Ethical and Fairness Implications of Model Multiplicity

Kacper Sokol  
Kacper.Sokol@rmit.edu.au  
ARC Centre of Excellence for Automated Decision-Making and Society, RMIT University  
Melbourne, Australia

Jeffrey Chan  
Jeffrey.Chan@rmit.edu.au  
ARC Centre of Excellence for Automated Decision-Making and Society, RMIT University  
Melbourne, Australia

Meelis Kull  
Meelis.Kull@ut.ee  
Institute of Computer Science, University of Tartu  
Tartu, Estonia

Flora Dilys Salim  
Flora.Salim@rmit.edu.au  
ARC Centre of Excellence for Automated Decision-Making and Society, RMIT University  
Melbourne, Australia

ABSTRACT

While predictive models are a purely technological feat, they may operate in a social context in which benign engineering choices entail unexpected real-life consequences. Fairness – pertaining both to individuals and groups – is one of such considerations; it surfaces when data capture protected characteristics of people who may be discriminated upon these attributes. This notion has predominantly been studied for a fixed predictive model, sometimes under different classification thresholds, striving to identify and eradicate its undesirable behaviour. Here we backtrack on this assumption and explore a novel definition of fairness where individuals can be harmed when one predictor is chosen ad hoc from a group of equally well performing models, i.e., in view of model multiplicity. Since a person may be classified differently across models that are otherwise considered equivalent, this individual could argue for a model with a more favourable outcome, possibly causing others to be adversely affected. We introduce this scenario with a two-dimensional example based on linear classification; then investigate its analytical properties in a broader context; and finally present experimental results on data sets popular in fairness studies. Our findings suggest that such unfairness can be found in real-life situations and may be difficult to mitigate with technical measures alone, as doing so degrades certain metrics of predictive performance.

KEYWORDS

Model View, Individual Fairness, Machine Learning

1 A NEW NOTION OF FAIRNESS

Data-driven predictive models are making great strides across numerous domains, leading to their quick proliferation throughout businesses and society, either supporting decision-making or outright automating relevant tasks. This speedy adoption of Artificial Intelligence (AI) and Machine Learning (ML) algorithms, however, outpaces research investigating their potential harmful impact on society. While excluding people from the decisive process bears the promise of faster and more precise outputs that lack implicit human biases, some (historical) patterns concealed in training data may easily overshadow these benefits. The ubiquity of oftentimes erratically tested and validated models can therefore contribute to and amplify problems with fairness, accountability and robustness of the predictive tasks being addressed. This is of particular concern when AI and ML models affect humans – with possibly long-term or legally binding decisions – which has been documented in various contexts including school admissions, job hiring, banking and judiciary rulings [1, 16]. It is thus critical to ensure fairness (among other desiderata) of the resulting predictions with respect to protected attributes such as ethnicity or gender when dealing with individuals and parts of the population (groups).

While mitigating disparate treatment of groups and individuals in relation to their protected attributes is of paramount importance, other notions of algorithmic fairness should not be neglected. Here we explore a novel viewpoint where instead of analysing bias of a single model we deal with their collection [5, 14] characterised by equal (or comparable) predictive performance according to a chosen evaluation strategy and metric [13]. For example, see Figure 1a where three distinct predictors from the family of linear classifiers all achieve 100% accuracy with respect to the plotted validation data. While models perfect in this respect do not make any observable mistakes, they may still suffer from disputable regions in which they disagree, giving rise to possible claims of unfair treatment for unseen individuals landing in these spaces. Whenever we cannot guarantee perfect classification on the designated validation set, however, these considerations become more immediate as selected individuals from within these data may be treated differently across seemingly equivalent models, as is the case in Figures 1b and 1c. Notably, as we start to employ more expressive families of predictive algorithms, their complexity and parameterisation space grow, possibly making the observed phenomenon more prominent, especially for workflows relying on stochastic components.

Since from the viewpoint of predictive performance – including an identical confusion matrix in certain cases – there may be no obvious reason to favour one model over another, a predictor could possibly be selected at random. Such an arbitrary choice, however, can be easily dismissed by (adversely affected) individuals who are classified differently across this collection of models since they can argue to be treated with the predictor granting them a favourable outcome instead (especially if the selection procedure was ad hoc in the first place). Nonetheless, as the models under consideration are intrinsically different, many choices will inevitably result in some people gaining at the expense of others, which in any case is difficult to justify. Notably, even if a collection of models boasts perfect performance with respect to a designated validation set, we still have to account for any disputable regions that are not covered by
these data, hence unnoticed, yet allow individuals who are placed there to claim unfair treatment. This line of reasoning becomes particularly important in the context of criminal justice, where a person might be entitled to be treated by the most advantageous model as long as it achieves certain performance for the (validation) population as a whole. Arguably, this concept of fairness can be linked to the Blackstone’s ratio [4] or the “presumption of innocence” [2], where lack of convincing evidence (or a unanimous vote) warrants the most favourable treatment – “It is better that ten guilty persons escape than that one innocent suffers.”

To the best of our knowledge such considerations have found limited recognition in fairness literature – albeit they have been reported in ML [13] – therefore our findings aim to establish solid foundations and provide initial analysis of individual fairness under model multiplicity [5]. With an abundance of predictive pipelines built from various data processing techniques and model families, deriving bounds on the least and most favourable treatment of each individual in a dedicated fairness validation data set may not be feasible. Understanding the stability of each prediction under certain modelling assumptions could nonetheless guarantee fairness, trustworthiness and accountability of important data-driven decisions or hand them over to a human supervisor. Additionally, such a purview encourages incorporating diverse selection heuristics and criteria that consider properties beyond predictive performance, e.g., overall complexity or coverage of a model, to increase chances of its uniqueness. Another possible research question concerns cases where only very few classifiers from a given performance band present an individual with a favourable outcome. For example, consider such a curated ensemble of predictive models taking the role of jurors where an overwhelming majority offers the less desirable decision. While the unfavourable ratio can be ignored with just a single model creating a precedence for a positive prediction, considering the proportion of classifiers and individual reasons for their respective outputs may provide important insights.

In particular, this paper investigates the scope and implications of individual fairness concerns under model multiplicity for crisp binary classifiers where one outcome is preferred over the other, e.g., grant or decline parole. This is especially important for data-driven decisions that influence humans, where an arbitrary choice of a model from such a group may disparately affect certain people, posing ethical dilemmas for engineers building and deploying predictive pipelines. While these observations are explicit for instances contained in a dedicated fairness evaluation set – as well as for the data used for performance validation when dealing with non-perfect models – they are just as much concerning, yet less pronounced, in a broader scope determined by disputable regions, which pertain even to seemingly perfect models as demonstrated in Figure 1. Specifically, this paper identifies, introduces and formally defines a novel notion of individual fairness stemming from disputable regions recognised under two distinct types of utility-based model multiplicity (Section 2), exemplified by and discussed for linear classification in two dimensions. It further offers analytical treatment of the problem (Section 3), investigating the influence of the expressiveness (flexibility) of predictive models on fairness and consequences of granting each individual the best possible prediction. Next, in Section 4 we propose a novel visualisation approach to help discover and analyse the degree of unfairness across the multiplicity spectrum, both as a high-level overview and a more detailed instance-specific perspective, which we apply to real-life data. Before we conclude our work and outline future research directions in Section 6, we discuss relevant literature in Section 5.

These contributions underlie concepts fundamental to our notion of individual fairness. To introduce them we first investigate diverse binary classification scenarios for a two-dimensional synthetic data set. In the process we uncover caveats and assumptions relevant to the modelling tasks and the approaches used to evaluate predictive performance; these findings help us to identify principles of fairness under utility-based model multiplicity and devise strategies to address consequences of this phenomenon. Our preliminary results show that allowing an individual to choose the (most favourable) model can have detrimental effects on the overall predictive performance of the classification system, particularly so when the employed AI or ML model is relatively expressive, hence flexible with respect to its decision boundary. This observation inspired an approach to limit the number of admissible predictors by imposing upon them (modeling) restrictions consistent with the natural process responsible for generating the underlying data, thus also making these models more interpretable [17]. Our graphical

Figure 1: Multiplicity of linear models with predictive performance (accuracy) measured on validation data (plotted). Perfect classifiers – Panel (a) – may still yield unfair decisions for (initially unobserved) instances falling into disputable spaces.
investigative tools corroborate these findings; we first explain them for the two-dimensional running example and next apply them to three real-life data sets popular in fairness studies: Credit Approval, German Credit and Adult. These experiments are based on top-performing classification workflows submitted to OpenML [19] for each of the aforementioned data sets.

While model multiplicity has been studied in ML as a metric of accountability – hence by extension robustness and fairness – of predictive systems [13], these considerations have not attracted much attention in the fairness community. Such a viewpoint is radically different from the most prominent types of fairness analysis published in the literature, which presuppose a fixed algorithmic predictor that is either assessed for discriminatory behaviour or modified to mitigate its biases [3]. In particular, there are two general categories of disparate impact: group fairness, where an algorithm is expected to perform similarly on distinct sub-populations (e.g., partitioned on protected attributes) according to a selected metric; and individual fairness, where the same logic is applied to any given data point. The latter may be seen through the lens of causality, thus rely on a counterfactual analysis [11], whereas the former can be framed as adjusting group-based classification thresholds in search of parity [9]. Notably, enforcing fairness in certain cases may require trading off a degree of utility (i.e., predictive performance) [8], and some fairness metrics are inherently incompatible with each other [10]. A quite distinct and complementary view on algorithmic fairness considers social and population changes over time, dealing with the delayed impact of fair decisions [12]. A more detailed discussion of relevant literature is presented in Section 5.

2 NAVIGATING MODEL MULTIPLICITY

The notion of fairness dealt with in this paper is built upon the model multiplicity phenomenon understood here as existence of a collection of data-driven predictors that are indistinguishable in terms of their predictive performance under a certain (fixed) evaluation strategy [5, 13]. Furthermore, throughout this work we are interested in crisp binary classification in which one outcome is universally preferred to the other by individuals whose case is being decided. Therefore, a model \( f : X \to Y \) classifies an instance \( x \in X \) as \( f(x) = y \), where \( y \in Y \equiv \{0, 1\} \) and 1 represents a favourable outcome. Additionally, predictive performance of such a model is measured on a predetermined validation data subset \( \tilde{X} \subseteq X \) accompanied by annotations \( \tilde{Y} \subseteq Y \), using a selected metric \( m : Y \times Y \to \mathbb{R} \) calculated as \( m(f(\tilde{X}), \tilde{Y}) \). In our initial analysis we rely on linear, polynomial, \( k \)-nearest neighbours and decision tree classifiers applied to a synthetic two-dimensional data set, which helps us to demonstrate the principles of our findings through visual inspection and hand-crafted examples; nonetheless, all of our results can be generalised beyond this setting.

While in principle model multiplicity may span a diverse range of classifiers, we restrict our considerations to a confined family of predictive functions (akin to such a formalisation of curves). This is desirable as distinct data-driven algorithms exhibit unique characteristics that translate into different shapes of their respective decision boundaries, allowing for diverse mistakes to be made, as seen when comparing Figures 1c and 2. Therefore, a family of data-driven models \( \mathcal{F} \) consists of trained classifiers \( f \in \mathcal{F} \) based upon a single predictive algorithm, i.e., \( \mathcal{F} := \{ f : X \to Y \} \). It can further be constrained by imposing restrictions on the parameterisation space or optimisation approach, among other properties, used with the selected method, e.g., a collection of decision trees no deeper than 7. Notably, algorithms relying on a stochastic training procedure may yield distinct models from a single specification when run multiple times. A family of models can also, by extension, include more complex predictive workflows built with pre- or post-processing steps such as input normalisation or output calibration.

To address fairness concerns under model multiplicity we further identify a subset of classifiers from within a single family that is characterised by a fixed level of predictive performance (i.e., utility). To this end, we need to select an evaluation metric; for simplicity we employ accuracy throughout this paper, which is defined as \( m_{\text{acc}}(\tilde{Y}, \hat{Y}) = \frac{\sum_{y \in \hat{Y}} 1(\hat{y} = y)}{|\hat{Y}|} \), where \( \hat{Y} = f(X) \) are predictions and \( Y \) are annotations for a data set \( X \). Furthermore, we require a designated collection of instances \( \tilde{X} \) and their ground truth \( \hat{Y} \) to serve as a dedicated validation set of predictive performance. This adds to the existing training and validation data used to fit the model and tune its hyper-parameters, both of which constitute an integral part of AI and ML workflows. Notably, the performance validation set \( \tilde{X} \) can additionally be employed to evaluate our notion of fairness, however in certain cases – discussed in Section 3 – it may be more beneficial to separate the two.

Utility-based model multiplicity \( \mathcal{F}_u \) is therefore determined by a fixed level of predictive performance \( \epsilon \in \mathbb{R} \) among classifiers from within a single family of data-driven models \( \mathcal{F} \). In particular, there are two meaningful viewpoints on multiplicity:

- **population-based** where we consider the performance of each model across the entire data space \( X \), which may be computationally infeasible given a possibly infinite number of qualifying models (e.g., slight alterations of a linear classifier); and
- **validation-based** where the performance is measured on a designated validation set \( X \), making the problem tractable.

In view of the former, the three models shown in Figure 1a are distinct, whereas based on the latter they are indistinguishable, i.e., they belong to a single multiplicity class \( \mathcal{F}_u \). Either of the two is distinct from a purely theoretical non-uniqueness of models within
Figure 3: Theoretical multiplicity of decision trees.

...a single family, where the same decision boundaries – thus unobserved changes – can be achieved with different realisations of a predictive pipeline; for example, see Figure 3 depicting two structurally different decision trees classifying the entire data space in the same fashion. Throughout this paper we are predominantly concerned with the validation-based multiplicity, which is detailed in Definition 2.1. Such a setting simplifies our considerations and allows us to avoid working with a possibly infinite number of models. Notably, this notion can be extended by allowing a certain variation δ in the desired level of predictive performance ε, i.e., \( \mathcal{F}_{\varepsilon+\delta} \) which we discuss in Appendix A.

**Definition 2.1. Utility-based model multiplicity \( \mathcal{F}_\varepsilon \) is a collection of classifiers from across a single family of models \( \mathcal{F} \) that exhibit a fixed level of predictive performance ε for a chosen metric \( m \):

\[
\mathcal{F}_\varepsilon := \left\{ f \in \mathcal{F} : m\left( f(\bar{X}), \bar{Y} \right) = \varepsilon \right\},
\]

where \( \bar{X} \) is a designated validation data set and \( \bar{Y} \) are its corresponding labels.

It is important to observe that the choice of the performance metric determines which models are considered equivalent. For example, all of the classifiers shown in Figure 1c make two mistakes, hence are indistinguishable with respect to accuracy; however, a different metric, e.g., precision, recall or specificity, would make the violet, orange and green models distinct. Since all of the predictive performance metrics for classification tasks are derived from confusion tables, instead of relying on calculated numerical values we can refer directly to the underlying contingency matrices. Table 1 lists specific errors made for the red and blue classes by the models shown in Figure 1c, where the green predictors incorrectly classify two blue points, the violet model makes two red errors and the orange classifier mistakes one blue and one red instance. However, even such a fine-grained approach is insufficient to capture the same type of a mistake made on different data points, i.e., individual predictions, motivating our notion of individual fairness under utility-based model multiplicity. We are therefore interested in classifiers that are indistinguishable in terms of performance measured on a dedicated population but provide a distinct class assignment for certain individuals (who are not necessarily included in this data set).

A generalisation of such individual mistakes – inspired by the population-based model multiplicity \( \mathcal{F}_\varepsilon \) viewpoint – are disputable spaces \( \bar{\mathcal{X}}_{\mathcal{F}_\varepsilon} \), an example of which is shown in Figure 1a. This extension – outlined by Definition 2.2 – is useful when the designated fairness validation data set \( \bar{X} \subseteq X \) is sparse and cannot capture inconsistency of predictions at the desired level of detail, i.e., \( \bar{X} \not\subseteq \bar{\mathcal{X}}_{\mathcal{F}_\varepsilon} \).

**Definition 2.2. A disputable space (or region) \( \bar{\mathcal{X}}_{\mathcal{F}_\varepsilon} \subseteq X \) for utility-based model multiplicity \( \mathcal{F}_\varepsilon \) is given by:

\[
\bar{\mathcal{X}}_{\mathcal{F}_\varepsilon} := \left\{ x \in X : \exists f_i, f_j \in \mathcal{F}_\varepsilon \text{ s.t. } f_i(x) \neq f_j(x) \right\}
\]

for a chosen predictive metric \( m \) and labelled (performance) validation set \( (\bar{X}, \bar{Y}) \), where \( \forall f \in \mathcal{F}_\varepsilon, m\left( f(\bar{X}), \bar{Y} \right) = \varepsilon \) (see Definition 2.1 for more details).

If the main property considered while choosing a classifier \( f \in \mathcal{F}_\varepsilon \) is predictive performance, then from the perspective of utility-based model multiplicity \( \mathcal{F}_\varepsilon \) all such predictors may be viewed as equivalent, without any particular preference for a given classifier. Lacking some further, well-defined selection criteria, an arbitrary choice can nonetheless lead to unfair treatment of individuals given the existence of an equally suitable model that may provide these people with a more favourable outcome. For example, consider the green classifiers shown in Figure 1c; both of them have identical confusion matrices (Table 1c) yet only one of the two borderline blue individuals may be assigned a desirable outcome – the red class – depending on the model choice. Framing such a scenario as an automated decision between granting (red) or denying (blue) a parole in an (algorithmic) judicial hearing, performance-based indistinguishability of models becomes an important matter. While ideally the task would be to minimise the breadth of any disputable regions, i.e., to deal with \( \bar{\mathcal{X}}_{\mathcal{F}_\varepsilon} \), in this paper we focus on a designated fairness validation set \( X \subseteq X \) as given by Definition 2.3, which outlines this type of model-based disparate treatment. Notably, such a notion of fairness can be expanded from individuals to groups by comparing the impact of multiplicity, e.g., ratios of affected instances, across them.

**Definition 2.3. A classifier \( f \in \mathcal{F}_\varepsilon \) is fair towards individuals \( x \in \bar{X} \subseteq X \) in view of utility-based model multiplicity \( \mathcal{F}_\varepsilon \) iff \( \forall f' \in \mathcal{F}_\varepsilon \forall x \in \bar{X}, f(x) = f'(x) \).

Lack of fairness in such a setting entails treating each individual with the most beneficial model \( f \in \mathcal{F}_\varepsilon \). Therefore, we can define a...
new, universally fair model \( f^* \) by combining all of the classifiers from their collection determined by utility-based model multiplicity \( \mathcal{F}_e \). This method may be considered as a simple model ensemble where we are interested in the best, rather than the average, result. An example of such a composition of classifiers is shown in Figure 4 for the models depicted in Figure 1b, assuming that the red class (1) is preferred to the blue class (0). (This particular fair classifier is composed of only the two green models as they maximise the space predicted with the favourable outcome.) Notably, it is possible that the fair model itself does not belong to the family, i.e., \( f^* \notin \mathcal{F}_e \), which is visible in the aforementioned example, as the decision boundary of \( f^* \) is not linear. The individually fair model \( f^* \) is formalised in Definition 2.4.

**Definition 2.4.** An individually fair classifier \( f^* \) under utility-based model multiplicity \( \mathcal{F}_e \) is defined as:

\[
f^*(x) := \max_{f \in \mathcal{F}_e} f(x)
\]

for any instance \( x \in \mathcal{X} \). (Recall that 1 is the preferred prediction in our binary classification setting.)

### 3 TOWARDS MODEL-BASED FAIRNESS

Operating in the utility-based model multiplicity purview creates various challenges for the proposed notion of fairness. While in theory a collection of such classifiers is not directly affected by the composition of the underlying training or test data, the corresponding performance and fairness validation sets are respectively key to model equivalence and identification of unfair behaviour. The representativeness, density and relative distribution of the fairness validation sets should therefore be carefully considered, and potentially expanded or augmented over time. In addition to the modelled data space, special attention ought to also be placed on auxiliary information pertinent to the classifiers themselves. Overall stability of the model family, which may depend on the stochasticity of its training procedure, as well as proximity of instances to decision boundaries – for example, conveyed by prediction uncertainty and (over- or under-) confidence – and behaviour in sparse data regions all play a role in the volatility of automated decision-making. Moreover, the inherent expressiveness and flexibility of the chosen model, and more broadly the data modelling workflow that can be composed upon it by incorporating new steps such as data preprocessing or feature engineering, can also be problematic. The operationalisation of a fair model given by Definition 2.4 should be scrutinised as well since it may lead to a drop in predictive performance, which can be noted in Figure 4 where \( f^* \) misclassifies two instances whereas the base models \( f \in \mathcal{F}_e \) make just one mistake. Importantly, all of these observations create opportunities (predominantly on technical grounds) for individuals adversely affected by an automated prediction to easily challenge its fairness under utility-based model multiplicity.

One conceivable attack on our notion of individual fairness can come from under-specification of the selected family of classifiers \( \mathcal{F} \) or a predictive workflow built around it. Expressive classifiers may be flexible enough to single out individual instances and assign them an arbitrary prediction, for example due to their inherent complexity or parameterisation scope. Such a freedom can adversely influence model-based fairness by permitting any two instances to swap predictions – regardless of their placement in the data space – while maintaining the desired level of performance \( \epsilon \), i.e., operating within the designated utility-based model multiplicity class \( \mathcal{F}_e \). This can be easily achieved for classifiers with a considerable parameterisation space such as deep neural networks, but it is also relevant to simple models, e.g., \( k \)-nearest neighbours, when they are misconfigured as shown in Figure 5.

Flexibility of a model \( f \) from a given family \( \mathcal{F} \) can be defined through a proxy such as complexity \( \Omega(f) \) or Vapnik–Chervonenkis dimension [20] and placed as an additional constraint next to the desired level of predictive performance \( \epsilon \). The precise specification of such a metric may be unique to each model family, for example, the number of non-zero parameters for linear models, the highest coefficient degree for polynomial classifiers and depth, width or number of instances per leaf for decision trees. More generally, expressiveness, hence flexibility, of models and workflows built around them may be restricted by ensuring diversity of training data, providing a lower bound on confidence of each decision, imposing restrictions on model parameterisation, enforcing regularisation such as pruning for trees and LASSO for linear models, or limiting overfitting in any other way. Otherwise, with a certain budget of mistakes determined by the required level of predictive performance, an excessive number of individuals could claim unfair treatment.

The aforementioned strategies tasked with constraining the model space can be complemented by (use case-specific) operational requirements. For example, instead of a single performance metric used to determine model equivalence, their hierarchy can be implemented, first measuring accuracy, followed by precision and recall to address any ties. More broadly, such an approach could be applied directly to the classification confusion matrix by sequentially imposing restrictions on its individual entries. Arguably, one could optimise exclusively for recall and specificity to maximise the number of favourable decisions; however, any such strategy is just a stopgap as it is likely to lack a solid foundation supported by convincing arguments. A more meaningful heuristic should therefore rely on well-defined properties of the employed model [5], for

---

**Figure 4:** Combining the two green models \( f \in \mathcal{F}_e \) shown in Figure 1b gives an individually fair classifier \( f^* \) under utility-based model multiplicity \( \mathcal{F}_e \). (Note that towards top left and out of view there is an additional segment of the violet model from Figure 1b.)
example, monotonicity of a particular feature with respect to the prediction, which is consistent with some definitions of explainability in AI and ML [15, 17, 18].

Lacking a (close to) unique classifier — obtained by imposing desiderata to narrow down the scope of the chosen family of predictive models, making any alternative difficult or impossible to find — we may need to rely instead on the fair classifier $f^*$ outlined by Definition 2.4. Treating each individual with the best available model in this fashion may however degrade the overall predictive performance for the task at hand, making the fair application of data-driven decisions arguably on a par with less effective classifiers to begin with. For example, this phenomenon can be observed for the fair model $f^*$ plotted in Figure 4. Notably, any model family $\mathcal{F}_\mathcal{F}$ subject to population- or validation-based multiplicity, i.e., with meaningful alternatives, is bound to suffer from disputable spaces. Depending on the density of data in these regions and their size, the fair model $f^*$ will boast a no worse recall but no better specificity than any individual model $f \in \mathcal{F}_\mathcal{F}$ as stated by Corollary 3.1. This change in predictive performance is especially prominent (reaching its limit) when the chosen family of models $\mathcal{F}$ is expressive enough to single out individual data points, in which case the fair classifier $f^*$ will assign the most favourable class to every instance — 0% specificity and 100% recall — as given by Corollary 3.2.

**Corollary 3.1.** Specificity $m_\mathcal{F}$ of an individually fair classifier $f^*$ under utility-based model multiplicity $\mathcal{F}_\mathcal{F}$ is no better than that of any other classifier $f \in \mathcal{F}_\mathcal{F}$:

$$\forall f \in \mathcal{F}_\mathcal{F} \ \ m_\mathcal{F}(f^*(X), Y) \leq m_\mathcal{F}(f(X), Y),$$

where $(X, Y)$ is the labelled data space. Additionally, recall $m_r$ of $f^*$ is no worse than that of any $f \in \mathcal{F}_\mathcal{F}$:

$$\forall f \in \mathcal{F}_\mathcal{F} \ \ m_r(f^*(X), Y) \geq m_r(f(X), Y).$$

The result above (Corollary 3.1) can be interpreted geometrically as expanding the space predicted with the more favourable outcome (red in our case) with the entirety of disputable regions — see Figure 4. It follows directly from observing individual predictions under $f^*$:

- Each data point from the preferred class (red) that is predicted as such, i.e., true positives, will not be affected; but misclassified positive instances, i.e., false negatives, may be corrected by $f^*$ — a possible improvement in recall.
- Each data point from the undesirable class (blue) that is predicted as such, i.e., true negatives, may be classified as positive by $f^*$; but misclassified negative instances, i.e., false positives, will not be affected — a possible decrease in specificity.

Furthermore, the fair classifier $f^*$ for models $f \in \mathcal{F}_\mathcal{F}$ from a family that is flexible enough to predict any individual with an arbitrary class will assign the preferred output to all data, from which Corollary 3.2 follows. These observations should encourage the developers of ML models to strive for high predictive performance evaluated on a comprehensive and representative validation data set to minimise the scope of any disputable regions.

**Corollary 3.2.** When a family of classifiers $\mathcal{F}_\mathcal{F}$ under utility-based model multiplicity is expressive enough to assign a selected class $c \in \mathcal{Y}$ to any data point $x \in X$, i.e.,

$$\forall x \in X \ \ \exists f \in \mathcal{F}_\mathcal{F} \text{ s.t. } f(x) = c,$$

the corresponding individually fair classifier $f^*$ achieves 0% specificity $(m_\mathcal{F})$ and 100% recall $(m_r)$:

$$m_\mathcal{F}(f^*(X), Y) = 0 \quad m_r(f^*(X), Y) = 100,$$

therefore every point is assigned the favourable outcome.

Given the benefits of finding a classifier $f \in \mathcal{F}_\mathcal{F}$ that is favourable for a concerned individual, the concept of fairness under utility-based model multiplicity can be framed as an adversarial challenge. In such a setting people disparately affected by an automated decision can confront the owner of the underlying predictive model by building an equivalent classifier (performance-wise) that provides the desired outcome instead. Therefore, “adversaries” attempt to identify a predictor $f^\prime$ from within the employed family of models $\mathcal{F}_\mathcal{F}$ — ensuring that it complies with all of the restrictions imposed by $\mathcal{F}_\mathcal{F}$, including the predetermined level of predictive performance $\varepsilon$ — which assigns a selected individual the most favourable class. The party responsible for building and deploying the challenged model, on the other hand, ought to minimise the possible number of such claims by reducing the size of any disputable regions. This
(a) **Stability profile** counts unique prediction vectors, i.e., class assignments across the entire fairness validation set, for a collection of models from a chosen family $\mathcal{F}$ grouped by predictive performance (accuracy) measured on a dedicated validation set. For example, the orange section informs us that there are 10 models with 99% accuracy spanning 6 unique class assignments (horizontal segments).

(b) **Fairness profile** captures the behaviour of each predictive model for individual instances from the fairness validation set; the colour intensity differentiates the two possible prediction types (sorted for the clarity of presentation). The plot shows 21 models (rows) belonging to three performance bands.

(c) Summary of the **fairness profile** – Panel (b) with values vertically sorted for each column in each performance band – may be more convenient to read for large fairness validation data sets, especially when they cannot be easily sorted.

Figure 6: **Stability and fairness** profiles for the two-dimensional data and models used across the paper. For simplicity, the performance validation set is also used to evaluate model-based individual fairness – note that the number of mistakes that can be read for each row in Panel (b) agrees with the corresponding accuracy (given that the data are sorted on their true labels). While such plots are intended for a single family of models $\mathcal{F}$, the panels above span diverse types of classifiers – linear, polynomial, decision trees and $k$-nearest neighbours – depicted in Figures 1, 2, 3 and 5 for illustrative purposes.

tug-of-war is bound to have a positive effect on the investigated predictive model by iteratively improving its accountability, robustness and overall quality.

Facilitating this workflow, however, requires releasing (a subset of) relevant training data as well as performance and fairness validation sets (in addition to specification of the utilised model family $\mathcal{F}$), which may be problematic due to inherent trade secrets. While distributing training data cannot be easily sidestepped, predictive performance of a model may be assessed without access to evaluation data as the classifier itself can instead be submitted for testing – akin to how ML competitions are run. Notably, having a comprehensive performance validation set that faithfully represents all of the individual cases is advantageous for the model creators as it restricts the number of admissible classifiers. Lastly, publishing a fairness validation set may also be beneficial to the owner of the model under investigation, encouraging a broader community to identify unfair predictions – and, more widely, disputable spaces – as well as engendering trust in the deployed model itself. This decoupling of the two – predictive performance validation data and fairness validation set – may therefore be desirable, especially that the latter does not need to be labelled since we are only interested in disparate treatment of these instances across equivalent classifiers $f \in \mathcal{F}$ to assess model-based fairness.

4 IMPACT OF MODEL MULTIPLICITY

To study the real-life impact of model multiplicity on individual fairness we require a toolkit to identify data points that are classified inconsistently for a selected family of models $\mathcal{F}$ and a dedicated approach to effectively communicate this phenomenon. As we have observed earlier in Section 2, standard performance metrics derived from confusion matrices are insufficient to this end – see Table 1c for reference. To fill this gap we propose a visual analytic tool – called **stability profile** – that summarises the prediction structure for each individual performance band $\varepsilon$ of a model family $\mathcal{F}$ (a model...
We complement this approach with fine-grained inspection plots – fairness profiles – that detail precise classification results for all data points across chosen performance bands $\epsilon$ (a prediction view). We explain these tools for the two-dimensional example used throughout this paper and then apply them to real-life data sets popular in fairness studies. The model families $\mathcal{F}$ investigated in this section are based on predictive workflows published in the OpenML repository [19]. Additionally, Appendix A analyses the same selection of properties when the performance bars are relaxed, and Appendix B inspects the (drop in) utility (accuracy) of fair models $\mathcal{F}^*$ constructed for the aforementioned data sets.

Stability profiles are our first visual diagnostic tool that provides an overview of predictive volatility for the entire fairness validation set from the model’s perspective – see Figure 6a for an example. It captures consistency of predictions – measured by the number of unique class assignments for the entire data set – across models belonging to each performance band $\epsilon$. At a glance, a stability profile informs us of the number (horizontal segments) and frequency (width of each segment) of different prediction vectors, with the most advantageous, i.e., the most fair, shape amassing all models in the bottom part. The other visual diagnostic tool are fairness profiles, which shed light on the volatility of, hence fairness towards, individual instances. These can either be plotted accurately for each model in a given performance band – Figure 6b – or instead aggregated for improved readability – Figure 6c. Ideally, each column ought to be in a single shade (light or dark), which indicates a consistent prediction for models within a single and across different levels of predictive performance.

Equipped with the right tools, we are in a position to assess fairness under utility-based model multiplicity for real-life data sets popular with the research community; in particular, we look into Credit Approval, German Credit and Adult. In lieu of designing and testing our own classifiers, we turn to OpenML – a reproducibility repository for machine learning experiments [19]. Such an approach ensures that the predictive workflows used for our studies are realistic, especially that we opt for one of the top-performing models for each data set: histogram-based gradient boosting classification tree, random classification forest and decision tree-based AdaBoost respectively. The ML task handled in each case is (supervised) crisp classification run on 10-fold cross validation; for our experiments, unless stated otherwise, we select the model trained on folds 2–10 whose performance is evaluated on fold 1, which also serves as the fairness validation set. A summary of this setup given as IDs of OpenML data, tasks and flows is provided in Table 2.

A stability profile for each data set–predictive workflow pairing under investigation is shown in Figure 7. Given a large number of (accuracy-based) performance bands, the plots show only some of the best performing models. While for the Adult and German Credit data sets the top classifiers appear fair, this ceases to hold for models with a slightly lower accuracy. In comparison, the predictive workflow for the Adult data set seems to generate a large collection of classifiers whose accuracy differs beyond the second decimal point, highlighting potential fairness issues. One reason for this behaviour may be a significantly larger predictive performance and fairness validation set – 4,885 instances as compared to 68 and 100 for the other two – however we show that this is not entirely the case using our second inspection method. (See Appendix A for an in-depth analysis.)

To paint a more accurate picture we generate summary fairness profiles – shown in Figure 8 – for our experiments. A consistent colour shading for each column, i.e., data point, conveys a stable, thus fair, prediction. A vertical bar that changes its saturation in the middle of a performance band indicates the maximum disagreement view).
between the models. While the profiles for the Credit Approval and German Credit data sets appear relatively consistent, the one for Adult is more jittery (note that for legibility it only shows a subset of individuals, all of whom are treated unfairly). To further investigate this phenomenon we analyse the ambiguity [13] – proportion of instances treated unfairly in each performance band – of the predictive workflows using all the data folds as fairness validation sets. The results are shown in Figure 9 and corroborate the insights gathered thus far. Specifically, the models for Credit Approval and German Credit seem to behave consistently, whereas the one used for Adult does not show any discernible patterns (see Appendix C for more details). We suspect that the culprits are volatility and high expressiveness of the underlying predictor – decision tree-based AdaBoost – however we cannot confirm it without further analysis. Regardless, our toolkit appears to serve its purpose well.

5 RELATED WORK

Despite its relatively young age, research on AI and ML fairness has attracted considerable attention in the recent years following rapid proliferation of data-driven predictive systems in real-life applications. Two main themes dominate this field: individual and group-based fairness, both focusing on strategies to identify and mitigate disparate impact [3]. Nonetheless, complementary viewpoints also emerge, investigating this topic under substantially different assumptions, e.g., future impact of enforcing fair decisions [12]. Regardless, a nearly universal assumption found in the literature is a fixed predictive model, thus overlooking its provenance and any implicit choices related to it. This is at odds with the model multiplicity phenomenon [5] – likely existence of a collection of equally capable classifiers – which has been largely neglected by the fairness community [7, 13].

Individual fairness deals with disparate treatment of a single person based on selected comparison criteria such as relevant protected characteristics. This can be captured by a dedicated similarity metric – assuming an intuitive notion that similar people should be treated comparably – however defining such a metric is non-trivial [6]. An alternative approach to individual fairness is formulated via counterfactuals, whereby changing a (protected) attribute should not yield a different prediction. This notion is also very intuitive and relatively easy to test, however the concept of a “counterfactual
individual” may be ill-defined. For example, altering ethnicity while preserving the remaining features intact may not be representative given that all the other personal traits and attributes are influenced by this single personal characteristic. Notably, causality may be employed to partially overcome such shortcomings [11]. Group-based fairness, on the other hand, deals with disparate treatment of sub-populations determined, for example, by divisions across protected attributes, in which case parity may be achieved by tweaking the underlying model separately for each group [9]. Many such notions of fairness are inherently incompatible [10], and enforcing some of them may adversely affect predictive performance [8].

Multiplicity of data-driven predictive models is a well observed phenomenon – sometimes called the Rashomon effect of statistics – where a group of classifiers exhibits comparable utility despite intrinsic differences [5]. This collection of models may rely on different subsets of attributes or patterns present in the training data, with some arguing that exogenous information may be required to narrow down its scope [7, 14]. In relation to fairness, model multiplicity has been used to understand the dependence of (inaccessible) classifiers on selected (protected) attributes, which can help to robustly identify discriminatory practices [7]. More recently, model multiplicity was suggested as an additional metric for evaluating classifiers, showing how this phenomenon may be problematic from an ethical and fairness perspective when individuals receive conflicting predictions and offering to address this issue by granting them the most favourable decision, akin to our approach [13]. While the line of reasoning presented by Marx et al. [13] is close to ours, their study is limited to linear models based on Mixed-Integer Programming, whereas ours is rooted directly in fairness and provides a more fundamental and general treatment of model multiplicity supported by a comprehensive analysis of real-life predictive pipelines.

6 CONCLUSIONS AND FUTURE RESEARCH

In this paper we formalised the notion of individual fairness under utility-based model multiplicity – a scenario in which a number of classifiers, considered equivalent based on their predictive power, provides certain data points with different predictions. Such a situation may arise for distinct types of models, diverse parameterisation of the same model or over its multiple training attempts when the underlying process is inherently stochastic. Additionally, high-dimensional and sparse data may contribute to this phenomenon given the curse of dimensionality (everything is far away from each other). In particular, we built these concepts around a user-specified family of predictive models, taking into consideration its expressiveness, which in the extreme may lead to granting every individual the most favourable prediction. We then defined two meaningful cases of model multiplicity: one determined by the entire data space and another measured with respect to a designated fairness validation data set. Furthermore, we generalised the former into disputable spaces – regions where at least two models disagree on the prediction – which may not necessarily be identified even with a comprehensive fairness validation set.

Next, we showed how to combine equivalent (performance-wise) models to present each individual with the best possible decision; however, such an approach may adversely affect the overall predictive power of the classification task at hand, especially for unexpectedly expressive models. As an alternative, we discussed imposing restrictions on the model space to limit the number of viable alternatives, arguing for enforcing (operational) constraints meaningful to each individual modelling problem, e.g., prediction monotonicity with respect to a chosen attribute, which is also recognised as a strategy for introducing (ante-hoc) explainability where predictions are aligned with human values and can be easily justified. Otherwise, being able to determine disputable spaces – “grey areas” or “edge cases” from a classification standpoint – can allow to engage a human expert in the decisive process, thus partially mitigate the burden of unfair predictions. To help identify individual unfairness stemming from model multiplicity we introduced bespoke visualisation tools, explained them for a simple two-dimensional example and applied them to real-life data sets popular in fairness research using predictive pipelines sourced from the OpenML repository. All of our findings motivate the importance of considering utility-based model multiplicity as a new dimension of algorithmic fairness.

In future work, we will investigate numerical metrics and analytical tools to assess multiplicity-based unfairness in a broader context, extending the notion beyond crisp binary classification. Moreover, we will look into developing methods to derive bounds on the most and least favourable treatment of each individual from within a given data set for a selection of predictive models. This should allow us to systematically identify the disputable and stable regions, possibly leading to a new optimisation objective – minimise the former or maximise the latter – for training individually fair, robust and effective predictive models. Finally, we will explore the lower limit of utility when employing our fair classifier that is created by applying the most beneficial model with a given level of predictive performance.
ACKNOWLEDGMENTS
This research was supported by the ARC Centre of Excellence for Automated Decision-Making and Society, funded by the Australian Government through the Australian Research Council (project number CE200100005); the Estonian Research Council (projects number PUT1458 and PRG1604); and the Dora Plus programme, sponsored by the European Regional Development Fund and Estonian government.

REFERENCES
[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. ProPublica.
[2] UN General Assembly. 1948. Universal declaration of human rights. UN General Assembly 302, 2 (1948), 14–25.
[3] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2017. Fairness and Machine Learning: Limitations and Opportunities.
[4] William Blackstone. 1830. Commentaries on the Laws of England. Vol. 2. Collins & Hannay.
[5] Leo Breiman. 2001. Statistical modeling: The two cultures. Statist. Sci. 16, 3 (2001), 199–231.
[6] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In Proceedings of the 3rd innovations in theoretical computer science conference. 214–226.
[7] 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.
[8] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2021. The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making. Commun. ACM 64, 4 (2021), 136–143.
[9] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. Advances in neural information processing systems 29 (2016).
[10] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In 8th Innovations in Theoretical Computer Science Conference (ITCS 2017). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.
[11] Matt Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. Advances in neural information processing systems 30 (2017).
[12] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed impact of fair machine learning. Proceedings of the 35th International Conference on Machine Learning (2018).
[13] Charles Marx, Flavio Calmon, and Berk Ustun. 2020. Predictive multiplicity in classification. In International Conference on Machine Learning. PMLR, 6765–6774.
[14] James W McAllister. 2007. Model selection and the multiplicity of patterns in empirical data. Philosophy of Science 74, 5 (2007), 884–894.
[15] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence 267 (2019), 1–38.
[16] Cathy O’Neil. 2016. Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books.
[17] 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.
[18] Kacper Sokol and Peter Flach. 2021. Explainability Is in the Mind of the Beholder: Establishing the Foundations of Explainable Artificial Intelligence. arXiv preprint arXiv:2112.14466 (2021).
[19] Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. 2013. OpenML: Networked Science in Machine Learning. SIGKDD Explorations 15, 2 (2013), 49–60. https://doi.org/10.1145/2641190.2641198
[20] V. N. Vapnik and A. Ya. Chervonenkis. 1971. On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. Theory of Probability & Its Applications 16, 2 (1971), 264–286. https://doi.org/10.1137/1116005
A RELAXING PERFORMANCE BANDS

The notion of utility-based model multiplicity \( F_e \) outlined in Definition 2.1 can be relaxed to allow for variation in the desired level of predictive performance \( \epsilon \). We can argue in favour of such an approach given that the predictive performance is measured on a data subset, an expansion or reduction of which is likely to yield some deviation in utility. Therefore, we can specify the allowed tolerance through an additional parameter \( \delta \in \mathbb{R}^+ \) – where the performance band is denoted with \( \epsilon \pm \delta \) – as outlined in Definition A.1. An alternative operationalisation of this concept can be achieved by rounding the predictive performance \( \epsilon \) to a specified decimal point \( \delta \), written as \( \epsilon \approx \delta \), e.g., \( \epsilon \approx 10^{-2} \) indicates \( \epsilon \) rounded to the second decimal place. Throughout this appendix we use the latter approach given that it is easier to apply in our experiments.

**Definition A.1.** Utility-based model multiplicity \( F_{e,\delta} \) is a collection of classifiers from across a single family of models \( F \) that exhibit a level of predictive performance within a fixed range \( (\epsilon - \delta, \epsilon + \delta] \) for a chosen metric \( m \):

\[
F_{e,\delta} := \{ f \in F : \epsilon - \delta < m(f(X), \tilde{Y}) \leq \epsilon + \delta \},
\]

where \( \tilde{X} \) is a designated validation data set, \( \tilde{Y} \) are its corresponding labels and \( \delta \) is the tolerance of the predictive performance \( \epsilon \).

Table 3 shows the influence of using non-zero tolerance on the quantity of predictive performance bands (measured with accuracy) for the three real-life data sets introduced in Section 4. While this procedure has no effect on Credit Approval and German Credit – given that their performance validation sets have only up to 100 instances – it affects Adult with its 4,885 data points. This observation suggests that for large validation sets we may need to relax the performance bands to get digestible and meaningful

| Data Set          | Test Instances | Performance Bands |
|-------------------|----------------|-------------------|
| Credit Approval   | 69             | 12                |
| German Credit     | 100            | 26                |
| Adult             | 4,885          | 185               |

Table 3: The number of performance bands is likely to decrease when the desired level of predictive performance \( \epsilon \) is relaxed (here rounded to the 3\textsuperscript{rd} and 2\textsuperscript{nd} decimal place).

Figure 11: Ambiguity of the Adult data set increases from 0–25% (Figure 9c) to 0–35% range when the predictive performance bands are rounded to the 3\textsuperscript{rd} decimal place (\( \epsilon \approx 10^{-3} \)).
results, however such an approach makes it easier to find alternative models and challenge individual fairness. In particular, we revisit the Adult data set (Figures 7c and 8c) by rounding the predictive performance to the second decimal place, with the corresponding stability profile shown in Figure 10 and fairness profile displayed in Figure 12. Based on the former we can see an overwhelming number of unique prediction vectors – visible as stacked, 1-wide segments – in each performance band, signifying unfair treatment of many instances; the latter plot paints a similar picture. In general, this procedure is likely to degrade individual fairness as it combines predictions from across “strict” performance bands, creating a more diverse sample that possibly disagrees on a larger subset of data points. This effect can be observed on the ambiguity measurement for Adult, where the metric values are stretched from 0–25% range without rounding – Figure 9c – to 0–35% range when rounded to the third decimal place – Figure 11.

**B PERFORMANCE OF THE FAIR MODEL**

We analyse the predictive utility – measured with accuracy – of individually fair models \( f^\star \) (Definition 2.4) for all performance bands of each fold in the real-life data sets introduced in Section 4. In addition to “strict” model multiplicity \( \mathcal{F}_\varepsilon \) (Definition 2.1), we investigate relaxed performance bands \( \mathcal{F}_\varepsilon \approx \delta \) (see Appendix A, Definition A.1) for Adult, rounding accuracy to the third \( (\varepsilon \approx 10^{-3}) \) and second \( (\varepsilon \approx 10^{-2}) \) decimal point. The experimental results – displayed in Figure 13 – show a nearly universal drop in predictive performance. An interesting exception is the German Credit data set – Figure 13b - for which the proportion of the preferred class (ground truth) is 70% in each fold. This class imbalance improves the accuracy of individually fair models for performance bands (x-axis) below 70%, whereas after this mark the accuracy drops as expected.

**C AMBIGUITY AND DISCREPANCY**

In addition to ambiguity – the proportion of instances treated unfairly in a given performance band – reported earlier in Section 4 (Figures 9 and 11), we look into discrepancy – the proportion of instances for which predictions change between any two models from within a given performance band – which is another multiplicity metric proposed by Marx et al. [13]. Specifically, discrepancy is reported as violin plots, showing the minimum, maximum and overall distribution of the number of individuals treated unfairly when switching between two arbitrary models within a performance band. For the Credit Approval and German Credit data sets we work with “strict” model multiplicity \( \mathcal{F}_\varepsilon \) (Definition 2.1), whereas Adult is investigated with relaxed performance bands \( \mathcal{F}_\varepsilon \approx \delta \) (see Appendix A, Definition A.1) that are rounded to the second \( (\varepsilon \approx 10^{-2}) \) decimal point. The results – shown in Figure 14 – are augmented by bar plots illustrating the number of workflow runs (on a logarithmic scale) for all performance bands across the data folds; they help to assess the reliability of ambiguity and discrepancy scores. For all the data sets, ambiguity raises quite sharply as we move away from an optimal model, reaching between 40 and 90% at its peak, which brings to attention a surprisingly high degree of individual unfairness for these classifiers. Discrepancy, on the other hand, shows that while in many cases it is possible to switch between two (performance-wise) equivalent models without influencing individual predictions, it is more likely to treat differently a non-negligible proportion of instances.
Figure 13: Predictive performance – accuracy – of fair models $f^*$ (y-axis) compared with individual models $f \in \mathcal{F}_\epsilon$ (x-axis), as well as $f \in \mathcal{F}_{\epsilon,\delta}$ in Panels (d) and (e) for the Adult data set. The diagonal line represents unchanged performance, with points above indicating an improvement and points below a decrease in the accuracy of fair models. The horizontal lines indicate the performance achieved by always predicting the preferred class for each fold separately; most of these overlap in Panels (a) and (b), and they are not visible (24%) in Panels (c), (d) and (e).
Figure 14: Ambiguity (top), discrepancy (middle) and count of predictive workflow runs per performance band (bottom, logarithmic scale) for each data set. Credit Approval – Panel (a) – and German Credit – Panel (b) – extend Figures 9a and 9b respectively. Adult – Panel (c) – is plotted with the performance bands relaxed by rounding them to the 2nd decimal place, i.e., $\epsilon \approx 10^{-2}$, for comprehensibility; additionally, the discrepancy for each performance band is based on up to 500 (randomly selected) workflow runs given the overwhelming number of their pairs for the full collection. The $\times$ symbol in the discrepancy plots indicates a single run of a predictive workflow, and the $-$ marker is a flat violin plot, i.e., multiple workflows with identical predictions.