The Effect of Diversity in Counterfactual Machine Learning Explanations

Rick Tompkins , Ronal Singh , Tim Miller

School of Computing and Information Systems
The University of Melbourne, Parkville, VIC, 3010, Australia
rtompkins@student.unimelb.edu.au, rr.singh@unimelb.edu.au, tmiller@unimelb.edu.au

Abstract

In this paper, we test the assumption that diversity in counterfactual explanations positively impacts understanding and trust in a machine learning model, finding limited support for this assumption. Counterfactual explanations are a well-advocated technique for machine learning explainability. A counterfactual shows the closest possible world in which an alternative decision would be produced. Several researchers make a case for multiple and diverse counterfactuals as a means to explain better —and better understand—the decisions made by machine learning models. We conduct a human-subject experiment through Amazon Mechanical Turk with 208 participants using task prediction and qualitative scales to test these assumptions for the credit scoring domain. Our results show that having fewer (two) rather than more (four or eight) counterfactuals or just a single counterfactual improved understanding, and those who received a high number of counterfactuals reported a preference for fewer. Surprisingly, we find that diversity had no positive impact on either understanding or trust of our participants. These results call into question the assumption that diverse counterfactuals are useful for understanding.

1 Introduction

The models powering contemporary systems for many high stakes decisions, such as credit risk [Danenas and Garsva, 2015], resume to job matching [Osipovs, 2019], and pretrial bail eligibility [Zavrsnik, 2020], can be complex and difficult to understand. Explainable Artificial Intelligence (XAI) seeks to solve this through methods that provide human-understandable explanations for AI decisions. There exist several techniques for generating explanations [Adadi and Berrada, 2018; Gilpin *et al.*, 2018; Guidotti *et al.*, 2018; Linardatos *et al.*, 2021], including feature-based [Ribeiro *et al.*, 2016], case-based [Lamy *et al.*, 2019], and more recently, counterfactual explanations [Wachter *et al.*, 2017; Russell, 2019; Ustun *et al.*, 2019; Joshi *et al.*, 2019; Pawelczyk *et al.*, 2020; Looveren and Klaise, 2019].

Counterfactuals have a strong history of support in philosophical and cognitive science [Byrne, 2019; Miller, 2019; Miller, 2021], and we take the definition by Lewis [1973] that a counterfactual is a "close possible world" that produces an alternate outcome for some decision. Existing literature focuses on developing effective technical solutions for generating diverse counterfactuals [Poyiadzi *et al.*, 2020; Karimi *et al.*, 2020a; Russell, 2019], but there has been no or limited human-subject evaluation of the impact on the number or diversity of counterfactuals on understanding. Given the significance of understanding [Mothilal *et al.*, 2020] and having trust in these machine learning models, and the ability of people to take effective action on available recourse [Wachter *et al.*, 2017; Ustun *et al.*, 2019], it is important to test the assumptions that underpin these methods.

In this paper, we undertake a human behavioural experiment via Amazon Mechanical Turk with 208 participants to investigate the impact of diversity of counterfactuals in the credit scoring domain. In particular, we test the impact of the diversity in terms of the number of counterfactuals, and the number of features changed by a counterfactual (*feature uniqueness*) to (1) improve participants' understanding of machine learning models; and (2) an increase or decrease the model's trust. We found that:

1. counterfactual explanations improve understanding of the model over a no-explanation baseline as measured by a task prediction exercise, and did not impact trust;
2. a small number of counterfactuals (two) improved participant understanding of the model more than larger numbers (four and eight); and
3. importantly, the diversity of counterfactuals had no impact on either understanding or perceived trust.

These results challenge the assumption that multiple, diverse counterfactuals are a good model for understanding.

2 Related Work

A counterfactual explanation describes a small change to feature values to change the outcome. Consider an example of a counterfactual explanation for a loan default prediction classifier:

> "An automated system predicted that you are likely to default on a home loan because your annual in-

come of $56,000 is too low. If your annual income was $70,000, the automated system would have predicted that you are not likely to default."

Counterfactual explanations described by Wachter *et al.* [2017] build on this idea. They focus on the closest possible world, the smallest possible change to feature values to change the outcome. Wachter *et al.* formalise this as follows. Assume that we have a machine learning model, f. Given an input x, the model predicts $f(x) = y$ as the outcome. In this context, a counterfactual explanation is a perturbation of the input, x, to generate a different output $y' = f(c)$ by f. This has been formalised as follows:

$$\arg\min_{c} \text{yloss}(f(c), y) + d(x, c), \tag{1}$$

where $d(\cdot, \cdot)$ is a distance measure, f the classifier function, $c \in C$ is a counterfactual from a set of counterfactuals C and y' the classifier responses we desire, yloss pushes the counterfactual c towards a different prediction than the original instance, and $d(x, c)$ keeps the counterfactual close to the original instance.

However, a single counterfactual may not be the most valuable or insightful to the recipient in cases where the recipient seeks recourse to an unfavourable decision [Wachter *et al.*, 2017; Russell, 2019; Karimi *et al.*, 2020b]. Wachter *et al.* [2017] discuss the obvious value from providing a *diverse* set of counterfactuals – including more possible paths for actionable recourse, the implicit knowledge gained when diversity is present, as well as an assumed net increase in understanding of the model [Mothilal *et al.*, 2020].

In addition to diversity, there are other properties desired of counterfactuals explanations, such as *validity*, *actionability* and *sparsity* [Wachter *et al.*, 2017; Verma *et al.*, 2020; Karimi *et al.*, 2020b]. However, in this paper, we focus only on diversity.

2.1 Diverse Counterfactual Explanations

Wachter *et al.* [2017] called for diverse counterfactuals and provided the basis for the distance functions that could facilitate the generation of diverse counterfactual explanations. Russell [2019] proposed an integer programming technique for generating diverse and coherent counterfactuals for classifications produced by linear models that use both continuous and categorical data types. Russell leans on the opinion shared by Wachter *et al.* that diverse counterfactuals help laypeople understand decisions made by automated systems. In terms of the distance function, $d(\cdot, \cdot)$ Russell [2019] used the weighted ℓ_1 norm, that is, $||\cdot||_{1,\text{MAD}}$. This measure is expected to generate counterfactuals that are sparse and robust to outliers [Wachter *et al.*, 2017].

Recently, there has been an interest in using counterfactual explanations to help users understand the deployed machine learning model or at least "guess" the decision boundary [Mothilal *et al.*, 2020]. Mothilal *et al.* provide several metrics to measure the diversity of a set of counterfactuals, as outlined below.

We can measure diversity by through determinantal point processes (DPP), as follows:

$$dpp\_diversity = det(\mathbf{K}), \tag{2}$$

where $\mathbf{K}_{i,j} = \frac{1}{1+dist(\boldsymbol{c}_i, \boldsymbol{c}_j)}$ and $dist(\boldsymbol{c}_i, \boldsymbol{c}_j)$ denotes a distance metric between the two counterfactual examples.

Another way to measure diversity is using the mean of the distances between each pair of examples, as follows:

$$\texttt{Diversity} : \Delta = \frac{1}{C_k^2} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \text{dist}(\boldsymbol{c}_i \boldsymbol{c}_j) \tag{3}$$

$\mathcal{C}$ is the set of k counterfactual examples. Furthermore, Mothilal *et al.* [2020] define different distance functions, dist, for continuous or categorical features.

A third way to measure diversity is through the fraction of features that are different between any two pairs of counterfactual examples, as follows:

$$\texttt{Count-Diversity} : \frac{1}{C_k^2 d} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \sum_{l=1}^{d} 1_{[\boldsymbol{c}_i^l \neq \boldsymbol{c}_j^l]} \tag{4}$$

where d is the number of input features and k is the number of counterfactuals.

While there have been significant technical contributions on generating diverse counterfactuals, the effects of diversity and number of counterfactuals on model understanding and trust have not been evaluated through human experiments. This paper aims to fill this gap and reiterate the importance of human-subject evaluations.

3 Experiment Design

While diverse counterfactuals may deliver on some capabilities [Wachter *et al.*, 2017], there is no empirical analysis on how a user's understanding of a model is impacted specifically by the diversity of counterfactuals presented to them or its effect on trust. These aspects are critical to adopting machine learning systems and the likelihood of a recipient understanding the machine learning model or using the information provided to achieve recourse.

This section outlines an experimental study design to evaluate the effect of diverse counterfactual explanations. Our study tests two hypotheses: H1) diverse counterfactuals improve human understanding of a model and that there is some good range of them to be useful; H2) diverse counterfactuals improve the trust elicited from the predictions of a machine learning model. The intuition behind these hypotheses is that having a diverse set of counterfactuals provides people with a wider range of 'behaviours' to summarise the model, and having some small collection that does not overwhelm users improves understanding and trust.

3.1 Diversity Metrics

In this paper, we test two diversity metrics, that is, *number of counterfactuals* and *feature uniqueness*. One is an existing metric and we propose a new metric inspired by the `Count-Diversity` metric discussed earlier.

Our first measure of diversity is the number of counterfactuals, N. In much of the existing literature, diversity is the number of counterfactuals generated according to a given distance measure, dist function. That is, having diverse counterfactuals is presenting several paths to the alternative classification. In our work, the N diverse counterfactual explanations are generated using Russell's method, which uses the weighted ℓ_1 norm for the distance function, $||\cdot||_{1,\text{MAD}}$.

Our second measure of diversity, D, is based on *feature uniqueness*. This diversity metric assesses how significantly different the counterfactuals are in terms of the number of features they modify. To compute D, we count the number of *unique features* modified by a counterfactual explanation, relative to the set of counterfactuals. Formally, feature uniqueness, K, of a set counterfactuals C is defined as:

$$K = \frac{|\cup_{c \in C} Unique(c)|}{|\cup_{c \in C} FeaturesChanged(c)|}$$

$$Unique(c) = \{c_j \in c \mid v(c_j) \neq v(x_j) \wedge \nexists c' \in C \setminus \{c\} \cdot v(c'_j) \neq v(x_j)\}$$

where x is the original instance, $c \in C$ is a counterfactual from the set of N counterfactuals, C, c_j is the j^{th} feature of an input, and $v(c_j)$ gives the value of the j^{th} feature of instance c. $Unique(c)$ gives the set of features that only counterfactual $c \in C$ modifies (compared to all other counterfactuals in C). $FeaturesChanged(c)$ gives the set of all features modified by the counterfactual $c \in C$ relative to the initial instance:

$$FeaturesChanged(c) = \{c_j \in c \mid v(c_j) \neq v(x_j)\}$$

Feature uniqueness, K, is the number of unique modified features over all counterfactuals, normalised by the total number of features (unique and not unique) targeted by all counterfactuals. We note that our *feature uniqueness* measure is similar to the *sparsity* measure introduced in Mothilal *et al.* [2020] that captures the number of features that are different, and the *feature diversity* in Smyth and Keane [2021] that measures the percentage of features that are different. Our measure is different from `Count-Diversity` of Mothilal *et al.* and the feature diversity from Smyth and Keane. They both consider the number of features that are change between pairs of counterfactuals while we propose a stricter definition that also insists the features must be unique within the entire set of counterfactuals.

Using the *feature uniqueness*, we separate explanations categorically into $high$ and low diversity. We acknowledge that the difference between two counterfactuals can be expressed in several ways, and as such, we use a simple technique of calculating feature uniqueness to determine D. We use this straightforward measure rather than fully exploring the effects of and best definition of the difference between counterfactuals. We define the diversity $D(C)$ of a set of counterfactuals C as:

$$D(C) = \left\{ \begin{array}{ll} low, & \text{for } K \leq 0.67 \\ high, & \text{for } K > 0.8 \end{array} \right\}$$

Our experiment excludes any data points between 0.67 and 0.8 to allow for a clear difference in their feature representation.

Together, these two metrics, N and D, fundamentally capture the essence of existing diversity metrics, that is, diversity by providing multiple counterfactual explanations and diversity along the lines of the number of features that need to change to arrive at an alternative decision.

3.2 Human Study

We use FICO's Home Equity Line of Credit[1] (HELOC) data set provided for their Explainable Machine Leaning Challenge[2]. We train a logistic regression model (prediction accuracy of 73%) and generate our counterfactuals using the integer programming technique proposed by Russell [2019]. We then visualise these counterfactuals in a table to be presented to participants alongside the applicant feature set. We frame the questions asked to participants as if they are an intermediary (e.g. a loan officer) between the model producing decisions, and the applicant who requested the product.

Our study design is inspired by techniques discussed by Hoffman *et al.* [2019]. We use task prediction (predicting the result produced by a model given some input) as a proxy for insight into the mental model formed by an explainee. In task prediction, participants are first trained on explanations and then are asked to predict outputs of the model on unlabelled examples. A higher success rate indicates a more sound interpretation of the decision process used by the model. To identify whether there is any change in perceived trust elicited by the diversity of counterfactuals, we use the 5-point Trust Scale based on the Cahour and Forzy's [2009] and Jian *et al.* [2000]'s trust scales.

Our experiment uses a human-subject survey on the Qualtrics[3] platform. Before the experiment, we received ethics approval from our institution. The survey had three phases: an introductory phase (explained in detail later), a training phase and a testing phase. We seek to build the participants' mental model in the training phase by presenting them with HELOC product applicants and then test this mental model in the testing phase. The applicants were presented individually as a table showing the features and their values used by the classifier.

A total of 251 participants were recruited through Amazon Mechanical Turk, a crowd-sourcing platform popular for obtaining data for human-subject experiments [Buhrmester *et al.*, 2011]. The opportunity to join the study is restricted to those with over 10,000 HITs completed with at least 95% acceptance rate. The survey presented to the participant includes some simple qualifying questions to filter out automated software respondents. Participants were from the United States, United Kingdom, Canada, Australia and New Zealand. Though we did not collect participant demographics for our sample, we provide the details of who might have participated in the study using the information about the population of MTurkers, provided via the *mturk-tracker* online

[1]FICO xML Challenge found at community.fico.com/s/xml

[2]https://community.fico.com/s/explainable-machine-learning-challenge

[3]https://www.qualtrics.com

service[4] developed by Difallah *et al.* [2018]. The population from which our sample was drawn had approximately 72% participants from the United States of America and 28% from the four countries. There were around 60% males and 40% females. In terms of age, 1% were between 60-70 years old, 5% were between 50-60, 10% between 40-50, 19% between 30-40, and 65% between 18-30 years.

Procedure: We test the number of counterfactuals N at intervals of 0, 1, 2, 4, and 8, where $N \in \{0, 1\}$ are the two baselines that present participants no counterfactuals and one counterfactual respectively. These baselines are included to grade performance of a participant who is not exposed to any counterfactuals and to only one counterfactual. We also test two levels of diversity, D: low and $high$ (as explained earlier). This leads to a total of eight conditions (note that diversity is zero for zero and one counterfactual(s)).

The survey was divided into three phases, introductory, training, and testing. In the first phase, the participants first received a plain language statement and a consent form. If the participants agreed to all items in the consent form, they were asked a question to filter out automated agents or bots. If the bot-check question was answered correctly, we provided the participants a tutorial. In the tutorial, we walked the participants through what they were required to do with a sample applicant record, what types of questions we were going to ask them and what types of responses we were expecting. Following the tutorial, the participants were randomly allocated to one of the eight conditions to start the training phase. Participants were paid USD $8.5 per hour for participating in the study and a bonus of 20 cents for every correct answer.

In the training phase, we presented the participants with eight different applicant records from the HELOC data set. Four of these applicants were approved, and the other four had their loans denied. Participants in all eight conditions were exposed to the same eight applicants. For each applicant, we present the 23 applicant features, F, their values, and the classification produced by the model. Alongside the values used by the model, we present N counterfactuals. These counterfactuals are presented in a sparse table where values are only present in cells for features that require a change. Otherwise, a hyphen denotes no change to the feature value. We present a partial example in Table 1 that shows an approved applicant with two counterfactuals. The first counterfactual requires changing one feature (*External Risk Estimate*), and the second requires changing two.

Existing research [Chi *et al.*, 1994] suggests that self-explanation improves understanding. Therefore, we encouraged the participants to interpret the features and their impact on the classification by prompting them with the following question:

> *The applicant has asked what would have needed to be different to receive the alternative outcome. Please provide a description of why the applicant received the outcome they did.*

In the test phase, we assessed the participant's understanding of, trust in, and overall satisfaction with the model based on their interaction with the model and the counterfactual explanations seen during the training phase. We present the participants with eight new applicants. Participants in all eight conditions were exposed to the same eight applicants. We omit the classification produced by the model as well as any counterfactuals. The participant is requested to complete two prediction tasks, TP-A and TP-B. TP-A has the participant provide their own prediction of what classification the model would produce for the applicant. TP-B has the participant select the set of features they believe are most significant in changing the model's outcome, should their values change. We ask that the participants select these significant features to invoke a counterfactual thought process in the participant.

Once TP-A and TP-B are completed for each of the eight test applicants, participants outside of the baseline group complete a Likert scale indicating their preference for more or fewer counterfactuals. We ask they base this preference on whether they believe they would have performed better on the prediction tasks. We refer to this as their "preference for diversity". All participants are then presented with a trust scale [Hoffman *et al.*, 2019] regarding the underlying model used in producing the classifications shown in the training phase of the experiment. Concerning trust, we ask the participants the following four questions (shown in Figure 3 defined by Hoffman *et al.* [2019] on a 5-point Likert scale (Completely disagree (1), Somewhat disagree (2), Neutral (3), Somewhat agree (4), Completely agree (5)):

1. What is your confidence in the model? Do you have a feeling of trust in it?;
2. Are the actions of the model predictable?;
3. Is the model reliable? Do you think it is safe?;
4. Is the model efficient at what it does?

Measures: To evaluate task performance for a participant p, we consider their TP-A and TP-B results. TP-A produces a simple measure of their success as a prediction score, *SP*. This value is simply the number of correct predictions out of the eight test participants. TP-B produces a total feature score:

$$TotalFeatureScore(p) = \sum_{a \in A} FeatureScore(p_a), \quad (5)$$

where

$$FeatureScore(a) = \frac{\sum_{f \in F'} w_f}{|F'|}. \quad (6)$$

This is the sum of the individual feature scores for the set of eight presented applicants A. The feature score for each applicant a is the summation of the weights w of the features F' selected by the participant for individual classification normalised by the number of features selected. This simple measure allowed us to judge whether participants were selecting highly weighted features.

When evaluating trust, we consider both the participant's preference for diversity and their answers to the trust scale. These values are simple means to identify whether there is any significant change in result depending on the number or

[4] http://demographics.mturk-tracker.com/#/gender/all

	Prediction: Good		CF1	CF2
History	External Risk Estimate	78	68	-
	Length of Credit History	13 years 11 months	-	-
	Time Since Most Recent Line of Credit Opened	0 years 0 months	-	-
	Average Line of Credit History Length	5 years 6 months	-	-
	Total No. Lines of Credit	56	-	-
	No. Lines of Credit Opened in Last 12 Months	3	-	5
	Percentage of Lines of Credit that are Instalment Loans	29	-	37

Table 1: Example applicant data presented to participants. This example shows only some of the 23 features and two counterfactuals.

diversity presented. We performed the DunnTest[5], which performs the post-hoc pairwise Kruskal-Wallis test with the p-values adjusted with the Bonferroni method.

4 Results

Before performing any analysis, we undertook a manual review of answers to the open-ended question from the 251 participants in phase one to filter those who may not have been paying attention. We removed those participants who provide nonsensical answers such as random strings of words or content totally irrelevant to the experiment (e.g writing about a completely different problem). After this process, we were left with 208 participants, and all results are based on these remaining 208 participants. The 208 participants were distributed into eight conditions as follows (number of counterfactuals-diversity: number of participants): 0 counterfactuals: 39 participants; 1: 25, 2-H: 30, 2-L: 27, 4-H: 21, 4-L: 22, 8-H: 24, 8-L: 20. Participants took on average 33.3 $minutes$ $(SD = 12.7)$.

We present our null and alternative hypotheses for **H1** - $H_0 : \mu_b = \mu_2 = \mu_4 = \mu_8$; $H_1 = \mu_b < \mu_{\{2,4,8\}}$ and **H2** – $H_0 : \tau_b = \tau_h = \tau_l$; $H_1 = \tau_b < \tau_{\{h,l\}}$. We condition these on the average scores (μ) of TP-A and TP-B and the trust score (τ) respectively, and the baselines are zero and one counterfactual explanation. We reject the null hypothesis if we identify a significant difference in performance for an interval of number or diversity.

Cond.	Avg Feature Score	Avg Prediction Score
0	2.39 (SD=0.79)	5.64 (SD=0.90)
1	3.74 (SD=1.34)	5.60 (SD=1.29)
2H	3.61 (SD=1.20)	6.00 (SD=1.02)
2L	4.21 (SD=1.28)	6.30 (SD=0.99)
4H	3.36 (SD=1.24)	5.62 (SD=1.16)
4L	2.81 (SD=0.87)	5.64 (SD=1.53)
8H	2.37 (SD=0.79)	5.46 (SD=1.53)
8L	2.34 (SD=0.70)	5.55 (SD=1.43)

Table 2: Average Feature and Prediction Scores by Condition

4.1 Task Prediction

Table 2 shows the averages of TP-A (Prediction Score) and TP-B (Feature Score) together with the standard deviations.

[5]https://rdrr.io/cran/DescTools/man/DunnTest.html

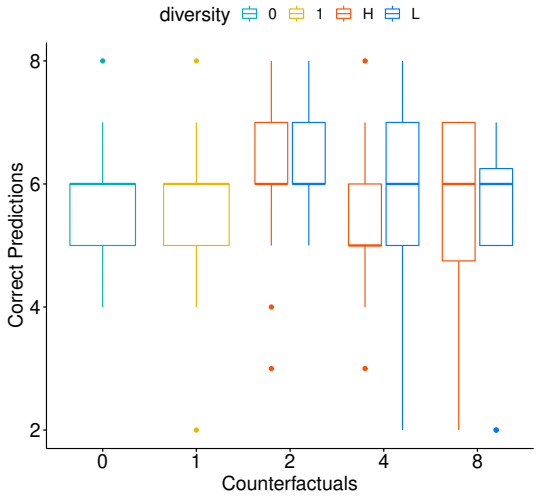

Figure 1: Quartiles and median scores for participant performance on TP-A: accuracy of prediction.

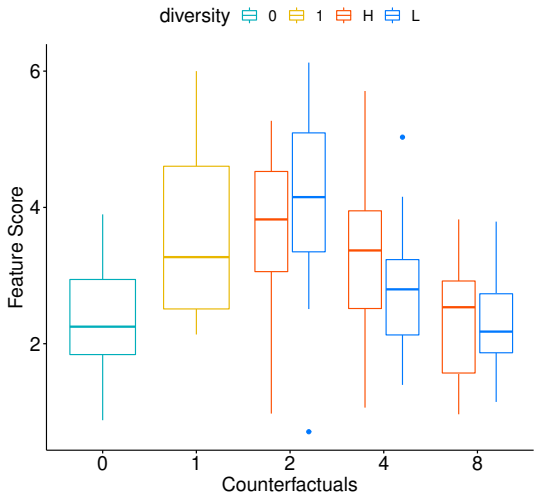

Figure 2: Quartiles and median scores for participant performance on TP-B: feature selection score.

Figure 1 shows participant performance in TP-A, that is, the ability of the participants to correctly state the classification produced by the machine learning model for the eight test

applicants. A Kruskal Wallis test revealed a not-significant effect of the number of counterfactuals on the mean performance of TP-A ($p = 0.09$) and diversity ($p = 0.43$). Kruskal-Wallis pair-wise tests confirm that there was no significant difference between the baseline and other groups. We also performed tests on the results for the number of times a participant selected "not sure" as an answer while completing TP-A and find no significance in the difference for number ($p = 0.09$) and diversity ($p = 0.73$). These results show that participants' ability to judge the classifier's decision boundary is not significantly improved by having more counterfactual explanations. In fact, we observed the best performance with only 2 counterfactual explanations.

Figure 2 shows participant performance in TP-B, that is, the ability of the participants to select the set of features that are most significant in changing the model's outcome. A Kruskal Wallis test revealed a significant effect of the number of counterfactuals on the mean performance of TP-B ($\chi^2(2) = 39.52, p < 0.001, \eta^2 = 0.26$), but not for diversity ($p = 0.90$). Pair-wise tests (DunnTest) revealed a significant difference between feature scores for all combinations ($p < 0.01$) ($\mu_0 = 2.39$ ($SD = 0.79$), $\mu_0 = 3.74$($SD = 1.33$), $\mu_2 = 3.89$ ($SD = 1.26$), $\mu_4 = 3.08$ ($SD = 1.09$), $\mu_8 = 2.35$ ($SD = 0.74$)) except for the following four pairs: when comparing $N = 0$ with $N = 4$ and $N = 8$, and when comparing $N = 1$ with $N = 2$ and $N = 4$. While results indicate a clear improvement in performance for those in the $N = 2$ group over the baseline of $N = 0$. The difference between the pair $N = 4$ and $N = 8$ and the pair $N = 2$ and $N = 8$ indicates once again that there is a point where an increase in N reduces performance and can make it comparable to the baseline. We point out the importance of this, given the $N = 0$ group having significantly more challenge in identifying features given no counterfactuals.

The above results for TP-A and TP-B lead to rejecting the null hypothesis (H_0) for our H1. This means that more diverse counterfactuals based on the $\ell 1$-distance (i.e. $||\cdot||_{1,\text{MAD}}$) and our feature uniqueness measure did not help participants improve their understanding of the machine learning model.

4.2 Trust

Our trust scale asked participants four questions [Cahour and Forzy, 2009]. Kruskal-Wallis tests did not identify any significance for number of counterfactuals or the diversity in any of the groups. The mean score of each found that answers for all questions generally indicate some positive level of trust in the model regardless of number or diversity (see Figure 3 for analysis of the responses based on the level of diversity. The results are similar for the number of counterfactuals and with baseline of 1 counterfactual). The results show that more diverse counterfactuals has not impact on trust in the model.

4.3 Preference for Number of Counterfactuals

We also asked the participants if they wanted more or fewer counterfactuals (Figure 4). Kruskal-Wallis tests revealed significance for the number of counterfactuals ($\chi^2(3) = 16.86, p < 0.001, \eta^2 = 0.08$). The eight counterfactuals group preference leans toward fewer counterfactuals while those presented with 2 and 4 prefer roughly the same number

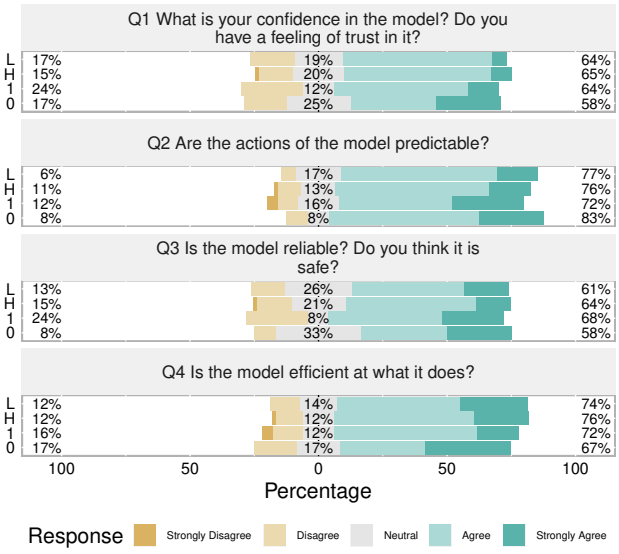

Figure 3: Trust scale based on diversity. L=low, H-high, and two baselines = zero and one counterfactuals

of explanations. The four counterfactuals groups leaned very slightly towards fewer explanations. The results with regards to TP-A and TP-B indicate that presenting two counterfactuals in our case already hits a threshold for "more is better". This shows that those in group 4 and 8's intuition of receiving fewer counterfactuals explanations about improving their performance are not unfounded.

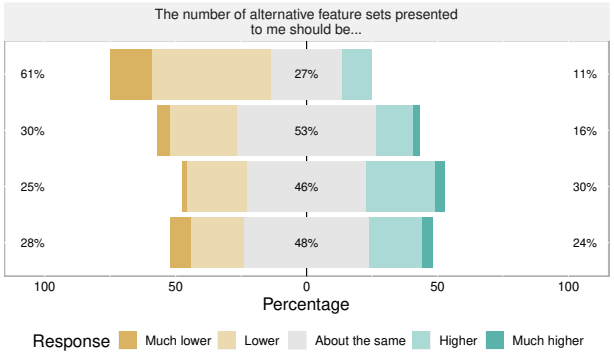

Figure 4: Preference for number of counterfactuals

4.4 Limitations

The primary limitation of this work so far is that we experimented with a single domain of credit scoring. More experiments are needed in other domains to comprehend the impact of diverse counterfactuals fully. However, even just the single experiment is a counterexample against the general assumption that diversity improves understanding. Further, we used non-expert participants on a task prediction domain, while we may find different results using loan officers making credit decisions, or in other domains. Despite this, it is important to note that our participants are representative of

loan applicants in this scenario. The results presented in this paper are limited to counterfactuals generated by considering the $\ell 1$-distance. There are other methods, such as [Mothilal *et al.*, 2020] and [Poyiadzi *et al.*, 2020]. Other diversity measures or other metrics may elicit better understanding and improved trust. In terms of the trust level, we measured the perceived trust through limited exposure to the machine learning model. We do not know how the trust level may change with many applicants and exposure to the decisions over time, nor whether trust as reliance would yielf conflicting results.

Having highlighted the above limitations, we note that our results suggest that more metrics, including those on diversity, need to be sufficiently validated to assess that the counterfactuals will have some utility for the intended users. However, we expect that other diversity measures would have limited impact, given that exposure to four or more counterfactuals was not useful for our participants.

5 Conclusion

This paper explored the impact that diverse counterfactuals have on elicited understanding and trust in a machine learning model. We found that two counterfactuals mildly impacted understanding while presenting participants with eight counterfactuals did not provide understanding, potentially due to increased cognitive load that hindered their understanding. We did not find a link between diversity and an improvement in understanding. Further, the number of counterfactuals and feature uniqueness had no impact on trust in the underlying model.

While these results call into question the assumption that multiple diverse counterfactuals improve understanding, it is important to note that there are other reasons why having multiple diverse counterfactuals may be desirable. In particularly, diverse counterfactuals could be important for *actionability* [Wachter *et al.*, 2017], as giving a small set of or less diverse counterfactuals may lead to situations in which all counterfactual states are infeasible.

References

[Adadi and Berrada, 2018] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160, 2018.

[Buhrmester *et al.*, 2011] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, pages 3–5, 2011.

[Byrne, 2019] Ruth MJ Byrne. Counterfactuals in explainable artificial intelligence (xai): Evidence from human reasoning. In *IJCAI*, pages 6276–6282, 2019.

[Cahour and Forzy, 2009] Béatrice Cahour and Jean-François Forzy. Does projection into use improve trust and exploration? an example with a cruise control system. *Safety science*, 47(9):1260–1270, 2009.

[Chi *et al.*, 1994] Michelene T.H. Chi, Nicholas De Leeuw, Mei-Hung Chiu, and Christian Lavancher. Eliciting self-explanations improves understanding. *Cognitive Science*, 18(3):439 – 477, 1994.

[Danenas and Garsva, 2015] Paulius Danenas and Gintautas Garsva. Selection of support vector machines based classifiers for credit risk domain. *Expert Syst. Appl.*, 42(6):3194–3204, April 2015.

[Difallah *et al.*, 2018] Djellel Difallah, Elena Filatova, and Panos Ipeirotis. Demographics and dynamics of mechanical turk workers. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 135–143, 2018.

[Gilpin *et al.*, 2018] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.

[Guidotti *et al.*, 2018] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

[Hoffman *et al.*, 2019] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. Metrics for explainable ai: Challenges and prospects, 2019.

[Jian *et al.*, 2000] Jiun-Yin Jian, Ann Bisantz, and Colin Drury. Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, 4:53–71, 10 2000.

[Joshi *et al.*, 2019] Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *CoRR*, abs/1907.09615, 2019.

[Karimi *et al.*, 2020a] Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera. Model-agnostic counterfactual explanations for consequential decisions. In *International Conference on Artificial Intelligence and Statistics*, pages 895–905. PMLR, 2020.

[Karimi *et al.*, 2020b] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. *arXiv preprint arXiv:2010.04050*, 2020.

[Lamy *et al.*, 2019] Jean-Baptiste Lamy, Boomadevi Sekar, Gilles Guezennec, Jacques Bouaud, and Brigitte Séroussi. Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. *Artificial Intelligence in Medicine*, 94:42–53, 3 2019.

[Lewis, 1973] David Lewis. *Counterfactuals*. Blackwell, 1973.

[Linardatos *et al.*, 2021] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18, 2021.

[Looveren and Klaise, 2019] Arnaud Van Looveren and Janis Klaise. Interpretable counterfactual explanations guided by prototypes. *CoRR*, abs/1907.02584, 2019.

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

[Miller, 2021] Tim Miller. Contrastive explanation: A structural-model approach. *Knowledge Engineering Review*, 36:E14, 2021.

[Mothilal *et al.*, 2020] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617, 2020.

[Osipovs, 2019] Pavels Osipovs. Classification tree applying for automated cv filtering in transport company. *Procedia Computer Science*, 149:406–414, 01 2019.

[Pawelczyk *et al.*, 2020] Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. Learning model-agnostic counterfactual explanations for tabular data. In Yennun Huang, Irwin King, Tie-Yan Liu, and Maarten van Steen, editors, *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 3126–3132. ACM / IW3C2, 2020.

[Poyiadzi *et al.*, 2020] Rafael Poyiadzi, Kacper Sokol, Raúl Santos-Rodríguez, Tijl De Bie, and Peter A. Flach. FACE: feasible and actionable counterfactual explanations. In Annette N. Markham, Julia Powles, Toby Walsh, and Anne L. Washington, editors, *AIES '20: AAAI/ACM Conference on AI, Ethics, and Society, New York, NY, USA, February 7-8, 2020*, pages 344–350. ACM, 2020.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. Association for Computing Machinery, 2016.

[Russell, 2019] Chris Russell. Efficient search for diverse coherent explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 20–28, 2019.

[Smyth and Keane, 2021] Barry Smyth and Mark T. Keane. A few good counterfactuals: Generating interpretable, plausible and diverse counterfactual explanations. *CoRR*, abs/2101.09056, 2021.

[Ustun *et al.*, 2019] Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM Press, 2019.

[Verma *et al.*, 2020] Sahil Verma, John P. Dickerson, and Keegan Hines. Counterfactual explanations for machine learning: A review. *CoRR*, abs/2010.10596, 2020.

[Wachter *et al.*, 2017] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

[Zavrsnik, 2020] Ales Zavrsnik. Criminal justice, artificial intelligence systems, and human rights. *ERA Forum*, 20, 02 2020.