The Recourse Problem under Model Multiplicity

Junqi Jiang *
Imperial College London
London , UK
junqi.jiang@imperial.ac.uk

Antonio Rago *
Imperial College London
London , UK
a.rago@imperial.ac.uk

Francesco Leofante
Imperial College London
London, UK
f.leofante@imperial.ac.uk

Francesca Toni
Imperial College London
London, UK
ft@imperial.ac.uk

ABSTRACT

Model Multiplicity (MM) arises when multiple, equally performing machine learning models can be trained to solve the same prediction task. Recent studies show that models obtained under MM may produce inconsistent predictions for the same input. When this occurs, it becomes challenging to provide counterfactual explanations (CEs), a common means for offering recourse recommendations to individuals negatively affected by models' predictions. Ensembling techniques have been proposed to deal with MM, but existing proposals fail to be robust to MM in that they cannot ensure that CEs are valid. In this paper, we identify desirable properties for supporting recourse under MM and show experimentally, on a number of datasets, that baselines from the literature do not satisfy them in general, thus advocating the need for novel ensembling techniques.

KEYWORDS

Explainable AI, Model Multiplicity, Counterfactual Explanations, Finance, Law

ACM Reference Format:
Junqi Jiang, Antonio Rago, Francesco Leofante, and Francesca Toni. 2023. The Recourse Problem under Model Multiplicity. In *Proceedings of XAI-FIN Workshop @ ICAI2023 (XAI-FIN'23).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The Model Multiplicity (MM) problem arises when multiple, equally performing (e.g. in terms of accuracy) models can be trained on the same data to solve a prediction task. These models may greatly differ in their internals [2, 28] and thus may produce inconsistent predictions for the same inputs at deployment time. A common way to overcome this problem is to resort to ensembling techniques, where the prediction outcomes of several models are aggregated to produce a single outcome. Aggregation can be performed in different ways as discussed in [1, 2], e.g. by *majority voting* [2], which is known to be an effective strategy to mediate conflicts between models and is routinely used in practical applications. However, existing solutions to the MM problem tend to ignore a crucial problem when deploying (individual or aggregated) prediction models, namely the need to provide avenues for recourse to users negatively impacted by the models' outputs, typically via the provision of *counterfactual explanations* (CEs) for the predictions (see [18, 30] for recent overviews of methods for generating CEs).

Dealing with MM while also taking CEs into account is non-trivial. Indeed, standard algorithms designed to generate CEs for single models typically fail to produce recourse recommendations that are valid across equally performing models, (e.g. see analysis in [25]), leading to a lack of what we may term *CE robustness under MM*.

In this paper we frame the recourse problem under MM in terms of a number of properties. We consider two natural extensions, to accommodate generation of CEs, of standard ensembling methods to deal with MM and show experimentally that they may violate some of the properties we define. Our analysis is supported by experiments on three datasets in the legal and financial contexts: loan approval [16], recidivism prediction [24], and credit risk [20]; even though the MM problem may occur with any models, we focus on neural network prediction models for these datasets, since, due to neural networks' sensitivity to randomness at training time, they suffer severely from the MM problem. Our analysis indicates that novel techniques for recourse-aware solutions to the MM problem are needed.

2 RELATED WORK

With the recent focus in AI on transparency and trustworthiness, the concept of model multiplicity (MM) in machine learning is attracting considerable attention [2]. MM amounts to the phenomenon that there usually exist multiple models that are equally accurate for a prediction task but behave differently for single inputs or groups of inputs, casting questions for model selection prior to or during development. MM is also known as the Rashomon Effect [4], or as predictive multiplicity [28].

Several works have shown evidence of the existence and severity of the problem. In particular, among equally accurate models, there could exist different fairness characteristics [8, 9, 14, 35, 36, 45], levels of model interpretability [7, 38, 39], and model robustness evaluations [10]. In terms of XAI, explanations of the equally-accurate

*Equal contributions.

XAI-FIN'23, 27 November 2023, New York, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

models have been found to be inconsistent [3, 12, 17, 26, 27, 29]. Most of the existing studies target deep neural networks, with [10] and [5] identifying under-specification as a common cause for the existence of various near-optimal models and distinct explanations.

Another line of work proposes useful metrics to measure the extent of MM. Specifically, [28] propose ambiguity and discrepancy metrics to characterise the severity of MM for classification tasks. Also, [44] extend similar measures to a probabilistic classification setting. This is also the target for [21] who introduce the concept of Rashomon Capacity. Similarly, [39] proposes a metric, Rashomon ratio, for a learning problem, to indicate the existence of simpler machine learning models.

Attempts have been made to address the MM problem. [2] proposes directions for possible solutions, specifying some meta-rules (e.g., accuracy, fairness evaluation thresholds) to filter out models, and using ensemble methods to aggregate models, or randomly selecting models. The selective ensembling method of [1] embeds statistical testing into the ensembling process, such that when the numbers of candidate models predicting an input to the top two classes are close, an abstention signal can be flagged for stakeholders. [46] propose an algorithm to enumerate all obtainable models under MM for decision trees, which can be useful for downstream tasks. [37] propose a model reconciling procedure to resolve the conflicts between two disagreeing models.

In the context of XAI, some have investigated feature attribution explanations under MM: specifically [26] construct ensembles by applying slight perturbations to model weights or by using mode connectivity to increase consistency of feature attribution explanations, and [27] finds more consistent feature attribution explanations by constructing uncertainty sets.

In this paper, we focus on solutions for the MM problem simultaneously accommodating recourse, in the form of counterfactual explanations (CEs). As a means to provide algorithmic recourse for data subjects, a CE for a prediction result of an input by a machine learning model was initially proposed as a data point that minimally modifies the input such that the prediction would change to a desired label [40, 43]. To improve the quality of CEs for real-world applications, several further desirable properties have been proposed. [42] advocate that CEs should account for the immutability of features. [11] propose that CEs should be lying on the data manifold and they should not be local outliers from the training data distribution. [31] generate a diverse set of CEs for each input instead of one data point. We refer readers to [18] for a comprehensive overview on CEs and their properties.

More closely related to the MM problem, [33] point out that CEs on data manifold are more likely to be valid for models under MM than minimum-distance CEs, whereas [25] propose an algorithm for a given ensemble of feed-forward neural networks to compute robust CEs that are provably valid for all models in the ensemble.

Another line of work focuses on the robustness of CEs against model changes (see [30] for an overview), e.g. parameter updates due to model retraining on the same or slightly shifted data distribution [3, 6, 13, 15, 19, 22, 23, 32, 41]. These studies usually evaluate the robustness of CEs over retrained models and demonstrate improved CE validity, which closely connects to the MM problem. However, as indicated in [34, 41], by making CEs more robust the closeness between CEs and the input will be undesirably sacrificed, meaning that the data subject would need more real-world efforts to achieve the recourse suggestions. Also, determining robust CEs is more computationally expensive.

Overall, though a few studies have investigated the validity of recourse under MM and closely related notions, little focus has been placed on providing recourse recommendations to individuals while reaching consistent predictions under MM.

3 MODEL MULTIPLICITY AND COUNTERFACTUAL EXPLANATIONS

Given a (non-empty) set of classification *labels* $\mathcal{L}$, a *model* is a mapping $M : \mathbb{R}^n \to \mathcal{L}$ ($n > 0$); we denote that M classifies an *input* $\mathbf{x} \in \mathbb{R}^n$ as ℓ iff $M(\mathbf{x}) = \ell$. Then, a *counterfactual explanation (CE)* for $\mathbf{x}$, given M, is some $\mathbf{c} \in \mathbb{R}^n \setminus \{\mathbf{x}\}$ such that $M(\mathbf{c}) \neq M(\mathbf{x})$, which may be optimised by some distance metric between the inputs (see [18, 30] for recent overviews of methods for generating CEs).

Models are typically obtained by training on some dataset. Model Multiplicity (MM) may occur when different techniques (e.g. XGBoost and neural networks) are used on the same dataset, but also when the same technique is used and there is inherently high randomness at training time, e.g. neural networks suffer severely from the MM problem due to their sensitivity to randomness at training time. This may result in a set of models $\mathcal{M}$, all equally performing on test sets, e.g. by exhibiting the same level of accuracy, but performing inconsistently at deployment time on some input of interest, due to the internal differences between models.

Example 3.1. Consider the commonly studied scenario of a loan application, where an applicant modelled by input $\mathbf{x}$ with features *unemployed* status, *33* years of age and *low* credit rating applies for a loan. Assume the bank has trained a set of models $\mathcal{M} = \{M_1, M_2, M_3\}$ to predict whether the loan should be granted (ℓ_1) or not (ℓ_0, i.e. $\mathcal{L} = \{\ell_0, \ell_1\}$). Each M_i may exhibit similarly good performance at training time but a conflict may arise when the models are deployed on $\mathbf{x}$, e.g. we observe that M_1 and M_2 would reject the loan ($M_1(\mathbf{x}) = M_2(\mathbf{x}) = \ell_0$), while M_3 would accept it ($M_3(\mathbf{x}) = \ell_1$).

A commonly used approach to tackle the MM problem amounts to *majority voting*, which may be recast in terms of the following notion of *naive ensembling* adapted from [2].

Definition 3.2. Consider an input $\mathbf{x}$, a (non-empty) set $\mathcal{L}$ of labels and a (non-empty) set $\mathcal{M}$ of models. We say that $\mathcal{M}$ *classifies* $\mathbf{x}$ *as* $\ell_i \in \mathcal{L}$ *by naive ensembling* iff[1]

$$\ell_i = argmax_{\ell_j \in \mathcal{L}} |\{M \in \mathcal{M} | M(\mathbf{x}) = \ell_j\}|.$$

We use $\mathcal{M}^{NE}(\mathbf{x}) = \ell_i$ to denote the aggregated classification. With an abuse of notation, we also let $\mathcal{M}^{NE}$ denote the set of models that agree on the aggregated classification, referred to as the set of *models obtained by naive ensembling* (for $\mathbf{x}$).

Example 3.3 (Example 3.1 continued). Given that M_1 and M_2 agree on the label ℓ_0 to be generated for $\mathbf{x}$, both disagreeing with M_3, in this simple case $\mathcal{M}^{NE}(\mathbf{x}) = \ell_0$ and $\mathcal{M}^{NE} = \{M_1, M_2\}$.

The definition of naive ensembling does not account for the need to provide counterfactual explanations (CEs) for model predictions.

[1] If there are multiple such ℓ_i, we choose one at random.

Doing so however is problematic, as the validity of CEs depends on the specific individual model, rather than the ensembling.

Example 3.4. Consider a set of models $\mathcal{M} = \{M_1, M_2, M_3, M_4, M_5\}$, for $\mathcal{L} = \{0, 1\}$, and an input $\mathbf{x}$ such that $M_1(\mathbf{x}) = M_2(\mathbf{x}) = M_3(\mathbf{x}) = 0$ and $M_4(\mathbf{x}) = M_5(\mathbf{x}) = 1$. Let $C = \{\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3, \mathbf{c}_4, \mathbf{c}_5\}$ be the set of CEs generated for $\mathbf{x}$ by the five models, i.e. $M_1(\mathbf{c}_1) = M_2(\mathbf{c}_2) = M_3(\mathbf{c}_3) = 1$, while $M_4(\mathbf{c}_4) = M_5(\mathbf{c}_5) = 0$. Applying naive ensembling yields $\mathcal{M}^{NE} = \{M_1, M_2, M_3\}$ and $\mathcal{M}^{NE}(\mathbf{x}) = 0$. Then, the set of CEs from agreeing models in $\mathcal{M}^{NE}$ is $C' = \{\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3\}$. Assume that $\mathbf{c}_1$ is invalid for M_2 (i.e. $M_2(\mathbf{c}_1) = 0$), $\mathbf{c}_2$ is invalid for M_1, $\mathbf{c}_3$ is invalid for M_2, and all three CEs are otherwise valid for all models in $\mathcal{M}^{NE}$. Then, which of the three CEs in C' should be given as a CE for $\mathcal{M}^{NE}(\mathbf{x}) = 0$?

We formalise the problem of generating recourse under MM and analyse a set of properties that are important in this setting. We then discuss possible extensions of naive ensembling to this setting and show how these extension may may fail to solve the recourse generation problem satisfactorily, thus motivating the need for a more expressive framework to deal with recourse under MM.

Below, we first frame the *recourse-aware MM problem* and then characterize its solutions in terms of formal properties. In the next section, we will then present two possible solutions, naturally building upon naive ensembling, and assess them against these properties.

Definition 3.5. Consider an input $\mathbf{x}$, a (non-empty) set $\mathcal{L}$ of labels, a (non-empty) set $\mathcal{M} = \{M_1, \ldots, M_k\}$ of models and a (non-empty) set $C = \{\mathbf{c}_1, \ldots, \mathbf{c}_k\}$ such that for each $i \in \{1, \ldots, k\}$, $\mathbf{c}_i$ is a CE for $\mathbf{x}$, given M_i. Then:

- a *candidate solution to the recourse-aware MM problem (for* $\mathbf{x}$*)* amounts to any $\mathcal{M}' \cup C'$ such that $\mathcal{M}' \subseteq \mathcal{M}$, $C' \subseteq C$;
- a *solution to the recourse-aware MM problem (for* $\mathbf{x}$*)* amounts to a cardinality-maximal candidate solution $\mathcal{M}' \cup C'$ to the recourse-aware MM problem (for $\mathbf{x}$) such that
  - $\mathcal{M}' \neq \emptyset$ and $C' \neq \emptyset$ (*non-emptiness*);
  - $\forall M_i, M_j \in \mathcal{M}'$, $M_i(\mathbf{x}) = M_j(\mathbf{x})$ (*model agreement*);
  - $\forall M_i \in \mathcal{M}'$ and $\forall \mathbf{c}_j \in C'$, $M_i(\mathbf{c}_j) \neq M_i(\mathbf{x})$ (*CE validity*);
  - $\forall i \in \{1, \ldots, k\}$, $M_i \in \mathcal{M}'$ iff $\mathbf{c}_i \in C'$ (*coherence*).

Thus, a candidate solution is any set of models and CEs from within the given pools $\mathcal{M}$ and C. Actual solutions need to satisfy additional, natural requirements. The first and most basic functional requirement for solutions of the recourse-aware MM problem is *non-emptiness*, requiring that the selected solution includes at least one model and at least one counterfactual . Clearly, as in the case of naive ensembling, the models in a solution cannot disagree among themselves, as required by *model agreement*. Further, we require that the generated recourse is robust. Previous work [25] considered a very conservative notion of robustness whereby CEs are required to be valid for all models in $\mathcal{M}$. While this might be desirable in some cases, we highlight that satisfying this property may not always be feasible in practice. We therefore propose a relaxed notion of robustness (*CE validity*), which requires that selected CEs are valid only for the selected models. The *coherence* requirement ensures that solutions are justifiable to end users. Indeed, intuitively, coherence requires that a CE is included in a solution iff it is supported by a model therein, guaranteeing strong justification as to why a given recourse is suggested, since selected models and their reasoning (represented by their CEs) are assessed in tandem.

Further requirements on candidate solutions to the recourse-aware MM problem can also be identified. The following is one such requirements, reflecting the spirit of naive ensembling.

Definition 3.6. Let $\mathcal{M}' \cup C'$ be a candidate solution to the recourse-aware MM problem (for $\mathbf{x}$). Assume that $\exists \ell_i = M_j(\mathbf{x})\ \forall M_j \in \mathcal{M}'$. Then $\mathcal{M}' \cup C'$ satisfies the property of *majority vote* iff $\nexists \ell_j \in \mathcal{L} \setminus \{\ell_i\}$ such that $|\{M_l \in \mathcal{M} | M_l(\mathbf{x}) = \ell_j\}| > |\{M_l \in \mathcal{M} | M_l(\mathbf{x}) = \ell_i\}|$.

Note that the existence of ℓ_i satisfying the condition '$\ell_i = M_j(\mathbf{x})\ \forall M_j \in \mathcal{M}'$' is automatically guaranteed if the candidate solution satisfies model agreement as per Definition 3.5.

We define two natural strategies leveraging naive ensembling to obtain candidate solutions to the recourse-aware MM problem.

Definition 3.7. Consider an input $\mathbf{x}$, a (non-empty) set $\mathcal{L}$ of labels, a (non-empty) set $\mathcal{M} = \{M_1, \ldots, M_k\}$ of models and a (non-empty) set $C = \{\mathbf{c}_1, \ldots, \mathbf{c}_k\}$ such that for each $i \in \{1, \ldots, k\}$, $\mathbf{c}_i$ is a CE for $\mathbf{x}$, given M_i. Let $\mathcal{M}^{NE} \subseteq \mathcal{M}$ be the set of models obtained by naive ensembling (for $\mathbf{x}$). Let

(S1) the set of *CEs from agreeing models* be

$$C_a = \{\mathbf{c}_i \in C \mid M_i \in \mathcal{M}^{NE}\};$$

(S2) the set of *valid-for-agreeing CEs from agreeing models* be

$$C_{va} = \{\mathbf{c}_i \in C \mid M_i \in \mathcal{M}^{NE} \wedge \forall M_j \in \mathcal{M}^{NE}, M_j(\mathbf{c}_i) \neq M_j(\mathbf{x})\}.$$

Then, $\mathcal{M}^{NE} \cup C_a$ is the ***S1**-driven candidate solution to the recourse-aware MM problem* and $\mathcal{M}^{NE} \cup C_{va}$ is the ***S2**-driven candidate solution to the recourse-aware MM problem.*

Intuitively, **S1** suggests taking all the CEs in C corresponding to the models that ended up being selected in $\mathcal{M}^{NE}$, whereas **S2** extends **S1** by enforcing the additional constraint that CEs are selected only if they are valid for all models in $\mathcal{M}^{NE}$. Note that the two candidate solutions given by strategies **S1**, **S2** above satisfy the property of majority vote by definition. Thus, for brevity, we will refer to these candidate solutions as *NE-ca* and *NE-cva* respectively. Note also that both NE-ca and NE-cva satisfy the model agreement requirement from Definition 3.5. We now provide an illustrative example to clarify the results produced by NE-ca and NE-cva.

Example 3.8 (Example 3.4 continued). $\mathcal{M}^{NE} = \{M_1, M_2, M_3\}$ is the result of naive ensembling. Then, the set of CEs from agreeing models (under **S1**) is $C_a = \{\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3\}$, and NE-ca=$\mathcal{M}^{NE} \cup C_a$. Given that $\mathbf{c}_1$ is invalid for M_2 (i.e. $M_2(\mathbf{c}_1) = 0$), $\mathbf{c}_2$ is invalid for M_1, $\mathbf{c}_3$ is invalid for M_2, and all three CEs are otherwise valid for all models in $\mathcal{M}^{NE}$, the set of valid-for-agreeing CEs from agreeing models (under **S2**) is empty, i.e. $C_{va} = \emptyset$, and NE-cva=$\mathcal{M}^{NE}$.

This example shows that, while satisfying by definition model agreement and majority vote, the strategy-driven candidate solutions may not satisfy some of the other properties required for actual solutions, e.g. NE-cva satisfies neither non-emptiness nor coherence in the setting of the example. Next, we assess how NE-ca and NE-cva behave empirically on three datasets in legal and financial settings.

	acc.	size M/C	c val. (fail)	acc.	size M/C	c val. (fail)	acc.	size M/C	c val. (fail)
	heloc			compas			credit		
$\|\mathcal{M}\| = 10$	.709±3e-3			.856±7e-4			.664±8e-3		
NE-ca	.709	.943/.943	.657 (.00)	.858	.980/.980	.572 (.00)	.697	.817/.817	.757 (.00)
NE-cva	.709	.943/.309	1.00 (.34)	.858	.980/.174	1.00 (.51)	.697	.817/.457	1.00 (.44)
$\|\mathcal{M}\| = 20$	.710±3e-3			.855±9e-4			.663±4e-3		
NE-ca	.717	.940/.940	.626 (.00)	.859	.978/.978	.544 (.00)	.708	.810/.810	.734 (.00)
NE-cva	.717	.940/.230	1.00 (.37)	.859	.978/.111	1.00 (.60)	.708	.810/.351	1.00 (.62)
$\|\mathcal{M}\| = 30$	.710±3e-3			.855±6e-4			.663±4e-3		
NE-ca	.718	.940/.940	.620 (.00)	.859	.976/.976	.530 (.00)	.710	.807/.807	.727 (.00)
NE-cva	.718	.940/.205	1.00 (.41)	.859	.976/.087	1.00 (.56)	.710	.807/.311	1.00 (.70)

Table 1: Quantitative evaluations of the S1-driven candidate solution to the recourse-aware MM problem (NE-ca) and the S2-driven candidate solution to the recourse-aware MM problem (NE-cva) on three dataset (heloc, compas, and credit). The acc. entries in the rows starting with $|\mathcal{M}|$ are the average single model accuracies. The size M/C columnns report the average size of the generated sets of models and CEs, respectively, measured as percentages of $|\mathcal{M}|$. The c val. columnns report the average validity of CEs in the generated sets over the generated models. The fail columns report the percentage of test inputs for which a method fails to produce valid CEs.

4 EXPERIMENTS

We focus on three datasets: loan approval (heloc) [16], recidivism prediction (compas) [24], and credit risk (credit) [20]. Even though the MM problem may occur with any models, we focus on neural network prediction models for these datasets, since, due to neural networks' sensitivity to randomness at training time, they suffer severely from the MM problem.

For each dataset, we train 150 neural classifiers using 80% of the dataset. We use the remaining 20% of each dataset as test inputs (or 500 inputs if the test set is larger). The 150 neural networks are trained using different random seeds for parameter initialisation and different train-test splits, forming a pool of possible models under MM from which we sample multiple sets $\mathcal{M}$ of models. At each run, we randomly sample $\mathcal{M}$ of sizes 10, 20, and 30 from the pool of possible models and apply NE-ca and NE-cva over each of the test inputs. For each size, we repeatedly run five times with different selections of model sets, and we record the mean and standard deviation of the results over five runs.

We evaluate NE-ca and NE-cva against the following metrics:

- prediction accuracy over the test set (*acc.*);
- average size of the generated set of models and set of CEs, measured as percentages of $|\mathcal{M}|$ (*size M/C*);
- average validity of CEs (a valid CE counts 1, an invalid CE counts 0) in the generated sets over the generated models (**c** *val.*);
- percentage of test inputs for which a method fails to produce CEs (*fail*).

The *size M/C* measure is identical for models and CEs for the NE-ca method, by design, as this guarantees coherence in Definition 3.5. In the case of NE-cva, this measure provides an indication of how often the requirement of coherence is satisfied. The **c** *val.* measure provide indications on how often the generated CEs satisfy the requirement of CE validity in Definition 3.5. The *fail* measure provides an indication of the violation of non-emptiness in Definition 3.5.

We report the results of the evaluation in Table 1. These results are averaged over all the test inputs except for the failure cases. We also report the average test set accuracies of the models in $\mathcal{M}$.

Clearly neither of the two methods provide fully-fledged solutions to the recourse-aware MM problem for all inputs, as, in particular, they cannot guarantee validity of generated CEs/existence of valid CEs in the generated sets.

5 CONCLUSIONS

We have proposed a novel framing of the model multiplicity problem in machine learning that accounts for the need of accommodating recourse for negatively affected users. Our framework envisages that, when deploying, on previously unseen inputs, models that are similarly performing at training times, one needs to select simultaneously models and counterfactual explanations for the inputs, guided by natural requirements on the chosen models and counterfactual explanations. We have defined two natural extensions of the popular majority vote solution to model multiplicity [2] so as to accommodate recourse. We have also shown experimentally, on datasets in the legal and financial contexts, that these two methods may fail to satisfy the envisaged properties, thus indicating that further work in this space is very much needed towards convincing recourse-aware model multiplicity solutions.

REFERENCES

[1] Emily Black, Klas Leino, and Matt Fredrikson. 2022. Selective Ensembles for Consistent Predictions. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022.* https://openreview.net/forum?id=HfUyCRBeQc

[2] Emily Black, Manish Raghavan, and Solon Barocas. 2022. Model Multiplicity: Opportunities, Concerns, and Solutions. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022.* 850–863. https://doi.org/10.1145/3531146.3533149

[3] Emily Black, Zifan Wang, and Matt Fredrikson. 2022. Consistent Counterfactuals for Deep Models. In *ICLR 2022.* https://openreview.net/forum?id=St6eyiTEHnG

[4] Leo Breiman. 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science* 16, 3 (2001), 199–231.

[5] Marc-Etienne Brunet, Ashton Anderson, and Richard Zemel. 2022. Implications of Model Indeterminacy for Explanations of Automated Decisions. *Advances in Neural Information Processing Systems* 35 (2022), 7810–7823.

[6] Ngoc Bui, Duy Nguyen, and Viet Anh Nguyen. 2022. Counterfactual Plans under Distributional Ambiguity. In *10th Int. Conf. on Learning Representations, ICLR*.

[7] Chaofan Chen, Kangcheng Lin, Cynthia Rudin, Yaron Shaposhnik, Sijia Wang, and Tong Wang. 2018. An interpretable model with globally consistent explanations for credit risk. *arXiv preprint arXiv:1811.12615* (2018).

[8] A Feder Cooper, Ellen Abrams, and Na Na. 2021. Emergent unfairness in algorithmic fairness-accuracy trade-off research. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 46–54.

[9] Amanda Coston, Ashesh Rambachan, and Alexandra Chouldechova. 2021. Characterizing fairness over the set of good models under selective labels. In *International Conference on Machine Learning*. PMLR, 2144–2155.

[10] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. 2022. Underspecification presents challenges for credibility in modern machine learning. *The Journal of Machine Learning Research* 23, 1 (2022), 10237–10297.

[11] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. 2018. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. *Advances in neural information processing systems* 31 (2018).

[12] Jiayun Dong and Cynthia Rudin. 2019. Variable importance clouds: A way to explore variable importance for the set of good models. *arXiv preprint arXiv:1901.03209* (2019).

[13] Sanghamitra Dutta, Jason Long, Saumitra Mishra, Cecilia Tilli, and Daniele Magazzeni. 2022. Robust Counterfactual Explanations for Tree-Based Ensembles. In *39th Int. Conf. on Machine Learning, ICML*.

[14] Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush Varshney. 2020. Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. In *International conference on machine learning*. PMLR, 2803–2813.

[15] Andrea Ferrario and Michele Loi. 2022. The Robustness of Counterfactual Explanations Over Time. *IEEE Access* 10 (2022), 82736–82750. https://doi.org/10.1109/ACCESS.2022.3196917

[16] FICO. 2018. Explainable Machine Learning Challenge. https://community.fico.com/s/explainable-machine-learning-challenge

[17] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. 2019. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *J. Mach. Learn. Res.* 20, 177 (2019), 1–81.

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

[19] Faisal Hamman, Erfaun Noorani, Saumitra Mishra, Daniele Magazzeni, and Sanghamitra Dutta. 2023. Robust Counterfactual Explanations for Neural Networks With Probabilistic Guarantees. *arXiv preprint arXiv:2305.11997* (2023).

[20] Hans Hofmann. 1994. Statlog (German Credit Data). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5NC77.

[21] Hsiang Hsu and Flavio Calmon. 2022. Rashomon Capacity: A Metric for Predictive Multiplicity in Classification. *Advances in Neural Information Processing Systems* 35 (2022), 28988–29000.

[22] Junqi Jiang, Jianglin Lan, Francesco Leofante, Antonio Rago, and Francesca Toni. 2023. Provably Robust and Plausible Counterfactual Explanations for Neural Networks via Robust Optimisation. *CoRR* abs/2309.12545 (2023). https://doi.org/10.48550/arXiv.2309.12545 arXiv:2309.12545

[23] Junqi Jiang, Francesco Leofante, Antonio Rago, and Francesca Toni. 2023. Formalising the Robustness of Counterfactual Explanations for Neural Networks. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*. AAAI Press, 14901–14909.

[24] Surya Mattu Julia Angwin, Jeff Larson and Lauren Kirchner. 2016. There's software used across the country to predict future criminals. And it's biased against blacks. https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis

[25] Francesco Leofante, Elena Botoeva, and Vineet Rajani. 2023. Counterfactual Explanations and Model Multiplicity: a Relational Verification View. In *Proceedings of the 20th International Conference on Principles of Knowledge Representation and Reasoning, KR 2023, Rhodes, Greece, September 2-8, 2023*, Pierre Marquis, Tran Cao Son, and Gabriele Kern-Isberner (Eds.). 763–768. https://doi.org/10.24963/kr.2023/78

[26] Dan Ley, Leonard Tang, Matthew Nazari, Hongjin Lin, Suraj Srinivas, and Himabindu Lakkaraju. 2023. Consistent Explanations in the Face of Model Indeterminacy via Ensembling. *arXiv preprint arXiv:2306.06193* (2023).

[27] Charles Marx, Youngsuk Park, Hilaf Hasson, Yuyang Wang, Stefano Ermon, and Luke Huan. 2023. But are you sure? an uncertainty-aware perspective on explainable ai. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 7375–7391.

[28] Charles T. Marx, Flávio P. Calmon, and Berk Ustun. 2020. Predictive Multiplicity in Classification. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*. 6765–6774. http://proceedings.mlr.press/v119/marx20a.html

[29] Johannes Mehrer, Courtney J Spoerer, Nikolaus Kriegeskorte, and Tim C Kietzmann. 2020. Individual differences among deep neural network models. *Nature communications* 11, 1 (2020), 5725.

[30] Saumitra Mishra, Sanghamitra Dutta, Jason Long, and Daniele Magazzeni. 2021. A survey on the robustness of feature importance and counterfactual explanations. *arXiv preprint arXiv:2111.00358* (2021).

[31] Ramaravind Kommiya Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual expl.. In *Conf. on Fairness, Accountability, and Transparency, FAT*. https://doi.org/10.1145/3351095.3372850

[32] Tuan-Duy H Nguyen, Ngoc Bui, Duy Nguyen, Man-Chung Yue, and Viet Anh Nguyen. 2022. Robust Bayesian Recourse. In *38hth Conference on Uncertainty in AI, UAI*.

[33] Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. 2020. On Counterfactual Explanations under Predictive Multiplicity. In *Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence, UAI 2020, virtual online, August 3-6, 2020 (Proceedings of Machine Learning Research, Vol. 124)*, Ryan P. Adams and Vibhav Gogate (Eds.). AUAI Press, 809–818. http://proceedings.mlr.press/v124/pawelczyk20a.html

[34] Martin Pawelczyk, Teresa Datta, Johannes van-den Heuvel, Gjergji Kasneci, and Himabindu Lakkaraju. 2022. Probabilistically Robust Recourse: Navigating the Trade-offs between Costs and Robustness in Algorithmic Recourse. *arXiv preprint arXiv:2203.06768* (2022).

[35] Kit T Rodolfa, Hemank Lamba, and Rayid Ghani. 2021. Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy. *Nature Machine Intelligence* 3, 10 (2021), 896–904.

[36] Kit T Rodolfa, Erika Salomon, Lauren Haynes, Iván Higuera Mendieta, Jamie Larson, and Rayid Ghani. 2020. Case study: predictive fairness to reduce misdemeanor recidivism through social service interventions. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 142–153.

[37] Aaron Roth, Alexander Tolbert, and Scott Weinstein. 2023. Reconciling Individual Probability Forecasts. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 101–110.

[38] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* 1, 5 (2019), 206–215.

[39] Lesia Semenova, Cynthia Rudin, and Ronald Parr. 2022. On the existence of simpler machine learning models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1827–1858.

[40] Gabriele Tolomei, Fabrizio Silvestri, Andrew Haines, and Mounia Lalmas. 2017. Interpretable Predictions of Tree-based Ensembles via Actionable Feature Tweaking. In *23rd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/3097983.3098039

[41] Sohini Upadhyay, Shalmali Joshi, and Himabindu Lakkaraju. 2021. Towards robust and reliable algorithmic recourse. *Adv. in Neural Inf. Processing Systems 34: NeurIPS* (2021).

[42] Berk Ustun, Alexander Spangher, and Yang Liu. 2019. Actionable Recourse in Linear Classification. In *Conf. on Fairness, Accountability, and Transparency, FAT*. https://doi.org/10.1145/3287560.3287566

[43] Sandra Wachter, Brent D. Mittelstadt, and Chris Russell. 2017. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harv. JL & Tech.* 31 (2017), 841.

[44] Jamelle Watson-Daniels, David C Parkes, and Berk Ustun. 2023. Predictive multiplicity in probabilistic classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 10306–10314.

[45] Michael Wick, Jean-Baptiste Tristan, et al. 2019. Unlocking fairness: a trade-off revisited. *Advances in neural information processing systems* 32 (2019).

[46] Rui Xin, Chudi Zhong, Zhi Chen, Takuya Takagi, Margo Seltzer, and Cynthia Rudin. 2022. Exploring the whole rashomon set of sparse decision trees. *Advances in Neural Information Processing Systems* 35 (2022), 14071–14084.