provide realistic guidelines as to what actions an individual can take to achieve a desired outcome. Desiderata of counterfactual explanations have been extensively discussed in previous literature [18, 21, 48, 49]. To be of practical use, a counterfactual explanation should at least satisfy the following characteristics:

- **Validity:** By definition, a counterfactual example must change the original black-box outcome to a desired one.
- **Sparsity:** Counterfactuals should be close to the original example where a minimal number of features are modified.
- **Actionability:** Counterfactual explanations should only suggest actionable or feasible changes. In particular, changes should be made on mutable features e.g., Work Experience or SAT scores, while leaving immutable features unchanged e.g., Gender or Ethnicity.
- **Diversity:** Diverse explanations are preferable to capture different preferences from the same user so that they can freely explore multiple options to select the best fit.

¹Counterfactual explanation is sometimes referred to as algorithmic recourse.

- *Plausibility*²: Plausible or realistic counterfactuals are to obey the input domain and constraints within/among features. For example, Age cannot decrease or be above 200.
- *Scalability*: Inference should be done simultaneously and efficiently for multiple input examples.

Among these desiderata, *diversity* emerges as a non-trivial property to address. Given an instance, a diverse counterfactual engine returns a set of different counterfactual profiles that should all lead to the desired outcome. Ensuring that the entire explanation set satisfies *validity* while dealing with constraints given by *actionability* and *plausibility* poses a computational challenge. *Scalability* becomes another important consideration mainly due to the fact that most of the existing approaches process counterfactuals separately for each input data point. Furthermore, strongly enforcing *sparsity* results in a smaller subset of features that can be changed. This hence can compromise *diversity* since we expect counterfactual states to differ from one to another substantially. On the other hand, there has been a growing concern over the privacy risks of model explanation [32, 43, 45]. Aïvodji et al. [2] points out that diverse counterfactual explanations make the system more vulnerable as the released examples reveal the model decision boundaries and could disclose sensitive information such as health conditions or financial data. A *linkage attack* is one such malicious attempt, which refers to the action of recovering the identity (i.e., re-identifying) of an anonymized record in the published dataset using background knowledge. It is often done by linking records to an external dataset of the population based on the combination of several attributes [12, 28, 41]. Netflix \$1M Machine Learning Contest is a notorious data breach, in which the company disclosed a dataset of 100 million subscribers with their movie ratings and preferences. Narayanan and Shmatikov [34] revealed a successful attack of 68% that was easily achieved by cross-referencing the users' dates and precise ratings of 2 movies with a non-anonymous dataset published by IMDb (Internet Movie Database).

Despite an overwhelming number of counterfactual explanation approaches, only a few works tackle diverse counterfactual generation [5, 22, 33, 38, 40]. However, the trade-offs between *diversity* and the aforementioned constraints, including privacy protection, have not been well studied in previous papers (See Table 1 for comparison). Filling this gap, our work proposes a novel learning-based framework that effectively addresses all the above desiderata while mitigating the re-identification risk. From a methodological perspective, **our method diverges markedly from existing approaches in the following ways:**

Firstly, we reformulate the combinatorial search task into a stochastic optimization problem to be solved via gradient descent. Unlike most previous methods that perform optimization per-input basis, we employ amortized inference to generate diverse counterfactual explanations efficiently. Amortization has been previously adopted wherein a counterfactual generative distribution is modelled via Markov Decision Processes [49] or Variational Autoencoders [9, 30, 36]. On one hand, none of these amortized methods addresses *diversity*. On the other, we here take a different approach: we construct a learnable generation module that directly models

the conditional distributions of individual features such that they form a valid counterfactual distribution when combined.

Another point of difference of ours lies in the usage of Bernoulli sampling to ensure *sparsity*. In prior works, standard metrics such as L1 or L2 are often used to penalize the distance between the counterfactual and original data point. Verma et al. [48] criticizes this approach as unnatural, especially for categorical features. Avoiding the use of distance measures, we optimize along a feature selection module to output the likelihood of the feature being mutated. This module can be adapted to any user-defined constraints about the mutability of features.

Finally, we go beyond existing approaches by tackling the constraint of privacy preservation exposed to diverse explanations. The key strategy is to discretize continuous features and operate the counterfactual generation engine in the categorical feature space. Discretization is closely related to the generalization technique used in privacy-preserving data mining (PPDM) [12, 31]. It is also treated as a subroutine to analyze the composition of differential privacy algorithms [14, 17]. The idea is that it is by nature easier to uniquely identify a profile based on continuous features, so discretization is expected to increase the quantities of profiles linked back to a certain group of attributes. Another defense effort against linkage attack can be found in [16]. The paper proposes an algorithm named CF-K that heuristically searches for an equivalence class for each counterfactual instance such that every record is indistinguishable from at least $k - 1$ others. CF-K is viewed as an add-on that theoretically can be implemented on top of any counterfactual generative system. However, in practice, this strategy is extremely expensive since it requires repetitively querying the model explainer for a possibly larger number of counterfactuals than requested. Though sharing the same motivation, we here contribute a counterfactual explanation model with a built-in privacy preservation functionality.

Our contributions can be summarized as follows:

- We introduce **Learning to Counter (L2C)** - a stochastic feature-based approach for learning counterfactual explanations that address the counterfactual desirable properties in a single end-to-end differentiable framework.
- Through extensive experiments on real-world datasets, L2C is shown to balance the counterfactual trade-offs more effectively than the existing methods and achieve diverse explanations with the lowest re-identifiability risk. To the best of our knowledge, L2C is the first amortized engine that supports diverse counterfactual generations with privacy-preservation capability.

2 RELATED WORKS

Recent years have seen an explosion in the literature on counterfactual explainability, from works that initially focused on one or two characteristics or families of models to those that can deal with multiple constraints and various model types. There have been many attempts to summarize major themes of research and discuss open challenges in great depth. We therefore refer readers to [18, 21, 48] for excellent surveys of methods in this area. We here focus on reviewing algorithms that can support diverse (or at least multiple) local counterfactual generations.

²The terms *plausibility* and *feasibility* are often used interchangeably.

Table 1: Desiderata comparison of related counterfactual explanation methods. *Privacy refers to whether the output data is protected against linkage attack or re-identification risk. L2C satisfies all of these critical constraints.

Method	Sparsity	Actionability	Diversity	Plausibility	Scalability	Privacy*
L2C (Ours)	✓	✓	✓	✓	✓	✓
DICE [33]	✓	✓	✓			
COPA [5]	✓	✓	✓			
MCCE [38]	✓	✓	✓		✓	
Coherent CF [40]	✓		✓			
MACE [20]	✓	✓	✓			
MOC [8]	✓	✓		✓		
CERTIFAI [42]		✓				
Feasible-VAE [30]	✓	✓		✓	✓	
FastAR [49]	✓	✓		✓	✓	
CRUDS [9]		✓		✓	✓	
C-CHVAE [36]		✓			✓	
CF-K [16]						✓

Dealing with the combinatorial nature of the task, earlier works commonly adopt mixed integer programming [40], genetic algorithms [42], or SMT solvers [20]. Another recent popular approach is gradient-based optimization [5, 33], which involves iteratively perturbing the input data point according to an objective function that incorporates desired constraints. The whole idea of *diversity* is to explore different combinations of features and feature values that can counter the original prediction while accommodating various user needs. To support *diversity*, Russell [40] in particular enforces hard constraints on the current generations to be different from the previous ones. Such a constraint will however be removed whenever the solver cannot be satisfied. Meanwhile, Mothilal et al. [33] and Bui et al. [5] add another loss term for *diversity* using Determinantal Point Processes [25], whereas the other works only demonstrate the capacity to generate multiple counterfactuals via empirical results. All of the aforementioned algorithms are computationally expensive in that input data points are handled singly and individual runs are additionally required to produce several counterfactuals. Reducing computational costs, Redelmeier et al. [38] attempts to model the conditional likelihood of mutable features given the immutable features using the training data. They then adopt Monte Carlo sampling to generate counterfactuals from this distribution and filter out samples that do not meet counterfactual constraints. Given such a generative distribution, sampling of counterfactuals can therefore be done straightforwardly. Amortized optimization is another strategy to improve inference speed [9, 30, 36, 49].

In response to the privacy warning about model explanations [32, 43, 45], several defense strategies have been introduced to alleviate the risks. With strong theoretical guarantees, differential privacy [11] stands out as the promising solution to preventing *member inference attack* and *model stealing attack* [2, 32, 44]. With regards to linkage attacks, CF-K [16] is the only work we are aware of that tackles linkage attack in counterfactual explanations.

3 STOCHASTIC FEATURE-BASED COUNTERFACTUAL LEARNING

3.1 Problem setup

Let \mathcal{X} denote the input space where $\mathbf{x} = [x_i]_{i=1}^N$ is an input vector with N features of both continuous and categorical types. As discussed previously, we discretize the continuous features into equal-sized buckets, which gives us an input of N categorical features wherein each feature x_i has c_i levels. We apply one-hot encoding on each feature and flatten them into a single input vector $\mathbf{z} \in \{0, 1\}^D$ where $D = \sum_{i=1}^N c_i$. Concretely, feature x_i is now represented by the vector $\mathbf{z}_i \in \mathbb{O}_{c_i}$ where the set of one-hot vectors \mathbb{O}_{c_i} is defined as $\{0, 1\}^{c_i} : \sum_{j=1}^{c_i} z_{ij} = 1$.

Let f be the black-box classifying function and $y = f(\mathbf{x})$ be the decision outcome on the input \mathbf{x} . A valid counterfactual example $\tilde{\mathbf{x}}$ associated with \mathbf{x} is one that alters the original outcome y into a desired outcome $y' \neq y$ with $y' = f(\tilde{\mathbf{x}})$. Let $\tilde{\mathbf{z}}$ denote the corresponding one-hot representation of $\tilde{\mathbf{x}}$.

Actionability indicates that some features can be *mutable* (i.e., changeable), while others should be kept *immutable* (i.e., unchangeable). Without loss of generality, let us impose an ordering on the set of N features such that the first K features are mutable features (i.e., the ones that can be modified) and denote $\mathbb{K} := \{1, \dots, K\} \subset \{1, \dots, N\}$. For each mutable feature (i.e., x_i or the one-hot vector \mathbf{z}_i with $i \in \mathbb{K}$), we aim to learn a local feature-based perturbation distribution $P(\tilde{\mathbf{z}}_i | \mathbf{z})$ where $\tilde{\mathbf{z}}_i \in \mathbb{O}_{c_i}$, while leaving the immutable features unchanged.

It is worth noting that our method functions equally well on heterogeneous data where only categorical features are one-hot encoded while continuous features are retained at their original values. However, we believe that performing data discretization (or generalization in terms of PPDm) initially and deploying the classifiers in the discrete feature space would provide better privacy protection. For the purpose of comparing our prototype with existing approaches, we follow the standard practice of explaining classification models trained on the mixed dataset of continuous and categorical features. To make it compatible with the discretization subroutine of our framework, we represent the prediction on a transformed (fully categorical) input vector with the prediction

on the input where the categorical values associated with mutable continuous features are substituted with the middle point of the corresponding intervals. We refer to this mechanism as **one-hot decoding**, which will be detailed shortly.

3.2 Methodology

We now detail how L2C works and addresses each counterfactual constraint. The explanation pipeline of L2C is depicted in Figure 1.

For each mutable feature z_i with $i \in \mathbb{K}$, we learn a **local feature-based perturbation distribution** $P(\tilde{z}_i | z)$ (i.e., $\tilde{z}_i \in \mathbb{O}_{c_i}$), which is a categorical distribution $\text{Cat}(\mathbf{p}_i | z)$ with category probability $\mathbf{p}_i = [p_{i1}, p_{i2}, \dots, p_{ic_i}]$. We form a counterfactual example \tilde{z} by concatenating $\tilde{z}_i \sim \text{Cat}(\mathbf{p}_i | z)$ for the mutable features and z_i for the immutable features. To achieve *validity*, we learn the local feature-based perturbation distribution by maximizing the chance that the counterfactual examples \tilde{z} counter the original outcome on \mathbf{x} . Additionally, learning local feature-based perturbation distributions over the mutable features allows us to conduct a global counterfactual distribution $P(\tilde{z} | z)$ over the counterfactual examples \tilde{z} defined above. Sampling from this distribution naturally leads to multiple counterfactual generations efficiently, and we also expect that individual samples \tilde{z}_i together can form diverse combinations of features, thereby promoting *diversity* within the generative examples.

As previously discussed, too much of *diversity* can compromise *sparsity*. Dealing with this constraint, for each mutable feature z_i , we propose to learn a **local feature-based selection distribution** that generates a random binary variable $s_i \sim \text{Bernoulli}(\pi_i | z)$ wherein we replace z_i by $\tilde{z}_i \sim \text{Cat}(\mathbf{p}_i | z)$ if $s_i = 1$ and leave $\tilde{z}_i = z_i$ if $s_i = 0$. Therefore, the formula to update \tilde{z}_i is

$$\tilde{z}_i = (1 - s_i)z_i + s_i\tilde{z}_i.$$

The benefit of having $\boldsymbol{\pi} = [\pi_i]_{i \in \mathbb{K}}$ is thus to control *sparsity* by adding one more channel to decide if we should modify a mutable feature z_i . Appendix D presents an ablation study showing that without the selection distribution, the perturbation distribution alone can generate diverse counterfactuals yet requires changing plenty of mutable features. Meanwhile, optimizing the selection distribution jointly helps harmonize the trade-off between *diversity* and *sparsity*.

3.3 Optimization Objective

In this section, we explain how to design the building blocks of our framework L2C. As shown in Figure 1, our framework consists of two modules: a **counterfactual generator** \mathcal{G} and a **feature selector** \mathcal{S} . The **counterfactual generator** \mathcal{G} is used to model the feature-based perturbation distribution, while **feature selector** \mathcal{S} is employed to model the feature-based selection distribution.

Specifically, given a one-hot vector representation \mathbf{z} of a data example \mathbf{x} , we feed \mathbf{z} to \mathcal{G} to form $\mathcal{G}(\mathbf{z}) = [\mathcal{G}_i(\mathbf{z})]_{i \in \mathbb{K}}$. We then apply the softmax activation function to $\mathcal{G}_i(\mathbf{z})$ to define the feature-based local distribution (i.e., $\text{Cat}(\mathbf{p}_i | z)$) for z_i as

$$p_{ij}(z) = \frac{\exp\{\mathcal{G}_{ij}(z)\}}{\sum_{k=1}^{c_i} \exp\{\mathcal{G}_{ik}(z)\}}, \forall j = 1, \dots, c_i. \quad (1)$$

The module \mathcal{S} takes \mathbf{z} to form $\mathcal{S}(\mathbf{z}) = [\mathcal{S}_i(\mathbf{z})]_{i \in \mathbb{K}}$. We then apply the Sigmoid function to $\mathcal{S}_i(\mathbf{z})$ to define the feature-based selection distribution (i.e., $\text{Bernoulli}(\pi_i | z)$) for z_i as

$$\pi_i(z) = \frac{1}{1 + \exp\{-\mathcal{S}_i(z)\}}.$$

To encourage *sparsity* by reducing the number of mutable features chosen to be modified, we regularize \mathcal{S} through L1-norm $\|\boldsymbol{\pi}(\mathbf{z})\|_1$ with $\boldsymbol{\pi}(\mathbf{z}) = [\pi_i(\mathbf{z})]_{i \in \mathbb{K}}$.

To summarize, given a one-hot vector representation \mathbf{z} of a data example \mathbf{x} , we use \mathcal{G} to work out the local feature-based perturbation distribution $\text{Cat}(\mathbf{p}_i(\mathbf{z}))$ for every $i \in \mathbb{K}$. We then sample $\tilde{z}_i \sim \text{Cat}(\mathbf{p}_i(\mathbf{z}))$ for every $i \in \mathbb{K}$. Subsequently, we use \mathcal{S} to work out the local feature-based selection distribution $\text{Bernoulli}(\pi_i(\mathbf{z}))$ for every $i \in \mathbb{K}$. We then sample $s_i \sim \text{Bernoulli}(\pi_i | z)$ and update $\tilde{z}_i = (1 - s_i)z_i + s_i\tilde{z}_i$ for every $i \in \mathbb{K}$. Finally, we concatenate \tilde{z}_i for $i \in \mathbb{K}$ and z_i for $i \notin \mathbb{K}$ to form the counterfactual example \tilde{z} .

\mathcal{G} and \mathcal{S} are parameterized with neural networks over total parameters θ . For \tilde{z} to be a *valid* and *sparse* counterfactual associated with a desired outcome y' , we propose the following criterion

$$\min_{\theta} \left[\mathbb{E}_{\tilde{z}} [\text{CE}(f(\tilde{z}), y')] + \alpha \mathbb{E}_{\tilde{z}} [\|\boldsymbol{\pi}(\mathbf{z})\|_1] \right], \quad (2)$$

where f is the black-box function, CE is the cross-entropy loss, $\|\cdot\|_1$ is L1-norm, α is a loss weight.

One-hot decoding. Recall that \tilde{z} formed by concatenating many one-hot vectors is an incompatible representation to the classifier f , which in fact requires both continuous and one-hot features. We make a design choice of reconstructing the continuous features by taking the middle point of the range corresponding to the selected level. Specifically, the input to the one-hot decoder is $\tilde{\mathbf{z}} = [\tilde{z}_i]_{i=1}^N$. If the feature i originally is a categorical feature, we set $\tilde{x}_i = \tilde{z}_i$. Otherwise, we set $\tilde{x}_i = a_i + \frac{(2k-1)(b_i-a_i)}{2c_i}$, which is the middle point of the k -th interval $[a_i + (k-1)(b_i-a_i)/c_i, a_i + k(b_i-a_i)/c_i]$ where $[a_i, b_i]$ is the original value range of the feature i and \tilde{z}_i corresponds to the level $k \in \{1, \dots, c_i\}$ (i.e., $\tilde{z}_{ik} = 1$ and $\tilde{z}_{ij} = 0$ if $j \neq k$). Note that one-hot decoding is only applied when a continuous feature is indicated by \mathcal{S} to be mutable ($s_i = 1$). Otherwise, we revert to the original continuous value. Formally, we rewrite

$$\tilde{x}_i = (1 - s_i)x_i + s_i \sum_{j=1}^{c_i} \tilde{z}_{ij} \left[a_i + \frac{(2j-1)(b_i-a_i)}{2c_i} \right]. \quad (3)$$

To assure model differentiability for training, the one-hot vector \tilde{z}_{ij} is relaxed into its continuous representation by using Gumbel-Softmax reparametrization trick, which is detailed in Section 3.4.

The final optimization objective is now given as

$$\min_{\theta} \left[\mathbb{E}_{\tilde{\mathbf{x}}} [\text{CE}(f(\tilde{\mathbf{x}}), y')] + \alpha \mathbb{E}_{\tilde{\mathbf{x}}} [\|\boldsymbol{\pi}(\mathbf{z})\|_1] \right]. \quad (4)$$

Plausibility. A counterfactual generative engine needs to ensure explanations are realistic for real-world applications. There are two common types of plausibility constraints: **Unary** and **Binary** monotonicity constraints. The former deals with individual features (e.g., Age cannot decrease) while the latter is concerned with the correlation of a pair of features (e.g., increasing in Education level increases Age). To handle binary constraints on heterogeneous data is not as straightforward as unary constraints. We thus delay the

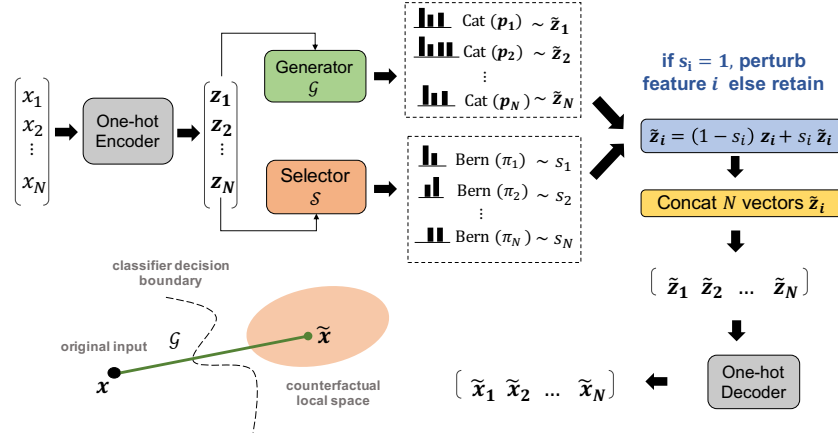


Figure 1: For illustration purposes only, all features are assumed mutable in the figure. We first discretize the continuous features of input x and one-hot encode all features into representations z . For every feature i , the generator \mathcal{G} learns a local perturbation distribution $\text{Cat}(p_i|z)$ so that together they form a distribution of diverse counterfactual representations \tilde{z} . Simultaneously, the selector \mathcal{S} learns to output the distribution Multi-Bernoulli($\pi|z$) capturing the probability of each feature i being modified. Every feature sample pair (\tilde{z}_i, s_i) is passed through an operation in the blue box, which decides whether to accept the change being made to the feature i given by \tilde{z}_i . The output is then decoded into the representations \tilde{x} compatible with the black-box system. \mathcal{G} and \mathcal{S} are jointly trained via back-propagation according to Eq. (4). Intuitively, \mathcal{G} aims to construct a “bridge” across the decision boundary travelling from the input to a local space of counterfactuals.

discussion on binary constraints until Section 5.4 and focus on unary constraints in the main analysis.

Dealing with such a constraint, one can simply eliminate any counterfactuals violating the constraints during inference time. This however creates additional computational overhead and may compromise some desiderata. There are only few attempts addressing feature constraints in counterfactuals, notably Mahajan et al. [30], Verma et al. [49]: Verma et al. [48] incorporate hard conditions in a Markov Decision process where a feature gets updated only if the corresponding action does not violate the constraints. Meanwhile, Mahajan et al. [30] includes a hinge loss into the loss function for unary features, while specifically learning a separate linear model for every feature pair subject to a binary constraint. For L2C, the learnable local distributions can be used for this purpose conveniently. Our proposed strategy is to impose rule-based unary constraints on related features in the optimization process. Technically, for every feature to be perturbed with a non-decreasing (non-increasing) constraint, we penalize the probabilities corresponding to lower (higher) levels towards zero by multiplying them with a positive infinitesimal quantity. Concretely, given a mutable feature i under monotonic constraints, let $l \in \{1, \dots, c_i\}$ denote the current state - the level corresponding to z_i (i.e., $z_{il} = 1$ and $z_{ij} = 0$ if $j \neq l$). Let us denote the restricted set of levels as $C_i = \{1, \dots, l-1\}$ if the feature is non-decreasing and $C_i = \{l+1, \dots, c_i\}$ if it is non-increasing. The perturbation distribution $\text{Cat}(p_i | z)$ for z_i given in Eq. (1) now becomes

$$p_{ij}(z) \propto \epsilon^{1_{C_i}(j)} \times \frac{\exp\{\mathcal{G}_{ij}(z)\}}{\sum_{k=1}^{c_i} \exp\{\mathcal{G}_{ik}(z)\}}, \forall j = 1, \dots, c_i, \quad (5)$$

where $1_{C_i}(\cdot)$ is the indicator function such that $1_{C_i}(j) = 1$ if $j \in C_i$ and $1_{C_i}(j) = 0$ otherwise, meaning that the probabilities

at the other levels are untouched. We here explicitly force the model to generate more samples at the higher (lower) levels while maintaining differentiability of the objective function. We choose $\epsilon = e^{-10}$ in our experiments, but any positive value arbitrarily close to zero would suffice.

3.4 Reparameterization for Continuous Optimization

Our L2C involves multiple sampling rounds back and forth to optimize the networks. To make the process continuous and differentiable for training, we adopt the reparameterization tricks [19, 29]:

1) *Sampling $\tilde{z}_i \sim \text{Cat}(p_i | z)$.* : To obtain differentiable counterfactual samples, we adopt the classic temperature-dependent Gumbel-Softmax trick [19, 29]. Given the categorical variable z_i with category probability $[p_{i1}, p_{i2}, \dots, p_{ic_i}]$. The relaxed representation is sampled from the Categorical Concrete distribution as $\tilde{z}_i \sim \text{Cat-Concrete}(\log p_{i1}, \dots, \log p_{ic_i})$ by

$$\tilde{z}_{ij} = \frac{\exp\{(\log p_{ij}(z) + G_j)/\tau\}}{\sum_{k=1}^{c_i} \exp\{(\log p_{ik}(z) + G_k)/\tau\}},$$

with temperature τ , random noises G_j independently drawn from Gumbel distribution $G_t = -\log(-\log u_t)$, $u_t \sim \text{Uniform}(0, 1)$. As discussed, we apply this mechanism consistently to the one-hot representations of all features. The continuous relaxation of Eq. (3) can be gained by simply using the one-hot relaxation \tilde{z}_i .

2) *Sampling $s_i \sim \text{Bernoulli}(\pi_i | z)$.* : We again apply the Gumbel-Softmax trick to relax Bernoulli variables of 2 categories. With temperature τ , random noises G_{i0} and $G_{i1} \sim G_t = -\log(-\log u_t)$, $u_t \sim \text{Uniform}(0, 1)$, the continuous representation s_i is sampled from

Binary Concrete distribution as $s_i \sim \text{Bin-Concrete}(\pi_i, 1 - \pi_i)$ by

$$s_i = \frac{\exp\{(\log \pi_i(z) + G_{i1})/\tau\}}{\exp\{(\log(1 - \pi_i(z)) + G_{i0})/\tau\} + \exp\{(\log \pi_i(z) + G_{i1})/\tau\}}.$$

4 EXPERIMENTAL SETUP

We experiment with 4 popular real-world datasets: German Credit [10], Adult Income [24] Graduate Admission [1] and Student Performance [7]. For each dataset, we select a fixed subset of immutable features based on our domain knowledge and suggestions from [49]. We reserve the privacy analysis for German Credit and Adult Income datasets, which contain personal financial information and various attributes through which data subjects can be re-identified [16]. While implementing the black-box classifiers and the baseline methods, we standardize numerical features to unit variance and one-hot encode categorical features. Note again that, for our method only, we discretize numerical features into equal-sized buckets and decode the numerical features back to their original representations whenever necessary to consult the black-box model. Appendix A describes our tasks and model design in greater detail. Our code repository can be anonymously accessed at <https://anonymous.4open.science/r/L2C-AD51/>.

Table 2: Description of quantitative evaluation metrics. \mathbb{C} denotes a set of counterfactual examples generated by an algorithmic recourse approach for a given input instance.

Desiderata	Metric	Description
Validity	Validity	Proportion of samples in \mathbb{C} can counter the original black-box decision outcome.
	Coverage	Coverage = 100% if there exists at least 1 valid counterfactual in \mathbb{C} .
Sparsity/ Actionability	Sparsity	Proportion of features kept unchanged, averaged over the number of samples in \mathbb{C} .
Diversity	Diversity	Hamming distance of a pair of counterfactual samples across all features where numerical features are discretized. The metric is averaged over all pairs of samples in \mathbb{C} .
Sparsity – Diversity Balance	Harmonic mean	F-measure of Diversity and Sparsity = $2 \cdot \text{Diversity} \cdot \text{Sparsity} / (\text{Diversity} + \text{Sparsity})$.
Plausibility	Unary	Proportion of examples \mathbb{C} meeting the unary monotonic constraints, averaged over the number of features subject to constraints.

Performance metrics. Following the past works Mothilal et al. [33], Redelmeier et al. [38], Verma et al. [49], Table 2 outlines the commonly used metrics for quantitatively assessing the desirability of counterfactual explanations. As for *diversity*, a widely adopted measure is the pairwise distance between counterfactual examples, with distance defined separately for numerical and categorical features [33, 38]. Though this approach is meaningful for interpreting categorical features, we however find it quite obscure for numerical features. This motivates us to discretize numerical features again when computing Diversity, which captures how often a feature gets altered as well as how much the change is - specifically via how often it switches to a different categorical level. The computation of Diversity only considers valid counterfactuals, so if valid counterfactuals are none, Diversity is set to zero. It is worth noting that there fundamentally exists a trade-off between *sparsity* and

diversity. To quantify how well a method can balance these two properties, we suggest taking Harmonic mean of Diversity and Sparsity, motivated by the development of F1-score in measuring Precision against Recall. For metrics used in privacy analysis, refer to Section 5.2.

Evaluation setup. We consider a general setting of binary classification where a counterfactual outcome y' is opposite to the original outcome y , whether y has label 1 or 0. From each method, we generate a set of 100 counterfactual explanations. During generation, most methods, including ours, require multiple iterations of searching for the optimal set of counterfactuals based on the optimization constraints. To assure a fair comparison on efficiency, a global maximum time budget of 5 minutes is imposed to search for a set of 100 counterfactuals per input sample. We compare our method top-performing baselines that support diverse counterfactual explanations: DICE [33], MCCE [38] and COPA [5]. DICE offers several search strategies: Random, KD tree, or Genetic algorithm. DICE-KDTree was consistently reported to fail across datasets [49], so we exclude it from our evaluation. We do not consider MACE [20] since it is extremely expensive on large datasets [49] and often fails to converge in our experiments, nor MOC [8] due to the lack of Python implementation.

5 RESULTS AND DISCUSSION

5.1 Counterfactual Explanation Desiderata

We first study whether an algorithmic recourse approach generates a set of diverse counterfactuals without sacrificing the other desiderata. Note that COPA has only been shown to work effectively on linear classifiers. Table 3 reports the average results over 5 model initializations. Appendix E provides several illustrative examples for qualitative assessment.

Under the same time budget, our method L2C succeeds in generating 100% valid counterfactuals with full coverage. Together with DICE, L2C first satisfies the most important criterion of a counterfactual explanation and resolves the trade-off against *validity*. Recall that we have specified a fixed set of immutable features for each dataset, based on which we can work out the minimum sparsity threshold a counterfactual explanation should adhere to (i.e., % immutable features). *Actionability* can then be assessed by comparing Sparsity with this level to determine if a method satisfies the mutability of features. An adequate explanation must achieve at least this level of sparsity. MCCE evidently fails to fulfill this constraint on Adult Income and Student Performance datasets.

Our reported results here are obtained under no other conditions than the constraints related to feature immutability and monotonicity described in Table 7. Nevertheless, we would like to highlight the flexibility of our framework in controlling the quality of counterfactual generations during inference. Users can freely specify any sparsity threshold or additional conditions of interest to filter out unsatisfactory examples without re-training or re-optimization as in methods like DICE. Specifically, DICE employs gradient search directly on each query according to a selected set of weighting hyperparameters for each term in the objective function. To get a less sparse or more diverse example than the current generation, one needs to activate a new search routine.

Too many constraints or too much sparsity clearly affects the diversity level of the counterfactual set. Maintaining a high Harmonic mean scores while satisfying almost all feature constraints demonstrates that L2C can effectively manage these trade-offs. The fact that L2C converges to valid counterfactuals with minimal violation in such a short inference time can be attributed to the practice of injecting hard constraints during optimization and global training does enhance the effect. It certainly helps circumvent the burden of heuristically eradicating violated samples. Notice also that our quantitative results align with the descriptions in Table 1. None of the diverse counterfactual explanation approaches address *plausibility* thoroughly whereas those reported to support the feasibility of features do not guarantee *diversity*. Appendix C provides empirical evidence for this claim, in which we compare L2C with popular amortized algorithmic recourse approaches and demonstrate our consistent superiority in generating diverse explanations efficiently without violating the required constraints (See Table 8).

5.2 Re-identification Risk Analysis

Preliminaries. We start by reviewing the fundamental concepts related to a public dataset:

- *Identifiers*: Attributes that uniquely identify an individual. Identifiers can be a person’s full name, government tax number or driver’s license number.
- *Quasi-identifiers*: Attributes that themselves do not uniquely identify a person, but when combined are sufficiently correlated to at least one individual record. For example, the combination of gender, birth dates and ZIP codes can re-identify 87% of American residents [47].
- *Sensitive attributes*: Attribute that are protected against unauthorized access. Sensitive data is confidential and if leaked could harm personal safety or emotional well-being. Examples are salary, medical conditions, salary, criminal histories, or phone numbers.
- *Equivalence class*: An equivalence class is a group of records with identical quasi-identifiers.

Every public dataset must first be anonymized by removing identifiers. However, the data may still be vulnerable to re-identification attacks due to the potential existence of quasi-identifiers. To quantify the level at which a dataset is susceptible to re-identification risk, the following 3 metrics are commonly used:

- *k-Anonymity* [41]: A dataset satisfies *k*-anonymity if for each record in the dataset, the quasi-identifiers are indistinguishable from at least $k - 1$ other people also in the dataset.
- *l-Diversity* [28]: A dataset has *l*-diversity if, for every equivalence class, there are at least l distinct values for each sensitive attribute.
- *k-Map* [12]: Given an auxiliary dataset used for re-identification (e.g., US Census or IMDb dataset in the Netflix example), so-called the ‘attack’ dataset, a dataset satisfies *k*-map if every equivalence class is mapped to at least k records in the ‘attack’ dataset.

Evaluation metrics. Suppose a company releases an API that permits users to query a set of counterfactual examples. We now analyze the level of privacy leakage associated with the output data,

by quantifying the percentage of successful attacks w.r.t the aforementioned metrics. Respectively, we measure (1) 1-Anonymity: % equivalence classes with only $k = 1$ member, (2) 1-Diversity: % equivalence classes with $l = 1$ value for a sensitive attribute, (3) 1-Map: % counterfactual examples exactly matched with any single record in an ‘attack’ dataset. Notice that given a *k*-anonymized dataset, the existence of a one-to-one mapping with the ‘attack’ dataset means the released dataset fails *k*-Map.

Experiments. By definition, violations w.r.t *k*-Anonymity and 1-Diversity are computed against the output set of examples, while *k*-Map requires an external dataset. For Adult Income, we choose the validation set as the ‘attack’ set. For German Credit, there exists multiple versions of this dataset across the literature. The one used in our main analysis is adopted from [49], which has been subject to pre-processing. We use another version published by Penn State University³ for re-identification. We assume these hold-out sets belong to some larger datasets of population available accessed by the public. Quasi-identifiers and sensitive attributes are given in Appendix A. Following Goethals et al. [16], we consider the black-box predicted label as part of the quasi-identifiers.

Let’s first look into the attacks on the raw output explanations presented in Table 4. We here assume that the attacker’s goal is to collect as many examples as possible without caring about which one is valid. If the attacker has no information about how the data is discretized, it is much less likely to find exact matches in the ‘attack’, thereby reducing the re-identifiability of L2C data. Now we assume the attacker gets access to both the API and our discretization mechanism. They therefore could correspondingly discretize the data in the ‘attack’ set and retrieve the matches. Our analysis assumes this worse scenario, meaning that for L2C only, we compute 1-Map against the discretized data. The entries N/A in Table 4 are due to the fact that no counterfactual example in the data of DICE-Genetic or COPA has matches, so their robustness remains unverifiable in our experiment. We must highlight that no match does not necessarily translate to zero privacy risk. We also note that such a case is different from L2C, whose result is nearly 0.00%. L2C in fact still returns matches for some records wherein we achieve *k*-Map of 2 – 3 specifically. Overall, L2C yields the lowest re-identifiability risk. Another interesting observation on German Credit is that although DICE performs well on *k*-Anonymity, the number of attacks on 1-Diversity is dramatically high. This sheds light on the limitation of *k*-Anonymity discussed in [28] about *Homogeneity attacks* and *Background knowledge attacks*. Basically, attacking a *k*-anonymized dataset with a high *k* can still reveal some private information of the data subjects because profiles in the same equivalence class are similar (very few distinct values in sensitive attributes), or the adversary has some background knowledge that any help narrow down possible values.

Privacy under CF-K. We here investigate the effectiveness of the idea behind CF-K proposed in [16]. CF-K searches for an equivalence class for each counterfactual example and suggests only publishing profiles of at least *k*-sized equivalence class. Since the authors do not publish their codes, plus it is hugely time-consuming to

³<https://online.stat.psu.edu/stat857/node/215/>

Table 3: Desirability of counterfactual explanation methods. ↓ Lower is better. ↑ Higher is better. **Bold / Underline** indicates the best / second-best performance for each dataset. Time records total inference time in seconds.

Method	Sparsity (%)↑	Diversity (%)↑	Harmonic Mean (%)↑	Validity (%)↑	Coverage (%)↑	Unary (%)↑	Time(s)↓
German Credit (Logistic Regression) - Min Sparsity: 20.00%							
L2C (Ours)	<u>61.35</u>	<u>37.31</u>	46.39	100.00	100.00	99.06	<u>18</u>
DICE-Random	88.23	15.29	26.06	100.00	100.00	<u>90.81</u>	1,150
DICE-Genetic	43.45	37.56	<u>40.29</u>	<u>62.87</u>	<u>90.24</u>	56.66	17,615
COPA	57.88	18.88	28.47	44.00	44.00	84.31	17,583
MCCE	28.76	33.40	30.91	48.74	100.00	58.76	2
Adult Income (Neural Network) - Min Sparsity: 30.77%							
L2C (Ours)	<u>45.70</u>	28.11	34.80	100.00	100.00	97.62	<u>444</u>
DICE-Random	89.26	9.05	16.44	100.00	100.00	<u>87.15</u>	12,332
DICE-Genetic	41.48	<u>26.27</u>	<u>32.14</u>	<u>92.64</u>	100.00	72.70	505,174
MCCE	24.93	4.58	7.74	30.63	74.76	45.79	98
Graduate Admission (Neural Network) - Min Sparsity: 14.29%							
L2C (Ours)	<u>42.23</u>	<u>37.90</u>	<u>39.94</u>	100.00	100.00	100.00	<u>4</u>
DICE-Random	66.25	30.93	42.15	100.00	100.00	<u>85.30</u>	412
DICE-Genetic	23.05	47.54	31.04	<u>92.91</u>	100.00	66.69	6,171
MCCE	17.39	22.98	19.51	43.79	84.60	79.11	1
Student Performance (Logistic Regression) - Min Sparsity: 38.57%							
L2C (Ours)	<u>55.32</u>	29.54	<u>38.51</u>	100.00	100.00	100.00	<u>6</u>
DICE-Random	87.60	13.64	23.60	100.00	100.00	<u>98.99</u>	2,518
DICE-Genetic	39.20	39.88	38.54	<u>84.83</u>	100.00	60.77	3,406
COPA	50.45	25.28	33.68	67.26	67.26	95.32	18,774
MCCE	25.97	24.97	25.46	60.98	<u>93.10</u>	67.70	1

run on models like DICE, we extend the above experiment and examine the effect when every output counterfactual set is 2-anonymized. Specifically, for every set of 100 generations, we remove records not belonging to any equivalence classes. Given now that the data is now 2-anonymized, we evaluate attacks against 1-Diversity and k-Map. We also measure % of valid counterfactuals left in the set, assuming that a user requests for 100 per instance. The more valid examples lost from a model explainer imply that it will be more costly to search for sufficient equivalence classes for every instance. Figure 2 depicts that in most cases, k -anonymization enhances the protection of the sensitive attributes but does greatly compromise the *validity* of the output explanations.

The purpose of k -anonymization is to ensure when attacked, an individual remains indistinguishable from at least $k - 1$ others. However, notice that k -anonymization does not prevent k-Map against which the number of successful attacks is still high for DICE and MCCE. The threat is thus no less severe when the attacker is interested in the re-identifiability of both datasets. If the released data contains the sensitive information that is missing from the ‘attack’ dataset, and given the fact that 1-diversity remains well above zero for all methods, the attacker could easily infer private information of every linked record. To this end, generalizing the data as done in L2C is proved to be useful to prevent such an inference attack. It is also observed that combining CF-K with L2C in particular significantly improves the anonymity of our counterfactual data. We therefore believe that the integration of L2C with other privacy

techniques in the cybersecurity area would yield a more effective safeguard.

Table 4: Successful attacks on counterfactual explanation methods. **Bold / Underline** indicates the lowest / second-lowest privacy risk for each dataset. 1-Diversity is evaluated on 2 most sensitive attributes as shown in the columns 2 & 3.

Method	1-Anoy.↓	1-Diversity↓	1-Map↓	
German Credit				
L2C (Ours)	62.15%	67.09%	71.60%	0.21%
DICE-Random	55.15%	82.75%	89.96%	23.67%
MCCE	62.83%	65.40%	76.95%	26.64%
DICE-Genetic	15.36%	90.19%	95.80%	N/A
COPA	87.61%	89.41%	89.24%	N/A
Adult Income				
L2C (Ours)	5.07%	14.90%	39.58%	0.00%
DICE-Random	72.21%	91.28%	93.07%	3.31%
MCCE	15.19%	66.42%	72.36%	1.77%
DICE-Genetic	15.11%	65.32%	87.54%	N/A

5.3 Discretization

Discretization is an important pre-processing step in data analysis in which the problem of optimal discretization with a minimum number of splits is proved to be NP-Hard [6, 35]. We here adopt the

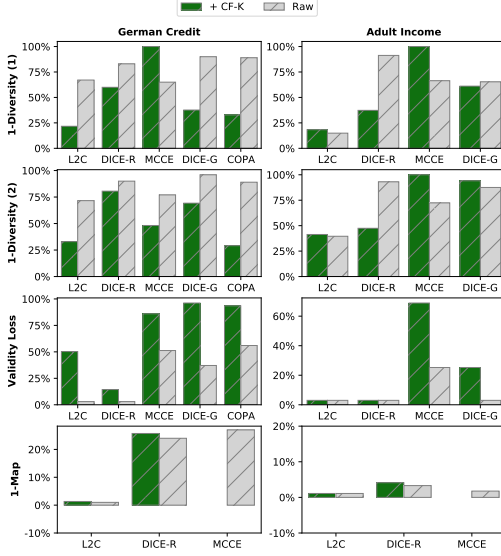


Figure 2: Privacy risk comparison between the raw output data and the data subject to 2-anonymization under the strategy of CF-K. For all metrics, lower is better. 1-Diversity is evaluated on 2 most sensitive attributes.

unsupervised *Equal-frequency* discretizer, which splits a continuous attribute into buckets with the same number of instances⁴. More concretely in this experiment, features are quantized using Python function `qcut`⁵, which requires specifying the maximum number of buckets/levels and later adjusts it depending on the input data distribution. We set the maximum buckets to be 4 such that every bucket averagely has 25% of total observations. Table 9 reports how the numerical features of each dataset are discretized. We argue that having very few buckets is likely to cause under-fitting since there are very few useful combinatorial patterns that can counter the original label. Whereas we need diversity for effective learning, too many buckets are undesirable since it can hurt generalization due to some following reasons : (1) each bucket would contain too little data and the chosen middle value may not represent the bucket well, and (2) the model has more combinations of features to explore, thus can converge to sub-optimal combinations that cannot generalize well on unseen test points. In this regard, we decide to split data into equal-sized buckets in the hope of balancing the trade-off.

Various discretization methods exist [37]. Table 5 analyzes the performance of our method L2C under 3 other discretization strategies. The first one is *Minimal entropy partitioning* (MDP) [13]. It is an old-school supervised approach and one of the most widely used. MDP determines the binary discretization for a value range by selecting the cut point that minimizes the class entropy. The algorithm can be applied recursively on sub-partitions induced by

⁴Except for the features Capital gain and Capital loss from Adult Income which we convert into binary variables to accurately reflect the semantics of their data.

⁵<https://pandas.pydata.org/docs/reference/api/pandas.qcut.html>

the initial cut point and the paper proposes using Minimum Description Length Principle⁶ [39] as a stopping condition. Another strategy is to apply *Domain knowledge* where feature values can be grouped based on common demographic or social characteristics. For example, Age could be translated into different age groups (e.g., Teenagers, Young Adults, etc.), or TOEFL scores are divided into proficiency levels. While the aforementioned methods are univariate, we also implement a multivariate approach using Decision tree. Following the motivation of MDP, we run CART [4] to search for the splits that minimize class information entropy. Note that we here consider the predicted labels from the black-box models as the target variable. The goal is to ensure the prediction on each combination of features is stable as possible. To avoid fine-grained intervals, we set the minimum number of samples for a split to be 30. Our experiment shows that we are still able to achieve 100% of Validity and Coverage in roughly the same amount of time. We therefore only present the remaining metrics in Table 5, which here demonstrates a comparable quality of explanations among different discretization strategies. Given that L2C performance is relatively insensitive to the choice of discretizers, we therefore suggest using *Equal-frequency* for which no labels or external knowledge is required.

Table 5: Desirability of L2C counterfactual explanations under various discretization strategies. *Proposed method.

Strategy	Sparsity (%)↑	Diversity (%)↑	Harmonic Mean (%)↑	Unary (%)↑
German Credit - Min Sparsity: 20.00%				
Equal Freq.*	61.35	37.31	46.39	99.06
MDP	61.58	35.00	44.63	100.00
CART	60.92	39.85	48.18	100.00
Domain Know.	61.73	37.56	46.70	100.00
Graduate Admission - Min Sparsity: 14.29%				
Equal Freq.*	42.23	37.90	39.94	100.00
MDP	58.20	41.34	41.56	100.00
CART	41.80	41.34	41.56	100.00
Domain Know.	42.17	42.30	42.22	100.00
Student Performance - Min Sparsity: 38.57%				
Equal Freq.*	55.32	29.54	38.51	100.00
MDP	55.57	28.50	37.67	100.00
CART	55.07	27.79	36.94	100.00
Domain Know.	55.47	31.57	40.23	100.00

5.4 Feature Correlational Constraints

While the treatment of unary constraints is straightforward for heterogeneous datasets, we argue that this is not the case for binary constraints. For example, suggesting that a person get a Master’s degree at precisely the age of 34 is unrealistically rigid. This issue indeed stems from the presence of continuous features. A direct solution is to allow for more flexible suggestions through discretization (e.g., suggesting an age range from 30 – 40 instead of an exact value at 34). This indeed aligns with the generative mechanism of our L2C, which sets us apart from existing works. However,

⁶<https://orange3.readthedocs.io/>

discretization is currently treated as a subroutine of internal processing, meaning that in the output examples, the values for the continuous features are still returned in the numerical format for the sake of consistency. Therefore, the best strategy would be to have the machine learning classifiers trained in the discretized feature space accordingly. Practitioners could then ignore the one-hot decoding stage and deploying L2C for this purpose would be effortless. In Appendix B, we demonstrate how L2C effectively addresses binary constraints in this scenario with success rates of 91.54% and 100.00% respectively on German Credit and Adult Income.

6 CONCLUSION AND FUTURE WORK

In this paper, we study the challenges facing algorithmic recourse approaches in generating diverse counterfactual explanations: how *diversity* can be tackled without compromising the other desiderata of an explanation while preserving privacy against linkage attacks. We analyze how existing engines fail to resolve the trade-offs among counterfactual constraints and fill the research gap with our novel framework L2C. Here we target a broad class of differentiable machine learning classifiers. To fit non-differentiable models in our framework, one could use policy gradient [46] or attempt to approximate such models as decision trees or random forests with a differentiable version [27, 51]. L2C is currently proposed to deal with re-identification risks of the released examples. Defense against *model stealing* and *membership inference attacks* however remains exigent. Integrating differential privacy in the framework of L2C is one interesting research avenue, which we leave for future works to explore.

REFERENCES

- [1] Mohan S Acharya, Asfia Armaan, and Aneeta S Antony. 2019. A comparison of regression models for prediction of graduate admissions. In *2019 international conference on computational intelligence in data science (ICCIDS)*. IEEE, 1–5.
- [2] Ulrich Aivodji, Alexandre Bolot, and Sébastien Gambs. 2020. Model extraction from counterfactual explanations. *arXiv preprint arXiv:2009.01884* (2020).
- [3] Ifeoma Ajunwa, Sorelle Friedler, Carlos E Scheidegger, and Suresh Venkatasubramanian. 2016. Hiring by algorithm: predicting and preventing disparate impact. Available at SSRN 2746078 (2016).
- [4] Leo Breiman, Jerome H Friedman, Richard A Olshen, and Charles J Stone. 2017. *Classification and regression trees*. Routledge.
- [5] Ngoc Bui, Duy Nguyen, and Viet Anh Nguyen. 2022. Counterfactual Plans under Distributional Ambiguity. *arXiv preprint arXiv:2201.12487* (2022).
- [6] Bogdan S Chlebus and Sinh Hoa Nguyen. 1998. On finding optimal discretizations for two attributes. In *International Conference on Rough Sets and Current Trends in Computing*. Springer, 537–544.
- [7] Paulo Cortez and Alice Maria Gonçalves Silva. 2008. Using data mining to predict secondary school student performance. (2008).
- [8] Susanne Dandl, Christoph Molnar, Martin Binder, and Bernd Bischl. 2020. Multi-objective counterfactual explanations. In *International Conference on Parallel Problem Solving from Nature*. Springer, 448–469.
- [9] Michael Downs, Jonathan L Chu, Yaniv Yacoby, Finale Doshi-Velez, and Weiwei Pan. 2020. Cruds: Counterfactual recourse using disentangled subspaces. *ICML WHI* 2020 (2020), 1–23.
- [10] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. <http://archive.ics.uci.edu/ml>
- [11] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*. Springer, 265–284.
- [12] Khaled El Emam and Fida Kamal Dankar. 2008. Protecting privacy using k-anonymity. *Journal of the American Medical Informatics Association* 15, 5 (2008), 627–637.
- [13] Usama Fayyad and Keki Irani. 1993. Multi-interval discretization of continuous-valued attributes for classification learning. (1993).
- [14] Badih Ghazi, Pritish Kamath, Ravi Kumar, and Pasin Manurangsi. 2022. Faster Privacy Accounting via Evolving Discretization. In *International Conference on Machine Learning*. PMLR, 7470–7483.
- [15] Richard Gifford. 2018. Legal technology: Criminal justice algorithms: Ai in the courtroom. *Proctor, The* 38, 1 (2018), 32–33.
- [16] Sofie Goethals, Kenneth Sörensen, and David Martens. 2022. The privacy issue of counterfactual explanations: explanation linkage attacks. *arXiv preprint arXiv:2210.12051* (2022).
- [17] Sivakanth Gopi, Yin Tat Lee, and Lukas Wutschitz. 2021. Numerical composition of differential privacy. *Advances in Neural Information Processing Systems* 34 (2021), 11631–11642.
- [18] Riccardo Guidotti. 2022. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery* (2022), 1–55.
- [19] Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144* (2016).
- [20] Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera. 2020. Model-agnostic counterfactual explanations for consequential decisions. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 895–905.
- [21] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. 2020. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. *arXiv preprint arXiv:2010.04050* (2020).
- [22] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. 2021. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 353–362.
- [23] Jack Klys, Jake Snell, and Richard Zemel. 2018. Learning latent subspaces in variational autoencoders. *Advances in neural information processing systems* 31 (2018).
- [24] Ron Kohavi et al. 1996. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, Vol. 96. 202–207.
- [25] Alex Kulesza, Ben Taskar, et al. 2012. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning* 5, 2–3 (2012), 123–286.
- [26] Stefan Lessmann, Bart Baesens, Hsin-Vonn Seow, and Lyn C Thomas. 2015. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research* 247, 1 (2015), 124–136.
- [27] Ana Lucic, Harrie Oosterhuis, Hinda Haned, and Maarten de Rijke. 2022. FO-CUS: Flexible optimizable counterfactual explanations for tree ensembles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 5313–5322.
- [28] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramkrishnan Venkatasubramanian. 2007. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1, 1 (2007), 3–es.
- [29] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. 2016. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712* (2016).
- [30] Divyat Mahajan, Chenhao Tan, and Amit Sharma. 2019. Preserving causal constraints in counterfactual explanations for machine learning classifiers. *arXiv preprint arXiv:1912.03277* (2019).
- [31] Ricardo Mendes and João P Vilela. 2017. Privacy-preserving data mining: methods, metrics, and applications. *IEEE Access* 5 (2017), 10562–10582.
- [32] Smitha Milli, Ludwig Schmidt, Anca D Dragan, and Moritz Hardt. 2019. Model reconstruction from model explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 1–9.
- [33] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 607–617.
- [34] Arvind Narayanan and Vitaly Shmatikov. 2006. How to break anonymity of the netflix prize dataset. *arXiv preprint cs/0610105* (2006).
- [35] Son H Nguyen and Andrzej Skowron. 1995. Quantization of real value attributes-rough set and boolean reasoning approach. In *Proc. of the Second Joint Annual Conference on Information Sciences, Wrightsville Beach, North Carolina, Sept 28-Oct 1*. Citeseer.
- [36] Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. 2020. Learning model-agnostic counterfactual explanations for tabular data. In *Proceedings of The Web Conference 2020*. 3126–3132.
- [37] Sergio Ramírez-Gallego, Salvador García, Héctor Mourino-Talín, David Martínez-Rego, Verónica Bolón-Canedo, Amparo Alonso-Betanzos, José Manuel Benítez, and Francisco Herrera. 2016. Data discretization: taxonomy and big data challenge. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 6, 1 (2016), 5–21.
- [38] Annabelle Redelmeier, Martin Jullum, Kjersti Aas, and Anders Løland. 2021. MCCE: Monte Carlo sampling of realistic counterfactual explanations. *arXiv preprint arXiv:2111.09790* (2021).
- [39] Jorma Rissanen. 1978. Modeling by shortest data description. *Automatica* 14, 5 (1978), 465–471.
- [40] Chris Russell. 2019. Efficient search for diverse coherent explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 20–28.
- [41] Pierangela Samarati and Latanya Sweeney. 1998. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. (1998).
- [42] Shubham Sharma, Jette Henderson, and Joydeep Ghosh. 2020. Certifai: A common framework to provide explanations and analyse the fairness and robustness of black-box models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 166–172.
- [43] Reza Shokri, Martin Strobel, and Yair Zick. 2021. On the privacy risks of model explanations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 231–241.
- [44] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*. IEEE, 3–18.
- [45] Kacper Sokol and Peter A Flach. 2019. Counterfactual explanations of machine learning predictions: opportunities and challenges for AI safety. *SafeAI@ AAAI* (2019).
- [46] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. 1999. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems* 12 (1999).
- [47] Latanya Sweeney. 2000. Simple demographics often identify people uniquely. *Health (San Francisco)* 671, 2000 (2000), 1–34.
- [48] Sahil Verma, John Dickerson, and Keegan Hines. 2020. Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596* (2020).
- [49] Sahil Verma, Keegan Hines, and John P Dickerson. 2022. Amortized Generation of Sequential Algorithmic Recourses for Black-Box Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 8512–8519.
- [50] Austin Waters and Risto Miikkiläinen. 2014. Grade: Machine learning support for graduate admissions. *Ai Magazine* 35, 1 (2014), 64–64.
- [51] Yongxin Yang, Irene Garcia Morillo, and Timothy M Hospedales. 2018. Deep neural decision trees. *ICML Workshop on Human Interpretability in Machine Learning*. (2018).

A EXPERIMENTAL DETAILS

A.1 Dataset statistics

- **German Credit** [10]: This dataset includes information of customers taking credit at a bank. The task is to classify a customer as a good (label 1) or bad (label 0) credit risk.
- **Adult Income** [24]: The dataset was extracted from the US 1994 census data on adult incomes. The task is to classify if an individual's income exceeds \$50,000 per year (label 1) or not (label 0).
- **Graduate Admission** [1]: The set contains data of Indian students' applications to a Master's program. The original target variable is an ordinal variable on the scale of $[0 - 1]$ indicating the chance of a student being admitted, where 1 indicates the highest chance. We set a threshold of 0.7 and re-categorize students as either "having a higher chance" (≥ 0.70 -label 1) or "having a lower chance" (< 0.70 -label 0). The binary classification task is to determine if a student profile has a higher chance of being successful at their application.
- **Student Performance** [7]: This dataset records the performance of students at two schools Gabriel Pereira and Mousinho da Silveira. The task is to predict if a student achieves a final score above average (label 1) or not (label 0). The train and test splits contain data of students from these two schools separately [5].

Table 7 summarizes the experimental settings for every dataset. Regarding the underlying black-box models, we experiment with Logistic Regression for the linear classifier and Neural Network for the non-linear classifier. We train linear classifiers on German Credit and Student Performance, and non-linear classifiers on Graduate Admission (with 3 layers and 40-dimensional hidden units) and Adult Income (with 3 layers and 30-dimensional hidden units). For each task, we further sample 20% random observations of the training sets as validation sets and train 5 black-box models with the same architecture but with different initializations. Since we have specified a fixed subset of immutable features, we therefore can derive the minimum sparsity level an explanation method should obey as

$$\text{Min Sparsity} = \frac{\text{No. immutable features}}{\text{Total no. features}}$$

A.2 Model Design

We parameterize \mathcal{G} and \mathcal{S} with neural networks of 3 and 2 layers respectively. Each layer consists of a dense layer and a ReLU activation, except the last layer of \mathcal{S} takes Sigmoid activation to produce a probability vector. The final layer of \mathcal{G} is another dense layer that outputs a logit vector of the same dimension as the input, representing the counterfactual distribution. We set the sparsity loss coefficient $\alpha = 1e - 4$ and use the same architecture for all tasks. We train our model with Adam optimizer for 200 epochs, at $\tau = 0.2$ and a learning rate of $1e - 4$.

B FEATURE CONSTRAINTS OF DISCRETIZED BLACK-BOX MODELS

We here demonstrate how binary constraints can be effectively addressed within L2C framework for realistic explanations. We use the same model architecture and experimental setup, except that the data now is initially discretized and the black-box classifiers

Table 6: Desirability of counterfactual examples generated from L2C for explaining a discretized black-box classifiers.

Harmonic Mean (%)↑	Validity (%)↑	Unary (%)↑	Binary (%)↑
German Credit			
48.56%	100.00%	98.38%	91.54%
Adult Income			
23.89%	100.00%	100.00%	100.00%

are trained on the discretized feature space. Not only would this help produce more realistic suggestions but it would also enhance privacy protection. We reuse the *Equal-frequency* strategy for consistency, yet note that practitioners are highly recommended to consider generalizing the data in such a way that satisfies the privacy metrics outlined in Section 5.2. We then run L2C models on German Credit and Adult Income datasets with feature correlations specified in Table 7: respectively the constraints *Present residence* \rightarrow *Age* and *Education level* \rightarrow *Age*. To enforce the increasing constraint on the child variable (i.e. *Age*), we further track the perturbation of the parent variable (i.e. *Present residence*, *Education level*) and apply the Eq. (5) to update the distribution of samples where the parent variable is indicated (by generator \mathcal{G}) to increase. Table 6 reports the quality of explanations produced under both types of constraints, where Binary measures the proportion of examples \mathcal{C} meeting the binary constraints. Here we again substantiate the effectiveness of L2C in balancing counterfactual constraints.

C AMORTIZED BASELINES

Table 8 compares L2C with popular amortized approaches: Feasible-VAE [30], CRUDS [9] and FastAR [49] across desiderata. We now provide the experimental setup for the amortized baseline models. Whereas the non-amortized algorithms are run directly on the testing sets, for amortized methods, we train the base generative models on the training sets and use the testing sets only for evaluation. We tune the base generative models under various different hyper-parameter settings via grid search and report the best results. We determine the best settings via two metrics: Coverage and Diversity. When there is a trade-off, Coverage is chosen to be the deciding criterion.

Specifically, for Feasible-VAE [30], we tune the hidden dimensions of the VAE encoder within $\{10, 30, 50, 70, 90\}$ and regularization term on *Validity* within $\{42, 62, 82, 102, 122\}$. For CRUDS [9], the base model is a Conditional Subspace Variational Auto-encoder [23]. In the original paper, the network only has 1 hidden layer of 64 nodes, which we find to be of low capacity. We thus experiment with 2 layers and different hidden dimensions within $\{16, 32, 64\}$. For FastAR [49], the hyper-parameters include manifold distance λ and entropy loss coefficient. Across datasets, Verma et al. [49] shows the best Coverage under $\lambda = 0.1$. We thus set $\lambda = 0.1$ as well in our experiments, while focusing on tuning the latter hyper-parameter within $\{0.01, 0.05, 0.1, 0.5, 1.0\}$. For the remaining hyper-parameters, we adopt the best values reported by the authors. A pre-trained

Table 7: Dataset statistics. *Features subject to non-decreasing constraints.

Dataset	German Credit	Adult Income	Graduate Admission	Student Performance
Train/Dev/Test No. features	640/160/200 20	28942/7235/9045 13	320/80/100 7	339/84/226 14
Immutable features	Foreign worker No. liable people Personal status Purpose	Race Sex Native country Marital status	University rating	Mother’s edu. Father’s edu. Family edu. support First period grade
Feature constraints*	Age Present employment Present residence Duration	Age Education level	Research experience	Age
Feature correlations	Increasing Age increases Present residence	Increasing Education level increases Age		
Quasi-identifiers	Age, Job Foreign worker Personal status Present employment Present residence Property, Housing	Age Sex Race Relationship Marital status		
Sensitive features	Credit amount Savings account	Capital gain Capital loss		
Black-box model	Logistic Regression	Neural Network	Neural Network	Logistic Regression
Test accuracy	67.00%	85.53%	90.60%	94.69%
Min Sparsity	20.00%	30.77%	14.29%	38.57%

FastAR model can only interpret one decision outcome chosen as the desired one (often the positive label). We must therefore train separate FastAR models on the positive and negative subsets and combine the results. We further find that although it is straightforward to obtain multiple generations in a single model, FastAR algorithm is optimized for one optimal counterfactual state for a given input. Thus in the hope of achieving better *diversity*, we train 100 different model initializations and accordingly collect a set of 100 explanations for evaluation. The inference time accumulates as a result, which is the reason why our reported results on time efficiency for FastAR are different from what are reported in the authors’ paper.

D THE ROLE OF FEATURE SELECTOR

We now validate the importance of learning the local feature-based selection distribution via the feature selector \mathcal{S} . We first remove \mathcal{S} from L2C framework and replace the probability vector $\pi(\mathbf{z})$ with a binary mask vector $\mathbf{m} \in [0, 1]^N$ where $m_i = 1$ if $i \in \mathbb{K}$ (i.e., a mutable feature) and $m_i = 0$ otherwise. We thus use m_i in substitution of s_i to update the counterfactual representations \tilde{z}_i as previously done. We only optimize the generator \mathcal{G} to learn the feature-based perturbation distribution, and the training objective Eq. (4) excludes the regularization term for *sparsity* accordingly. Figure 3 investigates the performance of L2C under this alternative

setup, in comparison with the proposed method that jointly optimizes \mathcal{S} and \mathcal{G} . L2C still achieves 100% of Validity and Coverage, so we only report the relevant metrics.

One drawback of omitting the Selector component is that we lose the flexibility in tailoring the quality of counterfactual generations to potential user preferences. It is seen that the Selector introduces significant *sparsity* to gain an effective balance for the trade-off against *diversity*. Furthermore, these results support our claim about the role of the generator \mathcal{G} in that the perturbation distribution alone can yield impressively diverse explanations with *sparsity* remaining under the required maximum level. This is a benefit of learning an entire feature distribution in that sometimes an output sample falls into the original input value i.e., $\tilde{z}_i = z_i$ while combining adequately with other features.

E QUALITATIVE EXAMPLES

Table 10 - 13 illustrate some examples of our generated counterfactuals for each dataset. For illustration purposes only, we report the discretized values for numerical features where the edge values of each numerical interval are rounded to the nearest whole number. Immutable features are italicized.

Table 8: Desirability of amortized counterfactual explanation methods. ↓ Lower is better. ↑ Higher is better. **Bold / Underline** indicates the best / second-best performance for each dataset. Time records total inference time in seconds.

Method	Sparsity (%)↑	Diversity (%)↑	Harmonic Mean (%)↑	Validity (%)↑	Coverage (%)↑	Unary (%)↑	Time(s)↓
German Credit (Logistic Regression) - Min Sparsity: 20.00%							
L2C (Ours)	<u>61.35</u>	37.31	46.39	100.00	100.00	<u>99.06</u>	18
FastAR	95.93	0.68	1.33	<u>95.79</u>	<u>95.79</u>	99.87	10,605
F-VAE	45.93	1.59	3.06	100.00	100.00	74.41	<u>36</u>
CRUDS	29.72	<u>14.21</u>	<u>19.18</u>	60.00	60.00	71.31	42,920
Graduate Admission (Neural Network) - Min Sparsity: 14.29%							
L2C (Ours)	42.23	37.90	39.94	100.00	100.00	100.00	4
FastAR	<u>27.97</u>	1.32	2.35	<u>87.41</u>	<u>87.41</u>	<u>85.28</u>	5,405
F-VAE	7.71	1.48	1.84	100.00	100.00	44.60	<u>16</u>
CRUDS	11.31	<u>17.38</u>	<u>13.40</u>	60.28	76.00	49.00	21,460
Student Performance (Logistic Regression) - Min Sparsity: 38.57%							
L2C (Ours)	<u>55.32</u>	<u>29.54</u>	38.51	100.00	100.00	100.00	6
FastAR	81.86	1.44	2.77	<u>97.71</u>	97.71	<u>99.76</u>	16,370
F-VAE	27.37	7.68	11.92	100.00	100.00	69.68	<u>36</u>
CRUDS	24.06	33.64	<u>28.55</u>	62.12	100.00	100.00	39,571

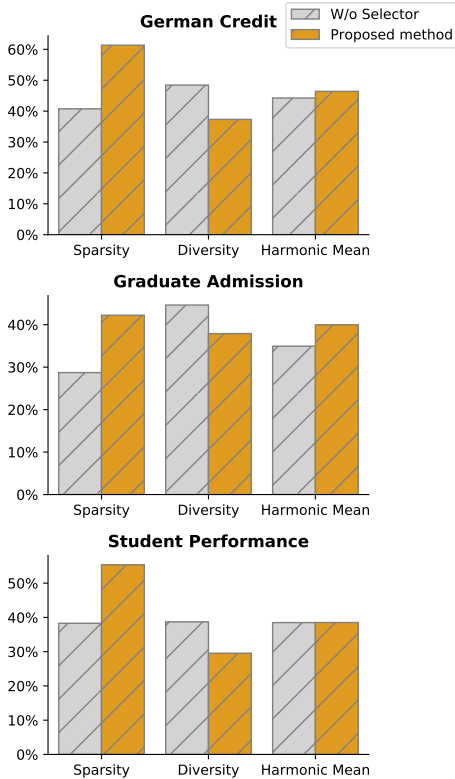
**Figure 3: Analysis of L2C performance when the selector S is removed.**

Table 9: Discretization of numerical features in each dataset.

Feature name	Bucket values	No. Data points
German Credit		
Duration (Months)	(4 , 12]	359
	(12 , 18]	187
	(18 , 24]	224
	(24 , 72]	230
Credit Amount	(247 , 1367]	250
	(1367 , 2320]	250
	(2320 , 3971]	250
	(3971 , 18425]	250
Age	(19 , 27]	291
	(27 , 33]	225
	(33 , 42]	249
	(42 , 75]	235
Adult Income		
Age	(18 , 30]	14,260
	(30 , 45]	17,727
	(45 , 60]	10,387
	(60 , 75]	2,848
Capital gain	≤ 0	41,432
	> 0	3,790
Capital loss	≤ 0	43,082
	> 0	2,140
Hours per week	(0 , 20]	3,602
	(20 , 40]	27,843
	(40 , 100]	13,777
Graduate Admission		
GRE score	(290 , 312]	189
	(312 , 322]	148
	(322 , 340]	163
TOEFL score	(92 , 104]	176
	(104 , 110]	175
	(110 , 120]	149
Undergraduate GPA	(7 , 8]	170
	(8 , 9]	163
	(9 , 10]	167
Student Performance		
Age	(15 , 16]	289
	(16 , 17]	179
	(17 , 22]	181
School absences	(-0 , 4]	466
	(4 , 32]	183
First period grade	(-0 , 10]	252
	(10 , 13]	245
	(13 , 19]	152
Second period grade	(-0 , 10]	228
	(10 , 13]	269
	(13 , 19]	152

Table 10: Counterfactual examples from German Credit dataset. *DM: Deutsche Mark

German Credit	Original input (Bad credit risk)	Counterfactuals (Good credit risk)				
Duration (months)	12 - 18	24 - 72	18 - 24	12 - 18	18 - 24	12 - 18
Credit amount (DM)	1367 - 2320	1367 - 2320	1367 - 2320	1367 - 2320	247 - 1367	247 - 1367
Age	19 - 27	19 - 27	19 - 27	19 - 27	19 - 27	42 - 75
Checking account (DM)	no account	200+	no account	200+	under 200	Under 0
Credit history	paid back duly	no credit taken	paid back duly	paid back duly	paid back duly	other credits
Purpose	furniture	furniture	furniture	furniture	furniture	furniture
Savings account (DM)	under 100	under 500	1000+	1000+	under 100	under 100
Present employment since	1 - 4	1 - 4	1 - 4	7+	1 - 4	1 - 4
Installment rate	Under 20	Under 20	Under 20	Under 20	Under 20	Under 20
Personal status	Male	Male	Male	Male	Male	Male
	divorce	divorced	divorced	divorced	divorced	divorced
Other debtors	none	none	guarantor	none	none	guarantor
Present residence since	7+ years	7+ years	7+ years	7+ years	7+ years	7+ years
Property	no property	no property	no property	insurance	real estate	no property
Other installment plans	none	none	none	none	none	none
Housing	rent	rent	rent	rent	own	own
No. existing credits	1	1	1	1	1	1
Job	skilled	skilled	skilled	skilled	skilled	skilled
No. people being liable	0 - 2	0 - 2	0 - 2	0 - 2	0 - 2	0 - 2
Telephone	No	Yes	Yes	No	No	No
Foreign worker	No	No	No	No	No	No

Table 11: Counterfactual examples from Adult Income dataset.

Adult Income	Original input (Low income)	Counterfactuals (High income)				
Age	18 - 30	18 - 30	18 - 30	18 - 30	18 - 30	18 - 30
Capital gain	No	Yes	Yes	No	No	Yes
Capital loss	No	Yes	Yes	Yes	Yes	Yes
Hours per week	20 - 40	20 - 40	20 - 40	20 - 40	20 - 40	20 - 40
Work class	Self-employed	Self-employed	Self-employed	Self-employed	Self-employed	Self-employed
Highest Edu. Level	12	15	14	12	14	12
Marital status	Married-civ-sps.	Married-civ-sps.	Married-civ-sps.	Married-civ-sps.	Married-civ-sps.	Married-civ-sps.
Occupation	Craft repair	Other service	Adm clerical	Armed forces	Protective service	Sales
Relationship	Not in family	Wife	Other relative	Wife	Own child	Wife
Race	Black	Black	Black	Black	Black	Black
Sex	Female	Female	Female	Female	Female	Female
Native country	Hungary	Hungary	Hungary	Hungary	Hungary	Hungary

Table 12: Counterfactual examples from Graduate Admission dataset.

Graduate Admission	Original input (Low chance)	Counterfactuals (High chance)				
GRE Score	290 - 312	290 - 312	312 - 323	290 - 312	290 - 312	290 - 312
TOEFL Score	92 - 104	104 - 110	104 - 110	110 - 120	110 - 120	92 - 104
Undergraduate GPA	7 - 8	8 - 9	9 - 10	9 - 10	8 - 9	9 - 10
<i>University Rating</i>	2 / 5	2 / 5	2 / 5	2 / 5	2 / 5	2 / 5
Statement of Purpose	2 / 5	2 / 5	5 / 5	2 / 5	2 / 5	2 / 5
Letter of Recommendation	2.5 / 5	4 / 5	2.5 / 5	2.5 / 5	2.5 / 5	2 / 5
Research Experience	No	No	No	Yes	No	Yes

Table 13: Counterfactual examples from Student Performance dataset.

Student Performance	Original input (Fail)	Counterfactuals (Pass)				
Age	16 - 17	16 - 17	16 - 17	16 - 17	16 - 17	16 - 17
School absences	0 - 4	0 - 4	0 - 4	4 - 32	0 - 4	4 - 32
<i>First period grade</i>	10 - 13	10 - 13	10 - 13	10 - 13	10 - 13	10 - 13
Second period grade	0 - 10	13 - 19	13 - 19	13 - 19	13 - 19	13 - 19
<i>Mother's education</i>	Primary	Primary	Primary	Primary	Primary	Primary
<i>Father's education</i>	None	None	None	None	None	None
Weekly study time	2 - 5 hours	2 - 5 hours	2 - 5 hours	2 - 5 hours	5 - 10 hours	4 - 10 hours
<i>Family educational support</i>	No	No	No	No	No	No
Wanting higher education	Yes	Yes	No	No	Yes	Yes
Internet access at home	Yes	No	Yes	Yes	No	No
In a romantic relationship	Yes	No	No	No	Yes	Yes
Free time	Moderate	High	Moderate	Moderate	Moderate	High
Going out frequency	Very often	Very often	Very often	Rarely	Very often	Very Often
Health status	Good	Good	Good	Very good	Good	Good