I Prefer Not To Say: Protecting User Consent in Models with Optional Personal Data

Tobias Leemann[1,2], **Martin Pawelczyk**[3], **Christian Thomas Eberle**[1], **Gjergji Kasneci**[2]

[1]University of Tübingen, Tübingen, Germany
[2]Technical University of Munich, Munich, Germany
[3]Harvard University, Cambridge, MA, USA
tobias.leemann@uni-tuebingen.de, martin.pawelczyk.1@gmail.com, ct.eberle@protonmail.ch, gjergji.kasneci@tum.de

Abstract

We examine machine learning models in a setup where individuals have the choice to share optional personal information with a decision-making system, as seen in modern insurance pricing models. Some users consent to their data being used whereas others object and keep their data undisclosed. In this work, we show that the decision not to share data can be considered as information in itself that should be protected to respect users' privacy. This observation raises the overlooked problem of how to ensure that users who protect their personal data do not suffer any disadvantages as a result. To address this problem, we formalize protection requirements for models which only use the information for which active user consent was obtained. This excludes implicit information contained in the decision to share data or not. We offer the first solution to this problem by proposing the notion of Protected User Consent (PUC), which we prove to be loss-optimal under our protection requirement. We observe that privacy and performance are not fundamentally at odds with each other and that it is possible for a decision maker to benefit from additional data while respecting users' consent. To learn PUC-compliant models, we devise a model-agnostic data augmentation strategy with finite sample convergence guarantees. Finally, we analyze the implications of PUC on challenging real datasets, tasks, and models.

1 Introduction

While the day-to-day impact of automated data processing is steadily growing, modern regulations such as the European Union's General Data Protection Regulation (GDPR) (GDPR 2016) or the California Consumer Privacy Act (CCPA) (OAG 2021) strive to give individuals more control over their personal data. In light of these regulations, we consider machine-learned classifiers in which individuals have the freedom to decide themselves on which data they would like to provide to an automated decision system.

Such systems are increasingly being deployed (Henning 2022): As a running example, we consider a realistic use-case of health insurance pricing: Suppose in an automated pricing model all potential customers are asked to fill out an application form where they enter certain *base features*, for instance information such as their state of residence and age.

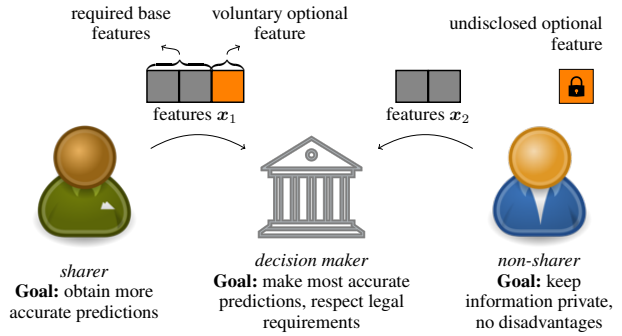


Figure 1: Overview of the relevant stakeholders. We consider a case where users can voluntarily provide information on optional features or choose to leave them undisclosed. The goals of sharers, non-sharers, and the decision maker have to be reconciled.

To improve the pricing model, the insurance offers an additional service, a "companion fitness app", through which additional health data about the customer's physical condition are collected. The customers decide whether to use the app or not; alternatively, customers can sign up for a policy without consenting to use the app. The health data that customers share may however influence the premium of the insurance policy they receive. We refer to data that provide additional, non-mandatory information beyond the base features as *optional features*. With fitness trackers and smartwatches rapidly gaining popularity (Reeder and David 2016; Zimmer et al. 2020; Statista 2023), such systems are already being deployed in practice, e.g., by major health insurance firms in Australia (Henning 2022).

The outlined scenario is challenging as there are three groups of stakeholders whose interests need to be reconciled: (1) The group of non-sharing individuals who do not want to provide additional information, for instance due to privacy concerns. We refer to them as *non-sharers*. For this group, the decision maker does not want to or cannot force them to provide the additional information for legal reasons. Consequently, the non-sharers do not want the additional information to be considered in the decision making process; in return, they are willing to sacrifice some accuracy, but they do not want to face other systematic disadvantages. (2)

On the other hand, individuals who voluntarily share data (*sharers*) explicitly want the additional information to be considered and want to obtain more accurate predictions. (3) Finally, the decision makers themselves desire the most accurate predictions with the lowest overall costs while respecting the users' privacy and legal requirements.

Among these requirements, it is crucial to the non-sharers to explicitly exclude the information contained in the decision to share or not to share. To see this, we note that smartwatch users are more likely to exercise in general than non-wearers (DeMarco 2023) which usually create lower costs for the insurance company as fitter customers take less sick days on average. Thus, only through observing the decision to share data, the insurance firm could make inferences about a person's fitness. This is problematic for two reasons: First, the company would unethically infer private data, that the non-sharers explicitly did not give consent to. Prior work (Wachter and Mittelstadt 2019) has argued for a "right to reasonable inferences". This rules out inferences from unrelated factors that are purely predictive and may infringe privacy, as they open the door for discriminatory and invasive decision-making (Mittelstadt et al. 2016). Second, this would lead to non-sharers being assigned a higher insurance premium than the estimate of the legacy model which only considered their base features. Many countries have laws that prohibit insurers from raising the base premium for users who do not share their data, as this is seen as a coercive and unfair practice. For example, the US only permits five factors to affect the premium, which are location, age, tobacco use, plan category, and dependent coverage (US Government. U.S. Centers for Medicare & Medicaid Services. 2023). It is however possible for insurers – and desired by many users – to award bonuses which reduce the premium based on participation in optional reward and incentive programs (Madison, Schmidt, and Volpp 2013; Henning 2022).

To summarize, we study machine learning models that can handle optional features and meet legal requirements and desiderata of three groups of stakeholders: the sharers, the non-sharers, and the decision makers. We consider it essential for these models to not make inferences based on the unavailability of a feature value for the non-sharers, a constraint that we term *Availability Inference Restriction (AIR)*. Finally, we are interested in obtaining models with optimal performance under this requirement.

Contribution. We address the problem of how to fairly and privately predict outcomes for users who share optional data and those who do not. We tackle this overlooked issue by making the following contributions:

- **Definition.** We introduce models with Protected User Consent (PUC), which are optimal under our protection requirement AIR. We derive performance guarantees, which formally show that it is possible to reconcile the decision maker's interest in improved predictions and the non-sharer's privacy preferences.
- **Algorithm.** We propose a PUC-inducing data augmentation (`PUCIDA`) technique that can be applied to any type of predictive architecture (e.g., tree or neural network) and any convex loss function (e.g., mean squared error or cross-entropy loss) to obtain such models
- **Analysis.** We prove that predictive models trained with `PUCIDA` satisfy PUC asymptotically, and provide finite sample convergence results that demonstrate that `PUCIDA` produces PUC-compliant models in practice.
- **Empirical evaluation.** We empirically show that without enforcing PUC, the average absolute prediction outcome (e.g., insurance quote) of users who do not share data can be almost 20 % worse than justified by their base data. We then evaluate our data augmentation technique on various ML models and show that PUC is achieved regardless of the model.

2 Related Work

In this Section, we review the most relevant streams of related work (see Appendix A.1 for additional references).

Classification with Missing Values. Classification models that can handle missing data have been studied previously with the goal of minimizing costs or increasing performance (Zhang et al. 2005; Aleryani, Wang, and De La Iglesia 2020), obtaining uncertainty estimates (Kachuee et al. 2020), or fulfilling classical fairness notions (Zhang and Long 2021; Jeong, Wang, and Calmon 2022; Wang and Singh 2021; Fernando et al. 2021). However, the mechanisms underlying missingness is different in this work, as missing values indicate explicit non-consent by the user, leading to different implications. In a related line of work, classification with noisy (Fogliato, Chouldechova, and G'Sell 2020) or missing labels (Kilbertus et al. 2020; Rateike et al. 2022) has been investigated, where the missingness is often a result of *selection bias*. The setting considered in this work is different in the sense that we are not concerned with fulfilling a fairness notion with respect to a sensitive attribute, but consider the interests of subjects that have and have not provided optional information.

Data Minimization. The principle of Data Minimization is anchored in the GDPR (GDPR 2016). Data Minimization demands minimal data collection. Several works are concerned with implementing (Goldsteen et al. 2021) or auditing compliance with this principle (Rastegarpanah, Gummadi, and Crovella 2021). Rastegarpanah et al. (Rastegarpanah, Crovella, and Gummadi 2020) consider decision systems that can handle optional features from a data minimization perspective where the decision maker decides which features are collected for each individual. This principle is distinct from the "right to be forgotten" (Biega et al. 2020), which enables individuals to submit requests to have their data deleted. In response to these regulations, several works consider the problem of updating an ML model without the need of retraining the entire model (Wu, Dobriban, and Davidson 2020; Ginart et al. 2019; Izzo et al. 2021; Golatkar, Achille, and Soatto 2020) or the effect of removals on model explanations (Rong et al. 2022; Pawelczyk et al. 2023). Our work differs from these works as our goal is to train a model where users decide themselves which data they deem relevant through sharing one or many optional features.

Algorithmic Fairness. A multitude of formal fairness definitions have been put forward in the literature (Verma and Rubin 2018). Examples include statistical parity (Dwork et al. 2012), predictive parity (Chouldechova 2017), equalized odds, equality of opportunity (Hardt, Price, and Srebro 2016), and individual fairness (Dwork et al. 2012). However, they are still a topic of discussion, for instance, because these definitions are known to be incompatible (Kleinberg, Mullainathan, and Raghavan 2016; Lipton, McAuley, and Chouldechova 2018). Additionally, there are a several definitions that rely on causal mechanisms to assess fairness, e.g., counterfactual fairness (Kusner et al. 2017), and the notion of unresolved discrimination (Kilbertus et al. 2017). While causal approaches to fairness might be preferable, they require information about the causal structure of the data generating process. Moreover, it has recently been shown that causal definitions may lead to adverse consequences, such as lower diversity (Nilforoshan et al. 2022). We discuss how existing fairness definitions could possibly be applied to the setting with optional features, but we find that none of the fairness definitions aligns with our desiderata theoretically and experimentally (see Appendix A.2).

Strategic Classification. In an even broader context, this work also relates to the field of strategic classification (Hardt et al. 2016). However, it is worth noting that in strategic classification research, the focus primarily revolves around users strategically manipulating their features for optimal outcomes, which may also involve information withholding (Krishnaswamy et al. 2021). In contrast to our work, privacy concerns are neglected in this research stream. As far as we are aware, there are no prior works on the specific problem of balancing the interests of *all three* groups of stakeholders (the non-sharers, sharers, and the decision makers).

3 Problem Formulation

3.1 Formalization and Notation

In this work, each data instance contains a realization of a number of base features $\mathbf{b} \in \mathcal{X}^b$, where $\mathcal{X}^b \subseteq \mathbb{R}^n$ is the space of the base features. Furthermore, let there be some optional information $z \in \mathcal{X}^z$, where $\mathcal{X}^z \subseteq \mathbb{R}$ is the value space of the optional feature.[1] It is the users' choice to decide if they want to disclose z to the system, which results in an availability variable $a \in \{0, 1\}$. Accordingly, only imputed samples $z^* = \{z \text{ if } a{=}1, \text{ else } \texttt{N/A}\}$ are observed, where a value of `N/A` indicates that a user did not reveal the optional information, e.g., did not use the companion app. In summary, the data observations are tuples $\mathbf{x} = (\mathbf{b}, a, z^*)$ that reside in $\mathcal{X} = \mathcal{X}^b \times \{0, 1\} \times (\mathcal{X}^z \cup \{\texttt{N/A}\})$. Each training sample comes with a label $y \in \mathcal{Y}$. Further, there is a data generating distribution $\mathbf{p}$ with support $\mathcal{X} \times \mathcal{Y}$ and we have access to an i.i.d. training sample $(\mathbf{x}, y) \sim \mathbf{p}$. Figure 2 shows such a data sample. We denote the random variables for the respective quantities by $\boldsymbol{B}, A, Z, Z^*, Y$. The label is probabilistically determined through the base features $\mathbf{B}$ and the hidden feature Z but the sharing decision does not influence the true label for a given $\mathbf{B}, Z$, such that $Y \perp\!\!\!\perp A | \mathbf{B}, Z$.

[1] We extend our definitions to integrate multiple optional features a later section.

base features $\mathbf{b}$		opt. feat. z^*	a	label y
state	plan	fitness score	avail.	treatment costs
New South Wales	basic	87 %	1	3k$
Queensland	gold	`N/A`	0	17k$
New South Wales	basic	92 %	1	5k$
New South Wales	basic	`N/A`	0	64k$
Victoria	premium	56 %	1	22k$

Figure 2: Samples for the insurance use-case. We have two base features $\mathbf{b}$ and one optional feature z^*, which either takes an observed value z, or it takes a value of `N/A` if unobserved. The variable $a \in \{0, 1\}$ indicates the availability of the feature. The goal is to predict the label y.

In many applications, the goal is to find a function $f : \mathcal{X} \to \mathcal{Y}$ that models the observed data. In particular, $f : \mathcal{X} \to [0, 1]$ may predict a probability of a positive outcome or $f : \mathcal{X} \to \mathbb{R}$ may return a numerical score. The test data for which the model will be used come from the same distribution $\mathbf{p}$, though with the label y unobserved, and we suppose that the information provided is always correct. We consider a convex loss function $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, e.g., mean-squared-error (MSE) or binary cross entropy (BCE), for which we minimize the expected loss for a sample from the data distribution. For instance, using the common MSE loss $\mathcal{L}(f(\mathbf{x}), y) = (f(\mathbf{x}) - y)^2$, an optimal predictor is given by $f^*_{\mathcal{L}}(\mathbf{x}) = \arg\min_{f(\mathbf{x})} \mathbb{E}_{\mathbf{p}(Y|\mathbf{x})} \left[(f(\mathbf{x}) - Y)^2\right] = \mathbb{E}[Y|\mathbf{x}]$, the conditional expectation. However, this notion can be generalized to other loss functions: An optimal predictor $f^*_{\mathcal{L}}(\mathbf{x})$ for the loss function $\mathcal{L}$ fulfills $\forall \mathbf{x}$:

$$f^*_{\mathcal{L}}(\mathbf{x}) = \mathbb{F}^{\mathcal{L}}_{\mathbf{p}}[Y|\mathbf{x}] \coloneqq \underset{f(\mathbf{x})}{\arg\min}\, \mathbb{E}_{\mathbf{p}(Y|\mathbf{x})} \left[\mathcal{L}(f(\mathbf{x}), Y)\right]. \quad (1)$$

We use $\mathbb{F}^{\mathcal{L}}[Y|\mathbf{x}]$ to denote a generalized expected value that minimizes the expected loss conditioned on $\mathbf{x}$. To ease our derivations, we suppose this minimum to be unique and finite. Intuitively, it represents the best guess of Y given $\mathbf{x}$. For the MSE-Loss, $\mathbb{F}^{\mathcal{L}}$ is equivalent to the expectation operator $\mathbb{E}$. In the following statements, the reader may thus mentally replace $\mathbb{F}^{\mathcal{L}}$ with an expectation $\mathbb{E}$ without further ramifications in order to get the high level intuition. Finally, we introduce two key terms, namely, *base feature model* and *full feature model*. The former refers to a model trained on the base features only, while the latter refers to a model trained on all features where some strategy is used to replace unavailable feature values. Typically these strategies are called *imputation* and replace unavailable values by zeros, a feature's mean or median (Emmanuel et al. 2021).

3.2 Desiderata

Our goal is to learn models $f : \mathcal{X} \to \mathcal{Y}$ that comply with the desideratum of *Availability Inference Restriction*, which we briefly introduced in Section 1, to protect the interests of the non-sharers. Under this constraint, the model should provide the best predictive performance to reflect the need

of the sharers and the decision maker for most accurate predictions.

Desideratum 1: Availability Inference Restriction. We start by considering the intricate case of individuals who *do not want to share optional information*. In this case, the model should compute the prediction based on the information the user gave their consent to. In particular, (a) the model should only use the base features *and* (b) should not use information that could be derived from the unavailability of the optional features to compute the prediction to avoid violating the user's consent.

For (a), this requires that the predictor does not use the information as an explicit input, i.e., the predictor should behave as if it only used base features $\mathbf{b}$ via some function $g : \mathcal{X}^b \to \mathcal{Y} : f_{|_{a=0}}(\mathbf{b}, a, z^*) = g(\mathbf{b})$. For (b), although a=0 is not an explicit input to g, a sufficiently complex function may still be implicitly adapting to the group $a = 0$ and thus incorporate information that the user did not give their consent to. We would like to make sure that the predictions of g cannot use more information than contained in the overall conditional distribution, given the base features $\mathbf{b}$. This overadaption can be prevented by constraining the model's loss on the population of non-sharers to match the loss of the optimal base model $f^*_{\mathcal{L}}$ on this population. The reasoning behind this rationale is that all models that would beat the performance of this model must implicitly use some additional side knowledge about this group that was not provided by the users.

Definition 1 (Availability Inference Restriction)**.** *For individuals that choose not to provide the optional feature (a=0), only the provided data $\mathbf{b}$ is used to compute the outcome in the decision process, i.e., $f_{|_{a=0}}(\mathbf{b}, a, z^*)$=$g(\mathbf{b})$, where $g : \mathcal{X}^b \to \mathcal{Y}$ is a base feature model. Further, we require*

$$\mathbb{E}\left[\mathcal{L}\left(g(\mathbf{B}), Y\right)|A = 0\right] \geq \mathbb{E}\left[\mathcal{L}\left(f^*_{\mathcal{L}}(\mathbf{B}), Y\right)|A = 0\right]. \quad (2)$$

This definition summarizes our intuition that the information encoded through the unavailability of feature information should neither be used explicitly (a) nor implicitly (b). We show how this constraint can analogously be derived from information-theoretic considerations in Appendix B.3.

Desideratum 2: Optimality. Our Definition 1 restricts the information that the predictor can use when the optional information is unavailable. To meet the interests of the decision maker and the sharers, we also want to find models with optimal performance, i.e., lowest loss, under this constraint.

4 Protecting User Consent

We are therefore looking for an *optimal* model within the class of predictors that comply with Availability Inference Restriction. In this Section, we derive a novel notion called Protected User Consent (PUC) that fulfills this purpose.

4.1 One-Dimensional PUC

The next result encodes an intuitive notion of protection for the users that do not want to share data on the optional features (a=0): Their prediction under f is then constrained to the best estimate for a user with the same base characteristics, no matter if additional data was provided. Contrarily, when additional information through the optional feature is provided, the predictor returns the best estimate using the available optional information:

Theorem 1 (1D-PUC)**.** *Let $f : \mathcal{X} \to \mathcal{Y} \subseteq \mathbb{R}$ be a full feature model (i.e., including optional features). Among all predictors compatible with the Availability Inference Restriction, a model f with minimal loss is given by:*

$$f^*_{PUC}(\mathbf{b}, a, z^*) = \begin{cases} \mathbb{F}^{\mathcal{L}}[Y|\mathbf{b}], & \text{if } a = 0 \\ \mathbb{F}^{\mathcal{L}}[Y|\mathbf{b}, A = 1, Z^* = z^*] & \text{if } a = 1. \end{cases}$$

We defer all proofs in this work to Appendix D. PUC is different from existing notions of group fairness, that do not fulfill the two desiderata in general (see Appendix A.2 for a discussion). Under the mentioned requirements, there is no model that can outperform f^*_{PUC}. We stress that 1D-PUC-compliant models have performance guarantees. These models match or improve upon an optimal base feature model $f^*_{\mathcal{L}}(\mathbf{B}) = \mathbb{F}_{\mathcal{L}}[Y|\mathbf{b}]$. This model can be seen as an upper bound for practical models obtained after model selection. Therefore, models that can beat its performance may offer improvements even after extensive hyper-parameter tuning and model selection, a property which we refer to as Predictive Non-Degradation (PND): a model f fulfills PND if its loss is smaller than that of the base feature model:

$$\mathbb{E}[\mathcal{L}\left(Y, f^*_{\mathcal{L}}(\mathbf{B})\right)] \geq \mathbb{E}[\mathcal{L}\left(Y, f(\mathbf{B}, A, Z^*)\right)]. \quad (3)$$

We prove the following result:

Corollary 1 (Predictive Non-Degradation of f^*_{PUC})**.** *For any density $\mathbf{p}$, a PUC-compliant model f^*_{PUC} fulfills Predictive Non-Degradation, i.e., it has a loss upper-bounded by the optimal base feature model $f^*_{\mathcal{L}}$.*

This is a remarkable result as it testifies that the decision maker can benefit from additional information in terms of loss, while protecting the privacy of users. This highlights that the interests of the different stakeholders are not contradictory and models that benefit all stakeholders do exist.

4.2 PUC under Strategic Considerations and Monotonicity Constraints

We have initially considered the case where the users desire the highest possible accuracy under data usage restrictions. However, in some cases such as our initial insurance example, the motivation to receive a lower premium might be a more important concern to some users than receiving an accurate prediction or their privacy concerns. If all users have full information (i.e., they see premiums with and without their optional data) and act strategically by sharing the value of z only if it would decrease their premiums, we obtain the following result.

Theorem 2 (Optimality of f^*_{PUC} under strategic actions)**.** *Let $\mathbf{p}'(\mathbf{B}, Z, Y)$ be any prior density on base features, true optional features and labels and let $f(\mathbf{b}, a = 0, z) = \mathbb{F}^{\mathcal{L}}[Y|\mathbf{b}]$, i.e., the decision maker uses the base feature model when no optional data is available. Further suppose that users*

strategically choose to share the optional feature z only if $f(\mathbf{b}, a = 1, z) \leq f(\mathbf{b}, a = 0, \texttt{N/A})$. Under these conditions, the model f^*_{PUC} (Theorem 1) has minimal loss among all predictors.

This result underlines that PUC models remain optimal if the decision maker cannot increase the premiums beyond the predictions of the current base model for the non-sharers. This is reasonable in many cases, where legal constraints mandate that the decision maker cannot implicitly force users to share data by inflating the base premium, as outlined in the introduction. The sharing decision can also be automated for the users by simply dropping the optional feature if it does not lead to a decrease in premiums. This would result in the aforementioned bonus systems, where sharing more data cannot increase the premium. We show that among the class of models with such a monotonicity constraint, the outlined PUC-model with automatic sharing decisions is still optimal under the same conditions as in Theorem 2 in Appendix D.5.

4.3 r-dimensional PUC

Next, we generalize our notion such that r features can be provided optionally. For example, the insurance firm might also accept voluntary results from prior medical examinations or diagnostic tests. Therefore, let there now be r optional features such that $\boldsymbol{z} \in \mathcal{X}_1^z \times \cdots \times \mathcal{X}_r^z$ and $\boldsymbol{a} \in \{0,1\}^r$, where $\mathcal{X}_i^z$ are the respective supports of each optional feature. By $\mathcal{I} \subseteq [r] = \{1, \ldots, r\}$, we denote an index set that contains all feature indices present, i.e., $\mathcal{I}(\boldsymbol{a}) = \{i \mid \boldsymbol{a}_i = 1, i = 1, \ldots, r\}$. When we index vectors with this set, e.g., $\boldsymbol{Z}_{\mathcal{I}}$, we refer to the subvector that only contains the indices in $\mathcal{I}$.

Definition 2 (Protected User Consent, PUC)**.** *Let $f : \mathcal{X} \to \mathcal{Y} \subseteq \mathbb{R}$ be a full feature model. The model f^*_{PUC} that fulfills Protected User Consent is given by*

$$f^*_{PUC}(\mathbf{b}, \boldsymbol{a}, \boldsymbol{z}^*) = \mathbb{F}^{\mathcal{L}}_{(\mathbf{B}, \boldsymbol{A}, \boldsymbol{Z}^*) \sim \mathbf{p}} \left[Y \middle| \mathbf{B} = \mathbf{b}, \boldsymbol{A}_{\mathcal{I}(\boldsymbol{a})} = \mathbf{1}, \boldsymbol{Z}_{\mathcal{I}(\boldsymbol{a})} = \boldsymbol{z}^*_{\mathcal{I}(\boldsymbol{a})} \right],$$

where $\boldsymbol{A}_{\mathcal{I}(\boldsymbol{a})} = \mathbf{1}$ means that each element that is set to 1 in $\boldsymbol{a}$ needs to be one in $\boldsymbol{A}$ as well.

For a single feature (r=1), the index set can either be $\mathcal{I} = \emptyset$ or $\mathcal{I} = \{1\}$ and the definition corresponds to 1D-PUC. The conditional expectation with $\boldsymbol{A}_{\mathcal{I}(\boldsymbol{a})} = \mathbf{1}$ effectively constrains the features in $\mathcal{I}$ to be available, but marginalizes over samples with or without further information.

5 Implementing Protected User Consent

In this section, we derive a model-agnostic approach called *PUC-inducing data augmentation* (PUCIDA) to achieve protected user consent. By using theoretical analysis, we establish that PUCIDA will result in exact protected user consent. Furthermore, we establish performance guarantees that provide an upper bound on the deviation between practical, finite sample-based PUC-compliant models and their theoretical infinite sample limits.

	state	plan	score	costs
	NSW	basic	87 %	3k$
⊕	**NSW**	**basic**	**N/A**	**3k$**
	NSW	basic	92 %	5k$
⊕	**NSW**	**basic**	**N/A**	**5k$**
	NSW	basic	N/A	64k$

Figure 3: Explaining PUCIDA. Our data augmentation procedure expands each instance with optional information into two samples: The original instance and a synthetic sample (⊕). The synthetic samples retain the base features and the labels, but the information on the optional features is dropped (fitness score → N/A). The model sees samples with the same base features with a missing value and will thus base its decision only on the base features. In this example, given the base features ("NSW", basic) and no optional statements, the model would estimate the costs to be 24k$, which is the dataset average conditioned on these values.

5.1 PUCIDA: PUC-inducing Data Augmentation

Intuitively, we want to prevent the model from making inference from a feature's missingness patterns. The core insight is to leverage synthetic samples that make the *distribution of the labels given missingness equal to the overall label distribution*. Thereby, we prevent the derivation of predictive information from the missingness itself (see Table 3).

For a single optional feature, extensively enumerating all samples as in the table is possible while for multiple features this may be intractable. Therefore, we do not list all samples but propose a stochastic, multifeature variant of the algorithm: **(1)** Instead of drawing samples with uniform probability from the distribution $\mathbf{p}$, we use non-normalized weights w:

$$w(\mathbf{x}) = w(\mathbf{b}, \boldsymbol{a}, \boldsymbol{z}^*) = 2^{|\mathcal{I}(\boldsymbol{a})|}. \tag{4}$$

This step corresponds to the expansion of an instance into $2^{|\mathcal{I}(\boldsymbol{a})|}$ synthetic ones; e.g., a sample with a single optional feature is assigned a weight of two (cf. Figure 3). Training instances are drawn with a probability proportional to these weights. This results in data instances with optional information being more frequently sampled. **(2)** We require a sample modification where optional features are randomly dropped from the samples. For each sampled item, we drop each available optional feature with probability p=0.5:

$$\mathbf{q}_i \sim \text{Bern}(0.5), i = 1, \ldots, r; \quad \overline{\boldsymbol{a}} = \mathbf{q} \odot \boldsymbol{a}; \tag{5}$$

$$\overline{\boldsymbol{z}}^*_i = \{\boldsymbol{z}^*_i \text{ if } \overline{\boldsymbol{a}}_i = 1, \text{ else N/A}\}, i = 1, \ldots, r. \tag{6}$$

(3) We train the predictive model on the modified samples $(\overline{\mathbf{x}}, y) = ((\mathbf{b}, \overline{\boldsymbol{a}}, \overline{\boldsymbol{z}}^*), y) \sim \overline{\mathbf{p}}$ derived through this procedure.

5.2 Theoretical Analysis

We summarize PUCIDA in pseudo-code in Appendix D.8 and provide the following theorem to demonstrate that PUCIDA leads to PUC-compliant models.

Theorem 3. *The loss-minimal model $f(\mathbf{b}, \boldsymbol{a}, \boldsymbol{z}^*) = \mathbb{F}^{\mathcal{L}}_{\overline{\mathbf{p}}}[Y | \mathbf{b}, \boldsymbol{A} = \boldsymbol{a}, \boldsymbol{Z}^* = \boldsymbol{z}^*]$ on the modified distribution $\overline{\mathbf{p}}$*

fulfills Protected User Consent with respect to $\mathbf{p}$*, i.e.,*

$$\mathbb{F}^{\mathcal{L}}_{\mathbf{p}}[Y|\mathbf{B}=\mathbf{b}, \boldsymbol{A}=\boldsymbol{a}, \boldsymbol{Z}^*=\boldsymbol{z}^*] =$$
$$\mathbb{F}^{\mathcal{L}}_{\mathbf{p}}\left[Y\middle|\mathbf{B}=\mathbf{b}, \boldsymbol{A}_{\mathcal{I}(\boldsymbol{a})}=\mathbf{1}, \boldsymbol{Z}_{\mathcal{I}(\boldsymbol{a})}=\boldsymbol{z}^*_{\mathcal{I}(\boldsymbol{a})}\right] = f^*_{PUC}(\mathbf{b}, \boldsymbol{a}, \boldsymbol{z}^*).$$

This result is remarkable in its generality as it enables PUC-compliant models using standard optimization procedures by modifying the distribution of the data; i.e., *PUCIDA can be combined with any existing model and training pipeline*. Next, 'we' study the theoretical convergence behavior for PUCIDA on finite samples. To this end, we define the PUC-Gap as the expected squared deviation from PUC:

$$\text{PUC-Gap}^2(f, \mathbf{p}) = \tag{7}$$
$$\mathbb{E}_{(\mathbf{B},\boldsymbol{A},\boldsymbol{Z}^*)\sim\mathbf{p}}\left[\left(f(\mathbf{B}, \boldsymbol{A}, \boldsymbol{Z}^*) - f^*_{\text{PUC}}(\mathbf{B}, \boldsymbol{A}, \boldsymbol{Z}^*)\right)^2\right].$$

We will restrict ourselves to $\mathcal{L} \equiv$ MSE and thus $\mathbb{F}^{\mathcal{L}} \equiv \mathbb{E}$, and study a *baseline conditional expectation estimator* $\hat{\mu}$ which averages the labels conditional on all observations with the same features $\mathbf{x}$. For brevity, we refer to Appendix D.7 (Eqn. 51) for a formal definition of this estimator. Since we usually cannot compute the exact expectation from Theorem 3, we are interested in the number of samples required from $\overline{\mathbf{p}}$ to obtain a fixed average estimation error for which we establish the following result.

Theorem 4 (Finite Sample Convergence)**.** *Let* $\mathcal{X} = \mathcal{X}^b \times (\mathcal{X}^z \cup \{N/A\})$ *be finite feature space and let* $\mathcal{Y} \subseteq \mathbb{R}$ *be the label space. All conditional expectations* $\mu(\mathbf{x}) := \mathbb{E}_{\overline{\mathbf{p}}}[y|\mathbf{x}]$ *and the conditional variances* $\sigma^2(\mathbf{x}) := \text{Var}_{\overline{\mathbf{p}}}[y|\mathbf{x}]$ *exist and are finite. Then there exists a baseline non-parametric regressor* $\hat{\mu} : \mathcal{X} \mapsto \mathbb{R}$ *from a finite number of* N *independent, identically distributed observations* $(\overline{\mathbf{x}}_i, y_i)_{i=1\dots N}$ *from* $\overline{\mathbf{p}}$ *with a convergence rate of* $\mathcal{O}(N^{-1})$*; more specifically*

$$\text{PUC-Gap}^2(\hat{\mu}, \mathbf{p}) = \mathbb{E}_{\mathbf{X}\sim\mathbf{p}}\left[\left(\hat{\mu}(\mathbf{X}) - \mu(\mathbf{X})\right)^2\right]$$
$$\leq \frac{2^r|\mathcal{X}|^2(\sigma^2_{\max} + \mu^2_{\max})}{N} + \mathcal{O}\left(\frac{1}{N^2}\right),$$

with $\sigma^2_{\max} := \max_{\mathbf{x}\in\mathcal{X}} \sigma^2(\mathbf{x})$ *and* $\mu^2_{\max} := \max_{\mathbf{x}\in\mathcal{X}} \mu^2(\mathbf{x})$.

In conjunction with Theorem 3, this result provides a bound on the expected gap to perfect protected user consent that is dependent of the sample size, which decreases with a rate of $\mathcal{O}\left(N^{-1}\right)$. Several remarks are in place: We obtain a multiplicative constant which depends on the number of optional features r and the size of the feature space $|\mathcal{X}|$. The square of this quantity enters the result because the number of samples available to estimate each conditional mean is not independent, as they need to sum up to N. For large feature spaces, however, they are almost independent and we expect the constant to scale almost linearly in $|\mathcal{X}|$. The growth of 2^r is attributed to the re-sampling strategy which might assign a very low probability to certain inputs, which may only be well approximated with a high number of samples. As the number of optional features is typically limited in realistic use-cases it will be well outgrown by N. Note that more powerful model (e.g., Tree based model + PUCIDA) usually outperform this baseline.

data	base model	Full feature model	PUCIDA
diab.(C)	33.84% ±2.47	31.44% ±2.19	34.01% ±1.71
compas (C)	44.47% ±0.37	41.47% ±1.09	44.54% ±0.54
adult (C)	13.37% ±0.07	12.84% ±0.28	13.41% ±0.12
water (C*)	10.65% ±1.64	10.00% ±1.58	10.97% ±1.21
colic (C*)	13.81% ±0.82	11.34% ±0.46	15.05% ±0.68
income (R)	109.56 ±1.00	109.11 ±1.29	110.73 ±1.29
calif. (R)	15.79 ±0.10	15.16 ±0.28	16.18 ±0.06
insurance (R)	283.47 ±0.53	279.78 ±0.42	285.31 ±0.39

Table 1: Availability Inference Restriction is violated by full feature models (Random Forests). As expected, the full feature models always have lower losses than the base-models, indicating that Availability Inference Restriction is violated while PUCIDA fullfils Availability Inference Restriction. We report misclassification error rates for classification models and MSE loss ($\times$ 100) for regression models.

Practical considerations. For smaller datasets, an alternative approach to random sampling is to use all possible samples to approximate the distribution $\overline{\mathbf{p}}$ by a method we call "exhaustive augmentation". This involves enumerating all possible variations of the original samples, including any optional features, to form a larger dataset $\mathcal{D}'$. The model is then trained on this expanded dataset.

6 Experimental Evaluation

Here, we empirically validate the effectiveness of our methods using eight real-world datasets and one synthetic dataset. In particular, we highlight that (a) full feature models violate the Availability Inference Restriction and make it harder for non-sharers to obtain the positive outcome, (b) PUCIDA results in PUC-compliant models as suggested by our theory, and that (c) the reduction in terms of model performance due to using PUC are moderate relative to deploying a full feature model.

Common datasets. We use eight real-world datasets commonly found in the related literature. For classification (C), the Diabetes (diab) and the horse colic dataset (colic) study the prediction of diseases, the COMPAS dataset is concerned with estimating likelihood of recidivism and UCI Adult income dataset requires to predict whether individuals have an income of over 50k$. The water treatment dataset (water) predicts the operational state of a facility. We also study the regression tasks (R) of house price estimation in California (calif), income prediction (income), and inferring information from insurance claims (insurance) to link to our initial example. Details about preprocessing, dataset sources and model hyperparameters are provided in Appendix F.2.

Availability. The colic and the water dataset come with inherent missing values that we use (indicated through $*$). For six more datasets we introduce availability dependent on a feature's value. We compute the probability of feature unavailability $\mathbf{p}(A_i = 0|z_i)$ by applying a sigmoid function centered at the feature mean and sample the availability a from the respective conditional distribution. We additionally

				Full feature model		PUCIDA	
task	data	optional	Base feature model	pred.	change	pred.	change
C	diab.	Glucose	60.27%	45.19%	-15.08% ±2.01	61.20%	0.93% ±0.93
C	compas	#priors	51.19%	32.86%	-18.33% ±0.89	51.34%	0.15% ±0.59
C	adult	edu-num	13.86%	11.44%	-2.42% ±0.07	13.92%	0.06% ±0.05
C*	water	oxygen. dem.	87.10%	84.52%	-2.58% ±2.81	87.42%	0.32% ±1.58
C*	colic	abdom. app.	6.39%	1.24%	-5.15% ±0.92	7.01%	0.62% ±1.64
R	income	WKHP	100.0%	81.2%	-18.8% ±0.61	101.2%	1.2% ±0.19
R	calif.	m_income	100.0%	94.4%	-5.6% ±0.67	103.8%	3.8% ±0.42
R	insurance	experience	100.0%	94.8%	-5.2% ±0.09	100.1%	0.1% ±0.05

Table 2: Measuring the average predictions for non-sharers. For classification tasks we report the positive outcomes (in %), and for regression tasks, we report relative predictions to the base feature model (set to 100 %). The non-sharers face disadvantages for not providing the voluntary information and are assigned less favorable prediction outcomes by the full feature models. This discrepancy vanishes when PUCIDA is applied.

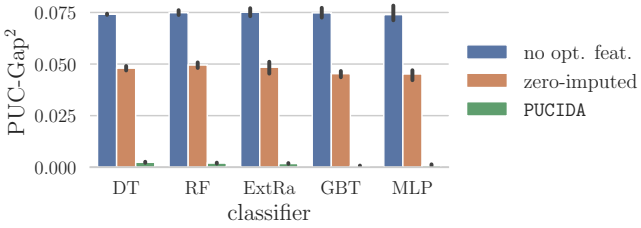


Figure 4: PUCIDA is model-agnostic. The PUC-gaps are close to zero when applying our technique across a variety of common models on the simulated dataset.

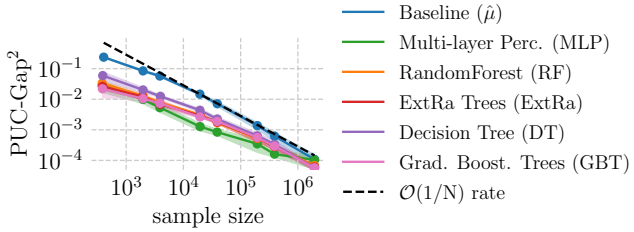


Figure 5: Convergence rate of models under PUCIDA. The estimate of PUC converges to the true value at a rate of $\mathcal{O}(\frac{1}{N})$ for the baseline estimator $\hat{\mu}$ and other commonly used models.

study these datasets in the setting of strategic withholding.

6.1 Evaluating PUCIDA

Availability Inference Restriction is violated by full feature models. First, we demonstrate the effect that full feature models have on Availability Inference Restriction. We follow common practices and use zero-imputation to deal with unavailable feature values (Emmanuel et al. 2021). Then, we train a Random Forest model on all features of the dataset where we have introduced stochastic availability into one feature (see previous paragraph). We also train a base feature model that fully drops the optional feature from the dataset. We consider the subset of individuals with unavailable feature values (i.e., a=0) and report the average loss and absolute prediction of the positive class for both models in Table 2. We observe that the full feature models use the information contained in the missingness to obtain a lower loss. This can reduce the chance of obtaining the positive outcome from the full feature model compared to the base feature model by significant margin of up to 18 % for non-sharers. Hence, these results impressively show how the full feature model implicitly infers information from missingness and thereby violates protection requirements. This stays the same when applying established fairness constraints on the models (see Appendix F.1). In contrast, when applying PUC using PUCIDA this gap vanishes or is significantly reduced. We show that the same effect can be observed independently of the imputation techniques, the model class, and the model hyperparameters in Appendix F.3.

6.2 Evaluating the Theoretical Bounds

PUCIDA guarantees Predictive Non-Degradation. Usually model performance degrades when training models with additional constraints (e.g., see Corbett-Davies et al. (2017)). To measure model performance, we use the misclassification rate for classification tasks (ROC-AUC scores lead to qualitatively similar results, see Appendix F.4) and the MSE for regression tasks. The results in Table 3a confirm that PUCIDA (using exhaustive augmentation) improves over the base feature model, suggesting that PUCIDA models benefits from using optional information. This is the case even under under strategic actions where users only provide data if it improves their outcome, and aligns with our theoretical result in Corollary 1. Under non-strategic actions, the performance figures show the same characteristics (Appendix F.4). As expected, PUC-compliant models fare moderately worse than full feature models which have no protection requirements.

We now compare two different PUCIDA variants on multiple optional features: the first strategy ensures a fixed dataset size, i.e., the number of samples is equivalent to the original dataset size. The second strategy, which uses ex-

task	data	opt. feature	base model	PUCIDA	Full feature model
C	diab.	Glucose	29.30% ±0.62	**26.61%** ±0.56	23.41% ±0.69
C	compas	#priors	42.89% ±0.10	**40.85%** ±0.15	36.67% ±0.36
C	adult	edu-num	16.05% ±0.03	**15.94%** ±0.05	14.86% ±0.06
R	income	WKHP	85.07 ±0.17	**80.22** ±0.15	73.25 ±0.16
R	calif.	m_income	15.62 ±0.14	**14.79** ±0.08	13.40 ±0.03
R	insurance	experience	262.43 ±0.21	**254.35** ±0.39	236.92 ±0.42

(a) One dimensional case, strategic withholding. Metrics: C: (1-Acc)×100, R: MSE×100

		Fair models				Full feature model
task	data (# opt.)	Base feature model	PUCIDA (f)	PUCIDA (e)	(×)	zero-imputed
C	diab. (2)	29.74 ±2.92	26.23 ±4.42	**25.58** ±3.69	2.2	24.16 ±4.18
C	compas (5)	40.83 ±0.56	37.65 ±0.23	**37.21** ±0.71	7.6	36.86 ±1.20
C	adult (5)	17.98 ±0.37	15.35 ±0.36	**15.27** ±0.25	7.9	15.15 ±0.33
R	income (3)	52.40 ±0.92	**49.47** ±1.71	51.21 ±0.86	3.4	46.15 ±1.60
R	calif. (4)	6.64 ±0.79	6.83 ±0.32	**6.36** ±0.08	5.1	5.69 ±0.22
R	insurance (3)	271.72 ±4.14	**242.99** ±4.47	260.77 ±2.74	3.2	232.59 ±2.39

(b) r-dimensional case. Metrics: C: (1-Acc)×100, R: MSE×100

Table 3: PUC-compliant models leverage optional information to improve predictive performance relative to base feature models This is in line with Corollary 1. In the bottom table, two strategies are considered to achieve PUC: *fixed-size (f)* and *exhaustive (e)* PUCIDA. When using exhaustive PUCIDA, the predictive performance is always better than the performance of the base feature model, and often similar to the performance of the full feature models.

haustive data augmentation, leads to an increased dataset size. The factor by which the dataset size is increased is indicated by (×) along with the results in Table 3b. We observe that competitive results can often be obtained without any dataset increase; fixed-size PUCIDA even outperforms the exhaustive variant on the larger income and the insurance dataset, whereas the exhaustive augmentation leads to a more reliable performance increase. We study the performance for sharers in Table 6 (Appendix) and find that it remains on par with the full feature model. Overall, our results demonstrate that optional information can be leveraged in a conscious way through PUC-inducing data augmentation without suffering from prohibitive performance decrease for the decision maker and the sharers.

Convergence of PUCIDA. Finally, we study the convergence behavior of PUCIDA. As a measure of approximation quality, we use the PUC-Gap[2] defined in Equation (7), which measures the squared deviation from perfect PUC. As this notion requires the knowledge of the ground truth distribution, we use a synthetic dataset for this experiment. The dataset consists of eight binary features (five base, three optional). All features in this dataset are sampled independently. Labels are induced via a logistic distribution, and availability of the optional information depends on the label. For experiments on a second synthetic dataset with five continuous features (two base, three optional) and more details, see Appendix F.5.

First, we observe that PUCIDA is model agnostic, i.e., it works with a variety of state-of-the-art models leading to negligible PUC-gaps (see Figure 4). Second, we verify that the PUC-Gaps converge to zero at the rate of $\mathcal{O}(\frac{1}{N})$ as the sample size increases (Figure 5), confirming what we derived in Theorem 4. While common models (e.g., RandomForest, MLP) have a lower error than the baseline estimator $\hat{\mu}$ the models approach the baseline estimator with larger datasets and the gap closes at the suggested rate.

7 Conclusion and Future Work

In this work, we studied machine learning predictions where users have the option to disclose optional information. To comply with legal regulations and respect user consent, we introduced the notion of Protected User Consent (PUC) that strikes a balance between the interests of sharers, non-sharers, and decision-makers. We demonstrated that leveraging optional information from consenting users through PUC results in superior performance compared to models that disregard the optional information entirely.

Our work gives raise to several follow-up questions. It would be interesting to study possible long-term effects of PUC and how PUC incentivizes improvements. Furthermore, we have only considered users that act entirely strategic or on privacy grounds. Modeling heterogeneous users, who might be willing to accept a certain increase in costs in return for their privacy could be a meaningful extension.

Additional Material

An extended version of this work including technical appendices is available online[2]. We also publish our code as an open-source project[3].

[2] https://arxiv.org/abs/2210.13954

[3] https://github.com/tleemann/protectedconsent

References

Abernethy, J. D.; Awasthi, P.; Kleindessner, M.; Morgenstern, J.; Russell, C.; and Zhang, J. 2022. Active Sampling for Min-Max Fairness. In *International Conference on Machine Learning*, 53–65. PMLR.

Agarwal, A.; Beygelzimer, A.; Dudík, M.; Langford, J.; and Wallach, H. 2018. A reductions approach to fair classification. In *International Conference on Machine Learning*, 60–69. PMLR.

Aleryani, A.; Wang, W.; and De La Iglesia, B. 2020. Multiple imputation ensembles (MIE) for dealing with missing data. *SN Computer Science*, 1(3): 1–20.

Amiri, S.; Belloum, A.; Nalisnick, E.; Klous, S.; and Gommans, L. 2022. On the impact of non-IID data on the performance and fairness of differentially private federated learning. In *2022 52nd Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W)*, 52–58. IEEE.

Bagdasaryan, E.; Poursaeed, O.; and Shmatikov, V. 2019. Differential privacy has disparate impact on model accuracy. *Advances in neural information processing systems*, 32.

Biega, A. J.; Potash, P.; Daumé, H.; Diaz, F.; and Finck, M. 2020. Operationalizing the legal principle of data minimization for personalization. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 399–408.

Bird, S.; Dudík, M.; Edgar, R.; Horn, B.; Lutz, R.; Milan, V.; Sameki, M.; Wallach, H.; and Walker, K. 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. Technical Report MSR-TR-2020-32, Microsoft.

Chai, J.; and Wang, X. 2022. Fairness with adaptive weights. In *International Conference on Machine Learning (ICML)*, 2853–2866. PMLR.

Chouldechova, A. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2): 153–163.

Corbett-Davies, S.; Pierson, E.; Feller, A.; Goel, S.; and Huq, A. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, 797–806.

Cribari-Neto, F.; Garcia, N. L.; and Vasconcellos, K. L. 2000. A note on inverse moments of binomial variates. *Brazilian Review of Econometrics*, 20(2): 269–277.

DeMarco, J. 2023. Nearly 70% of Americans Would Wear a Fitness Tracker/Smartwatch for Discounted Health Insurance.

Ding, F.; Hardt, M.; Miller, J.; and Schmidt, L. 2021. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34: 6478–6490.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

Dwork, C.; McSherry, F.; Nissim, K.; and Smith, A. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, 265–284. Springer.

Emmanuel, T.; Maupong, T.; Mpoeleng, D.; Semong, T.; Mphago, B.; and Tabona, O. 2021. A survey on missing data in machine learning. *Journal of Big Data*, 8(1): 1–37.

Fernando, M.-P.; Cèsar, F.; David, N.; and José, H.-O. 2021. Missing the missing values: The ugly duckling of fairness in machine learning. *International Journal of Intelligent Systems*, 36(7): 3217–3258.

Fogliato, R.; Chouldechova, A.; and G'Sell, M. 2020. Fairness evaluation in presence of biased noisy labels. In *International Conference on Artificial Intelligence and Statistics*, 2325–2336. PMLR.

Ganev, G.; Oprisanu, B.; and De Cristofaro, E. 2022. Robin Hood and Matthew Effects: Differential Privacy Has Disparate Impact on Synthetic Data. In *International Conference on Machine Learning*, 6944–6959. PMLR.

GDPR. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council. *Official Journal of the European Union*.

Geurts, P.; Ernst, D.; and Wehenkel, L. 2006. Extremely randomized trees. *Machine learning*, 63: 3–42.

Ginart, A.; Guan, M. Y.; Valiant, G.; and Zou, J. 2019. Making AI Forget You: Data Deletion in Machine Learning. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada*.

Golatkar, A.; Achille, A.; and Soatto, S. 2020. Eternal Sunshine of the Spotless Net: Selective Forgetting in Deep Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Goldsteen, A.; Ezov, G.; Shmelkin, R.; Moffie, M.; and Farkash, A. 2021. Data minimization for GDPR compliance in machine learning models. *AI and Ethics*, 1–15.

Hardt, M.; Megiddo, N.; Papadimitriou, C.; and Wootters, M. 2016. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, 111–122.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Henning, L. 2022. Wellness apps and fitness trackers: Why insurers love your smartwatch. *Sydney Morning Herald*.

Izzo, Z.; Anne Smart, M.; Chaudhuri, K.; and Zou, J. 2021. Approximate Data Deletion from Machine Learning Models. In Banerjee, A.; and Fukumizu, K., eds., *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 130. PMLR.

Jeong, H.; Wang, H.; and Calmon, F. P. 2022. Fairness without imputation: A decision tree approach for fair prediction with missing values. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 9558–9566.

Kachuee, M.; Karkkainen, K.; Goldstein, O.; Darabi, S.; and Sarrafzadeh, M. 2020. Generative imputation and stochastic prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Kilbertus, N.; Rodriguez, M. G.; Schölkopf, B.; Muandet, K.; and Valera, I. 2020. Fair decisions despite imperfect predictions. In *International Conference on Artificial Intelligence and Statistics*, 277–287. PMLR.

Kilbertus, N.; Rojas Carulla, M.; Parascandolo, G.; Hardt, M.; Janzing, D.; and Schölkopf, B. 2017. Avoiding discrimination through causal reasoning. *Advances in neural information processing systems*, 30.

Kleinberg, J.; Mullainathan, S.; and Raghavan, M. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Krishnaswamy, A. K.; Li, H.; Rein, D.; Zhang, H.; and Conitzer, V. 2021. Classification with strategically withheld data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 5514–5522.

Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. *Advances in neural information processing systems*, 30.

Li, P.; and Liu, H. 2022. Achieving fairness at no utility cost via data reweighing with influence. In *International Conference on Machine Learning*, 12917–12930. PMLR.

Li, Y.; and Vasconcelos, N. 2019. Repair: Removing representation bias by dataset resampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9572–9581.

Lipton, Z.; McAuley, J.; and Chouldechova, A. 2018. Does mitigating ML's impact disparity require treatment disparity? *Advances in neural information processing systems*, 31.

Madison, K.; Schmidt, H.; and Volpp, K. G. 2013. Smoking, obesity, health insurance, and health incentives in the Affordable Care Act. *Jama*, 310(2): 143–144.

Mittelstadt, B. D.; Allo, P.; Taddeo, M.; Wachter, S.; and Floridi, L. 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2): 2053951716679679.

Mohan, K.; and Pearl, J. 2021. Graphical models for processing missing data. *Journal of the American Statistical Association*, 116(534): 1023–1037.

Mohan, K.; Thoemmes, F.; and Pearl, J. 2018. Estimation with incomplete data: The linear case. In *Proceedings of the International Joint Conferences on Artificial Intelligence Organization*.

Nilforoshan, H.; Gaebler, J. D.; Shroff, R.; and Goel, S. 2022. Causal conceptions of fairness and their consequences. In *International Conference on Machine Learning*, 16848–16887. PMLR.

OAG, C. 2021. CCPA regulations: Final regulation text. *Office of the Attorney General, California Department of Justice*.

Pawelczyk, M.; Leemann, T.; Biega, A.; and Kasneci, G. 2023. On the Trade-Off between Actionable Explanations and the Right to be Forgotten. In *International Conference on Learning Representations (ICLR)*.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830.

Poole, D.; Mehr, A. M.; and Wang, W. S. M. 2020. Conditioning on "and nothing else": Simple Models of Missing Data between Naive Bayes and Logistic Regression. In *ICML Workshop on the Art of Learning with Missing Values (Artemiss)*.

Rastegarpanah, B.; Crovella, M.; and Gummadi, K. P. 2020. Fair inputs and fair outputs: The incompatibility of fairness in privacy and accuracy. In *Adjunct Publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, 260–267.

Rastegarpanah, B.; Gummadi, K.; and Crovella, M. 2021. Auditing black-box prediction models for data minimization compliance. *Advances in Neural Information Processing Systems*, 34: 20621–20632.

Rateike, M.; Majumdar, A.; Mineeva, O.; Gummadi, K. P.; and Valera, I. 2022. Don't Throw it Away! The Utility of Unlabeled Data in Fair Decision Making. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1421–1433.

Reeder, B.; and David, A. 2016. Health at hand: A systematic review of smart watch uses for health and wellness. *Journal of biomedical informatics*, 63: 269–276.

Roh, Y.; Lee, K.; Whang, S.; and Suh, C. 2021. Sample selection for fair and robust training. *Advances in Neural Information Processing Systems*, 34: 815–827.

Romano, Y.; Bates, S.; and Candes, E. 2020. Achieving equalized odds by resampling sensitive attributes. *Advances in Neural Information Processing Systems*, 33: 361–371.

Rong, Y.; Leemann, T.; Borisov, V.; Kasneci, G.; and Kasneci, E. 2022. Evaluating feature attribution: An information-theoretic perspective. In *International Conference on Machine Learning*, 18770 – 18795.

Rubin, D. B. 1976. Inference and missing data. *Biometrika*, 63(3): 581–592.

Statista. 2023. Wearable Shipments Worldwide.

Suriyakumar, V. M.; Papernot, N.; Goldenberg, A.; and Ghassemi, M. 2021. Chasing your long tails: Differentially private prediction in health care settings. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 723–734.

US Government. U.S. Centers for Medicare & Medicaid Services. 2023. How insurance companies set health premiums.

Verma, S.; and Rubin, J. 2018. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (Fairware)*, 1–7. IEEE.

Wachter, S.; and Mittelstadt, B. 2019. A right to reasonable inferences: re-thinking data protection law in the age of big data and AI. *Colum. Bus. L. Rev.*, 494.

Wang, Y.; and Singh, L. 2021. Analyzing the impact of missing values and selection bias on fairness. *International Journal of Data Science and Analytics*, 12(2): 101–119.

Wu, Y.; Dobriban, E.; and Davidson, S. 2020. DeltaGrad: Rapid retraining of machine learning models. In *International Conference on Machine Learning (ICML)*, 10355–10366. PMLR.

Zafar, M. B.; Valera, I.; Gomez-Rodriguez, M.; and Gummadi, K. P. 2019. Fairness constraints: A flexible approach for fair classification. *The Journal of Machine Learning Research*, 20(1): 2737–2778.

Zhang, S.; Qin, Z.; Ling, C. X.; and Sheng, S. 2005. " Missing is useful": Missing values in cost-sensitive decision trees. *IEEE transactions on knowledge and data engineering*, 17(12): 1689–1693.

Zhang, Y.; and Long, Q. 2021. Assessing Fairness in the Presence of Missing Data. *Advances in neural information processing systems*, 34: 16007–16019.

Zimmer, M.; Kumar, P.; Vitak, J.; Liao, Y.; and Chamberlain Kritikos, K. 2020. 'There's nothing really they can do with this information': unpacking how users manage privacy boundaries for personal fitness information. *Information, Communication & Society*, 23(7): 1020–1037.

A Related Work and other Fairness Notions

A.1 Additional Related Work

Estimation of causal effects in the presence of missing data. The works by Mohan, Thoemmes, and Pearl (2018); Mohan and Pearl (2021) introduce graphical models for incomplete data and study the consistent estimation of causal effects amidst missing values. Our work differs as we are not concerned with estimating true causal effects but focus on building a definition of fairness in the presence of optional data.

Implementing Fairness in ML Systems. There are different strategies to implement fairness and mitigate bias in practical decision-making systems. This can be done by adding additional constraints to the optimization problem (e.g., Zafar et al. (2019)). To solve the such an optimization problem, one can employ the reductions approach (Agarwal et al. 2018), where the fairness constraint is reduced to a series of classification problems with different costs assigned to each sample. Furthermore, another line of work consists of preprocessing approaches to obtain models that are compliant with classical fairness notions. These work through sample selection (Roh et al. 2021; Abernethy et al. 2022) and reweighting approaches (Chai and Wang 2022; Li and Vasconcelos 2019; Li and Liu 2022) or through resampling of the sensitive attribute (Romano, Bates, and Candes 2020). While these approaches can help fulfill common fairness notions, they cannot easily be applied to obtain PUC.

Trade-offs between Privacy and Fairness. Possible trade-offs between classical notions of privacy such as Differential Privacy (DP, Dwork et al. 2006) have been previously studied (Bagdasaryan, Poursaeed, and Shmatikov 2019; Ganev, Oprisanu, and De Cristofaro 2022; Amiri et al. 2022), showing that imposing DP may lead to disparate outcomes across sensitive groups or reinforce existing biases. Suriyakumar et al. (2021) recently found that imposing privacy constraints can lead to an undue influence of majoritiy groups over miniories, thus possibly impacting fairness. Although we are considering personal data in this work, this paper differs from classical privacy literature because we are not concerned with data leakage. Instead, we strive to give users a better choice of which data to provide in the first place.

A.2 Common fairness notions are not applicable

As many definitions of fair outcomes between an advantaged and a disadvantaged group exist, we investigate whether existing definitions can readily be applied or easily adapted to the optional feature setting considered in this work. In other words, here we study whether existing fairness notions comply with our desiderata of Optimality and Availability Inference Restriction. In the conventional fairness literature, the impact of a sensitive attribute on the prediction is restricted. However, in the optional feature setting the point of departure is different since the optional feature may contain discriminative information that we explicitly want to use in come cases (recall that the sharers would like to obtain the most accurate prediction given their information). If not stated otherwise, we consider the availability feature A (see Figure 2) to be the sensitive attribute. We denote the predicted label by $\hat{Y}$ and discuss binary labels $Y \in \{0, 1\}$ as in most of the original definitions.

Fairness through Unawareness. This notion demands that the availability indicator A is not used as an explicit input in the decision-making process. Removing explicit information on the availability can be done easily by dropping the feature A. This makes "Fairness through Unawareness" very easy to implement. However, the group information is still implicitly encoded in the optional feature through the value `N/A` (see Fig. 2). A sufficiently complex classifier can infer this group information and include it into its decision-making. Therefore, this fairness notion cannot be applied in the optional feature setting as it violates Availability Inference Restriction.

Predictive Parity. This notion of fairness constrains the False Discovery Rates to be equal across groups, i.e., $P(Y = 0|\hat{Y} = 1, A = 0) = P(Y = 0|\hat{Y} = 1, A = 1)$. We argue that this definition and other error rate-based ones will not work in our setup because they bound performance an thus violate Optimality. It is desired by the sharers and the decision maker that the predictions will be more accurate when the feature z is present ($A = 1$) because the information in z should explicitly be used in the decision-making process if users decide to share their data on the optional features. One can make an analogous argument for other error-rate based notions such as equalized odds and equal opportunity.

Equalized Odds and Equal Opportunity (Hardt, Price, and Srebro 2016). Equalized odds requires the predicted label $\hat{Y}$ and the the protected attribute A to be conditionally independent given the true label Y. Formally, this means $P(\hat{Y}|A, Y) = P(\hat{Y}|Y)$ for all values of $Y, A, \hat{Y}$. This effectively constrains the true and false positive rates to be equal across groups. However, by the desideratum of Optimality, it is required to use class-discriminative information in the optional feature, which will necessarily lead to lower missclassification rates for subjects with $A{=}1$. Another fairness notion is Equal Opportunity which is a relaxation of Equalized Odds that only demands $P(\hat{Y}|A{=}1, Y{=}1) = P(\hat{Y}|A{=}0, Y{=}1)$, thus constraining the true positive rates across groups. To fulfill this notion, for $A{=}1$, the true positive rate would have to be kept artificially low to match that of the case $A{=}0$, with less information. This would thus result in a lower $P(\hat{Y}{=}1|A{=}0, Y{=}1)$, than could be achieved otherwise. Let $Y{=}1$ be the desirable outcome (e.g., being assigned a low insurance quote); this means that less subjects are rewarded with the justified positive outcome. This is incompatible with our desideratum of Optimality.

Statistical Parity (Dwork et al. 2012; Kusner et al. 2017). This definition is satisfied by a classifier if subjects in both protected and non-protected groups have an equal probability of getting a positive classification outcome: $P(\hat{Y}{=}1|A{=}0) = P(\hat{Y}{=}1|A{=}1)$. If the set of people providing additional information has more favorable base features in general, this definition may lead to different thresholds where people that choose to provide information are getting a lower score to achieve parity. This

definition would even forbid using this base features' full distinctive power, because one has to equalize over both missingness classes, thus contradicting Optimality.

Individual fairness (Dwork et al. 2012). Fairness definitions in this category use a distance metric m to define similarities $m(\mathbf{x}_i, \mathbf{x}_j)$ between individuals x_i and x_j. Considering the application in mind, the sensitive attributes should not play a role in determining the distance. The classifier output distributions for $f(\mathbf{x}_j)$ and $f(\mathbf{x}_i)$ that are compared by some divergence D should not differ more than the distance between these individuals, i.e., $D(f(\mathbf{x}_j), f(\mathbf{x}_i)) \leq m(\mathbf{x}_i, \mathbf{x}_j)$ (Dwork et al. 2012). In the considered setting, following the proposition by Verma and Rubin (2018), we could define the distance to be 0, if individuals have the same base features $\mathbf{b}$. This would effectively constrain the classification outcome to be identical independently of the optional feature specified, effectively prohibiting its use. Even when defining other distance metrics, the classification outcome will still be constrained to a certain range, again contradicting our desideratum of Optimality.

(Conditional) Statistical Parity. Statistical parity (SP) is known to be notoriously unfair on an individual level (Dwork et al. 2012). Therefore, Corbett-Davies et al. (2017) define the notion of conditional statistical parity (CSP), which is an extension of SP, where some attributes are allowed to affect the decision. If we allow all base features $\mathbf{b}$, the resulting definition expressed in expectations would be $P[\hat{Y}|\boldsymbol{B} = \mathbf{b}, A = 0] = P[\hat{Y}|\boldsymbol{B} = \mathbf{b}, A = 1]$. While this definition can be compliant with Availability Inference Restriction, we show that CSP-compliant models cannot meet the desire of the sharers for most accurate predictions. They cannot assign most accurate predictions to sharers or would encounter prohibitively high costs due to the CSP constraint if they did so. Indeed, they can be worse than the performance of a base feature model, even when we assign the sharers the most accurate predictions. This is the case even under idealized conditions (i.e., known expectations) and when Incentivization is perfectly fulfilled. PUC-models do not suffer from this limitations and can assign sharers most accurate predictions while always matching the performance of the base feature model.

Lemma 1 (CSP-compliant models can degrade model performance over base feature model)**.** *There exists a density* $\mathbf{p}$ *for which a CSP-compliant model* $f^*_{CSP} : \mathcal{X} \to [0, 1]$ *which assigns the most accurate predictions to the sharers, i.e.,* $f^*_{CSP|_{a=1}}(\mathbf{b}, a, z^*) = \mathbb{F}_{\mathcal{L}}[Y|\mathbf{b}, A = 1, z^*]$ *leads to higher expected losses (for MSE and BCE losses) than an optimal base feature model* g^*_{BCE} *:*

$$\mathbb{E}_{\mathbf{p}}[\mathcal{L}(g^*_{BCE}(\mathbf{B}), Y)] < \mathbb{E}_{\mathbf{p}}[\mathcal{L}(f^*_{CSP}(\mathbf{B}, A, Z^*), Y)]. \tag{8}$$

A proof is provided in Appendix D.1. There, we give a density $\mathbf{p}$ that serves as such a counterexample. We argue that this fairness notion is incompatible with the desire of the sharers for accurate predictions and the decision makers desire for low overall costs. Thus, we have established that common fairness definitions fail to conform to our desiderata of Availability Inference Restriction or result in models with unreasonable performance characteristics.

B Intuition and Additional Examples

In this section, we provide a simple example to show the problem of possible unfairness and provide more intuition for our notion of Protected User Consent.

B.1 Standard losses may lead to unfair treatment

We revisit the example of college admission, to show how imputation leads to possibly unfair treatment. Suppose we are given the samples $\{(\mathbf{x}^i, y^i)\}_{i=1,..N}$ with $N = 5$ from Fig. 2. Using the standard Mean-Squared Error (MSE) loss, we solve the following empirical risk minimization problem:

$$f^* = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{N} (f(\mathbf{x}^i) - y^i)^2,$$

with a sufficiently expressive function class $\mathcal{F}$. For samples in the example data set, this will yield the outcome $f^*(\mathbf{x}) = \frac{1}{|\{\mathbf{x}^i = \mathbf{x}\}|} \sum_{\{\mathbf{x}^i = \mathbf{x}\}} y_i$, the empirical mean. Consider the samples $\mathbf{x}^1$ and $\mathbf{x}^4$ in Fig. 2. Candidate 1 chose to share an additional feature, while candidate 4 did not. Although they have the same base features, their classification will be different at test time (as the data in the training set) with $f^*(\mathbf{x}^1) = 3k\$$ and $f^*(\mathbf{x}^4) = 64k\$$.

We argue that in the case of candidate 4, the availability information was implicitly used to compute the score and resulted in a lower outcome. If only the base features had been available, i.e., f^* would have been trained on the data set $\{(\mathbf{b}^i, y^i)\}_{i=1..k}$, the model outcome would be $f^*(\mathbf{b}) = \frac{1}{|\{\mathbf{b}^i = \mathbf{b}\}|} \sum_{\{\mathbf{b}^i = \mathbf{b}\}} y^i$ with $f^*(\mathbf{b}^4) = 24k\$$ which is the dataset average for all customers from NSW with a basic coverage plan. In this work, we argue that the unavailability of certain features itself should not be used in the determination of the model outcome when no additional information is available.

B.2 Example: Missing at random (MAR) data.

For missing at random data (Rubin 1976), the likelihood of unavailability can be entirely accounted for by the observed base features $\mathbf{b}$ and is not affected by the partially observed z and the label y. Formally, for a single optional feature with random

availability A, $\mathbf{p}(A=0|\mathbf{b}) = \mathbf{p}(A=0|\mathbf{b}, z, y)$ for every $z \in \mathcal{X}^z, y \in \mathcal{Y}$. Therefore,

$$\mathbf{p}\left(y|\mathbf{b}, A=0\right) = \frac{\mathbf{p}(y, \mathbf{b}, A=0)}{\mathbf{p}(\mathbf{b}, A=0)} = \frac{\mathbf{p}(y)\mathbf{p}(\mathbf{b}|y)\mathbf{p}(A=0|\mathbf{b}, y)}{\sum_{y'} \mathbf{p}(y')\mathbf{p}(\mathbf{b}|y')\mathbf{p}(A=0|\mathbf{b}, y')} \tag{9}$$

$$= \frac{\mathbf{p}(y)\mathbf{p}(\mathbf{b}|y)\mathbf{p}(A=0|\mathbf{b})}{\sum_{y'} \mathbf{p}(y')\mathbf{p}(\mathbf{b}|y')\mathbf{p}(A=0|\mathbf{b})} = \frac{\mathbf{p}(y)\mathbf{p}(\mathbf{b}|y)\mathbf{p}(A=0|\mathbf{b})}{\mathbf{p}(A=0|\mathbf{b})\sum_{y'} \mathbf{p}(y')\mathbf{p}(\mathbf{b}|y')} \tag{10}$$

$$= \frac{\mathbf{p}(y)\mathbf{p}(\mathbf{b}|y)}{\sum_{y'} \mathbf{p}(y')\mathbf{p}(\mathbf{b}|y')} = \mathbf{p}(y|\mathbf{b}). \tag{11}$$

Therefore, we also have $\mathbb{E}\left[Y|B=\mathbf{b}, A=0\right] = \mathbb{E}_{\mathbf{p}}\left[Y|B=\mathbf{b}\right]$, indicating that the missingness does not affect the expected value of the label (or that of any other functional of $p(y|\mathbf{b})$) over the entire data distribution. Therefore, a perfect discriminative model with $f(\mathbf{x}) = \mathbb{F}^{\mathcal{L}}_{\mathbf{p}}\left[Y|\mathbf{x}\right]$ will fulfill Theorem 1, our definition of PUC, right away.

B.3 A probabilistic derivation of Non-Penalization

First, we require that the predictor does not use the information as an explicit input, i.e., the predictor should behave as if it only used base features $\mathbf{b}$ via some function $g : \mathcal{X}^b \to \mathcal{Y}$:

$$f_{|a=0}(\mathbf{b}, a, z^*) = g(\mathbf{b}). \tag{12}$$

For (b), although $a{=}0$ is not an explicit input to g, a sufficiently complex function may still be implicitly adapting to the group $a{=}0$ and thus incorporate information that the user did not give their consent to. Therefore, on its own, the constraint in eqn. (12) is insufficient to enforce Availability Inference Restriction, and we need to formally define which predictors g are not specific to the information provided by the group of non-consenting users. To make matters more concrete, we first consider the case of binary classification with $\mathcal{Y} \in \{0, 1\}$ from a probabilistic perspective and suppose $g(\mathbf{b})$ returns a numerical probability socre in $[0, 1]$. We let $\mathbf{p}_g(\hat{Y}|\mathbf{b})$ denote stochastic predictions $\hat{Y}$ defined through g where $\mathbf{p}_g(\hat{Y}=1|\mathbf{b}) \coloneqq g(\mathbf{b})$. We would like to constrain the information contained in the predictions $G(\mathbf{b})$ to not use any additional information that the users did not actively consent to. To come up with a suitable constraint, we consider two simple others predictors, where one should be allowed and the other should be ruled out.

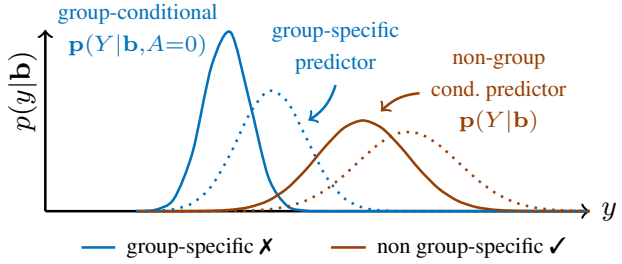


Figure 6: In our definition, predictors are called group-specific (blue) if they are closer to the group conditional distribution than to the overall label distribution (pink). Our requirement forbids the use of such group-specific predictors for the group of users with no additional information.

There are two canonical examples: a probabilistic estimator that is certainly not adapted to a specific group would be the one matching the ground truth overall conditional probability $\mathbf{p}\left(Y|\mathbf{b}\right)$. On the other extreme, the predictor $\mathbf{q}$ equivalent to $\mathbf{q}\left(Y|\mathbf{b}\right) \coloneqq \mathbf{p}\left(Y|\mathbf{b}, A{=}0\right)$ is fully leveraging the protected information and would thus be non-compliant. Generalizing this insight, we rule out all probabilistic predictors that are closer to the most non-compliant predictor than the overall conditional predictor, which we consider valid. These forbidden, group-specific predictors are visualized in Figure 6.

To this end, a suitable distance metric is required between the predictive distributions. A common choice is the Kullback-Leibler divergence $\mathcal{D}_{\text{KL}}$, which results in the following requirement for a predictor g:

$$\mathcal{D}_{\text{KL}}\left(\mathbf{p}(Y|\mathbf{B}, A{=}0)||\mathbf{p}_g(Y|\mathbf{B})\right) \geq \mathcal{D}_{\text{KL}}\left(\mathbf{p}(Y|\mathbf{B}, A{=}0)||\mathbf{p}(Y|\mathbf{B})\right). \tag{13}$$

The condition from eqn. (13) can be equivalently stated in terms of expected loss for binary classification and regression problems; i.e., the above condition allows to derive a generalized principle of Availability Inference Restriction (see Appendix D.4 for the proof).

C Effects of user choices in PUC-compliant models

In this section, we provide a brief discussion on the effect a user's choice to provide or not provide optional information has for the decision maker and for the affected end users.

Balancing the interests of consenting users, non-consenting users and decision makers. From the user's viewpoint, we would like to outline that a user's choice to provide optional information or not may depend on different factors:

(a) **Relevance of information:** How relevant does the individual deem the information that is asked for. The user may (correctly or incorrectly) deem certain information irrelevant for the decision and therefore they may not be willing to provide information on optional features.

(b) **Sensitivity of information:** How concerned is the individual about the optional information being unintentionally leaked or intentionally passed on to a third party.

(c) **Prior beliefs and expected outcomes:** The user's mental models of the decision system and the role their information plays in the system could be essential as a user may be more inclined to provide information which they deem beneficial for their prediction.

In summary, the *utility* of an individual for providing data is composed of several factors, including sensitivity, perceived relevance and anticipated outcomes.

From the perspective of the decision maker, PUC models become increasingly more accurate with more data being voluntarily provided. Therefore, from the perspective of the decision maker, it is important that users can also benefit from providing optional information. This desiderata is captured by optimality requirement, which allows the decision maker to make the most of all voluntarily provided information and allows the users to obtain more accurate decisions when additional information is provided.

Can less information lead to more favorable predictive outcomes for users? A user's predictive outcome depends on which information was provided by the user, and the predictive outcomes do not need to be monotonic in the number of optional features being provided; i.e., *providing more information on optional features does not necessarily lead to a better outcome for the user*.

To see why this behavior is necessary and desired, we consider the two extreme cases on either side of the spectrum, where (a) optional information may not impact the prediction outcome, and where (b) predictions can only get worse when providing strictly less features. In case (a), where no changes to the predictions occur, the setup becomes trivial and results in the base feature model. If this is the goal, then the collection of any additional information is useless for the decision and one should refrain from collecting these data, directly following the principle of Data Minimization. In case (b), where users can only get worse predictions with less information, a machine learning model would always have to treat users, who did not provide information, as if the worst possible value of a feature was provided. This is to make sure that the prediction outcomes of consenting users remains higher than the outcomes of non-consenting users with identical features. In the real-world setting of college admission we considered throughout the main text, this would lead to severe penalization of users who did not share optional test score results. This could de-facto rule them out entirely. *We argue that this behavior is not desired as it implicitly forces users to provide their data*. If this is desired, then the decision maker should make this choice explicit and the feature should then be mandatory. To conclude, for the setting of optional features with a decision maker who is interested in providing users a real choice, any sensible notion of fairness must allow for differences in the outcomes depending on the provided information.

D Proofs

D.1 Counterexample: Conditional statistical parity can be inferior to the base model

Expressed in terms of expectations, the notion of conditional statistical parity $\mathbb{E}[\hat{Y}|\boldsymbol{B}=\mathbf{b}, A=0] = \mathbb{E}[\hat{Y}|\boldsymbol{B}=\mathbf{b}, A=1]$ requires the prediction averages conditioned on $\mathbf{b}$ to be equal among groups that provided the optional features and those that did not. We now consider a non-probabilistic prediction function $\hat{Y} = f(\mathbf{b}, a, z^*)$. Plugging in the functional form would result in the following definition: $f_{|\boldsymbol{a}=0}(\mathbf{b}, \boldsymbol{a}=0, z^*) = \mathbb{E}_{Z^*\sim\mathbf{p}(Z^*|\mathbf{b},\boldsymbol{a}=1)}\left[f(\mathbf{b}, \boldsymbol{a}=1, Z^*)\right], \forall \mathbf{b}$. In the case $a=0$, z^* is constrained to be `N/A` so we can ignore its value. The subscript is used to indicate the restriction of f on the set of points with $\boldsymbol{a}$=0. This definition constrains the output $f_{|m=0}$, when no additional features provided, to match the average output of the individuals that provided features.

We follow the requirement most accurate predictions by the sharers which requires $f_{|a=1}$ to be the best approximation of $\mathbb{E}\left[Y|\boldsymbol{B}, A=1, Z^*\right]$. Thus, we would have to set $f_{|a=0}$ to be $f_{|a=0}(\mathbf{b}, a, z^*) = \mathbb{E}_{Z^*\sim\mathbf{p}(z^*|\mathbf{b},a=1)}\left[\mathbb{E}\left[Y|\boldsymbol{B}=\mathbf{b}, A=1, Z^*\right]\right] = \mathbb{E}\left[Y|B=\mathbf{b}, A=1\right]$ when marginalizing over Z^*. Overall, this derivation results in a function f_{csp} of the following form:

$$f_{\text{csp}}(\mathbf{b}, a, z^*) = \begin{cases} \mathbb{E}[Y|\mathbf{b}, A=1] & \text{if } a=0 \\ \mathbb{E}[Y|\mathbf{b}, A=1, Z^*=z^*], & \text{if } a=1 \end{cases}. \tag{14}$$

In this section we present a simple example to show that this function f_{csp} derived from notion of conditional statistical parity may lead to an increased Mean-Squared-Error (MSE) and Binary Cross Entropy (BCE) loss compared to the base model (not

using the optional feature) even when the estimators of the conditional means are perfect. Note that in a binary space $\mathcal{Y} = \{0, 1\}$ for both losses, predicting the conditional expectation is optimal.

For the example, we take any value $\mathbf{b}$ and suppose $p(y|\mathbf{b}, A, Z)$ depends on A but that Z is useless and does not contribute any new information, i.e. $\forall z \in \mathcal{X}^z, A \in \{0, 1\} : p(y|\mathbf{b}, A, z) = p(y|\mathbf{b}, A)$. Furthermore, we set the outcome to be deterministic of A:

$$\mathbb{E}[Y|\mathbf{b}, A = 0] = 0 \tag{15}$$
$$\mathbb{E}[Y|\mathbf{b}, A = 1] = 1 \tag{16}$$
$$\mathbb{E}[Y|\mathbf{b}] = \mathbf{p}(A = 1|\mathbf{b})\mathbb{E}[Y|\mathbf{b}, A = 0] + \mathbf{p}(A = 0|\mathbf{b})\mathbb{E}[Y|\mathbf{b}, A = 1] = \mathbf{p}(A = 1|\mathbf{b}) \coloneqq \alpha. \tag{17}$$

Let $\mathbf{p}(A = 1|\mathbf{b}) = \alpha$ be in the range $0 < \alpha < 1$. The optimal base feature model g^* would predict:

$$g^*(\mathbf{b}) = \alpha, \tag{18}$$

whereas the model based on CSP is given by:

$$f_{csp}(\mathbf{b}) = 1, \tag{19}$$

independently of the realization of A (because it is not allowed to use this information). The expected MSE Loss is given by:

$$L_{base,MSE} = \mathbf{p}(A = 0|\mathbf{b})(\alpha - 0)^2 + \mathbf{p}(A = 1|\mathbf{b})(\alpha - 1)^2 \tag{20}$$
$$= (1 - \alpha)\alpha^2 + \alpha(1 - \alpha)^2 = (1 - \alpha)\alpha(\alpha + 1 - \alpha) = (1 - \alpha)\alpha, \tag{21}$$

$$L_{base,BCE} = -\mathbf{p}(A = 0|\mathbf{b}) \log(1 - \alpha) - \mathbf{p}(A = 1|\mathbf{b}) \log \alpha < \infty. \tag{22}$$

If the notion derived from Conditional statistical parity is used, we would use $f_{csp}(\mathbf{b}) = \mathbb{E}[Y|\mathbf{b}, A = 1] = 1$ to predict in both cases and obtain:

$$L_{csp,MSE} = \mathbf{p}(A = 0|\mathbf{b})(0 - 1)^2 + \mathbf{p}(A = 1|\mathbf{b})(1 - 1)^2 = p(A = 0|\mathbf{b}) = 1 - \alpha, \tag{23}$$

$$L_{csp,BCE} = -\mathbf{p}(A = 0|\mathbf{b}) \log(1 - 1) - \mathbf{p}(A = 1|\mathbf{b}) \log 1 = p(A = 0|\mathbf{b}) = \infty. \tag{24}$$

For the BCE, we already see that the loss is unbounded in the case of CSP. One can construct the same example with non-infinite losses by adding a slight probability of the other outcome, i.e., setting $\mathbb{E}[Y|\mathbf{b}, A = 1] = 1 - \epsilon$ with some small $\epsilon > 0$ and obtain an analogous result.

For the MSE, in every case with $\alpha = \mathbf{p}(A = 1|\mathbf{b}) < 1$ this results in:

$$L_{csp} = 1 - \alpha > (1 - \alpha)\alpha = L_{base}. \tag{25}$$

We have now shown that for an arbitrary $\mathbf{b}$, the loss can be higher that of the base feature model. We can complete the example to the overall loss over a distributions of $\mathbf{b}$'s by supposing $\mathbf{p}(\mathbf{B} = \mathbf{b}) = 1$, which would however be a degenerate distribution. As a broader alternative, one can assume the above for a set of $\mathbf{b} \in \mathcal{B}$ and suppose any probability distribution with support in $\mathcal{B}$, i.e., $\mathbf{p}(\mathbf{B} \notin \mathcal{B}) = 0$.

D.2 Proof: The notion of Protected User Consent is optimal in the set of predictors conforming to the two desiderata

We can consider both predictors (for the case with and without optional features) independently. On the one hand, the notion of Availability Inference Restriction demands that the base predictor $f_{|_{a=0}}(\mathbf{b}) = g(\mathbf{b})$ should not outperform the optimal base predictor $f^*_{\mathcal{L}}$ trained on the full data set,

$$\mathbb{E}_{\mathbf{p}}[\mathcal{L}(f^*_{\mathcal{L}}(\mathbf{B}), Y)|A = 0] \leq \mathbb{E}_{\mathbf{p}}[\mathcal{L}(g(\mathbf{B}), Y)|A = 0]. \tag{26}$$

This directly provides us with one predictor $g(\mathbf{b})$, that is optimal in terms of loss for these individual namely, $g \equiv f^*_{\mathcal{L}}$ where $f^*_{\mathcal{L}}(\mathbf{b}) = \mathbb{F}_{\mathcal{L}}[Y|\mathbf{b}]$

On the other hand, for the group of individuals with optional information, we face no constraints and thus use the best predictor possible, i.e.,

$$f_{|_{a=1}}(\mathbf{b}, a, z^*) = \mathbb{F}_{\mathcal{L}}[Y|\mathbf{b}, A = 1, z^*] = \underset{f(\mathbf{b},a,z^*)}{\arg\min}\, \mathbb{E}_{\mathbf{p}(Y|\mathbf{b},A=1,z^*)}[\mathcal{L}(f(\mathbf{b}, a, z^*), Y)]. \tag{27}$$

Together, this results in the given definition of PUC. □

D.3 Proof: 1D-PUC obeys Predictive Non-Degradation

For the case of optional features ($A = 1$), we have:

$$f^{\text{PUC}}_{|a=1}(\mathbf{b}, a, z^*) = \mathbb{F}_{\mathcal{L}}\left[Y|\mathbf{b}, A = 1, z^*\right] = \underset{f(\mathbf{b},a,z^*)}{\arg\min}\, \mathbb{E}_{\mathbf{p}(Y|\mathbf{b},A=1,z^*)}\left[\mathcal{L}(f(\mathbf{b}, a, z^*), Y)\right]. \tag{28}$$

As is is the optimal predictor, its loss on these samples is smaller than that of any model, including the optimal model on the base features. Therefore, for each $\mathbf{b}, z^*$, we have:

$$\mathbb{E}_{\mathbf{p}(Y|\mathbf{b},A=1,z^*)}\left[\mathcal{L}(f^{\text{PUC}}(\mathbf{b}, a{=}1, z^*), Y)\right] \leq \mathbb{E}_{\mathbf{p}(Y|\mathbf{b},A=1,z^*)}\left[\mathcal{L}\left(f^*_{\mathcal{L}}(\mathbf{b}), Y\right)\right]. \tag{29}$$

Averaging over the entire class of samples with $A = 1$, we obtain:

$$\mathbb{E}_{\mathbf{p}}\left[\mathcal{L}\left(f^{\text{PUC}}(\mathbf{B}, A, Z^*), Y\right)\middle|A = 1\right] \leq \mathbb{E}_{\mathbf{p}}\left[\mathcal{L}\left(f^*_{\mathcal{L}}(\mathbf{B}), Y\right)|A = 1\right]. \tag{30}$$

On the other hand, the definition of PUC demands that the predictor in case $A = 0$ is equivalent to the optimal predictor on the base features. Thus they have equal loss and:

$$\mathbb{E}_{\mathbf{p}}\left[\mathcal{L}\left(f^{\text{PUC}}(\mathbf{B}, A, Z^*), Y\right)\middle|A = 0\right] = \mathbb{E}_{\mathbf{p}}\left[\mathcal{L}\left(f^*_{\mathcal{L}}(\mathbf{B}), Y\right)|A = 0\right]. \tag{31}$$

In total, we have

$$\begin{aligned}
\mathcal{L}^{\text{PUC}} &= \mathbb{E}_{\mathbf{p}}\left[\mathcal{L}\left(f^{\text{PUC}}(\mathbf{B}, A, Z^*), Y\right)\middle|A = 0\right]\mathbf{p}(A{=}0) + \mathbb{E}_{\mathbf{p}}\left[\mathcal{L}\left(f^{\text{PUC}}(\mathbf{B}, A, Z^*), Y\right)\middle|A = 1\right]\mathbf{p}(A{=}1) && (32)\\
&\leq \mathbb{E}_{\mathbf{p}}\left[\mathcal{L}\left(f^*_{\mathcal{L}}(\mathbf{B}), Y\right)|A = 0\right]\mathbf{p}(A{=}0) + \mathbb{E}_{\mathbf{p}}\left[\mathcal{L}\left(f^*_{\mathcal{L}}(\mathbf{B}), Y\right)|A = 1\right]\mathbf{p}(A{=}1) && (33)\\
&= \mathbb{E}_{\mathbf{p}}\left[\mathcal{L}\left(f^*_{\mathcal{L}}(\mathbf{B}), Y\right)\right] = \mathcal{L}^{\text{base}}. && (34)
\end{aligned}$$

□

D.4 The Generalized Principle of Availability Inference Restriction

We can reformulate the probabilistic definition of Availability Inference Restriction in terms of loss functions, which allows for generalization. We define $\mathbf{p}_0 = \mathbf{p}(\mathbf{B}|A = 0)$. We start by the notion given in the definition:

$$\mathcal{D}_{KL}\left(\mathbf{p}(Y|\mathbf{B}, A{=}0)||\mathbf{p}_g(Y|\mathbf{B})\right) \geq \mathcal{D}_{KL}\left(\mathbf{p}(Y|\mathbf{B}, A{=}0)||\mathbf{p}(Y|\mathbf{B})\right) \tag{35}$$

$$\mathbb{E}_{\mathbf{b}\sim\mathbf{p}_0}\mathcal{D}_{KL}\left(\mathbf{p}(Y|\mathbf{b}, A{=}0)||\mathbf{p}_g(Y|\mathbf{b})\right) \geq \mathbb{E}_{\mathbf{b}\sim\mathbf{p}_0}\mathcal{D}_{KL}\left(\mathbf{p}(Y|\mathbf{b}, A{=}0)||\mathbf{p}(Y|\mathbf{b})\right). \tag{36}$$

The Kullback-Leibler divergence can be decomposed as $\mathcal{D}_{KL}(\mathbf{p}||\mathbf{q}) = H(\mathbf{p}) + CE(\mathbf{p}||\mathbf{q})$, which results in:

$$\begin{aligned}
&\iff \mathbb{E}_{\mathbf{b}\sim\mathbf{p}_0}\text{CE}\left(\mathbf{p}(Y|\mathbf{b}, A{=}0)||\mathbf{p}_g(Y|\mathbf{b})\right) + \mathbb{E}_{\mathbf{b}\sim\mathbf{p}_0}H\left[Y|\mathbf{b}, A{=}0\right] && (37)\\
&\quad\geq \mathbb{E}_{\mathbf{b}\sim\mathbf{p}_0}\text{CE}\left(\mathbf{p}(Y|\mathbf{b}, A{=}0)||\mathbf{p}(Y|\mathbf{b})\right) + \mathbb{E}_{\mathbf{b}\sim\mathbf{p}_0}H\left[Y|\mathbf{b}, A{=}0\right] && (38)\\
&\iff \mathbb{E}_{\mathbf{b}\sim\mathbf{p}_0}\text{CE}\left(\mathbf{p}(Y|\mathbf{b}, A{=}0)||\mathbf{p}_g(Y|\mathbf{b})\right) \geq \mathbb{E}_{\mathbf{b}\sim\mathbf{p}_0}\text{CE}\left(\mathbf{p}(Y|\mathbf{b}, A{=}0)||\mathbf{p}(Y|\mathbf{b})\right) && (39)\\
&\iff \mathbb{E}_{\mathbf{b}\sim\mathbf{p}_0}\mathbb{E}_{Y\sim\mathbf{p}(Y|\mathbf{b},A=0)}\left[-\log\mathbf{p}_g(Y|\mathbf{b})\right] \geq \mathbb{E}_{\mathbf{b}\sim\mathbf{p}_0}\mathbb{E}_{Y\sim\mathbf{p}(Y|\mathbf{b},A=0)}\left[-\log\mathbf{p}(Y|\mathbf{b})\right] && (40)
\end{aligned}$$

The inner expectation is equivalent to the BCE loss for a specific $\mathbf{b}$. Averaged over all $\mathbf{b} \sim \mathbf{p}_0$ we obtain.

$$\Rightarrow \mathbb{E}_{\mathbf{p}}\left[\text{BCE}\left(g(\mathbf{B}), Y\right)|A = 0\right] \geq \mathbb{E}_{\mathbf{p}}\left[\text{BCE}\left(f^*_{\text{BCE}}(\mathbf{B}), Y\right)|A = 0\right]. \tag{41}$$

This notion allows for generalization by replacing BCE with some general loss function $\mathcal{L}$. Doing so results in

$$\mathbb{E}_{\mathbf{p}}\left[\mathcal{L}\left(g(\mathbf{B}), Y\right)|A = 0\right] \geq \mathbb{E}_{\mathbf{p}}\left[\mathcal{L}\left(f^*_{\mathcal{L}}(\mathbf{B}), Y\right)|A = 0\right], \tag{42}$$

the version of the desideratum of Availability Inference Restriction mentioned in the main paper. □

D.5 PUC under strategic withholding of data

To prove Theorem 2, we first note that the decision maker can only realize improvements over the base model in the setup of strategic interactions for individuals by offering them a lower premium than the prediction of the base model. Otherwise, they would strategically not provide their data.

It is only beneficial for the decision maker to do so if there exists an $y' \leq y_{\text{base}} := \mathbb{F}^{\mathcal{L}}[Y|\mathbf{b}]$ with a lower expected loss, i.e.,

$$\mathbb{E}_Y\left[\mathcal{L}(y', Y)|\mathbf{B} = \mathbf{b}, Z = z\right] \leq \mathbb{E}_Y\left[\mathcal{L}(y_{\text{base}}, Y)|\mathbf{B} = \mathbf{b}, Z = z\right] \tag{43}$$

Due to the convexity of the loss $\mathcal{L}$, this expected value will as well be convex in the prediction y' and we will also have $\mathbb{F}^{\mathcal{L}}[Y|\mathbf{b}, z] \leq \mathbb{F}^{\mathcal{L}}[Y|\mathbf{b}]$. The loss-minimal prediction would be $f(b, a = 1, z) = \mathbb{F}^{\mathcal{L}}[Y|\mathbf{b}, z]$, which will not be hindered through strategic actions. This however results in a PUC-model again, as $\mathbb{F}^{\mathcal{L}}[Y|\mathbf{b}, z] = \mathbb{F}^{\mathcal{L}}[Y|\mathbf{b}, A = 1, z]$, because the sharing decision does not influence the label given $\mathbf{B}, Z$. □

PUC with monotonicity constraints. A similar argument can be made when monotonicity constraints need to be enforced, i.e., the outcome can only decrease over the base model with more information provided. We can consider each optional feature value z separately for sharers. If the sample comes with a better average $\mathbb{F}^{\mathcal{L}}[Y|\mathbf{b}, z] \leq \mathbb{F}^{\mathcal{L}}[Y|\mathbf{b}]$ than the base prediction, we can confidently return this full-feature optimal prediction. In the contrary case, where $\mathbb{F}^{\mathcal{L}}[Y|\mathbf{b}, z] > \mathbb{F}^{\mathcal{L}}[Y|\mathbf{b}]$, the best prediction that the decision maker is allowed to make is the base feature models prediction (due to the convexity of the loss function). This is equivalent to dropping the optional feature in this case and using the corresponding PUC model.

D.6 Equivalence of Expectations for the Resampling model

In this section, we show that the resampling technique proposed in this work converges to the desired outcome. Therefore, we show that in the infinite sample-limit, the optimum reached when optimizing the loss over the modified distribution corresponds to the desired PUC model.

We introduce the usual mapping $\mathcal{I}(\boldsymbol{a}) \coloneqq \{i \mid \boldsymbol{a}_i = 1, i = 1, \ldots, r\}$ to denote the set of all indices that are 1 in the vector $\boldsymbol{a}$ but also use $\mathbb{I}_S$ to denote the binary indicator vector where all components corresponding to indices in S are set to 1 and to zero otherwise, i.e., $(\mathbb{I}_S)_i = \{1 \text{ if } i \in S, \text{else } 0\}$. Note that these operations invert each other such that $\mathbb{I}_{\mathcal{I}(\boldsymbol{a})} = \boldsymbol{a}$. We can show that the optimal prediction $\hat{y} = \hat{y}(\mathbf{b}, \boldsymbol{a}, \boldsymbol{z}^*)$ is given by:

$$\hat{y} = \mathbb{F}^{\mathcal{L}}_{((\mathbf{b},\boldsymbol{a},\boldsymbol{z}^*),y)\sim\overline{\mathbf{p}}}\left[Y|\mathbf{b}, \boldsymbol{A} = \boldsymbol{a}, \boldsymbol{Z}^* = \boldsymbol{z}^*\right] = \arg\min_{\hat{y}} \mathbb{E}_{((\mathbf{b},\boldsymbol{a},\boldsymbol{z}^*),y)\sim\overline{\mathbf{p}}}\left[\mathcal{L}(\hat{y}, y)|\mathbf{b}, \boldsymbol{A} = \boldsymbol{a}, \boldsymbol{Z}^* = \boldsymbol{z}^*\right] \tag{44}$$

$$= \arg\min_{\hat{y}} \sum_{\mathcal{I}(\boldsymbol{a})\subseteq S} \mathbf{p}(\boldsymbol{A}{=}\mathbb{I}_S|\mathbf{B}{=}\mathbf{b}, \boldsymbol{Z}_{\mathcal{I}(\boldsymbol{a})}{=}\boldsymbol{z}_{\mathcal{I}(\boldsymbol{a})})\mathbb{E}_{\mathbf{p}}\left[\mathcal{L}(\hat{y}, y)\middle|\boldsymbol{A} = \mathbb{I}_S, \mathbf{B} = \mathbf{b}, \boldsymbol{Z}_{\mathcal{I}(\boldsymbol{a})} = \boldsymbol{z}_{\mathcal{I}(\boldsymbol{a})}\right] \tag{45}$$

$$= \arg\min_{\hat{y}} \mathbb{E}_{\mathbf{p}}\left[\mathcal{L}(\hat{y}, y)\middle|\mathcal{I}(\boldsymbol{a}) \subseteq \mathcal{I}(\boldsymbol{A}), \mathbf{B} = \mathbf{b}, \boldsymbol{Z}_{\mathcal{I}(\boldsymbol{a})} = \boldsymbol{z}_{\mathcal{I}(\boldsymbol{a})}\right] \tag{46}$$

$$= \arg\min_{\hat{y}} \mathbb{E}_{\mathbf{p}}\left[\mathcal{L}(\hat{y}, y)\middle|\boldsymbol{A}_{\mathcal{I}(\boldsymbol{a})} = \mathbf{1}, \mathbf{B} = \mathbf{b}, \boldsymbol{Z}_{\mathcal{I}(\boldsymbol{a})} = \boldsymbol{z}_{\mathcal{I}(\boldsymbol{a})}\right] \tag{47}$$

$$= \mathbb{F}^{\mathcal{L}}_{\mathbf{p}}\left[Y|\mathbf{B} = \mathbf{b}, \boldsymbol{A}_{\mathcal{I}(\boldsymbol{a})} = \mathbf{1}, \boldsymbol{Z}_{\mathcal{I}(\boldsymbol{a})} = \boldsymbol{z}_{\mathcal{I}(\boldsymbol{a})}\right]. \tag{48}$$

In Equation (45), we use the fact that we can express the distribution $\overline{\mathbf{p}}$ for a subset of inputs with $\boldsymbol{A} = \boldsymbol{a}$ as a mixture of $\mathbf{p}$, averaged over all subsets of inputs S with more optional features than $\boldsymbol{a}$, weighted equally but with the optional information erased. This is a result of the data augmentation procedures that defines $\overline{\mathbf{p}}$. The total weight is just a factor and does not play a role in the $\arg\min$ operation. The following steps are just reformulations of the expression. □

D.7 Proof: Convergence of the sample approximation for a finite feature space

In this section we provide a general estimation of the error of a non-parametric regressor from a finite number of samples on a finite feature space $\mathcal{X}$ (e.g., finite, discrete features) and a label space $\mathcal{Y}$ that can be either continuous or discrete. Before we can prove the main result, we establish the following lemma.

Lemma 2. *The density $\overline{\mathbf{p}}$ that is obtained from $\mathbf{p}$ by applying the augmentation strategy described in the paper (`PUCIDA`) is related to the original density through the following relation:*

$$\forall \mathbf{x} \in \mathcal{X} : \overline{\mathbf{p}}(\mathbf{x}) \geq \frac{1}{2^r}\mathbf{p}(\mathbf{x}).$$

In particular, this implies that the support of $\overline{\mathbf{p}}$ is at least as big as the support of $\mathbf{p}$.

Proof. The resampling procedure consists of two steps. First, a reweighting is done. As we state in the main text, this reweighting from $\overline{\mathbf{p}}$ can be implemented through rejection sampling with samples from $\mathbf{x} = (\mathbf{b}, \boldsymbol{a}, \boldsymbol{z}^*) \sim \mathbf{p}$. Samples are passed on the the next stage with a probability of $\frac{2^{|\mathcal{I}(\boldsymbol{a})|}}{2^r}$. Using this scheme, we know that for a certain $\mathbf{x} = (\mathbf{b}, \boldsymbol{a}, \boldsymbol{z}^*)$, the probability of the sample to be observed after applying only the reweighting step is bounded by $\frac{2^{|\mathcal{I}(\boldsymbol{a})|}}{2^r}\mathbf{p}(\mathbf{x})$. To see this, we can consider the worst case, where all other samples are passed on with probability 1 and only the considered vector $\mathbf{x}$ is downweighted by a factor of $\frac{2^{|\mathcal{I}(\boldsymbol{a})|}}{2^r}$. If the other samples are also downweighted, this is a strict lower bound. In the second step, some optional features are dropped at random with a probability of $\frac{1}{2}$. We are interested in $\overline{\mathbf{p}}(\mathbf{x})$, the probability of obtaining the exact original sample with all its optional features still present. The probability that all optional features remain present with the Bernoulli distribution used, is given by $\frac{1}{2^{|\mathcal{I}(\boldsymbol{a})|}}$. Bringing it all together we obtain:

$$\forall \mathbf{x} \in \mathcal{X} : \overline{\mathbf{p}}(\mathbf{x}) \geq \frac{2^{|\mathcal{I}(\boldsymbol{a})|}}{2^r}\frac{1}{2^{|\mathcal{I}(\boldsymbol{a})|}}\mathbf{p}(\mathbf{x}) = \frac{1}{2^r}\mathbf{p}(\mathbf{x}). \tag{49}$$

Theorem 5 (Convergence of Finite Sample Approximation)**.** *Suppose a finite feature space $\mathcal{X}$ and a numerical label space $\mathcal{Y} \subseteq \mathbb{R}$. Suppose all conditional expectations $\mu_{\overline{\mathbf{p}}}(\mathbf{x}) \coloneqq \mathbb{E}_{\overline{\mathbf{p}}}[y|\mathbf{x}]$ and the conditional variances $\sigma^2_{\overline{\mathbf{p}}}(\mathbf{x}) \coloneqq \mathrm{Var}_{\overline{\mathbf{p}}}[y|\mathbf{x}]$ exist (and thus are finite). We can estimate a (discrete) non-parametric regressor $\hat{\mu}^{\mathcal{D}} : \mathcal{X} \mapsto \mathbb{R}$ from a finite number N of independent, identically distributed observations $\mathcal{D} = (\mathbf{x}_i, y_i)_{i=1\ldots N}$ from $\overline{\mathbf{p}}$ which satisfies:*

$$\mathbb{E}_{\mathbf{x}\sim\mathbf{p},\mathcal{D}\sim\overline{\mathbf{p}}}\left[\left(\hat{\mu}^{\mathcal{D}}(\mathbf{x}) - \mu_{\overline{\mathbf{p}}}(\mathbf{x})\right)^2\right] \leq \frac{2^r|\mathcal{X}|^2(\sigma^2_{\max} + \mu^2_{\max})}{N} + \mathcal{O}\left(\frac{1}{N^2}\right), \tag{50}$$

where $\sigma^2_{\max} \coloneqq \max_{\mathbf{x}\in\mathcal{X}} \sigma^2_{\overline{\mathbf{p}}}, (\mathbf{x})$ and $\mu^2_{\max} \coloneqq \max_{\mathbf{x}\in\mathcal{X}} \mu^2_{\overline{\mathbf{p}}}(\mathbf{x})$. The expected squared deviation to the optimal estimator converges at an order of $\mathcal{O}\left(\frac{1}{N}\right)$.

Proof. Before we proof the rate of convergence, we first define the estimator for which we establish this bound. We can draw N samples $\mathcal{D} \sim (\mathbf{x}_i, y_i) \sim \overline{\mathbf{p}}$. Then, we split these into $|\mathcal{X}|$ equal batches of size $M = \left\lfloor \frac{N}{|\mathcal{X}|} \right\rfloor$ samples. We can thus assign each possible feature value $\mathbf{x} \in \mathcal{X}$ a batch $\text{Batch}(\mathbf{x}) \subset [N]$, of samples, although the value the features $\mathbf{x}_i$ for i in the batch corresponding to $\mathbf{x}$ are still randomly distributed according to $\overline{\mathbf{p}}$. We only use M different samples to estimate each conditional mean. Denoting the true conditional mean by $\mu_{\mathbf{x}} \coloneqq \mu_{\overline{\mathbf{p}}}(\mathbf{x})$ and its estimate by $\hat{\mu}^{\mathcal{D}}_{\mathbf{x}} \coloneqq \hat{\mu}^{\mathcal{D}}(\mathbf{x})$ for the feature $\mathbf{x} \in \mathcal{X}$, we estimate:

$$\hat{\mu}^{\mathcal{D}}_{\mathbf{x}} = \frac{\sum_{(\mathbf{x}_i, y_i) \in \text{Batch}(\mathbf{x})} y_i \boldsymbol{\delta}_{\mathbf{x}_i = \mathbf{x}}}{1 + \sum_{(\mathbf{x}_i, y_i) \in \text{Batch}(\mathbf{x})} \boldsymbol{\delta}_{\mathbf{x}_i = \mathbf{x}}}, \tag{51}$$

where $\boldsymbol{\delta}_{\mathbf{x}_i = \mathbf{x}} = \{1 \text{ if } \mathbf{x}_i = \mathbf{x}, \text{else } 0\}$ denotes the indicator function. Depending on the number $b_{\mathbf{x}} = \sum_{(\mathbf{x}_i, y_i) \in \text{Batch}(\mathbf{x})} \boldsymbol{\delta}_{\mathbf{x}_i = \mathbf{x}}$ of samples with matching feature values that are used in the estimation of $\hat{\mu}_{\mathbf{x}}$, the estimator is slightly biased as $\mathbb{E}\left[\hat{\mu}^{\mathcal{D}}_{\mathbf{x}} \middle| b_{\mathbf{x}} = q\right] = \frac{q\mu_{\mathbf{x}}}{q+1}$ but the bias will vanish as $b_{\mathbf{x}} \to \infty$. Note that $b_{\mathbf{x}}$ is a random variable. The variance of the estimator on iid samples is $\text{Var}_{\overline{\mathbf{p}}}\left[\hat{\mu}^{\mathcal{D}}_{\mathbf{x}} \middle| b_{\mathbf{x}} = q\right] = \frac{q\sigma^2_{\mathbf{x}}}{(q+1)^2}$. Without loss of generality, we will suppose $\overline{p}_{\mathbf{x}} \coloneqq \overline{\mathbf{p}}(\mathbf{x}) > 0$: By Lemma 2, we obtain $\overline{\mathbf{p}}(\mathbf{x}) \geq \frac{1}{2^r}\mathbf{p}(\mathbf{x})$. Thus, $\overline{\mathbf{p}}(\mathbf{x}) = 0$ implies $\mathbf{p}(\mathbf{x}) = 0$ and the error of the estimator will not play a role in expected squared error we are interested in obtaining. By the well-known Bias-Variance decomposition, the square error of the single estimator $\mu_{\mathbf{x}}$ for a given $b_{\mathbf{x}} = q$ can be written as:

$$\mathbb{E}_{\mathcal{D} \sim \overline{\mathbf{p}}}\left[\left(\hat{\mu}^{\mathcal{D}}_{\mathbf{x}} - \mu_{\mathbf{x}}\right)^2 \middle| b_{\mathbf{x}} = q\right] = \left(\mathbb{E}_{\overline{\mathbf{p}}}\left[\hat{\mu}^{\mathcal{D}}_{\mathbf{x}} \middle| b_{\mathbf{x}} = q\right] - \mu_{\mathbf{x}}\right)^2 + \text{Var}_{\overline{\mathbf{p}}}\left[\hat{\mu}^{\mathcal{D}}_{\mathbf{x}} \middle| b_{\mathbf{x}} = q\right] \tag{52}$$

$$= \left(\frac{q\mu_{\mathbf{x}}}{q+1} - \mu_{\mathbf{x}}\right)^2 + \frac{q\sigma^2_{\mathbf{x}}}{(q+1)^2} = \left(\frac{1}{q+1}\right)^2 \mu^2_{\mathbf{x}} + \frac{q\sigma^2_{\mathbf{x}}}{(q+1)^2} \tag{53}$$

$$\leq \frac{1}{q+1}\mu^2_{\mathbf{x}} + \frac{(q+1)\sigma^2_{\mathbf{x}}}{(q+1)^2} = \frac{1}{q+1}\left(\mu^2_{\mathbf{x}} + \sigma^2_{\mathbf{x}}\right). \tag{54}$$

Due to the sampling procedure, the $b_{\mathbf{x}}$ are independently binomially distributed with $b_{\mathbf{x}} \sim \text{Bin}\left(M, \overline{p}_{\mathbf{x}}\right)$. Therefore, we can first aggregate the results for a single $\mathbf{x}$ and then average over the entire distribution over $\mathcal{X}$. We obtain:

$$\mathbb{E}_{\mathcal{D} \sim \overline{\mathbf{p}}}\left[\left(\hat{\mu}^{\mathcal{D}}_{\mathbf{x}} - \mu_{\mathbf{x}}\right)^2\right] = \sum_{q=0}^{M} p(b_{\mathbf{x}} = q) \mathbb{E}_{\mathcal{D} \sim \mathbf{p}}\left[\left(\hat{\mu}^{\mathcal{D}}_{\mathbf{x}} - \mu_{\mathbf{x}}\right)^2 \middle| b_{\mathbf{x}} = q\right] \tag{55}$$

$$\leq \sum_{q=0}^{M} \text{Bin}(q; M, \overline{p}_{\mathbf{x}}) \frac{1}{q+1}\left(\mu^2_{\mathbf{x}} + \sigma^2_{\mathbf{x}}\right) = \left(\mu^2_{\mathbf{x}} + \sigma^2_{\mathbf{x}}\right) \mathbb{E}_{b_{\mathbf{x}} \sim \text{Bin}(M, \overline{p}_{\mathbf{x}})}\left[\frac{1}{q+1}\right] \tag{56}$$

$$= \left(\mu^2_{\mathbf{x}} + \sigma^2_{\mathbf{x}}\right)\left(\frac{1}{\overline{p}_{\mathbf{x}}(M+1)}\right)\left(1 - (1 - \overline{p}_{\mathbf{x}})^{M+1}\right) \tag{57}$$

$$\leq \left(\mu^2_{\mathbf{x}} + \sigma^2_{\mathbf{x}}\right)\left(\frac{1}{\overline{p}_{\mathbf{x}}(M+1)}\right) < \left(\mu^2_{\mathbf{x}} + \sigma^2_{\mathbf{x}}\right)\left(\frac{1}{\overline{p}_{\mathbf{x}} M}\right), \tag{58}$$

where $\text{Bin}(q; M, \overline{p}_{\mathbf{x}}) = \binom{M}{q}(\overline{p}_{\mathbf{x}})^q (1 - \overline{p}_{\mathbf{x}})^{M-q}$ is the probability given by the binomial law and the equality in Equation (57) is provided in (Cribari-Neto, Garcia, and Vasconcellos 2000, p.271). We aggregate this result to an expected value over samples from the original distribution $\mathbf{p}$. The sample $\mathbf{x} \sim \mathbf{p}$ that the estimator is evaluated on and the data set $\mathcal{D} \sim \overline{\mathbf{p}}$ are independent, and we can derive an expected error for the distribution $\mathbf{p}$ over all features $\mathbf{x}$, as:

$$\mathbb{E}_{\mathbf{x} \sim \mathbf{p}, \mathcal{D} \sim \overline{\mathbf{p}}}\left[\left(\hat{\mu}(\mathbf{x}) - \mu(\mathbf{x})\right)^2\right] = \sum_{\mathbf{x} \in \mathcal{X}} p_{\mathbf{x}} \mathbb{E}_{\mathcal{D} \sim \mathbf{p}}\left[\left(\hat{\mu}^{\mathcal{D}}_{\mathbf{x}} - \mu_{\mathbf{x}}\right)^2\right] \tag{59}$$

$$< \sum_{\mathbf{x} \in \mathcal{X}} p_{\mathbf{x}}\left(\mu^2_{\mathbf{x}} + \sigma^2_{\mathbf{x}}\right)\left(\frac{1}{\overline{p}_{\mathbf{x}} M}\right) \leq \sum_{\mathbf{x} \in \mathcal{X}} 2^r\, \overline{p}_{\mathbf{x}}\left(\mu^2_{\mathbf{x}} + \sigma^2_{\mathbf{x}}\right)\left(\frac{1}{\overline{p}_{\mathbf{x}} M}\right) \tag{60}$$

$$\leq \sum_{\mathbf{x} \in \mathcal{X}} \frac{2^r(\mu^2_{\max} + \sigma^2_{\max})}{M} \tag{61}$$

$$= |\mathcal{X}| \frac{2^r(\mu^2_{\max} + \sigma^2_{\max})}{M} = \frac{|\mathcal{X}|^2(2^r(\mu^2_{\max} + \sigma^2_{\max}))}{M|\mathcal{X}|} \leq \frac{2^r|\mathcal{X}|^2(\mu^2_{\max} + \sigma^2_{\max})}{N - |\mathcal{X}| + 1} \tag{62}$$

$$= \frac{2^r|\mathcal{X}|^2(\sigma^2_{\max} + \mu^2_{\max})}{N} + \mathcal{O}\left(\frac{1}{N^2}\right), \tag{63}$$

where we use the fact that $2^r\, \overline{p}_{\mathbf{x}} \geq p_{\mathbf{x}}$ and the definitions of $\mu^2_{\max}, \sigma^2_{\max}$ as specified in the theorem.

□

Algorithm 1: PUC-SGD: SGD with Protected User Consent

Require: Data set $\mathcal{D}$, Loss function $\mathcal{L}$, predictor $f_{\boldsymbol{\theta}}$ with parameters $\boldsymbol{\theta}$
  $\mathbf{w} \leftarrow \{\text{Distribution over } \mathcal{D} \text{ with } \mathbf{w}(\mathbf{x}) \propto w(\mathbf{x})\}$
  while $r \neq 0$ **do**
    Sample batch $(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(k)}, y^{(k)}) \sim \mathbf{w}$
    for $j = 1, \ldots, k$ **do** ▷ $\mathbf{x}^{(j)}=(\mathbf{b}^{(j)}, \boldsymbol{a}^{(j)}, \boldsymbol{z}^{*(j)})$
      $\mathbf{q} \leftarrow \text{Bernoulli}(0.5)$ ▷ iid. Bernoulli vector
      $\overline{\boldsymbol{a}}^{(j)} = \mathbf{q} \odot \boldsymbol{a}^{(j)}$
      $\overline{\boldsymbol{z}}_i^{*(j)} = \left\{ \boldsymbol{z}_i^{*(j)} \text{ if } \overline{\boldsymbol{a}}_i^{(j)}{=}1, \text{ else } \texttt{N/A} \right\}, i \in [r]$
      $\overline{\mathbf{x}}^{(j)} \leftarrow (\mathbf{b}^{(j)}, \overline{\boldsymbol{a}}^{(j)}, \overline{\boldsymbol{z}}^{*(j)})$
    end for
    $\boldsymbol{d\theta} \leftarrow \nabla_{\boldsymbol{\theta}} \left(\frac{1}{k} \sum_{j=1}^{k} \mathcal{L} \left(f_{\boldsymbol{\theta}}(\overline{\mathbf{x}}^{(j)}), y^{(j)} \right) \right)$
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \gamma \boldsymbol{d}\theta$
  end while
  return $\boldsymbol{\theta}$

D.8 Algorithms

An example of how Protected User Consent through data augmentation can be incorporated in an SGD-type algorithm is provided in Algorithm 1.

E Protected User Consent on Simulated Distributions

In this section we introduce two types of parametric data distributions with optional information that we use in our experiments with simulated data. They allow to independently control the complexity and to obtain as many samples as needed to study the convergence behavior. The first family is based on a Naive Bayes model (Appendix E.1) with binary features, whereas the second one introduced in Appendix E.2 allows for continuous features with logistic distributions.

E.1 Naïve Bayes models revisited

We can also consider a Naïve Bayes models with binary features which can possibly be unavailable as in Poole et al. (Poole, Mehr, and Wang 2020). Suppose that we have a Naive Bayes model with independent availability mechanisms, i.e., the availability of feature i is only dependent on the label y and the corresponding feature value z_i and thus $\mathbf{p}(\mathbf{b}, \boldsymbol{a}, \boldsymbol{z}, y) = \left(\prod_{i=1}^{n} \mathbf{p}(b_i|y)\right) \left(\prod_{i=1}^{r} \mathbf{p}(z_i|y)\mathbf{p}(a_i|z_i, y)\right) \mathbf{p}(y)$. A graphical representation of this model can be found in Figure 7. In this case, we can express the odds ratio as:

$$\text{odds}(Y = 1|\mathbf{b}, \boldsymbol{z}_{\mathcal{I}}^*, \boldsymbol{A}_{\mathcal{I}} = \mathbf{1}) = \frac{\mathbf{p}(Y = 1, \mathbf{b}, \boldsymbol{z}_{\mathcal{I}}, \boldsymbol{A}_{\mathcal{I}} = \mathbf{1})}{\mathbf{p}(Y = 0, \mathbf{b}, \boldsymbol{z}_{\mathcal{I}}, \boldsymbol{A}_{\mathcal{I}} = \mathbf{1})} = \tag{64}$$

$$\left(\prod_{i=1}^{n} \frac{\mathbf{p}(b_i|Y = 1)}{\mathbf{p}(b_i|Y = 0)} \right) \left(\prod_{i \in \mathcal{I}} \frac{\mathbf{p}(z_i|Y{=}1)\mathbf{p}(A_i{=}1|z_i, Y{=}1)}{\mathbf{p}(z_i|Y{=}0)\mathbf{p}(A_i{=}1|z_i, Y{=}0)} \right) \frac{\mathbf{p}(Y = 1)}{\mathbf{p}(Y = 0)}. \tag{65}$$

As we furthermore suppose the features are binary, the odds are specified through the ratios $\frac{\mathbf{p}(b_i|Y=1)}{\mathbf{p}(b_i|Y=0)}$ for $b_i \in \{0, 1\}$ and $\frac{\mathbf{p}(z_i|Y=1)}{\mathbf{p}(z_i|Y=0)} \frac{\mathbf{p}(A_i=1|z_i,Y=1)}{\mathbf{p}(A_i=1|z_i,Y=0)}$ for $z_i \in \{0, 1\}$. This requires only $2r + 2n$ parameters to be specified in total.

E.2 A parametric familiy of distributions with logistic subset models

In this section, we describe a set of conditions that can be used to construct a family of densities that will have a logistic form when applying PUC. Formally, this means that for each $\mathcal{I} \subseteq [r]$ of optional features being present, there exists a $\mathbf{w} \in \mathbb{R}^n$, $\boldsymbol{\beta} \in \mathbb{R}^{|\mathcal{I}|}$, and $s \in \mathbb{R}$ that allow to represent the odds$(Y = 1|\mathbf{b}, \boldsymbol{z}_{\mathcal{I}}^*, \boldsymbol{A}_{\mathcal{I}} = \mathbf{1})$ in the form:

$$\frac{\mathbf{p}\left(Y = 1|\mathbf{B} = \mathbf{b}, \boldsymbol{Z}_{\mathcal{I}}^* = \boldsymbol{z}_{\mathcal{I}}^*, \boldsymbol{A}_{\mathcal{I}} = \mathbf{1}\right)}{\mathbf{p}\left(Y = 0|\mathbf{B} = \mathbf{b}, \boldsymbol{Z}_{\mathcal{I}}^* = \boldsymbol{z}_{\mathcal{I}}^*, \boldsymbol{A}_{\mathcal{I}} = \mathbf{1}\right)} = \exp\left[\mathbf{w}(\mathcal{I})^\top \mathbf{b} + \boldsymbol{\beta}(\mathcal{I})^\top \boldsymbol{z}_{\mathcal{I}}^* + s(\mathcal{I})\right].$$

This allows for complex dependencies (e.g., the base feature can influence availability and value of the optional features), while also allowing to compute the ground truth PUC model relatively easy. Formally, we suggest the following assertions and show

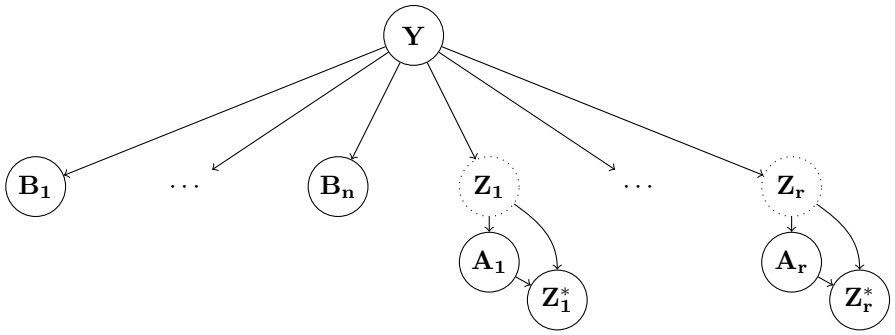


Figure 7: The Naive Bayes model with independent availability mechanisms. We observe the Label Y, the base features B_1 to B_n and the possibly unavailable features $Z_i^* = A_i \cdot Z_i$

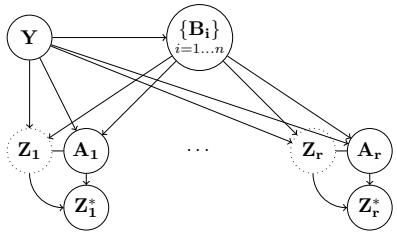


Figure 8: The relaxed graphical model with independent missingness mechanisms given the the Label Y and the base features B_1 to B_n. The observed, possibly missing features are $Z_i^* = M_i \cdot Z_i$.

that they will result in logistic models for each set of features present:

C1: Base model is logistic:	$\mathbf{p}(Y = 1\|\mathbf{b}) = \sigma(\mathbf{w}^\top \mathbf{b} + t)$
C2: Availability is cond. independent	$\forall \mathcal{I} \subseteq [r] : \mathbf{p}(\boldsymbol{A}_\mathcal{I} = \mathbf{1}\|\mathbf{b}, y) = \prod_{i \in \mathcal{I}} \mathbf{p}(A_i = 1\|\mathbf{b}, y)$
C3: Mut. independence when present:	$\forall \mathcal{I} \subseteq [r] : \mathbf{p}(\boldsymbol{z}_\mathcal{I}\|\mathbf{b}, y, \boldsymbol{A}_\mathcal{I} = \mathbf{1}) = \prod_{i \in \mathcal{I}} \mathbf{p}(z_i\|\mathbf{b}, y, A_i{=}1)$
C4: Availability is sigmoidal:	$\mathbf{p}(A_i = 1\|\mathbf{b}, Y = 1) = N(\mathbf{b})\sigma\left(\mathbf{u}_i^\top \mathbf{b} + \lambda_i\right)$
	$\mathbf{p}(A_i = 1\|\mathbf{b}, Y = 0) = N(\mathbf{b}) - \mathbf{p}(A_i = 1\|\mathbf{b}, Y = 1)$
C5: Base-dependent Normal distributions:	$\mathbf{p}(z_i\|\mathbf{b}, y, A_i = 1) \sim \mathcal{N}\left(\mathbf{v}_i^\top \mathbf{b} + \tau_i(y), \eta^2\right).$

Intuitively, after ensuring that the base feature model has a logistic form (C1), the next two assumptions follow directly from the graphical dependency model (C2, C3), see Figure 8. C4 suggests the availability should sigmoidally depend on the base features with a different offset for each class. The last condition (C5) allows the z_i to depend on the base features $\mathbf{b}$ with the same $\mathbf{v}_i$ for both classes y. However, a different offset by the coefficient τ_i can be added for each class. The entire distribution can be specified through the parameters $\mathbf{w}$, t, and $\boldsymbol{u}_i$, $\boldsymbol{v}_i$, λ_i, $\tau_i(0)$, $\tau_i(1)$ and s_i for each missing feature in $i = 1 \dots r$.

In this special case, we can show that each of the models required will have the form of logistic regression again. We start by determining some density ratios that will arise later:

$$\frac{\mathbf{p}(A_i = 1|\mathbf{b}, Y{=}1)}{\mathbf{p}(A_i = 1|\mathbf{b}, Y{=}0)} = \frac{N(\mathbf{b})\sigma\left(\mathbf{u}_i^\top \mathbf{b} + \lambda_i\right)}{N(\mathbf{b})\left(1 - \sigma\left(\mathbf{u}_i^\top \mathbf{b} + \lambda_i\right)\right)} = \exp\left(\mathbf{u}_i^\top \mathbf{b} + \lambda_i\right), \tag{66}$$

where the identity $\frac{\sigma(x)}{1-\sigma(x)} = \exp(x)$ was used. Furthermore,

$$\frac{\mathbf{p}(z_i|\mathbf{b}, Y=1, A_i=1)}{\mathbf{p}(z_i|\mathbf{b}, Y=0, A_i=1)} = \frac{\exp\left(-\frac{\left(z_i-\mathbf{v}_i^\top\mathbf{b}-\tau_{i1}\right)^2}{2\eta^2}\right)}{\exp\left(-\frac{\left(z_i-\mathbf{v}_i^\top\mathbf{b}-\tau_{i0}\right)^2}{2\eta^2}\right)} \tag{67}$$

$$= \exp\left[-\frac{\left(z_i-\mathbf{v}_i^\top\mathbf{b}-\tau_{i1}\right)^2-\left(z_i-\mathbf{v}_i^\top\mathbf{b}-\tau_{i0}\right)^2}{2\eta^2}\right] \tag{68}$$

$$= \exp\left[\frac{\left(z_i-\mathbf{v}_i^\top\mathbf{b}\right)^2-2\left(z_i-\mathbf{v}_i^\top\mathbf{b}\right)\tau_{i1}+\tau_{i1}^2-\left(z_i-\mathbf{v}_i^\top\mathbf{b}\right)^2+2\left(z_i-\mathbf{v}_i^\top\mathbf{b}\right)\tau_{i0}-\tau_{i0}^2}{-2\eta^2}\right] \tag{69}$$

$$= \exp\left[\frac{2\left(\tau_{i1}-\tau_{i0}\right)\left(z_i-\mathbf{v}_i^\top\mathbf{b}\right)+\tau_{i0}^2-\tau_{i1}^2}{2\eta^2}\right] \tag{70}$$

$$= \exp\left[\underbrace{\eta^{-2}\left(\tau_{i1}-\tau_{i0}\right)}_{\beta_i} z_i - \underbrace{\eta^{-2}\left(\tau_{i1}-\tau_{i0}\right)\mathbf{v}_i^\top}_{\boldsymbol{\gamma}_i^\top}\mathbf{b} + \underbrace{\frac{1}{2}\eta^{-2}\left(\tau_{i0}^2-\tau_{i1}^2\right)}_{\theta_i}\right]. \tag{71}$$

Let $\mathcal{I} \subseteq [r]$ be the set index set of of present features once again. We can insert the previous results and obtain:

$$\text{odds}(Y=1|\mathbf{b}, \boldsymbol{Z}_\mathcal{I}, \boldsymbol{A}_\mathcal{I}{=}\mathbf{1}) = \frac{\mathbf{p}(Y=1, \mathbf{b}, \boldsymbol{Z}_\mathcal{I}, \boldsymbol{A}_\mathcal{I}{=}\mathbf{1})}{\mathbf{p}(Y=0, \mathbf{b}, \boldsymbol{Z}_\mathcal{I}, \boldsymbol{A}_\mathcal{I}{=}\mathbf{1})} \tag{72}$$

$$= \frac{\mathbf{p}(Y=1|\mathbf{b})}{\mathbf{p}(Y=0|\mathbf{b})}\frac{\mathbf{p}(\boldsymbol{A}_\mathcal{I}=\mathbf{1}|\mathbf{b}, Y{=}1)}{\mathbf{p}(\boldsymbol{A}_\mathcal{I}=\mathbf{1}|\mathbf{b}, Y{=}0)}\frac{\mathbf{p}(\boldsymbol{Z}_\mathcal{I}|\mathbf{b}, Y{=}1, \boldsymbol{A}_\mathcal{I}{=}\mathbf{1})}{\mathbf{p}(\boldsymbol{Z}_\mathcal{I}|\mathbf{b}, Y{=}0, \boldsymbol{A}_\mathcal{I}{=}\mathbf{1})} \tag{73}$$

$$= \frac{\mathbf{p}(Y=1|\mathbf{b})}{\mathbf{p}(Y=0|\mathbf{b})}\left(\prod_{i\in\mathcal{I}}\frac{\mathbf{p}(z_i|\mathbf{b}, Y{=}1, A_i{=}1)\mathbf{p}(A_i{=}1|\mathbf{b}, Y{=}1)}{\mathbf{p}(z_i|\mathbf{b}, Y{=}0, A_i{=}1)\mathbf{p}(A_i{=}1|\mathbf{b}, Y{=}0)}\right) \tag{74}$$

$$= \exp(\mathbf{w}^\top\mathbf{b} + t + \sum_{i\in\mathcal{I}}\underbrace{(\mathbf{u}_i-\boldsymbol{\gamma}_i)^\top}_{\boldsymbol{\omega}_i^\top}\mathbf{b} + \beta_i z_i + \underbrace{\lambda_i+\theta_i}_{s_i}). \tag{75}$$

As this derivation shows, each subset model will again be of the logistic form. On a sidenote, the probability of a true model with no fairness constraints can be estimated as:

$$\text{odds}(Y=1|\mathbf{b}, \boldsymbol{Z}, \boldsymbol{A}) = \frac{\mathbf{p}(Y=1, \mathbf{b}, \boldsymbol{Z}_\mathcal{I}, \boldsymbol{A}_\mathcal{I}{=}\mathbf{1}), \boldsymbol{A}_{\overline{\mathcal{I}}}{=}\mathbf{0})}{\mathbf{p}(Y=0, \mathbf{b}, \boldsymbol{Z}_\mathcal{I}, \boldsymbol{A}_\mathcal{I}{=}\mathbf{1}, \boldsymbol{A}_{\overline{\mathcal{I}}}{=}\mathbf{0})} \tag{76}$$

$$= \frac{\mathbf{p}(Y=1|\mathbf{b})}{\mathbf{p}(Y=0|\mathbf{b})}\frac{\mathbf{p}(\boldsymbol{A}_\mathcal{I}=\mathbf{1}|\mathbf{b}, Y{=}1)}{\mathbf{p}(\boldsymbol{A}_\mathcal{I}=\mathbf{1}|\mathbf{b}, Y{=}0)}\frac{\mathbf{p}(\boldsymbol{Z}_\mathcal{I}|\mathbf{b}, Y{=}1, \boldsymbol{A}_\mathcal{I}{=}\mathbf{1})}{\mathbf{p}(\boldsymbol{Z}_\mathcal{I}|\mathbf{b}, Y{=}0, \boldsymbol{A}_\mathcal{I}{=}\mathbf{1})}\frac{\mathbf{p}(\boldsymbol{A}_{\overline{\mathcal{I}}}{=}\mathbf{0}|\mathbf{b}, Y{=}1)}{\mathbf{p}(\boldsymbol{A}_{\overline{\mathcal{I}}}{=}\mathbf{0}|\mathbf{b}, Y{=}0)}. \tag{77}$$

F Additional Experimental Results and Details

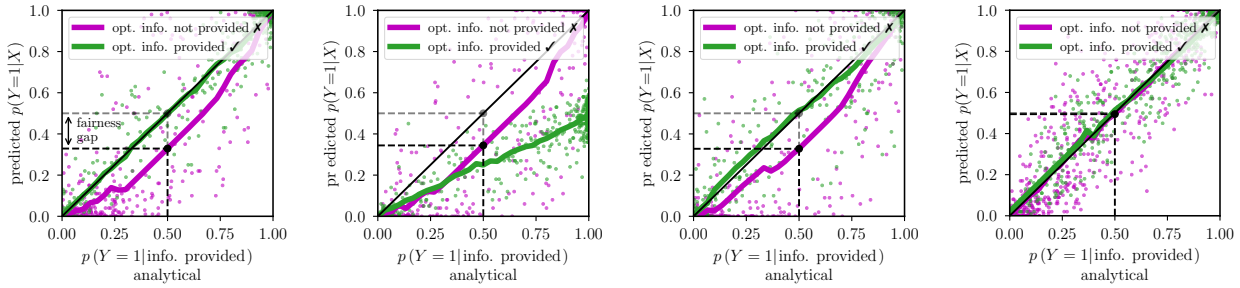
F.1 Comparing PUC to existing fairness notions

In this section, we visually show the effect of not compensating for information contained in the decision to share data. We refer Figure 9, were we compute probabilities for positive outcomes for a standard model ("fairness through unawareness") and other fairness-constrained models. The figure shows that all models apart from PUC, are not calibrated *with respect to the data explicitly provided*. The data set used to create the Figure was sampled according to the logistic family described in Appendix E.2. The feature value distributions follow a logistic form. There were two base features and one optional feature. The availability and the values of this feature was dependent on the label and the value of the base features as described in the mentioned section. Specifically, the following parameters were used to instantiate the logistic family described in Appendix E.2:

base features	n=2, $\mathbf{b} \sim \mathcal{N}(\mathbf{0}, 5\boldsymbol{I})$, $\boldsymbol{w}{=}(-1.5, 1.0)^\top$, $t{=}0$
opt. feature 1	$\boldsymbol{u}_1 = (0.8, 0.4)^\top$, $\boldsymbol{v}_1 = (0, 1)^\top$, $\lambda_1{=}0.7$, $\tau_1(0){=}-0.25$, $\tau_1(1){=}0.25$

The models used in these experiments were `sklearn` RandomForests with default parameters. To incorporate the Fairness constraints of Statistical Parity and Equalized Odds, we leverage the `fairlearn`[4] library (Bird et al. 2020), which implements the ExponentiatedGradient algorithm by Agarwal et al. (2018). Although this algorithms only returns an approximate solution, we verified that the corresponding fairness gaps for Statistical Parity and Equalized Odds were substantially improved.

[4] https://fairlearn.org/



(a) Fairness through unawareness

(b) Statistical parity

(c) Equalized odds

(d) Protected User Consent (ours)

Figure 9: **Standard models treat users who do not share optional information not according to the data they provided**. In this work, users can provide information on optional features and only the provided information should be used in the decision making. We show calibration curves for a model without fairness considerations (a) and with common fairness constraints enforcing statistical parity (b) and equalized odds (c) with respect to a model that uses only the explicity provided information (base feature model in case of no optional information, full feature model in case of optional information). The first three models can penalize users not sharing the optional information (fairness gap in left panel), whereas a model trained with Protected User Consent through `PUCIDA` (d) exhibits no systematic bias. Models are probabilistic Random Forests trained on a synthetic data set (see Appendix F.1).

F.2 Data Sets and Preprocessing

The diabetes data set[5] was collected by the National Institute of Diabetes and Digestive and Kidney Diseases. It contains diagnostic measurements of female patients that are at least 21 years old. The target variable "Outcome" describes whether or not a person has diabetes.

The COMPAS data set[6] was originally collected by ProPublica and contains features describing criminal defendants in Broward County, Florida. It also contains their respective recividism score provided by the COMPAS algorithm and whether or not they reoffended within the following two years. For our analysis, we only kept features relevant for the prediction of recividism within the next two years and dropped irrelevant features such as name or date. Furthermore, we turned the categorical features race, sex and charge degree into numerical features by encoding the categories with integers.

The UCI adult data set[7] is one of the most poplar tabular data sets and has appeared in over 300 publications (Ding et al. 2021). The goal is to predict whether an individuals yearly income is above 50k$ (worth of 1994).

The California Housing data set[8] contains infomration and average prices of properties in certain areas in the state of California, USA. The regression target is to predict the value of a property. Because the income values range over several values of magnitudes, we apply log normalization to the label.

The ACSIncome data set ("income") is derived from US census data in the work of Ding et al. (2021). Code to download it is available online[9]. As for Adult, the goal is to predict an inviduals yearly income. It features are similar from the one used in the Adult data set, however the exact incomes of each person are reported, and the data set can therefore be used in a regression setting. Because the income values range over several values of magnitudes, we apply log normalization to the labels.

The Health Insurance ("insurance") dataset[10] contains insurance data from individuals. It is a regression dataset, where inferences about the number of hours worked are to be made (whrswk, hours worked per week). We use the experience (years of potential work experience) as optional feature in the task.

We furthermore use two datasets with natural missing features. The UCI horse colic dataset[11] ("colic") is a database of leasion surgeries on horses and contains a number of health attributes such as temperatures, pulse, respiratory rate and others. The target feature describes the outcome of the pathology. We use the feature abdominocentesis appearance as optional feature, which describes the appearance of fluid that is obtained from the abdominal cavity. This information is not available for each horse in the database and thus comes with natural missingness.

[5] https://www.kaggle.com/s/mathchi/diabetes-data-set
[6] https://www.kaggle.com/s/danofer/compass
[7] https://archive.ics.uci.edu/ml/datasets/Adult/
[8] https://www.kaggle.com/datasets/camnugent/california-housing-prices
[9] https://github.com/zykls/folktables/tree/main/folktables
[10] https://api.openml.org/d/44993
[11] https://archive.ics.uci.edu/ml/datasets/Horse+Colic

The water treatment dataset[12] contains features describing the operational state of a water treatment plant, which is to be classified as either postive or negative. We use the feature RD-DBO-P ("oxygen demand") as optional feature, which describes the Biological demand of oxygen in primary settler and comes with missing values.

Across all data sets, multi-value categorical features were one-hot encoded. We provide an overview of the characteristics of the different data sets in Table 4.

Data Sets	Label	Num. features	Num. samples (N)
diabetes	Outcome	8	768
compas	two_year_recid	9	7192
adult	ZFYA	5	21791
california housing	med_house_val	9	20640
income	income	10	19567
insurance	whrswk	11	22272
water	binaryClass	36	527
colic	pathology_cp_data	26	368

Table 4: Characteristics of the data sets studied in this work.

Stochastic Availability: We make values available by the following scheme over continuous features $z_i \in \mathcal{X}^z$:

$$\mathbf{p}(A_i = 0|z_i) = \text{sigmoid}\left(\lambda_i(z_i - \bar{z_i})\right) = \frac{1}{1 + \exp(-\lambda_i(z_i - \bar{z_i}))}, \tag{78}$$

where we denote the empirical feature mean by $\bar{z_i}$ and $\lambda_i \in \mathbb{R}$ denotes a parameter that specifies how quickly the probability of unavailability ($A_1 = 0$) increases with higher feature values (for positive values of λ_i). For negative values of λ_i, values of the feature that are lower than the mean are more likely to be unavailable. We chose λ_i such that values which were negatively influcing the prediction were more likely to be missing. We show the probabilities curves used of the feature distribution with the corresponding values of λ_i in Figure 10.

Adversarial Availability: We also experiment with adversarial sharing decision as discussed in the paper. To this end, we first train a full feature model (with no missing data) and a base feature model. We then modify the dataset and drop all optional feature values where the full feature model would lead to a lower regression score or chance of the positive outcome and retrain the corresponding classifiers on this dataset. As a final check, we replace all `PUCIDA` prediction that are higher than the base model's predictions by the base model's prediction to arrive at the aformentioned bonus system.

Models. We use standard models from the `sklearn` library (Pedregosa et al. 2011). Across all experiments, we used these models with the following parameters:

model	parameters
RandomForestClassifier / RandomForestRegressor	default parameters
ExtraTreesClassifier / ExtraTressRegressor	min_samples_split=10
GradientBoostingClassifier / GradientBoostingRegressor	min_samples_split=10
DecisionTreeClassifier / DecisionTreeRegressor	min_samples_split=10
MLPClassifier / MLPRegressor	hidden_layer_sizes= [30,40], max_iter=500

If not stated otherwise, we report averages over 5 runs with a random 80/20 test split. Code to reproduce experiments is provided in the supplementary material and will be publicly released in case of acceptance.

F.3 Experiment 1: Protecting User consent on real-world data sets

This section provides additional results for Experiment 1 (Table 2) showing that Availability Inference Restriction is violated.

Ablation studies To test the robustness of the results shown in Table 2, we performed three ablation studies. For all alternative parameters tested, the results are not qualitatively different from the original ones.

Imputation Values. In Table 2, imputed data points are replaced by zeros. Alternatively, one could also use the mean or the median of the voluntary feature as imputation values, which does not lead to substantial changes as expemplarily shown for classification datasets in Table 8. We conclude that it is hard to stop Penalization through simple imputation. Note: Our current implementation of the data augmentation strategy is implicitly converts missing values to zero for all missing values, so the results are the same as in the main paper for `PUCIDA`.

Random forest hyperparameters. In Table 2, the default parameters of random forest are used (min_samples_split=2, n_estimators=100, max_depth=None). The ablation studies with different hyperparameters are shown in Tables 9 –11.

[12] https://api.openml.org/d/940

Different models. As an alternative to random forest, we test gradient boosting models (see Table 12) and ExtRaTrees by Geurts, Ernst, and Wehenkel (2006) (see Table 13). While the extent of change differs to some extent, for every model and hyperparameter configuration, the full feature model uses the information in the sharing decision and the individuals that do not have feature values are rated worse. Occasionally, PUC models can be non-significantly better than base models, but this is due to statistical errors (as indicated through the standard deviations).

task	data	opt. feature	Base feature model	PUC	Full feature model
C	diab.	Glucose	24.75% ±1.78	20.22% ±2.01	19.90% ±2.35
C	compas	#priors	42.02% ±0.20	36.89% ±0.42	36.81% ±0.51
C	adult	edu-num	18.75% ±0.08	17.95% ±0.07	17.85% ±0.12
R	income	WKHP	63.56 ±1.08	54.66 ±0.83	56.22 ±1.13
R	calif.	m_income	12.76 ±0.11	8.51 ±0.16	8.60 ±0.17
R	insurance	experience	245.00 ±0.47	223.53 ±0.45	230.95 ±0.34

Table 5: **Performance for sharers is maintained with PUCIDA.** For the setup corresponding to Table 2, we montior performance measures for the subgroup of sharers. We report missclassification rate (1-Acc) for classification task and MSE (×100) for regression tasks. We show that the performance in this group is close to the unconstrained model, an indication that their optional information is used.

task	data	opt. feature	base model	imputed	PUCIDA
C	diab.	Glucose	29.30% ±0.62	25.83% ±0.41	26.95% ±0.82
C	compas	#priors	42.89% ±0.10	38.51% ±0.59	39.55% ±0.37
C	adult	edu-num	16.05% ±0.03	15.38% ±0.11	15.62% ±0.09
R	income	WKHP	85.09 ±0.12	79.76 ±0.47	81.52 ±0.28
R	calif.	m_income	13.98 ±0.06	11.90 ±0.15	12.55 ±0.05
R	insurance	experience	262.43 ±0.21	249.31 ±0.27	254.84 ±0.13

Table 6: Costs for PUC with non-adversarial sharing decisions. Otherwise the setup is equivalent to Table 4a.

task	data	opt. feature	base model	imputed	PUCIDA
Non-Adversarial Sharing					
C	diab.	Glucose	23.96 ±0.11	20.35 ±0.43	21.85 ±0.69
C	compas	#priors	40.85 ±0.06	34.88 ±0.54	36.85 ±0.21
C	adult	edu-num	11.90 ±0.03	11.18 ±0.07	11.10 ±0.05
C	water	oxygen. dem.	3.97 ±0.42	3.47 ±0.25	3.19 ±0.25
C	colic	abdom. app.	12.07 ±0.31	9.08 ±0.30	9.13 ±0.28
Adversarial Sharing					
C	diab.	Glucose	23.96 ±0.11	18.25 ±0.85	19.91 ±0.46
C	compas	#priors	40.85 ±0.06	32.03 ±0.35	34.86 ±0.26
C	adult	edu-num	11.90 ±0.03	10.26 ±0.05	10.57 ±0.03

Table 7: Costs for PUC when using 100×(1-ROCScores) as cost functions for the classification models instead of accuray. Setup as in Table 4a.

F.4 Experiment 2: Validating Non-Degradation and costs of fairness with respect to optional information

This section contains additional details on the experiments leading up to Table 3.

Single optional feature. We first investigate the performance of the models in the setup corresponding to Table 2, i.e., with only a single optional feature. In Table 6 we show the cost setup when non-stategic sharing decisions are taking, which leads to qualitatively equivalent results as in the main paper. Table 7 shows the results for the classification models when using the area under the 1-ROC-curve as a cost function.

C: (1-Acc)×100, R: MSE×100

task	data	opt. feature	base model	Full feature model	PUCIDA
C	diab.	Glucose	35.00% ±1.57	32.24% ±1.67	34.67% ±1.07
C	compas	#priors	44.55% ±0.45	41.84% ±0.96	44.56% ±0.51
C	adult	edu-num	13.40% ±0.13	13.02% ±0.27	13.37% ±0.20
R	income	WKHP	109.09 ±1.05	107.80 ±1.09	110.12 ±1.27
R	calif.	m_income	17.83 ±0.25	16.41 ±0.34	19.04 ±0.18
R	insurance	experience	282.99 ±0.78	278.27 ±0.46	284.88 ±0.93

(a) Corresponding to Table 1, costs, mean imputation.

task	data	optional	Base feature model	Full feature model pred.	Full feature model change	PUCIDA pred.	PUCIDA change
C	diab.	Glucose	64.17%	51.17%	-13.00% ±3.51	63.12%	-1.05% ±2.92
C	compas	#priors	51.39%	33.77%	-17.63% ±0.84	51.18%	-0.21% ±0.14
C	adult	edu-num	13.77%	11.35%	-2.42% ±0.16	13.77%	0.01% ±0.03
R	income	WKHP	100.0%	81.5%	-18.5% ±0.48	101.4%	1.4% ±0.18
R	insurance	experience	100.0%	94.9%	-5.1% ±0.10	100.1%	0.1% ±0.05
R	calif.	m_income	100.0%	95.3%	-4.7% ±0.28	104.2%	4.2% ±0.42

(b) Corresponding to Table 2, absolute predictions, mean imputation.

C: (1-Acc)×100, R: MSE×100

task	data	opt. feature	base model	Full feature model	PUCIDA
C	diab.	Glucose	35.00% ±1.57	32.22% ±1.18	34.67% ±1.07
C	compas	#priors	44.55% ±0.45	41.70% ±0.97	44.56% ±0.51
C	adult	edu-num	13.40% ±0.13	12.99% ±0.22	13.37% ±0.20
R	income	WKHP	109.09 ±1.05	107.70 ±1.18	110.12 ±1.27
R	calif.	m_income	17.83 ±0.25	16.28 ±0.31	19.04 ±0.18
R	insurance	experience	282.99 ±0.78	278.29 ±0.64	284.88 ±0.93

(c) Corresponding to Table 1, costs, median imputation.

task	data	optional	Base feature model	Full feature model pred.	Full feature model change	PUCIDA pred.	PUCIDA change
C	diab.	Glucose	64.17%	50.83%	-13.34% ±4.10	63.12%	-1.05% ±2.92
C	compas	#priors	51.39%	33.53%	-17.86% ±0.92	51.18%	-0.21% ±0.14
C	adult	edu-num	13.77%	11.29%	-2.48% ±0.17	13.77%	0.01% ±0.03
R	income	WKHP	100.0%	81.9%	-18.1% ±0.60	101.4%	1.4% ±0.18
R	insurance	experience	100.0%	94.9%	-5.1% ±0.09	100.1%	0.1% ±0.05
R	calif.	m_income	100.0%	94.9%	-5.1% ±0.52	104.2%	4.2% ±0.42

(d) Corresponding to Table 2, absolute predictions, median imputation.

Table 8: Same setup as Table 2 using *mean imputation* (upper line) and *median imputation* (lower line). The differencees between the two imputation techniques are minimal.

C: (1-Acc)×100, R: MSE×100

task	data	opt. feature	base model	Full feature model	PUCIDA
C	diab.	Glucose	34.38% ±1.60	33.52% ±1.30	34.25% ±1.01
C	compas	#priors	42.53% ±0.40	39.21% ±0.60	42.92% ±0.43
C	adult	edu-num	12.04% ±0.11	11.75% ±0.19	11.99% ±0.19
R	income	WKHP	104.62 ±0.56	103.15 ±0.80	105.85 ±0.60
R	calif.	m_income	17.84 ±0.23	16.15 ±0.50	19.11 ±0.46
R	insurance	experience	260.05 ±0.29	256.41 ±0.19	262.31 ±0.54

(a) Corresponding to Table 1 (costs).

task	data	optional	Base feature model	Full feature model pred.	Full feature model change	PUCIDA pred.	PUCIDA change
C	diab.	Glucose	64.86%	52.13%	-12.73% ±2.15	66.11%	1.25% ±1.89
C	compas	#priors	51.89%	29.92%	-21.97% ±0.97	52.11%	0.22% ±0.57
C	adult	edu-num	12.08%	9.49%	-2.59% ±0.06	12.18%	0.10% ±0.09
R	income	WKHP	100.0%	81.4%	-18.6% ±0.36	101.3%	1.3% ±0.31
R	insurance	experience	100.0%	94.8%	-5.2% ±0.06	100.2%	0.2% ±0.07
R	calif.	m_income	100.0%	94.6%	-5.4% ±1.00	104.1%	4.1% ±0.75

(b) Corresponding to Table 2 (absolute predictions).

Table 9: **Availability Inference Restriction is violated by full feature models**. Same setup as Table 2 using a *Random Forest model* with $min\_samples\_split = 10$.

Having verified these results for a single feature, we now continue with the more challenging setup of multiple optionality.

Introducing multiple optionality. For the real data experiment, we apply the following preprocessing steps to induce stochastic availability:

- We identify the most discriminative numerical features by dropping each feature from the data set and reporting the decline in predictive performance of a model trained without the feature with respect to a model trained on all features. We rank the features starting with the one resulting in the highest performance loss.
- We select the r most discriminative features, such that on average, each subset of missing pattern has at least 150 samples out of the initial data set size of N to be fitted with, i.e.,

$$r = \inf \left\{ r' \in \mathbb{N} : \frac{N}{2^r} > 150 \right\}.$$

- We do not consider numerical features were the relation to the label is not clear (i.e., is there a positive or negative correlation). The optional features are listed in Table 17.
- We independently induce stochastic availability into each feature using the sigmoidal strategy. We use a $\lambda_i = \pm \frac{1}{\sqrt{\mathrm{Var}\left[f_i\right]}}$, which is effectively equivalent to applying a sigmoid over normalized feature values. The signs are determined by the context such that negative indicators are more likely to be not provided and are also reported in the Table 17.

We show the corresponding results of Table 3b using 1-ROC as cost function in Table 14. We provide ablations with the two other models in Table 15 and Table 16

F.5 Experiment 3: Convergence to analytical PUC

In this section, we provide additional details regarding the experiment where we study the gaps to analytical Protected User Consent on our simulated data sets.

C: (1-Acc)×100, R: MSE×100

task	data	opt. feature	base model	Full feature model	PUCIDA
C	diab.	Glucose	36.49% ±0.45	34.27% ±0.99	37.87% ±0.95
C	compas	#priors	43.99% ±0.70	37.34% ±0.45	50.43% ±0.96
C	adult	edu-num	13.29% ±0.23	13.96% ±0.13	13.18% ±0.10
R	income	WKHP	117.24 ±0.73	115.76 ±0.79	123.08 ±1.01
R	calif.	m_income	26.08 ±0.10	20.29 ±0.21	25.51 ±0.12
R	insurance	experience	251.99 ±0.13	251.21 ±0.12	258.73 ±0.16

(a) Corresponding to Table 1 (costs).

				Full feature model		PUCIDA	
task	data	optional	Base feature model	pred.	change	pred.	change
C	diab.	Glucose	73.66%	58.45%	-15.22% ±4.23	79.07%	5.41% ±2.30
C	compas	#priors	65.82%	27.97%	-37.85% ±4.26	79.98%	14.16% ±0.75
C	adult	edu-num	2.34%	1.54%	-0.80% ±0.41	2.54%	0.20% ±0.27
R	income	WKHP	100.0%	75.5%	-24.5% ±0.78	108.6%	8.6% ±0.38
R	insurance	experience	100.0%	94.6%	-5.4% ±0.09	103.5%	3.5% ±0.04
R	calif.	m_income	100.0%	83.9%	-16.1% ±0.30	99.4%	-0.6% ±0.21

(b) Corresponding to Table 2 (absolute predictions).

Table 10: **Availability Inference Restriction is violated by full feature models**. Same setup as Table 2 using a *Random Forest model* with $max\_depth = 4$.

C: (1-Acc)×100, R: MSE×100

task	data	opt. feature	base model	Full feature model	PUCIDA
C	diab.	Glucose	34.07% ±0.87	33.61% ±0.43	33.58% ±0.92
C	compas	#priors	44.64% ±0.20	41.40% ±0.61	44.78% ±0.42
C	adult	edu-num	13.31% ±0.06	12.83% ±0.14	13.31% ±0.05
R	income	WKHP	107.98 ±0.37	106.71 ±0.32	109.29 ±0.54
R	calif.	m_income	17.70 ±0.10	16.11 ±0.31	18.93 ±0.20
R	insurance	experience	281.96 ±0.12	277.03 ±0.51	283.79 ±0.19

(a) Corresponding to Table 1 (costs).

				Full feature model		PUCIDA	
task	data	optional	Base feature model	pred.	change	pred.	change
C	diab.	Glucose	63.30%	51.63%	-11.67% ±1.20	63.21%	-0.09% ±1.14
C	compas	#priors	51.31%	32.62%	-18.69% ±1.27	51.36%	0.05% ±0.43
C	adult	edu-num	13.93%	11.47%	-2.46% ±0.24	13.91%	-0.02% ±0.09
R	income	WKHP	100.0%	81.4%	-18.6% ±0.46	101.3%	1.3% ±0.12
R	insurance	experience	100.0%	94.8%	-5.2% ±0.08	100.1%	0.1% ±0.02
R	calif.	m_income	100.0%	94.1%	-5.9% ±0.86	103.8%	3.8% ±0.48

(b) Corresponding to Table 2 (absolute predictions).

Table 11: **Availability Inference Restriction is violated by full feature models**. Same setup as Table 2 using a *Random Forest model* with $n\_estimators = 500$.

Synthetic data sets We initially conduct a synthetic data experiment to verify our theory.

First Synthetic Data Set. For the data set used in Figure 5 and Figure 4, we create binary features according to the Naive Bayes scheme described in Appendix E.1. The probabilities of each feature pointing to the corresponding class were drawn randomly, we made the three features with the highest discriminatory power optional. We then drew probabilities of the feature values being missing also at random. FOr this example, the missingness did not depend on the feature value, but only on the label. The resulting parameters are given in Table 18 for the sake of completeness.

Second Synthetic Data Set. We create a more complicated data set with continuous features as describeb by the familiy in Appendix E.2. We create two normally distributed base features and three optional features to test interesting dependency combinations by using the parameters in Table 19. This distribution includes cases where:

- the availability distribution depends on the base features ($\boldsymbol{u} \neq \mathbf{0}$, feature 1)
- the availability distribution depends on the class value ($\lambda \neq 0$, feature 1, feature 2)
- the feature value depends on the base features and the class value ($\boldsymbol{v} \neq \mathbf{0}$, $\tau(0) \neq \tau(1)$, feature 1, feature 2, feature 3)

We draw increasing numbers of samples from the known distribution and fit the corresponding estimators. The test set on which the PUC-Gap^2 is estimated on 5000 independently drawn samples. For Figure 4, we use 50000 samples to train each model.

Additional Results We also conduct the approximation experiment on the more complicated continuous data set. The results can be found in Figure 11. Note that on this continuous data sets the models will not perfectly converge. However, we show that the PUC-Gap is in range of the irreducible, random estimation error, by computing the average squared estimation error without PUC on the unfair data set and adding the ranges of this error to the plot.

C: (1-Acc)×100, R: MSE×100

task	data	opt. feature	base model	Full feature model	PUCIDA
C	diab.	Glucose	34.07% ±0.87	33.61% ±0.43	33.58% ±0.92
C	compas	#priors	44.64% ±0.20	41.40% ±0.61	44.78% ±0.42
C	adult	edu-num	13.31% ±0.06	12.83% ±0.14	13.31% ±0.05
R	income	WKHP	107.98 ±0.37	106.71 ±0.32	109.29 ±0.54
R	calif.	m_income	17.70 ±0.10	16.11 ±0.31	18.93 ±0.20
R	insurance	experience	281.96 ±0.12	277.03 ±0.51	283.79 ±0.19

(a) Corresponding to Table 1 (costs).

				Full feature model		PUCIDA	
task	data	optional	Base feature model	pred.	change	pred.	change
C	diab.	Glucose	62.80%	52.96%	-9.84% ±2.46	63.75%	0.95% ±1.35
C	compas	#priors	62.82%	21.29%	-41.54% ±1.69	64.14%	1.32% ±0.22
C	adult	edu-num	9.73%	5.98%	-3.76% ±0.10	9.31%	-0.43% ±0.07
R	income	WKHP	100.0%	79.6%	-20.4% ±0.13	100.0%	0.0% ±0.09
R	insurance	experience	100.0%	94.4%	-5.6% ±0.05	101.2%	1.2% ±0.02
R	calif.	m_income	100.0%	91.1%	-8.9% ±0.36	102.9%	2.9% ±0.75

(b) Corresponding to Table 2 (absolute predictions).

Table 12: **Availability Inference Restriction is violated by full feature models**. Same setup as Table 2 using a *Gradient Boosting model*.

C: (1-Acc)×100, R: MSE×100

task	data	opt. feature	base model	Full feature model	PUCIDA
C	diab.	Glucose	35.27% ±0.89	33.20% ±1.46	35.68% ±1.54
C	compas	#priors	42.22% ±0.41	36.69% ±0.29	42.66% ±0.56
C	adult	edu-num	11.25% ±0.09	11.37% ±0.07	11.27% ±0.12
R	income	WKHP	104.79 ±0.72	100.90 ±0.75	106.02 ±0.66
R	calif.	m_income	16.99 ±0.10	15.50 ±0.22	17.74 ±0.34
R	insurance	experience	243.33 ±0.12	242.30 ±0.12	246.22 ±0.12

(a) Corresponding to Table 1 (costs).

				Full feature model		PUCIDA	
task	data	optional	Base feature model	pred.	change	pred.	change
C	diab.	Glucose	71.27%	48.73%	-22.54% ±5.84	68.78%	-2.49% ±1.27
C	compas	#priors	53.53%	29.53%	-24.00% ±1.16	53.39%	-0.14% ±0.18
C	adult	edu-num	12.27%	9.58%	-2.69% ±0.27	12.33%	0.06% ±0.06
R	income	WKHP	100.0%	80.1%	-19.9% ±0.60	101.0%	1.0% ±0.08
R	insurance	experience	100.0%	94.6%	-5.4% ±0.07	100.1%	0.1% ±0.02
R	calif.	m_income	100.0%	90.9%	-9.1% ±0.55	104.0%	4.0% ±0.28

(b) Corresponding to Table 2 (absolute predictions).

Table 13: **Availability Inference Restriction is violated by full feature models**. Same setup as Table 2 using a *Extra Trees model*

		Fair models				Full feature model
task	data (# opt.)	Base feature model	PUCIDA (f)	PUCIDA (e)	(×)	zero-imputed
C	diab. (2)	73.14 ±2.59	77.13 ±3.67	**77.22** ±3.92	2.3	78.42 ±2.61
C	compas (5)	61.32 ±0.92	61.90 ±0.96	**62.32** ±0.86	7.6	62.69 ±1.58
C	adult (5)	84.90 ±0.46	90.39 ±0.35	**89.68** ±0.33	7.4	90.57 ±0.21

Table 14: **PUC-compliant models improve predictive performance**. Same setup as in to Table 3b, but in this case we use *ROC-AUC as the performance metric*. A higher ROC-AUC is preferable.

		Fair models				Full feature model
task	data (# opt.)	Base feature model	PUCIDA (f)	PUCIDA (e)	(×)	zero-imputed
C	diab. (2)	29.87 ±2.25	28.70 ±2.37	**28.18** ±2.74	2.3	27.14 ±2.67
C	compas (5)	40.78 ±0.63	**37.71** ±0.63	37.79 ±0.90	7.6	35.62 ±0.60
C	adult (5)	17.84 ±0.41	13.43 ±0.49	**13.43** ±0.48	7.4	13.31 ±0.45
R	calif. (4)	11.01 ±1.89	9.45 ±0.19	**9.32** ±0.34	5.1	9.04 ±0.17
R	income (3)	46.31 ±2.02	45.00 ±2.42	**44.28** ±1.86	3.4	41.89 ±1.52
R	insurance (3)	230.00 ±0.72	212.30 ±1.73	**211.27** ±2.13	3.2	210.72 ±1.55

Table 15: **PUC-compliant models improve predictive performance**. Same setup as in to Table 3b, but in this case we use *Gradient Boosted Decision Trees*.

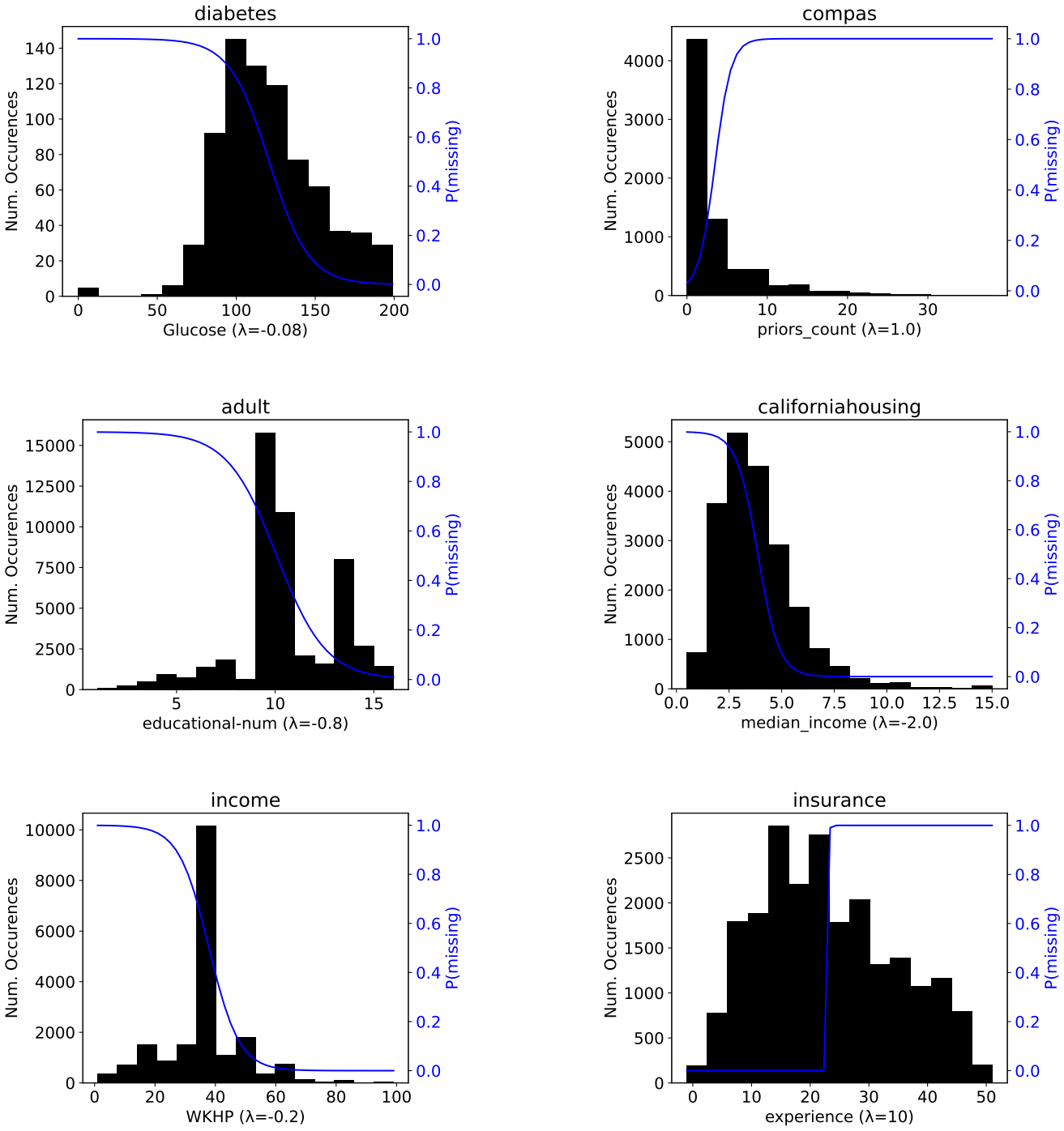


Figure 10: Value distribution of the respective optional features per data set and corresponding function $\mathbf{p}(A_i = 0|z_i)$ with parameter λ used to introduce stochastic availability.

		Fair models				Full feature model
task	data (# opt.)	Base feature model	PUCIDA (f)	PUCIDA (e)	(×)	zero-imputed
C	diab. (2)	27.79 ±4.22	28.18 ±2.48	**27.92** ±2.63	2.3	26.88 ±3.50
C	compas (5)	40.83 ±0.53	39.82 ±0.75	**39.33** ±1.38	7.7	39.74 ±1.39
C	adult (5)	18.00 ±0.37	**15.21** ±0.52	15.31 ±0.51	7.4	15.15 ±0.37
R	calif. (4)	5.83 ±0.27	9.21 ±0.64	**7.86** ±0.27	5.0	7.01 ±0.23
R	income (3)	48.99 ±1.60	47.32 ±1.87	**46.98** ±1.85	3.4	43.78 ±1.83
R	insurance (3)	251.47 ±2.57	**238.28** ±1.18	249.41 ±2.21	3.2	226.85 ±1.75

Table 16: **PUC-compliant models improve predictive performance**. Same setup as in to Table 3b, but in this case we use *Extra Trees*.

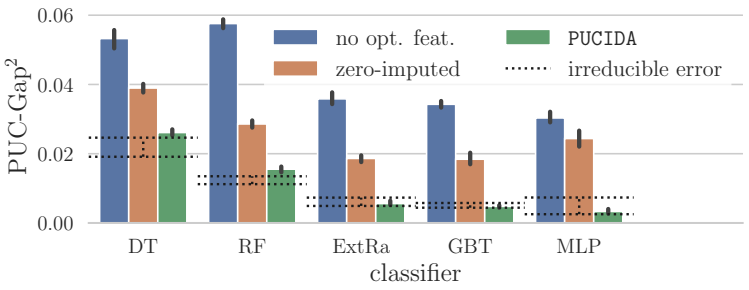


Figure 11: PUCIDA converges independently of the ML model on the second simulated data set with optional features. The fairness gaps are close to the irreducible model estimation error (due to imperfect models on this continuous data set) when applying our technique across a variety of common models on the continuous simulated data set.

data set	optional features
insurance	experience (−), kidslt6 (−), kids618 (−)
adult	age(−) , educational-num (−), hours-per-week (−), capital-gain (−), capital-loss (−)
compas	priors_count (+), age (−), c_days_from_compas (−), c_charge_degree (+), juv_misd_count (+)
diabetes	Glucose (+), age (+)
california housing	housing_median_age (+), population (−), households (−), median_income (−)
income	AGEP (−), SCHL (−), WKHP (−)

Table 17: Features made optional in the experiment with multiple optional features. Direction: (+) means higher values more likely to be unavailable, (−) indicates lower values to be more likely to be unavailable. The direction was chosen such that feature values that lead to more negative outcomes tend to be undisclosed more frequently.

feature	$p(x_i = 1 \mid y = 0)$	$p(x_i = 1 \mid y = 1)$	$p(a_i = 0 \mid y = 0)$	$p(a_i = 0 \mid y = 1)$
1	0.090	0.141	–	–
2	0.915	0.930	–	–
3	0.225	0.020	–	–
4	0.771	0.377	–	–
5	0.202	0.347	–	–
6	0.968	0.322	0.920	0.345
7	0.874	0.239	0.647	0.294
8	0.723	0.159	0.508	0.207

Table 18: Parametric distribution parameters used in the first synthetic data set. Features are all binary. Features 1–5 are base feature which are always available. Features 6 –8 are unavailable with a certain probability given the class label.

base features	n=2, $\mathbf{b} \sim \mathcal{N}(\mathbf{0}, 5\boldsymbol{I})$, $\boldsymbol{w}$=$(-1.5, 1.0)^\top$, t=0
opt. feature 1	$\boldsymbol{u}_1 = (0.8, 0.4)^\top$, $\boldsymbol{v}_1 = (0, 1)^\top$, λ_1=0.7, $\tau_1(0) = -0.25$, $\tau_1(1)$=0.25
opt. feature 2	$\boldsymbol{u}_2 = \mathbf{0}$, $\boldsymbol{v}_2 = (0, -0.15)^\top$, λ_2=1.0, $\tau_2(0)$=0.4, $\tau_2(1) = -0.4$
opt. feature 3	$\boldsymbol{u}_3 = \mathbf{0}$, $\boldsymbol{v}_3 = (0.1, 0.2)^\top$, λ_3=0.0, $\tau_3(0) = -0.2$, $\tau_3(1)$=0.2

Table 19: Parametric distribution parameters used in the synthetic data experiment. The used density covers all possible dependencies between availability, feature values and the base features that are allowed by the graphical model.