Is Your Model Predicting the Past?

Moritz Hardt* Michael P. Kim†

Abstract

When does a machine learning model predict the future of individuals and when does it recite patterns that predate the individuals? In this work, we propose a distinction between these two pathways of prediction, supported by theoretical, empirical, and normative arguments. At the center of our proposal is a family of simple and efficient statistical tests, called *backward baselines*, that demonstrate if, and to what extent, a model recounts the past. Our statistical theory provides guidance for interpreting backward baselines, establishing equivalences between different baselines and familiar statistical concepts. Concretely, we derive a meaningful backward baseline for auditing a prediction system as a black box, given only background variables and the system's predictions. Empirically, we evaluate the framework on different prediction tasks derived from longitudinal panel surveys, demonstrating the ease and effectiveness of incorporating backward baselines into the practice of machine learning.

1 Introduction

Proponents of predictive technologies for consequential decision-making emphasize the seeming ability of statistical models to anticipate individual actions. The ability to predict the future, so the argument goes, creates a rationale for adopting machine learning as policy: if a risk score charted the future trajectory of individuals, then intervening in a person's life on the basis of the risk score would be justified [KLMO15, OE16]. At the same time, critical scholars caution that predictive technologies reproduce historical patterns of injustice and social stratification. In this account, rather than predicting future outcomes, statistical risk assessment tools punish individuals for factors predating their own agency [Eub18, Ben19].

Does a statistical model predict individual agency or recite the past? The answer to the question is often not obvious. Consider the problem of loan default prediction, one of many tasks often framed as predicting future outcomes. One predictor might identify individual behavior detrimental to loan repayment and adjust the predicted likelihood of default accordingly. Another predictor might rely on historical associations between repayment and demographic factors, then predict based solely on the historical factors. Even if the models achieve the same accuracy, they derive predictive power along distinct pathways. In one solution, we rely on the effects of individual behavior on future outcomes. In the other, we reproduce patterns from the past that were

*Max Planck Institute for Intelligent Systems, Tübingen, and Tübingen AI Center
†Miller Institute for Basic Research in Science, University of California, Berkeley

determined before, and independently of, individual behavior. The latter form of prediction—resembling a kind of stereotyping—is core to many documented examples of bias and unfairness in the use of machine learning [DHP+12, ALMK16, Cho17, HKRR18, BG18, BCZ+16, CBN17].

The distinction we draw is fundamental to the theory of equality of opportunity. Dworkin partitions attributes of an individual into factors for which the individual is responsible and factors outside the individual's control [Dwo18a, Dwo18b]. Similarly, Roemer distinguishes between effort that an individual takes and the individual's *type*. A type groups individuals of the same *circumstances*, where "circumstances are those aspects of one's environment (including, perhaps, one's biological characteristics) which are beyond one's control" [Roe00, Roe02, RT16]. Dworkin and Roemer build on this fundamental moral distinction to define what it means to achieve equality in the allocation of resources and opportunity. Here, we focus on the consequences of the same distinction within the context of prediction.

Although the precise distinction is more subtle, we can approximate it with the help of time. Background variables in a prediction problem are those that were determined before the individual, such as, place and date of birth, or parents' educational attainment. Background variables generally influence both an individual's actions and the target of prediction. Individual factors are variables that the individual can exert direct—possibly not full—control over. Correspondingly, we coin the term *backward prediction* to describe the use of background variables in prediction, and we use *forward prediction* to refer to the use of individual factors.

Our contribution. In this work we formalize the distinction between forward and backward prediction. We build a theory for forward and backward prediction around a family of simple and effective statistical tests, we call *backward baselines*. Backward baselines quantify how much of a predictor's strength should be attributed to a given set of background variables. Applying our tools, we empirically find that in representative prediction problems involving longitudinal panel data, backward prediction contributes significantly to the strength of the predictor.

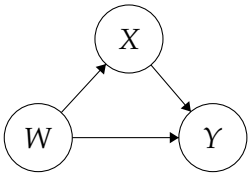
The strength of backward prediction has important consequences. When prediction draws on background factors primarily, it is misleading to interpret the predictor as an individualized risk score. After all, backward predictors are invariant under individual variation and only depend on the individual's background. The strength of backward prediction speaks to the social and environmental constitution of the target of prediction. Consequently, there is no relative advantage to targeting interventions at the individual level on the basis of backward predictors. Even if individual-level interventions are helpful, there is no added benefit in targeting them based on individual predictions compared with targeting based on background variables.

To give an example, we consider predicting an individual's year-round medical expenditure based on the longitudinal MEPS panel survey data. We find that a predictor trained only on background variables nearly matches the predictive performance of classifiers trained on all features. Our finding echoes scholarship about the social determinants of health and medical expenditure [Kri11].

We envision that backward baselines will form a useful component of the machine learning evaluation toolkit. Straightforward to apply, backward baselines provide valuable insights into the interpretation and validity of prediction in consequential settings.

Predicting the past. To introduce our discussion of backward prediction, we consider an explicit data generating process that moves through time. In Figure 1, we depict the temporal dynamics, in the form of a causal graph, with time evolving from left to right. We think of X as individual-level covariates measured today, and Y as an outcome of interest to be measured in the future. In addition to the standard supervised learning variables, we also model an additional *context* variable W—predating the measurement of the covariates or outcome—that may directly influence both X and Y. Concretely, X could represent a record of an individual's educational, personal, and financial history, used to predict income Y measured in 10 years, and W could represent specific demographic features from the past, like childhood household income.

This explicit temporal model elucidates the distinction between *forward* prediction and *backward* prediction. Forward predictors model how the present measurements X causally effect the future outcome Y. Backward predictors estimate the outcome by first inferring the past context W from X, then predicting Y based on W. In other words, backward prediction provides information about Y that could equally be explained by the past context W.



Generating $\mathcal{D}$:

1. $W \sim \mathcal{D}_W$
2. $X \sim \mathcal{D}_{X|W}$
3. $Y \sim \mathcal{D}_{Y|X,W}$

Figure 1: Example data generating process for covariates X, outcome Y, and context W. Time starts from the left with context W and evolves forward to the right, realizing X then Y.

Backward baselines. Machine learning practitioners often build models using any and every predictive pathway available, including the backward pathway. Our goal is to elucidate and disentangle the prediction pathways that a given predictor uses. Backward baselines provide a careful accounting of the predictor's use of the forward and backward predictive pathways. The baselines are lightweight to run, only requiring input-output access to the predictive model, and are built on simple, but rigorous statistical foundations. For instance, a key challenge in reasoning about backward prediction is that the context W is typically robustly encoded within an individual's covariates X. That is, even if we explicitly censor the attributes defining the context, backward prediction from X may still be possible. Backward baselines handle this statistical subtlety gracefully, providing guaranteed estimates of the forward and backward predictive power, regardless of how redundantly W is encoded in X.

Our work establishes backward baselines as an effective tool for investigating predictive models. Our perspective is *not* that the backward prediction pathway is inherently problematic. Rather, we advocate that investigators use backward baselines to understand and contextualize performance numbers in prediction tasks. Adding the baselines to the "report card" for supervised learning would add clarity about the underlying mechanisms used to predict. This clarity, in turn, may inform debate about whether machine learning is an appropriate tool for the task at hand. If model builders cannot find a predictor that improves significantly over backward baselines, we should hesitate before turning prediction into policy.

2 Backward baselines

We work over a data universe $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is a feature space and $\mathcal{Y}$ is a discrete set of labels in the case of classification problems. For regression problems, we take $\mathcal{Y}$ to be the real line $\mathbb{R}$. Fixing a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$, for a given predictor $h : \mathcal{X} \to \mathcal{Y}$, we measure the fit of the predictor in terms of its expected loss over a distribution $(X, Y) \sim \mathcal{D}$ supported on $\mathcal{X} \times \mathcal{Y}$:

$$\ell_{\mathcal{D}}(Y, h(X)) = \mathbf{E}[\ell(Y, h(X))]$$

Throughout, we assume ℓ is symmetric in its two arguments. We study both binary classification and regression, focusing on the zero-one loss $\mathbf{Pr}[Y \neq h(X)]$ with $\mathcal{Y} = \{0, 1\}$ and squared loss $\mathbf{E}[(Y - h(X))^2]$ where $\mathcal{Y} = [0, 1]$, respectively.

We extend this standard setup with a random variable W, jointly distributed with (X, Y) and supported on a discrete domain $\mathcal{W}$. The variable W represents a *context* of both the individual covariates and the outcome of interest. While we model them as separate random variables, at times, we assume that X encodes W, explicitly or implicitly. For instance, in Proposition 2(a), we assume that perfect reconstruction of W is statistically possible from X.

Backward prediction baseline. In our typical story of backward prediction from X, we imagine that a predictor first resolves W from X, then predicts Y from W. As such, if we are concerned that a predictor h is using the backward pathway, a natural baseline to compare against is predicting Y directly from the context W. Fixing a loss ℓ, we take $g^* : \mathcal{W} \to \mathcal{Y}$ to be the statistically optimal predictor of Y from W, and consider the following *backward prediction baseline*, $\ell_{\mathcal{D}}(Y, g^*(W))$.

$$g^*(w) = \operatorname*{argmin}_{\hat{y} \in \mathcal{Y}} \ \mathbf{E}[\ell(Y, \hat{y}) | W = w] \qquad \ell_{\mathcal{D}}(Y, g^*(W)) = \mathbf{E}[\ell(Y, g^*(W))]$$

The loss $\ell_{\mathcal{D}}(Y, g^*(W))$ provides a fundamental baseline for how predictable the outcome Y is from W. By comparing this baseline to $\ell_{\mathcal{D}}(Y, h(X))$, we can better contextualize the quality of predictions h produces. In particular, if h does not achieve significantly better loss than g^*, then h is not a very impressive predictor: rather than using machine learning to make decisions, you could get the same performance simply by stereotyping based on W.

Backward rounding baseline. While the optimal backward predictor g^* is fundamental, it only depends on the underlying relationship between W and Y, and does not depend on any predictor from X. Given such a predictor $h : \mathcal{X} \to \mathcal{Y}$, we may instead consider a baseline based on prediction of $h(X)$ from W. We consider the *backward rounding baseline*, defined by $g^h : \mathcal{W} \to \mathcal{Y}$, which we take to be the optimal predictor of $h(X)$ from W.

$$g^h(w) = \operatorname*{argmin}_{\hat{y} \in \mathcal{Y}} \ \mathbf{E}[\ell(h(X), \hat{y}) | W = w] \qquad \ell_{\mathcal{D}}(h(X), g^h(W)) = \mathbf{E}[\ell(h(X), g^h(W))]$$

Intuitively speaking, if the prediction $h(X)$ is itself predictable from W, then it seems h must be using the backward pathway. Contrapositively, if h is a forward predictor, then $h(X)$ cannot be predicted from W. An interesting aspect of this baseline is that g^h can be estimated even when true outcomes are unavailable, unobserved, or unreliable. Moreover, in settings where predictions are

performative, in the sense of influencing the distribution on outcomes [PZMDH20], the backward prediction baseline may not be applicable, while the backward rounding baseline is unaffected.

To understand what these two baselines measure exactly and how they relate, we need to formally define backward and forward prediction.

2.1 Distinguishing forward and backward prediction

We draw a distinction between two forms of prediction of Y from X: *forward* prediction models the mechanism by which X influences Y; *backward* prediction forecasts Y from X indirectly, by exploiting correlations through the context W. Because W may be redundantly encoded within X, we cannot simply remove W from the features to evaluate the predictive power along the forward pathway. Instead, we define forward and backward prediction based on conditional independence statements involving $Y, h(X)$, and W.

Definition 1 (Backward and forward prediction)**.** *A predictor $h : \mathcal{X} \to \mathcal{Y}$ is a (pure) backward predictor of Y if $h(X)$ is conditionally independent of Y given W.*

$$h(X) \perp Y | W$$

A predictor $h : \mathcal{X} \to \mathcal{Y}$ is a (pure) forward predictor of Y if $h(X)$ is independent of W.

$$h(X) \perp W$$

Most classifiers will be not be pure forward or pure backward predictors, but instead $h(X)$ will have some correlation with Y that goes through W and some correlation that is independent of W. By comparing the loss achieved by a classifier h to one of our backward baselines, we can understand how close to a backward predictor the classifier is.

Backward prediction as random targeting. Connecting Definition 1 to our motivating question, we see that using a backward predictor as the basis for intervention on *individuals* is fruitless. In particular, once we condition on a category defined by W, backward predictions $h(X)$ can be randomized across individuals with no loss.

Fact 1. *Suppose $h : \mathcal{X} \to \mathcal{Y}$ is a backward predictor. For any setting of the context $W = w$, consider a randomized prediction strategy, where R_w is an independent random variable distributed as $h(X)|W = w$. Then, the loss of h is equal to that of random prediction according to R_w.*

$$\mathbf{E}[\ell(Y, h(X))|W = w] = \mathbf{E}[\ell(Y, R_w)|W = w]$$

This fact follows immediately from the definition of backward prediction through conditional independence, but it gives a powerful conclusion. Given a predictor that uses the backward pathway through W, once we condition on a particular setting of $W = w$, then the predictions $h(X)$ may as well be randomly assigned. That is, using a backward predictor as the basis of intervention, is analogous to stereotyping according to the categories defined by W, and then targeting the intervention randomly within categories.

3 Properties of backward baselines

In this section, we develop basic theory for backward baselines, demonstrating how these baselines give us a lens into understanding backward and forward prediction. We study the basic properties of backward baselines, establish interpretations of these baselines in terms of familiar statistical quantities, and draw connections to concepts from the study of fairness in prediction. We defer proofs of all formal claims to Appendix A.

3.1 Basic properties

Here, we establish some basic properties about backward baselines. These properties are intuitive, but also reveal subtleties in what we can(not) conclude about backward and forward prediction from backward baselines. We start with three simple properties of backward baselines, that help us to compare the predictive power from X to the predictive power from W.

Proposition 2. *The following properties of backward baselines hold.*

(a) *When X encodes W, there exists a predictor $h^* : \mathcal{X} \to \mathcal{Y}$ that achieves loss at most the backward prediction baseline.*

$$\ell_{\mathcal{D}}(Y, h^*(X)) \leq \ell_{\mathcal{D}}(Y, g^*(W))$$

(b) *If $h : \mathcal{X} \to \mathcal{Y}$ is a backward predictor, then its loss is at least the backward baselines.*

$$\ell_{\mathcal{D}}(Y, g^*(W)) \leq \ell_{\mathcal{D}}(Y, h(X)) \quad \textit{and} \quad \ell_{\mathcal{D}}(h(X), g^h(W)) \leq \ell_{\mathcal{D}}(Y, h(X))$$

(c) *If $h : \mathcal{X} \to \mathcal{Y}$ is a forward predictor, then g^h is comparable to a constant predictor. Formally,*

$$\ell_{\mathcal{D}}(h(X), g^h(W)) \geq \operatorname*{argmin}_{\hat{y} \in \mathcal{Y}} \mathbf{E}[\ell(h(X), \hat{y})] \quad \textit{and} \quad \ell_{\mathcal{D}}(Y, g^h(W)) \geq \operatorname*{argmin}_{\hat{y} \in \mathcal{Y}} \mathbf{E}[\ell(Y, \hat{y})]\,.$$

These straightforward properties provide a foundation for reasoning about backward and forward prediction. Proposition 2(a) establishes that the backward prediction baseline is reasonable minimum standard for predictive accuracy from X. Proposition 2(b)-(c) can be viewed as one-sided tests that let us demonstrate that a predictor is not a (pure) backward or forward predictor.

For a backward predictor, the backward baselines lower bound the loss $\ell_{\mathcal{D}}(Y, h(X))$. On the other hand, for a forward predictor that achieves nontrivial loss (i.e., beating a constant), the backward baselines upper bound the loss. While Proposition 2(b)-(c) each provide one-sided tests, together they can tell a rich story. For instance, suppose a forward predcitor f and backward predictor b achieve similar loss $\ell_{\mathcal{D}}(Y, f(X)) \approx \ell_{\mathcal{D}}(Y, b(X))$. We may distinguish these cases by backward rounding the predictors to g^f and g^b. Rounding f to g^f will cause a significant deterioration in loss (to that of a constant predictor), but the rounded backward predictor g^b will maintain the predictive power of b. In this case, we may still decide to reject f if it achieves mediocre accuracy, but cannot reliably reject it on the basis of being a backward predictor.

3.2 Rounding recovers optimal backward prediction

As discussed, we can define backward baselines in terms of the optimal predictor g^* of Y from W, and also in terms of the backward-rounded predictor g^h of $h(X)$ from W. In generality, these two predictors realize different baselines; however, if $h(X)$ is an accurate predictor of Y, then intuitively, it would seem that the baselines over g^* and g^h might be similar. For instance, for classification according to the zero-one loss and regression according to the squared loss, these predictors have closed forms.

Zero-one	$g^*(w) = \text{argmax}_{\hat{y} \in \mathcal{Y}} \mathbf{Pr}[Y = \hat{y} \mid W = w]$	$g^h(w) = \text{argmax}_{\hat{y} \in \mathcal{Y}} \mathbf{Pr}[h(X) = \hat{y} \mid W = w]$
Squared	$g^*(w) = \mathbf{E}[Y \mid W = w]$	$g^h(w) = \mathbf{E}[h(X) \mid W = w]$

We introduce the following technical conditions, which are useful for analyzing various properties of backward baselines.

Definition 2 (Confidence). *A classifier $h : \mathcal{X} \to \mathcal{Y}$ is (over)confident on Y over W if*

$$\mathbf{Pr}[h(X) = g^*(W)] \geq \mathbf{Pr}[Y = g^*(W)].$$

Intuitively, confidence says that $h(X)$ does not underestimate the probability that Y takes it's most likely value within the context W. Such (over)confidence of classifiers is typically observed in practice [GPSW17].

Definition 3 (Weak calibration). *A predictor $h : \mathcal{X} \to [0, 1]$ is weakly calibrated*[1] *to Y over W if*

$$\mathbf{E}[Y \mid W] = \mathbf{E}[h(X) \mid W] \qquad \textit{and} \qquad \mathbf{E}[Yh(X) \mid W] = \mathbf{E}[h(X)^2 \mid W].$$

Weak calibration rules out predictors that blatantly ignore variation in Y based on the context W (including pure forward predictors). Definition 3 relaxes traditional notions of calibration [Daw85] and is implied by loss minimization, both in theory and our experiments. We show that under these conditions, backward rounding obtains optimal prediction of Y from W.

Proposition 3 (Informal). *For a confident classifier $h : \mathcal{X} \to \{0, 1\}$ or a weakly calibrated predictor $h : \mathcal{X} \to [0, 1]$, we have $g^h = g^*$ for the zero-one loss and squared loss, respectively.*

The interchangeability of g^* and g^h may be useful practically and conceptually. For instance, the analysis of Proposition 3 reveals that the backward rounding baseline lower bounds the backward prediction baseline, $\ell_{\mathcal{D}}(h(X), g^h(W)) \leq \ell_{\mathcal{D}}(Y, g^*(W))$ (which, in turn, gives a strengthening of Proposition 2(b) under confidence or weak calibration).

3.3 Measuring forward predictive power

A key motivation for our study of backward baselines was the observation that, given a predictor h, determining the extent of forward prediction may be challenging. We show that under natural

[1] This notion of weak calibration was introduced recently by [GKSZ22], who refer to it as degree-2 calibration.

conditions, the backward rounding baseline for g^h reveals insight into the forward predictive power of h. Conveniently, evaluating this baseline only requires black-box access to the predictive model and (X, W) samples—not labels Y. The lightweight nature of the baseline makes it an appealing option to audit for backward prediction, especially for proprietary predictive models. Concretely, we show that the backward rounding baseline gives insight into the covariance between $h(X)$ and Y after conditioning on W.

Proposition 4. *Suppose a classifier $h : \mathcal{X} \to \{0, 1\}$ is confident on Y over W. Let $\ell_W(h, g^h)$ denote the backward rounding baseline $\mathbf{Pr}[h(X) \neq g^h(W)|W]$ conditioned on W. Then,*

$$\mathbf{Cov}(h(X), Y|W) \leq \mathbf{Var}(h(X)|W) = \ell_W(h, g^h) \cdot (1 - \ell_W(h, g^h)) \leq \mathbf{Var}(Y|W)\,.$$

If a predictor $h : \mathcal{X} \to [0, 1]$ is weakly calibrated to Y over W, then

$$\mathbf{E}[(h(X) - g^h(W))^2] = \mathbf{E}[(Y - g^*(W))^2] - \mathbf{E}[(Y - h(X))^2] = \mathop{\mathbf{E}}_{W}[\mathbf{Cov}(Y, h(X)|W)]\,.$$

In other words, if $h(X)$ carries lots of information about Y, even after conditioning on W, then the backward rounding baseline will be large. The arguments to establish Proposition 4 are elementary, but the consequences are powerful. An auditor, who is given only black-box access to a classifier or predictor h, can reliably determine when h is a backward predictor by evaluating the backward rounding baseline without any labels Y from the true distribution. Concretely, the backward rounding baseline allows the auditor to establish an upper bound on the amount of information about Y contained in $h(X)$ that isn't explained by W.

In the classification setting, the bound obtained by the rounding baseline is an inequality, but is tighter than the bound given by the backward prediction baseline. In the regression setting, the rounding baseline also characterizes the difference between the backward prediction baseline and the expected loss of h, which would otherwise require labeled outcomes Y to evaluate. In Appendix A, we describe an additional backward baseline for classification, which use labels from Y to gives an exact characterization of the forward predictive power of h.

3.4 Backward baselines and demographic parity

When W is defined by demographic features that are considered to be sensitive attributes, forward prediction recovers the notion of *demographic parity* from the literature on fair machine learning [DHP+12]. While a natural desideratum for equal treatment under a decision rule, the shortcomings of demographic parity as a notion of fairness have been documented extensively [DHP+12, LDR+18]. As such, requiring pure forward prediction may result in unintended and undesirable consequences, just as blinding predictors of a sensitive attribute can.

Exploring the analogy between backward baselines and fair prediction sheds new light on demographic parity and stereotyping. In Appendix B, we formalize a duality between forward and backward prediction. Translating the duality into the language of fairness, the optimal unconstrained prediction decomposes into the optimal prediction under demographic parity plus the optimal "stereotyping" prediction that makes its judgments solely based on the sensitive attribute.

4 Empirical evaluation of backward baselines

The goal of our experiments is to empirically evaluate backward baselines. Toward this goal, we searched for datasets that meet at least four important criteria:

1. The outcome variable demonstrably lies in the future relative to the features.
2. The dataset contains general demographic background variables, as well as features specific to the prediction task.
3. Non-trivial prediction accuracy is possible.
4. Individual-level microdata are publicly available.

Many machine learning datasets are unclear about the temporality of the outcome variable, thus falling short of the first criterion. For example, several datasets about credit default prediction do not clarify whether data points correspond to individuals who have already defaulted, or individuals that ended up defaulting some specific time after feature collection.

Well-suited to our evaluation are longitudinal panel surveys. Each panel consists of some number of survey participants who are interviewed in multiple rounds (or waves). By taking features from one round to predict outcomes in a later round, we can create prediction problems where outcomes and features are temporally well-separated. We choose two major panel surveys relating to medical expenditure and income. Complementing these panel surveys, we also consider a notorious dataset from the criminal legal domain. Extended results and full details are in Appendix C and Appendix D. The code is available at:

https://github.com/socialfoundations/backward_baselines

4.1 Medical Expenditure Panel Survey (MEPS)

The Medical Expenditure Panel Survey (MEPS) is a set of large-scale surveys of families and individuals, their medical providers, and employers across the United States, aimed at providing insights into health care utilization. We work with the publicly available MEPS 2019 Full Year Consolidated Data File. The dataset we consider has 28,512 instances corresponding to all persons that were part of one of the two MEPS panels overlapping with calendar year 2019. Specifically, Panel 23 has Rounds 3–5 in 2019, and Panel 24 has rounds 1–3 in 2019. Round 3 of Panel 23 and Round 1 of Panel 24 are the first of each panel in 2019. The survey distinguishes between demographic variables and variables corresponding to survey questions in of the rounds of the two panels. We create a prediction task whose goal is to predict a full-year outcome from Round 3 of Panel 23 and Round 1 of Panel 24. The target variable measures the total health care utilization across the year. We create a roughly balanced binarization of the target variable. A precise definition and further details are in the appendix.

We compute backward baselines in terms of the features *age*, *race*, *age* and *race* together, as well as all variables designated as *demographic* by the survey documentation. These include additional variables relating to age, race and ethnicity, marital status, nationality, and languages spoken.

Figure 2 summarizes our findings. In particular, backward baselines trained on all demographic background variables match nearly all of the predictive performance of the classifiers trained on all features, similarly across three different prediction models. An extended set of figures is included in Appendix C.

4.2 Survey of Income and Program Participation (SIPP)

The Survey of Income and Program Participation (SIPP) is an import longitudinal survey conduced by the U.S. Census Bureau, aimed at capturing income dynamics as well as participation in government programs.

We consider Wave 1 and Wave 2 of the SIPP 2014 panel data. The target variable is based on the official poverty measure (OPM), a cash-income based measure of poverty. We compute this measure based on Wave 2 data. We again discretize the measure to obtain to roughly balanced classes for our binary prediction task. The goal is to predict this outcome based on features collected in Wave 1. After cleaning and preprocessing our data contains 39720 rows and 54 columns. We consider background variables *education*, *race*, *education* and *race* together, as well as all demographic variables, specifically, *age*, *gender*, *race*, *education*, *marital status*, *citizenship status*. In Figure 3, we restrict our attention to the logistic regression model. The other models perform similarly and the full set of results can be found in Appendix C.

4.3 ProPublica COMPAS Recidivism Scores

A proprietary recidivism risk score, called COMPAS, was the subject of a notorious investigation into racial bias by ProPublica [ALMK16] in 2016. As part of the investigation ProPublica released a dataset of COMPAS scores about defendants associated with two-year recidivism outcomes. The dataset released by ProPublica has significant and well-documented issues that make it inadequate for the development of new risk scores as well as fairness interventions [BZZ+21, Bar19]. In experimenting with the COMPAS data set, our primary goal is to demonstrate the effectiveness of backward baselines in auditing problematic risk predictors. The results of backward baselines echo earlier findings that the performance COMPAS scores can be achieved by simple models [RWC20, WHPR22].

Note that we do not have access to the training data used for producing the COMPAS scores as is common in algorithmic audit scenarios. This is, fortunately, not required for evaluating backward baselines. We only need the scores, as well as associated demographic information. Figure 4 evaluates backward baselines against the COMPAS scores. The results are rather striking in how well backward baselines do in comparison. In particular, a single feature (prior convictions) appears to account for all of the predictive power of the COMPAS score.

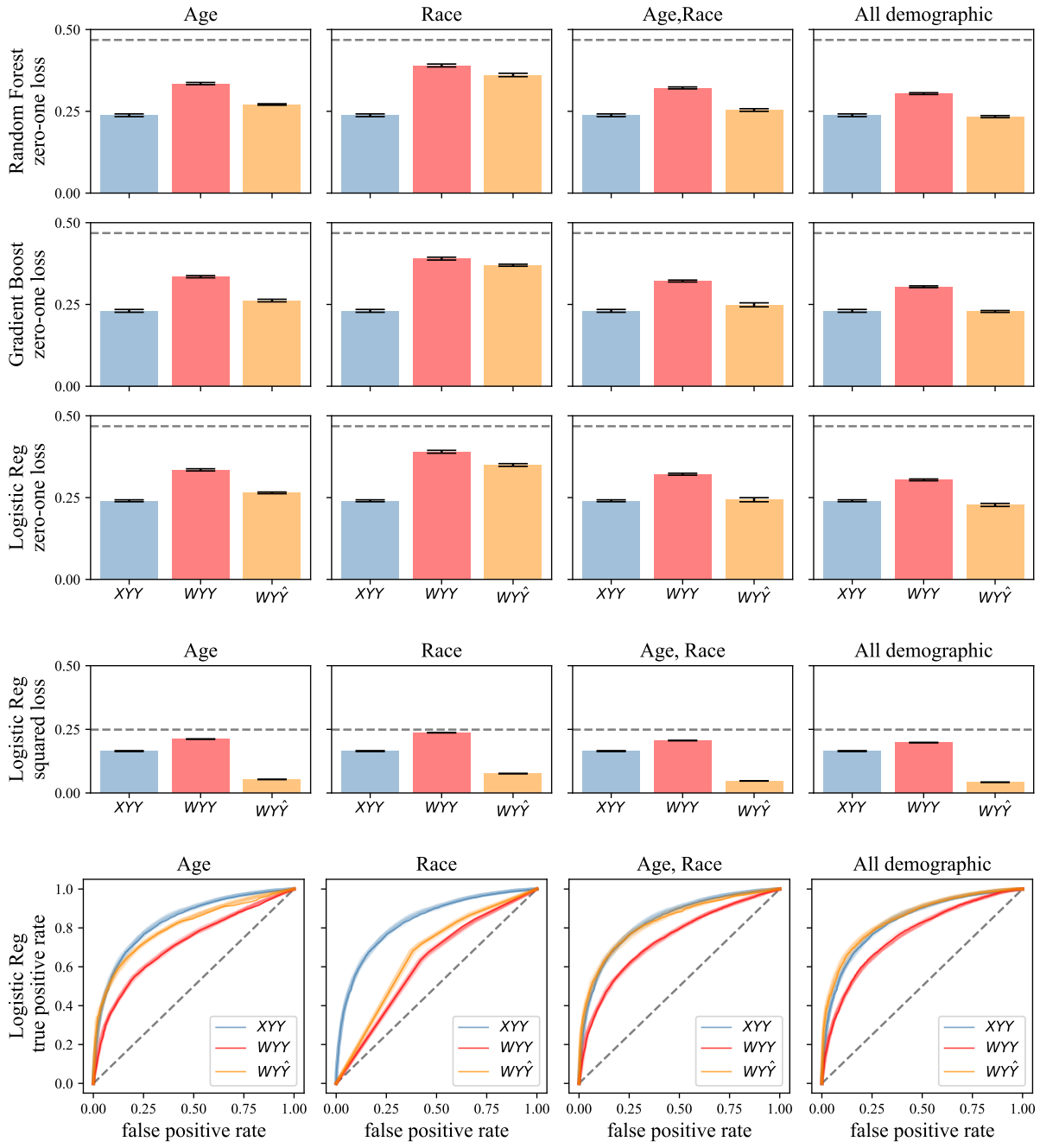


Figure 2: Backward baselines on MEPS, columns are different features, rows are different classifiers (random forest, gradient boosting, logistic regression) and metrics (zero-one loss, squared loss, ROC curves). Label XYY denotes standard training and testing, label WYY is the backward prediction baseline, label $WY\hat{Y}$ is the backward rounding baseline. Gray dashed line indicates performance of constant predictor. Error bars represent a standard deviation across 10 random seeds.

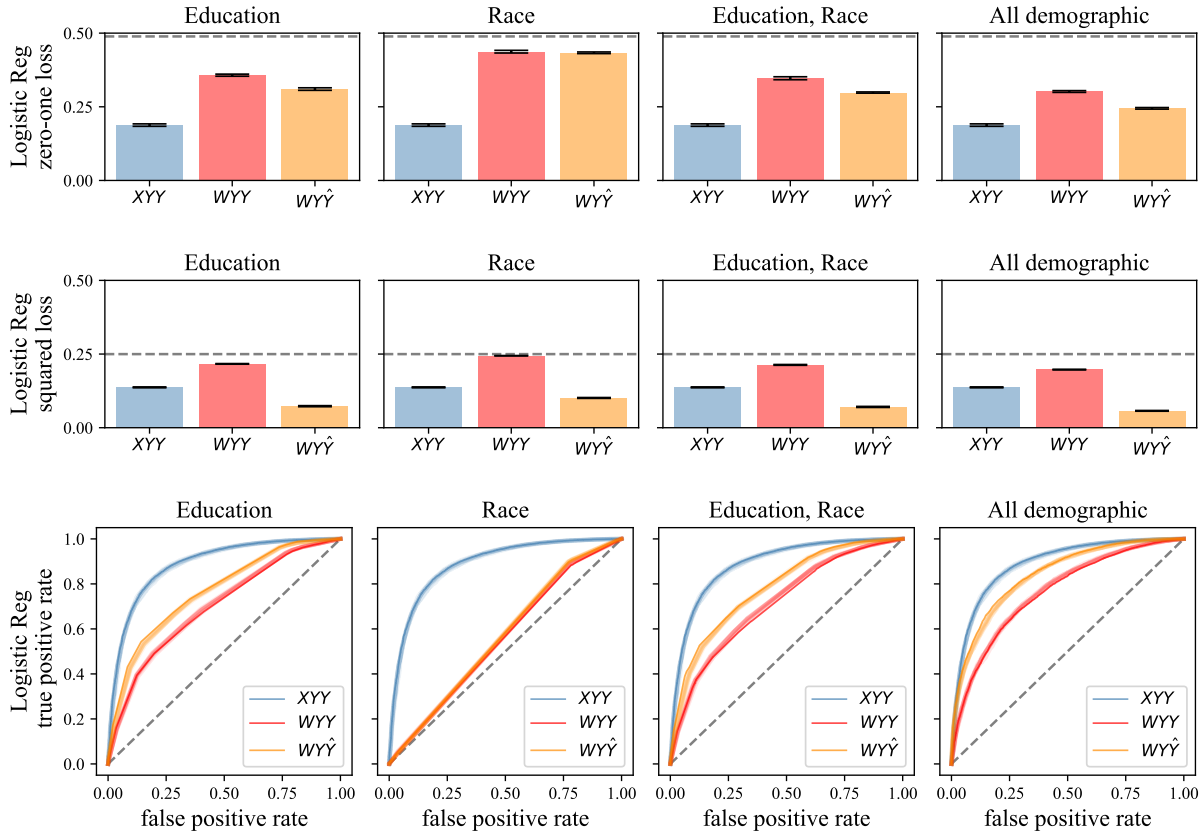


Figure 3: Backward baselines on SIPP.

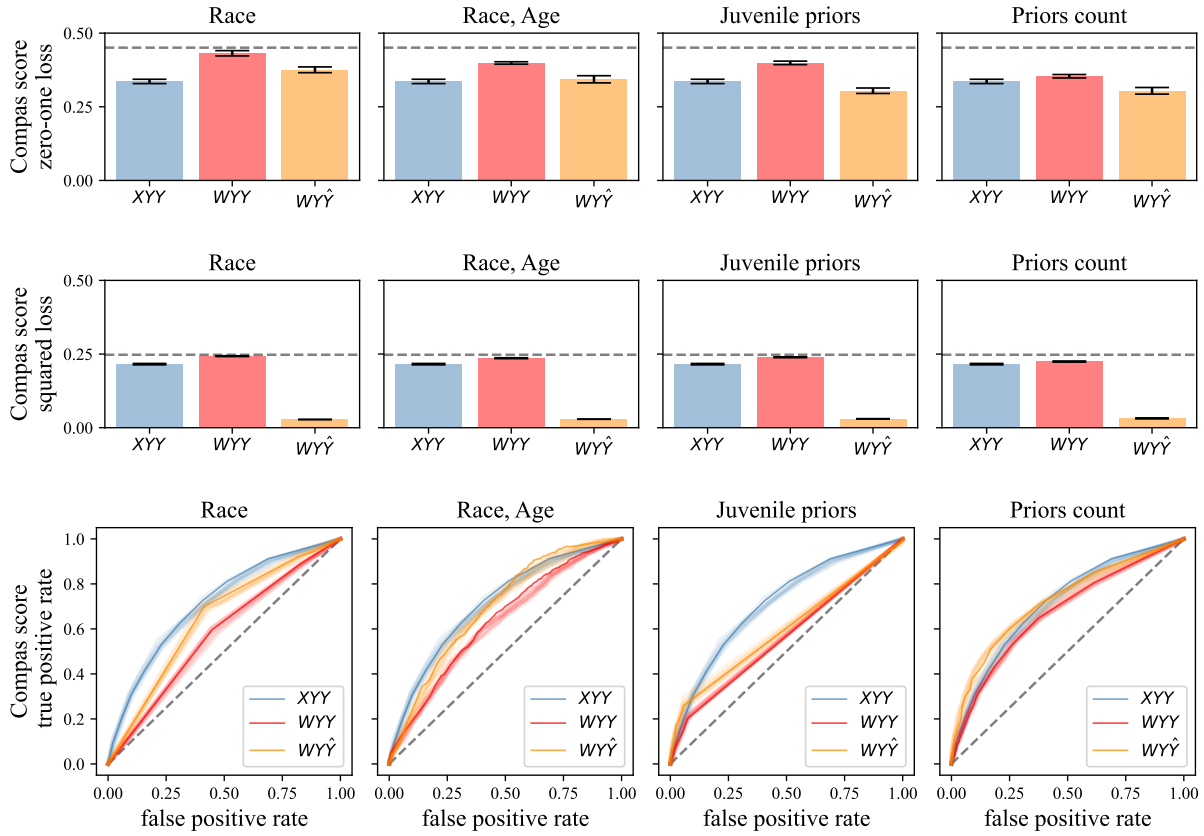


Figure 4: Backward baselines on COMPAS

5 Discussion

It might seem that the question we ask runs headlong into the centuries-old problem of induction: *How do we draw conclusions about the future based on past experience?* But our study addresses a distinct and more specific question: *Does a given predictor capitalize on individual behaviors that influence the outcome or on historical patterns that correlate with the outcome?* Backward baselines ask about induction *from what*. Initiating this investigation, we start from factors that apparently predate the point of prediction, such as demographic background variables, and test to what extent a predictor utilizes these factors. Our findings—that backward prediction often serves a significant role in forecasting individuals' outcomes—adds relevant evidence for the ongoing deliberation of the meaning of individual risk scores [San03, Daw17, DKR+21, Dwo21].

Our present investigation into backward baselines is limited to settings where the variables defining a past context W are measured and observed by the auditor. A fundamental question in evaluating backward baselines is which variables constitute the right choice for context W. We emphasized the role of time in deciding what is outside the individual's control. Some factors are obviously in the past, e.g., place of birth, and parents' educational attainment. Other factors, such as race, gender, and individual's educational attainment, involve the individual at present but are nonetheless socially constituted. Time alone is therefore an imperfect guiding principle in choosing what we count as a suitable background variable W. Choosing W appropriately is not a purely technical question, but rather is up for debate based on the context and scope of the prediction task.

5.1 Additional Related Works

On the level of techniques, backward baselines bear resemblance to a number of tools developed in the causal inference and machine learning communities. Backward baselines do not make any assumptions on the underlying causal structure between X, Y, and W. Still, the backward baseline toolkit is similar in ways to tools developed in settings for understanding the causal structure between variables, both for measuring confounding [McN03, Pea09] and mediation analysis [MFF07]. At first glance, backward baselines may also feel similar to the study of spurious correlation, which has received considerable attention in the ML literature (e.g., [SRKL20, SHL20, VDYE21]). We caution, however, that correlation with background (or individual) features should not be understood as "spurious". Instead, correlations with background features reveal important structure in the data distribution, how the predictor exploits this structure, and in turn, when intervening on the basis of prediction may be ineffective.

As discussed, backward baselines also add new color to concepts studied in algorithmic fairness. In particular, under a specific causal interpretation, forward prediction may be understood similarly as the notion of Counterfactual Fairness [KLRS17]. This connection can be made somewhat formal, as the latter notion has been shown to be closely related to the notion of Demographic Parity as well [RW22]. Indeed, in some accounts of fairness, understanding the causal pathways of prediction is essential [KRCP+17].

Finally, the findings of a recent empirical study of prediction systems in the American public school system are closely related to our theoretical work. [PBHA23] studies the Early Warning System

(EWS) of the Wisconsin Department of Public Instruction. The research concludes that, largely, the individual-risk prediction sytems is (a) effective at prediction, and (b) reliant predominantly on environmental features (e.g., what percent of an individual's school qualifies for free or reduced lunch). In this way, while the predictions are accurate, they are backward predictors, and do not provide an effective tool for intervention on individuals' educational plans.

5.2 Conclusions

Our contribution has a normative, a theoretical, and an empirical component. We argue that the distinction between predicting the future of an individual and reproducing the past is central to the debate around where and how we should use statistical methods to make consequential decisions. The effectiveness of backward prediction, when observed, should question support for prediction as policy, and instead redirect focus toward interventions that target the background conditions.

Theoretically, we begin to develop a statistical learning theory of backward baselines. The theory helps simplify the landscape of possible backward baselines, while clarifying how to interpret different backward baselines. A notable outcome of our theory is that it supports the use and interpretation of a backward baseline that requires no observed outcomes. At the outset, it was not obvious that a meaningful backward baseline without measurement of the target variable is possible. This finding enables *auditing without measured outcomes*: An investigator can probe a predictive system with access to only background variables and predictions.

On the empirical side, we show the strength and versatility of backward baselines on a variety of datasets. Utilizing multiple waves of longitudinal panel surveys, our evaluation is careful about the temporality of features and outcomes. Along the way, we contribute to a better empirical understanding of how machine learning leverages past contexts to predict future life outcomes. In conclusion, we propose backward baselines as a simple, broadly applicable tool to strengthen evaluation and audit practices in the use of machine learning.

Acknowledgments

We thank Rediet Abebe for insightful and formative interactions throughout the course of this work. We thank Juan C. Perdomo for helpful discussions and feedback. We thank Ricardo Sandoval for providing us with code for the SIPP data and the associated prediction task. **MPK** is supported by the Miller Institute for Research in Basic Science and the Simons Collaboration on the Theory of Algorithmic Fairness. Authors listed alphabetically.

References

[ALMK16] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. In *Ethics of Data and Analytics*, pages 254–264. Auerbach Publications, 2016.

[Bar19] Matias Barenstein. Propublica's compas data revisited. *arXiv preprint arXiv:1906.04711*, 2019.

[BCZ+16] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

[Ben19] Ruha Benjamin. Race after technology: Abolitionist tools for the new jim code. *Social forces*, 2019.

[BG18] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[BZZ+21] Michelle Bao, Angela Zhou, Samantha Zottola, Brian Brubach, Sarah Desmarais, Aaron Horowitz, Kristian Lum, and Suresh Venkatasubramanian. It's COMPASlicated: The messy relationship between rai datasets and algorithmic fairness benchmarks. *arXiv:2106.05498*, 2021.

[CBN17] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[Cho17] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[Daw85] A. Philip Dawid. Calibration-based empirical probability. *The Annals of Statistics*, pages 1251–1274, 1985.

[Daw17] Philip Dawid. On individual risk. *Synthese*, 194(9):3445–3474, 2017.

[DHP+12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM, 2012.

[DKR+21] Cynthia Dwork, Michael P Kim, Omer Reingold, Guy N Rothblum, and Gal Yona. Outcome indistinguishability. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1095–1108, 2021.

[Dwo18a] Ronald Dworkin. What is equality? part 1: Equality of welfare. In *The notion of equality*, pages 81–142. Routledge, 2018.

[Dwo18b] Ronald Dworkin. What is equality? part 2: Equality of resources. In *The notion of equality*, pages 143–205. Routledge, 2018.

[Dwo21] Cynthia Dwork. Pseudo-randomness and the crystal ball. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 1–2, 2021.

[Eub18] Virginia Eubanks. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, 2018.

[GKSZ22] Parikshit Gopalan, Michael P Kim, Mihir Singhal, and Shengjia Zhao. Low-degree multicalibration. *arXiv preprint arXiv:2203.01255*, 2022.

[GPSW17] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

[HKRR18] Úrsula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning, ICML*, 2018.

[KLMO15] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. Prediction policy problems. *American Economic Review*, 105(5):491–95, 2015.

[KLRS17] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

[KMR17] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference, ITCS*, 2017.

[KRCP+17] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. *Advances in neural information processing systems*, 30, 2017.

[Kri11] Nancy Krieger. *Epidemiology and the people's health: theory and context*. Oxford University Press, 2011.

[LDR+18] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pages 3150–3158. PMLR, 2018.

[McN03] Roseanne McNamee. Confounding and confounders. *Occupational and environmental medicine*, 60(3):227–234, 2003.

[MFF07] David P MacKinnon, Amanda J Fairchild, and Matthew S Fritz. Mediation analysis. *Annu. Rev. Psychol.*, 58:593–614, 2007.

[OE16] Ziad Obermeyer and Ezekiel J Emanuel. Predicting the future—big data, machine learning, and clinical medicine. *The New England journal of medicine*, 375(13):1216, 2016.

[PBHA23] Juan C Perdomo, Tolani Britton, Moritz Hardt, and Rediet Abebe. Difficult lessons on social prediction from Wisconsin public schools. *arXiv preprint arXiv:2304.06205*, 2023.

[Pea09] Judea Pearl. *Causality*. Cambridge university press, 2009.

[PZMDH20] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *International Conference on Machine Learning*, pages 7599–7609. PMLR, 2020.

[Roe00] John E. Roemer. *Equality of Opportunity*. Harvard University Press, 2000.

[Roe02] John E Roemer. Equality of opportunity: A progress report. *Social Choice and Welfare*, pages 455–471, 2002.

[RT16] John E Roemer and Alain Trannoy. Equality of opportunity: Theory and measurement. *Journal of Economic Literature*, 54(4):1288–1332, 2016.

[RW22] Lucas Rosenblatt and R Teal Witter. Counterfactual fairness is basically demographic parity. *arXiv preprint arXiv:2208.03843*, 2022.

[RWC20] Cynthia Rudin, Caroline Wang, and Beau Coker. The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review*, 2(1), 3 2020. https://hdsr.mitpress.mit.edu/pub/7z100269.

[San03] Alvaro Sandroni. The reproducible properties of correct forecasts. *International Journal of Game Theory*, 32(1):151–159, 2003.

[SHL20] Megha Srivastava, Tatsunori Hashimoto, and Percy Liang. Robustness to spurious correlations via human annotations. In *International Conference on Machine Learning*, pages 9109–9119. PMLR, 2020.

[SRKL20] Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pages 8346–8356. PMLR, 2020.

[VDYE21] Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations in text classification. *Advances in neural information processing systems*, 34:16196–16208, 2021.

[WHPR22] Caroline Wang, Bin Han, Bhrij Patel, and Cynthia Rudin. In pursuit of interpretable, fair and accurate machine learning for criminal recidivism prediction. *Journal of Quantitative Criminology*, pages 1–63, 2022.

A Omitted Proofs

Proposition (Restatement of Proposition 2)**.** *The following properties of backward baselines hold.*

(a) *When X encodes W, there exists a predictor $h^* : \mathcal{X} \to \mathcal{Y}$ that achieves loss at most the backward prediction baseline.*

$$\ell_{\mathcal{D}}(Y, h^*(X)) \leq \ell_{\mathcal{D}}(Y, g^*(W))$$

(b) *If $h : \mathcal{X} \to \mathcal{Y}$ is a backward predictor, then its loss is at least the backward baselines.*

$$\ell_{\mathcal{D}}(Y, g^*(W)) \leq \ell_{\mathcal{D}}(Y, h(X)) \quad \textit{and} \quad \ell_{\mathcal{D}}(h(X), g^h(W)) \leq \ell_{\mathcal{D}}(Y, h(X))$$

(c) *If $h : \mathcal{X} \to \mathcal{Y}$ is a forward predictor, then g^h is comparable to a constant predictor. Formally,*

$$\ell_{\mathcal{D}}(h(X), g^h(W)) \geq \operatorname*{argmin}_{\hat{y} \in \mathcal{Y}} \mathbf{E}[\ell(h(X), \hat{y})] \quad \textit{and} \quad \ell_{\mathcal{D}}(Y, g^h(W)) \geq \operatorname*{argmin}_{\hat{y} \in \mathcal{Y}} \mathbf{E}[\ell(Y, \hat{y})]\,.$$

Proof of Proposition 2. We prove each statement separately.

(a) By the assumption that X encodes W, i.e., that $I(W; X) = H(W)$, there exists a computable map $M : \mathcal{X} \to \mathcal{W}$ such that for any $(X, W, Y) \sim \mathcal{D}$, $M(X) = W$. Thus, the predictor $h^* : \mathcal{X} \to \mathcal{Y}$ defined as the composition of g^* and M,

$$h^*(X) = g^* \circ M(X) = g^*(W)$$

is feasible, and achieves loss $\ell_{\mathcal{D}}(Y, h^*(X)) = \ell_{\mathcal{D}}(Y, g^*(W))$.

(b) Suppose h is a backward predictor; that is, $h(X) \perp Y | W$. Consider the loss achieved by h on $\mathcal{D}$.

$$\begin{aligned} \ell_{\mathcal{D}}(Y, h(X)) &= \mathbf{E}[\ell(Y, h(X))] \\ &= \mathop{\mathbf{E}}_{W} \mathbf{E}[\ell(Y, h(X)) | W] \end{aligned}$$

Note that by the conditional independence of $h(X)$ and Y, we can take the expectation over X and Y conditioned on W separately. Then, the expected loss over the choice of $h(X) \in \mathcal{Y}$ is lower bounded by the optimal choice $\hat{y} \in \mathcal{Y}$.

$$\mathop{\mathbf{E}}_{h(X)|W} \mathop{\mathbf{E}}_{Y|W} [\ell(Y, h(X)) | W] \geq \min_{\hat{y} \in \mathcal{Y}} \mathop{\mathbf{E}}_{Y|W} [\ell(Y, \hat{y}) | W] = \mathop{\mathbf{E}}_{Y|W} [\ell(Y, g^*(W)) | W]$$

Thus, in all, we conclude that $\ell_{\mathcal{D}}(Y, g^*(W)) \leq \ell_{\mathcal{D}}(Y, h(X))$. The second inequality follows similarly, by lower bounding the expected loss over the draw of Y by the optimal $\hat{y} \in \mathcal{Y}$, which results in $g^h(W)$.

(c) Suppose h is a forward predictor; that is $h(X) \perp W$. Consider the definition of g^h,

$$\begin{aligned} g^h(W) &= \operatorname*{argmin}_{\hat{y} \in \mathcal{Y}} \mathbf{E}[\ell(h(X), \hat{y}) | W] \\ &= \operatorname*{argmin}_{\hat{y} \in \mathcal{Y}} \mathbf{E}[\ell(h(X), \hat{y})] \end{aligned}$$

where the equality between the conditional and unconditional expectation follows by independence. Thus, $g^h : \mathcal{W} \to \mathcal{Y}$ must be a constant predictor, and can only hope to compete with the best fixed prediction in predicting Y or $h(X)$. □

Proposition (Formal restatement of Proposition 3)**.** *Suppose a classifier $h : \mathcal{X} \to \{0,1\}$ is confident on Y over W. Then,*

$$g^*(W) = \operatorname*{argmax}_{\hat{y} \in \mathcal{Y}} \mathbf{Pr}[Y = \hat{y}|W] = \operatorname*{argmax}_{\hat{y} \in \mathcal{Y}} \mathbf{Pr}[h(X) = \hat{y}|W] = g^h(W).$$

Suppose a predictor $h : \mathcal{X} \to [0,1]$ is weakly calibrated to Y over W. Then,

$$g^*(w) = \mathbf{E}[Y|W] = \mathbf{E}[h(X)|W] = g^h(W).$$

Proof of Proposition 3. First, we prove the equality for classifiers. Note that minimization is entirely determined by which side of $1/2$ the probability that the outcome is 1. That is, $\operatorname{argmin}_{\hat{y} \in \mathcal{Y}} \mathbf{Pr}[Y \neq \hat{y}|W]$ returns the indicator of whether $\mathbf{Pr}[Y = 1|W] > 1/2$. Thus, as long as $\mathbf{Pr}[Y = 1|W]$ and $\mathbf{Pr}[h(X) = 1|W]$ are on the same side of $1/2$, then the equality follows. By confidence, if $g^*(W) = 1$, then $1/2 < \mathbf{Pr}[Y = 1|W] \leq \mathbf{Pr}[h(X) = 1|W]$, so $g^h(W) = 1$ as well. The statement holds analogously for the case $g^*(W) = 0$.

Next, we prove the equality for regression predictors. By the definition of weak calibration, we have that h matches expectations with Y conditional on W.

$$\mathbf{E}[h(X)|W] = \mathbf{E}[Y|W].$$

Thus, by the closed-form solution for g^* and g^h, we have the stated equality.

$$\begin{aligned} g^*(W) &= \mathbf{E}[Y|W] \\ &= \mathbf{E}[h(X)|W] \\ &= g^h(W) \end{aligned}$$

□

Proposition (Restatement of Proposition 4)**.** *Suppose a classifier $h : \mathcal{X} \to \{0,1\}$ is confident on Y over W. Let $\ell_W(h, g^h)$ denote the backward rounding baseline $\mathbf{Pr}[h(X) \neq g^h(W)|W]$ conditioned on W. Then,*

$$\mathbf{Cov}(h(X), Y|W) \leq \mathbf{Var}(h(X)|W) = \ell_W(h, g^h) \cdot (1 - \ell_W(h, g^h)) \leq \mathbf{Var}(Y|W)$$

If a predictor $h : \mathcal{X} \to [0,1]$ is weakly calibrated to Y over W, then

$$\mathbf{E}[(h(X) - g^h(W))^2] = \mathbf{E}[(Y - g^*(W))^2] - \mathbf{E}[(Y - h(X))^2] = \mathop{\mathbf{E}}_{W}[\mathbf{Cov}(Y, h(X)|W)].$$

Proof of Proposition 4. Suppose $h : \mathcal{X} \to \{0,1\}$ is confident on Y over W. First, the covariance $\mathbf{Cov}(h(X), Y|W)$ is upper bounded by the variance $\mathbf{Var}(h(X)|W)$. Then, we express the variance of this Bernoulli random variable in terms of the backward rounding baseline. Specifically, for either $\hat{y} \in \{0,1\}$:

$$\begin{aligned} \mathbf{Var}(h(X)|W) &= \mathbf{Pr}[h(X) = \hat{y}|W] \cdot \mathbf{Pr}[h(X) \neq \hat{y}|W] \\ &= \mathbf{Pr}[h(X) \neq g^h(W)|W] \cdot (1 - \mathbf{Pr}[h(X) \neq g^h(W)|W]). \end{aligned}$$

Further, by confidence and Proposition 3, we can bound the probabilities.

$$\begin{aligned}\mathbf{Pr}[h(X) \neq g^h(W)|W] &\leq \mathbf{Pr}[Y \neq g^h(W)|W] \\ &= \mathbf{Pr}[Y \neq g^*(W)|W] \\ &\leq 1/2\end{aligned}$$

By properties of variance of Bernoullis, $h(X)$ given W is more peaked than Y given W, so will have lower variance.

Given a weakly calibrated predictor $h : \mathcal{X} \to [0, 1]$, we expand the difference in squared loss as follows.

$$\mathbf{E}[(Y - g^h(W))^2] - \mathbf{E}[(Y - h(X))^2] = 2\,\mathbf{E}[Y(h(X) - g^h(W))] - \mathbf{E}[h(X)^2 - g^h(W)^2]$$

By the fact that $g^h(W) = \mathbf{E}[h(X)|W]$, the second term can be rewritten as the squared error between g^h and h.

$$\begin{aligned}\mathbf{E}[h(X)^2 - g^h(W)^2] &= \mathop{\mathbf{E}}_{W}\mathbf{E}[h(X)^2 - g^h(W)^2|W] \\ &= \mathop{\mathbf{E}}_{W}[\mathbf{E}[h(X)^2|W] - g^h(W)^2] \\ &= \mathop{\mathbf{E}}_{W}[\mathbf{E}[h(X)^2|W] - 2g^h(W)\,\mathbf{E}[h(X)|W] + g^h(W)^2] \\ &= \mathbf{E}[(h(X) - g^h(W))^2]\end{aligned}$$

The first term can be rewritten as the expected covariance between Y and $h(X)$ conditioned on W.

$$\begin{aligned}\mathbf{E}[Y((h(X) - g^h(W))] &= \mathop{\mathbf{E}}_{W}\mathbf{E}[Y((h(X) - g^h(W))|W] \\ &= \mathop{\mathbf{E}}_{W}[\mathbf{E}[Yh(X)|W] - \mathbf{E}[Y|W]g^h(W)] \\ &= \mathop{\mathbf{E}}_{W}[\mathbf{E}[Yh(X)|W] - \mathbf{E}[Y|W]\,\mathbf{E}[h(X)|W]] \\ &= \mathop{\mathbf{E}}_{W}[\mathbf{Cov}(Y, h(X)|W)]\end{aligned}$$

In sum, the difference in losses is equal to

$$\mathbf{E}[(Y - g^h(W))^2] - \mathbf{E}[(Y - h(X))^2] = 2\mathop{\mathbf{E}}_{W}[\mathbf{Cov}(Y, h(X)|W] - \mathbf{E}[(h(X) - g^h(W))^2]$$

Finally, if h is weakly calibrated to Y over W, then the expected covariance is equal to the squared of g^h to h,

$$\begin{aligned}\mathop{\mathbf{E}}_{W}[\mathbf{Cov}(Y, h(X)|W)] &= \mathop{\mathbf{E}}_{W}[\mathbf{E}[Yh(X)|W] - \mathbf{E}[Y|W]\,\mathbf{E}[h(X)|W]] \\ &= \mathop{\mathbf{E}}_{W}[\mathbf{E}[h(X)^2|W] - g^h(W)^2] \qquad (1) \\ &= \mathbf{E}[(h(X) - g^h(W))^2]\end{aligned}$$

where (1) follows by the assumption that h is weakly calibrated to Y over W. Thus, the difference in losses simplifies to the squared difference between g^h and h. □

An alternative backward baseline for classification covariance. We present an additional backward baseline for classifiers that may be of interest when an underlying score function is not available. In this baseline, we manipulate the distribution over outcomes, leaving the predictions fixed. Specifically, given a sample $(X, W, Y) \sim \mathcal{D}$, we resample the outcome $\tilde{Y} \sim \mathcal{D}_{Y|W}$, ensuring that $h(X)$ and $\tilde{Y}$ are conditionally independent given W. We show that for the zero-one loss, the difference between the backward baseline $\ell_{\mathcal{D}}(\tilde{Y}, h(X))$ and $\ell(Y, h(X))$ is proportional to the expected conditional covariance of Y and $h(X)$ given W.

Proposition 5. *For any classifier $h : \mathcal{X} \to \{0, 1\}$,*

$$\mathbf{Pr}[\tilde{Y} \neq h(X)] - \mathbf{Pr}[Y \neq h(X)] = 2 \mathop{\mathbf{E}}_{W}[\mathbf{Cov}(Y, h(X)|W)]$$

Proof. We expand the difference in zero-one loss by exploiting the identity that $\mathbf{Pr}[Y \neq h(X)] = \mathbf{E}[Y + h(X) - 2Yh(X)]$ for binary Y and $h(X)$, and using the fact that $\mathbf{E}[\tilde{Y}|W] = \mathbf{E}[Y|W]$.

$$\begin{aligned}
\mathbf{Pr}[\tilde{Y} \neq h(X)] - \mathbf{Pr}[Y \neq h(X)] &= 2\left(\mathbf{E}[Yh(X)] - \mathbf{E}[\tilde{Y}h(X)]\right) \\
&= 2 \mathop{\mathbf{E}}_{W}[\mathbf{E}[Yh(X)|W] - \mathbf{E}[\tilde{Y}h(X)|W]] \\
&= 2 \mathop{\mathbf{E}}_{W}[\mathbf{E}[Yh(X)|W] - \mathbf{E}[Y|W]\,\mathbf{E}[h(X)|W]] \qquad (2) \\
&= 2 \mathop{\mathbf{E}}_{W}[\mathbf{Cov}(Y, h(X)|W)]
\end{aligned}$$

where (2) follows by the fact that $\tilde{Y} \sim \mathcal{D}_{Y|W}$ is sampled conditionally independently from the distribution on Y given W. □

B Backward Prediction and Demographic Parity

While conceived from different vantages, backward baselines and fair machine learning share similarities in perspective and technical structure. On a technical level, pure forward prediction is equivalent to demographic parity, a notion of fairness introduced by [DHP+12]. Based on this observation, certain insights about backward baselines have an analogue in fair prediction, and vice versa. For instance, we note that in combination, Proposition 2 and Proposition 3 imply that forward predictors cannot be calibrated to Y over W. Translating this observation into the language of fairness in prediction, we recover a specific case of the well-known results on the incompatibility of calibration and parity-based definitions of fairness in prediction [KMR17, Cho17].

In addition to giving insight into the backward rounding baseline, Proposition 4 shows a formal sense in which forward and backward predictors are orthogonal to one another. In particular, for weakly calibrated regression predictors $h : \mathcal{X} \to [0, 1]$,

$$\mathbf{E}[(Y - g^h(W))^2] = \mathbf{E}[(Y - h(X))^2] + \mathbf{E}[(h(X) - g^h(W))^2]$$

is a sort of Pythagorean theorem, stating that the variation in Y after accounting for $g^h(W)$ can be broken into the variation in Y given $h(X)$ and the variation in $h(X)$ given $g^h(W)$.

Connecting the backward baselines framework to fairness in prediction suggests a simple algorithm for learning predictors satisfying demographic parity, that relies only on unconstrained

learning primitives. First, we learn to predict Y as $h(X)$; then, we learn to predict $h(X)$ as $g^h(W)$; finally, we return $f_\alpha(X, W)$ defined as

$$f_\alpha(X, W) = h(X) - \alpha g^h(W)$$

for $\alpha \in [0, 1]$. Taking $\alpha = 1$ achieves a relaxed first-order demographic parity. Specifically, $f_1(X, W)$ has a constant expectation over all W.

$$\mathbf{E}[f_1(X, W)|W] = \mathbf{E}[h(X) - g^h(W)|W] = \mathbf{E}[h(X)|W] - g^h(W) = 0$$

In effect, $f_1(X, W)$ predicts optimally according to X then removes all variation that can be accounted for through W. Other choices of α may be interesting to interpolate between forward and backward prediction modes.

C Details on Empirical Evaluation

In this section, we show all figures for all baselines, classifiers, and metrics that we considered. We also provide additional details on the data sources, feature engineering, and target variable creation.

In all bar plots, the height of the bar is the mean value from 10 different random seeds and error bars indiciate a standard deviation across 10 different random seeds. In the case of ROC curves, the plot shows 10 curves overlaid from 10 different random seeds. None of the experiments require significant compute resources.

Given features X, context W, a given predictor $\hat{Y}$, and target variable Y, our plots evaluate five different methods:

- XYY : Train on (X, Y), test model on (X, Y)
- WYY : Train baseline on (W, Y), test baseline on (W, Y) (backward prediction baseline)
- $W\hat{Y}Y$: Train baseline on $(W, \hat{Y})$, test baseline on (W, Y) (equivalent to backward prediction baseline)
- $WY\hat{Y}$: Train baseline on (W, Y), test baseline on $(W, \hat{Y})$ (equivalent to backward rounding baseline)
- $W\hat{Y}\hat{Y}$: Train baseline on $(W, \hat{Y})$, test baseline $(W, \hat{Y})$ (backward rounding baseline)

In the main body of the paper we included only two baselines and ommitted the equivalent ones.

Models. We use three standard models available in the Python `sklearn` package. We do no or only minimal hyperparameter tuning:

- Gradient boosting: `GradientBoostingClassifier()`

- Random Forests: `RandomForestClassifier()`
- Logistic regression:
  `make_pipeline(StandardScaler(), LogisticRegression(max_iter=1000, tol=0.1))`

It is possible that other model families achieve better accuracy. However, on the kind of tabular data we experiment with ensemble methods such as random forests or gradient boosting tend to achieve state-of-the-art performance. We include a reference implementation of these five methods in Section D.

C.1 Medical Expenditure Panel Survey (MEPS)

For extensive documentation and background on this survey, see: https://www.meps.ahrq.gov/mepsweb/

Data sources and use conditions. Our dataset is constructed from the 2019 MEPS data. The MEPS 2019 Full Year Consolidated Data File (HC-216) is available online at https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-216. The same website contains extensive documentation regarding features and data collection. The MEPS data use agreement is available online: https://meps.ahrq.gov/data_stats/download_data/pufs/h216/h216doc.shtml#DataA.

Features. Features for Round 3 of Panel 23 and Round 1 of Panel 24 have a suffix of 31 or 31X in the data. We include all of these in the dataset, as well as all demographic features: FCSZ1231, FCRP1231, RUSIZE31, RUCLAS31, FAMSZE31, FMRS1231, FAMS1231, REGION31, REFPRS31, RESP31, PROXY31, BEGRFM31, BEGRFY31, ENDRFM31, ENDRFY31, INSCOP31, INSC1231, ELGRND31, PSTATS31, SPOUID31, SPOUIN31, ACTDTY31, RTHLTH31, MNHLTH31, CHBRON31, ASSTIL31, ASATAK31, ASTHEP31, ASACUT31, ASMRCN31, ASPREV31, ASDALY31, ASPKFL31, ASEVFL31, ASWNFL31, IADLHP31, ADLHLP31, AIDHLP31, WLKLIM31, LFTDIF31, STPDIF31, WLKDIF31, MILDIF31, STNDIF31, BENDIF31, RCHDIF31, FNGRDF31, ACTLIM31, WRKLIM31, HSELIM31, SCHLIM31, UNABLE31, SOCLIM31, COGLIM31, VACTDY31, VAPRHT31, VACOPD31, VADERM31, VAGERD31, VAHRLS31, VABACK31, VAJTPN31, VARTHR31, VAGOUT31, VANECK31, VAFIBR31, VATMD31, VAPTSD31, VALCOH31, VABIPL31, VADEPR31, VAMOOD31, VAPROS31, VARHAB31, VAMNHC31, VAGCNS31, VARXMD31, VACRGV31, VAMOBL31, VACOST31, VARECM31, VAREP31, VAWAIT31, VALOCT31, VANTWK31, VANEED31, VAOUT31, VAPAST31, VACOMP31, VAMREC31, VAGTRC31, VACARC31, VAPROB31, VACARE31, VAPACT31, VAPCPR31, VAPROV31, VAPCOT31, VAPCCO31, VAPCRC31, VAPCSN31, VAPCRF31, VAPCSO31, VAPCOU31, VAPCUN31, VASPCL31, VASPMH31, VASPOU31, VASPUN31, VACMPM31, VACMPY31, VAPROX31, EMPST31, RNDFLG31, MORJOB31, HRWGIM31, HRHOW31, DIFFWG31, NHRWG31, HOUR31, TEMPJB31, SSNLJB31, SELFCM31, CHOIC31, INDCAT31, NUMEMP31, MORE31, UNION31, NWK31, STJBMM31, STJBYY31, OCCCAT31, PAYVAC31, SICPAY31, PAYDR31, RETPLN31, BSNTY31, JOBORG31, OFREMP31, CMJHLD31, MCRPD31, MCRPB31, MCRPHO31, MCDHMO31, MCDMC31, PRVHMO31, FSAGT31, HASFSA31, PFSAMT31, MCAID31, MCARE31, GOVTA31, GOVAAT31, GOVTB31, GOVBAT31, GOVTC31,

GOVCAT31, VAPROG31, VAPRAT31, IHS31, IHSAT31, PRIDK31, PRIEU31, PRING31, PRIOG31, PRINEO31, PRIEUO31, PRSTX31, PRIV31, PRIVAT31, VERFLG31, DENTIN31, DNTINS31, PMEDIN31, PMDINS31, PMEDUP31, PMEDPY31, AGE31X, MARRY31X, FTSTU31X, REFRL31X, MOPID31X, DAPID31X, HRWG31X, DISVW31X, HELD31X, OFFER31X, TRIST31X, TRIPR31X, TRIEX31X, TRILI31X, TRICH31X, MCRPD31X, TRICR31X, TRIAT31X, MCAID31X, MCARE31X, MCDAT31X, PUB31X, PUBAT31X, INS31X, INSAT31X, SEX, RACEV1X, RACEV2X, RACEAX, RACEBX, RACEWX, RACETHX, HISPANX, HISPNCAT, EDUCYR, HIDEG, OTHLGSPK, HWELLSPK, BORNUSA, WHTLGSPK, YRSINUS

Demographic features. The full list of demographic features we use is:

- AGE31X
- SEX
- RACEV1X
- RACEV2X
- RACEAX
- RACEBX
- RACEWX
- RACETHX
- HISPANX
- HISPNCAT
- EDUCYR
- HIDEG
- OTHLGSPK
- HWELLSPK
- BORNUSA
- WHTLGSPK
- YRSINUS

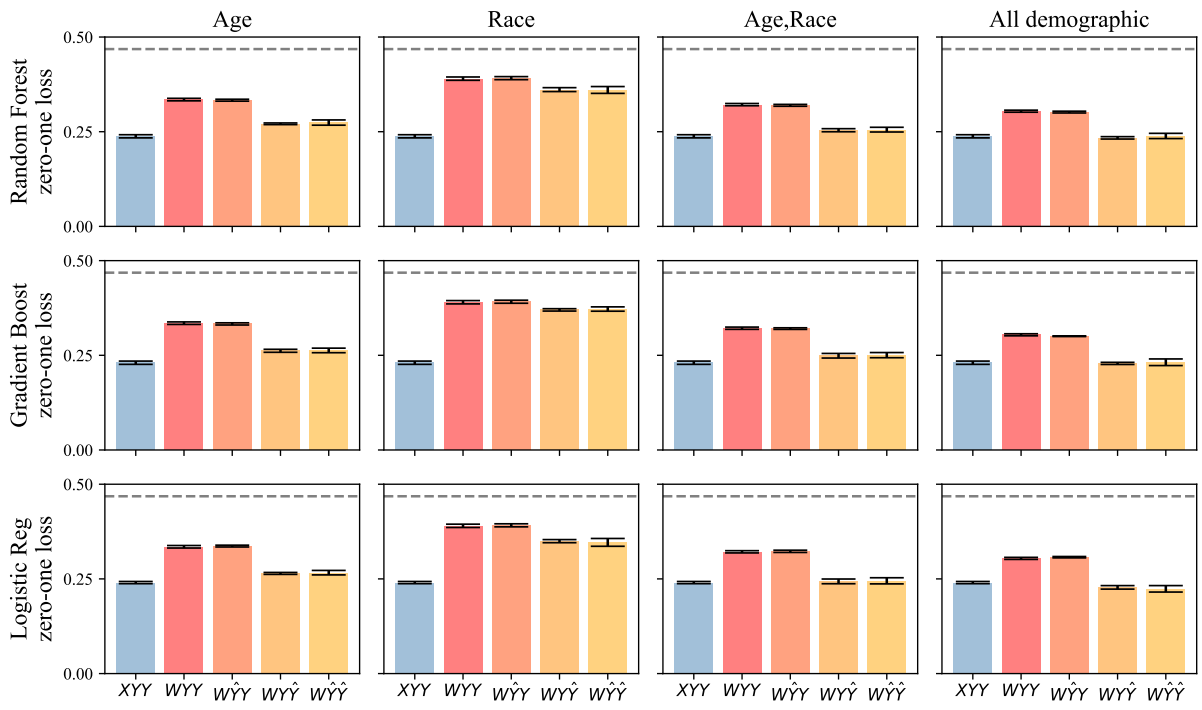


Figure 5: Baselines on MEPS for varying features and classifiers (zero-one loss)

Target variable. We construct the target variable by summing up the following features:

- OBTOTV19 — NUMBER OF OFFICE-BASED PROVIDER VISITS 2019
- OPTOTV19 — NUMBER OF OUTPATIENT DEPT PROVIDER VISITS 2019
- ERTOT19 — NUMBER OF EMERGENCY ROOM VISITS 2019
- IPNGTD19 — NUMBER OF NIGHTS IN HOSP FOR DISCHARGES, 2019
- HHTOTD19 — NUMBER OF HOME HEALTH PROVIDER DAYS 2019

We label all instances *positive* (1) where the sum is strictly greater than 3. We label all other instances *negative* (0). This leads to 53.17% positive instances. Hence, an all ones predictor achieves 46.83% classification error.

Full set of figures. Figure 5 shows all results for the zero-one loss, Figure 6 for the squared loss, and Figure 7 for ROC curves.

C.2 Survey of Income and Program Participation (SIPP)

Extensive documentation and background information on this survey is available from the websites of the US Census Bureau: https://www.census.gov/programs-surveys/sipp.html

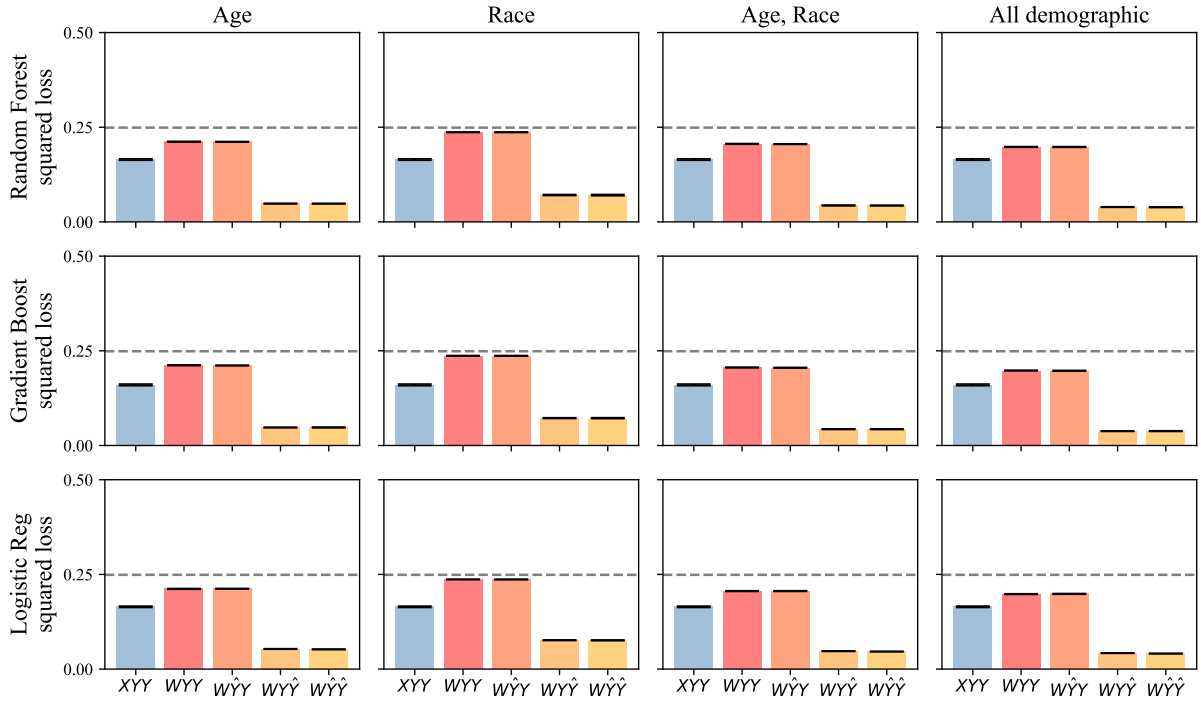


Figure 6: Baselines on MEPS for varying features and classifiers (squared loss)

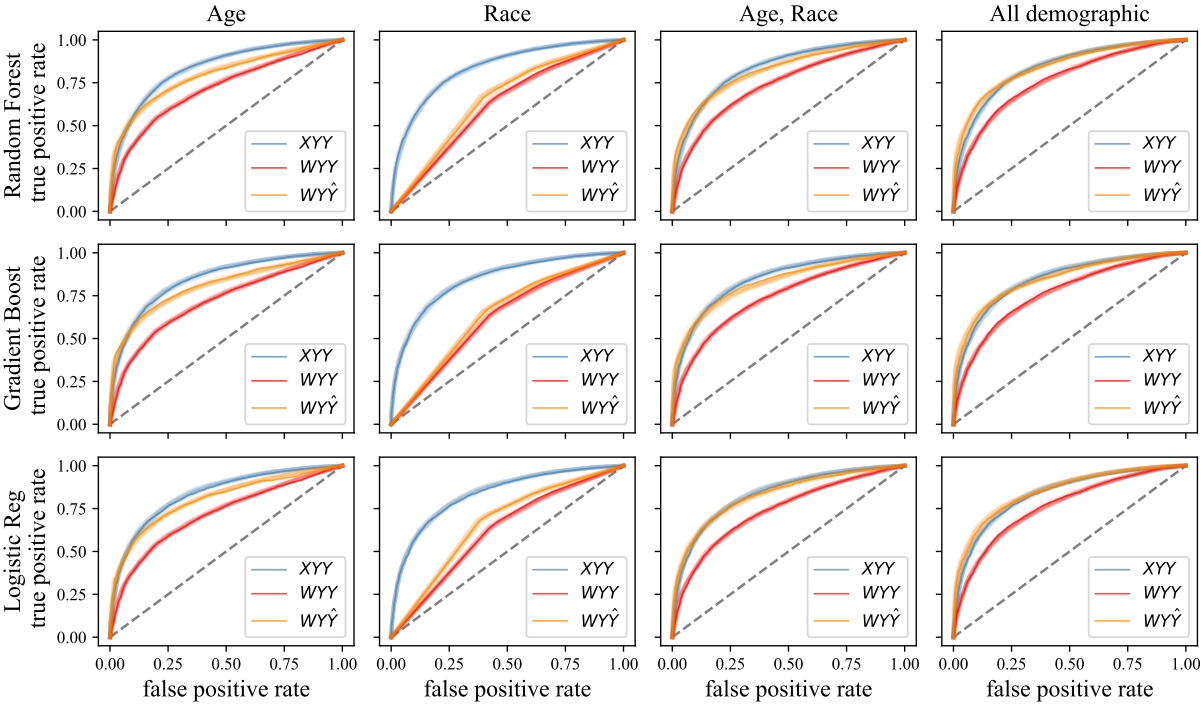


Figure 7: Baselines on MEPS for varying features and classifiers (ROC curves)

Data availability and conditions. The SIPP data provided by the US Census Bureau are in the public domain. We use the first two waves of the SIPP 2014 Panel data, available here:

- Wave 1: https://www.census.gov/programs-surveys/sipp/data/datasets/2014-panel/wave-1.html
- Wave 2: https://www.census.gov/programs-surveys/sipp/data/datasets/2014-panel/wave-2.html

Features. The dataset we derive from the 2014 SIPP panel data uses a set of 50 variables constructed from one or multiple variables appearing in the SIPP raw data in Wave 1. The list below shows each feature we use (in capital letters) followed by the original SIPP feature(s) it is derived from.

- LIVING_QUARTERS_TYPE : tlivqtr
- LIVING_OWNERSHIP : etenure
- SNAP_ASSISTANCE : efs
- WIC_ASSISTANCE : ewic
- MEDICARE_ASSISTANCE : emc
- MEDICAID_ASSISTANCE : emd
- HEALTHDISAB : edisabl
- DAYS_SICK : tdaysick
- HOSPITAL_NIGHTS : thospnit
- PRESCRIPTION_MEDS : epresdrg
- VISIT_DENTIST_NUM : tvisdent
- VISIT_DOCTOR_NUM : tvisdoc
- HEALTH_INSURANCE_PREMIUMS : thipay
- HEALTH_OVER_THE_COUNTER_PRODUCTS_PAY : totcmdpay
- HEALTH_MEDICAL_CARE_PAY : tmdpay
- HEALTH_HEARING : ehearing
- HEALTH_SEEING : eseeing
- HEALTH_COGNITIVE : ecognit
- HEALTH_AMBULATORY : eambulat

- HEALTH_SELF_CARE : eselfcare
- HEALTH_ERRANDS_DIFFICULTY : eerrands
- HEALTH_CORE_DISABILITY : rdis
- HEALTH_SUPPLEMENTAL_DISABILITY : rdis_alt
- AGE : tage
- GENDER : esex
- RACE : trace
- EDUCATION : eeduc
- MARITAL_STATUS : ems
- CITIZENSHIP_STATUS : ecitizen
- FAMILY_SIZE_AVG : rfpersons
- HOUSEHOLD_INC : thtotinc
- RECEIVED_WORK_COMP : ewc_any
- TANF_ASSISTANCE : etanf
- UNEMPLOYMENT_COMP : eucany
- SEVERANCE_PAY_PENSION : elmpnow
- FOSTER_CHILD_CARE_AMT : tfccamt
- CHILD_SUPPORT_AMT : tcsamt
- ALIMONY_AMT : taliamt
- INCOME_FROM_ASSISTANCE : tptrninc, tpscininc, tpothinc
- INCOME : tpprpinc, tptotinc
- SAVINGS_INV_AMOUNT : tirakeoval, tthr401val
- UNEMPLOYMENT_COMP_AMOUNT : tuc1amt, tuc2amt, tuc3amt
- VA_BENEFITS_AMOUNT : tva1amt, tva2amt, tva3amt, tva4amt, tva5amt
- RETIREMENT_INCOME_AMOUNT : tret1amt, tret2amt, tret3amt, tret4amt, tret5amt, tret6amt, tret7amt, tret8amt
- SURVIVOR_INCOME_AMOUNT : tsur1amt, tsur2amt, tsur3amt, tsur4amt, tsur5amt, tsur6amt, tsur7amt, tsur8amt, tsur11amt, tsur13amt

- DISABILITY_BENEFITS_AMOUNT : tdis1amt, tdis2amt, tdis3amt, tdis4amt, tdis5amt, tdis6amt, tdis7amt, tdis10amt
- FOOD_ASSISTANCE : efood_type1, efood_type2, efood_type3, efood_oth
- TRANSPORTATION_ASSISTANCE : etrans_type1, etrans_type2, etrans_type3, etrans_type4, etrans_oth
- SOCIAL_SEC_BENEFITS : esssany, esscany

These variables represent features derived from columns in the original data source via our own data cleaning and processing script. In particular, we discount columns that have more than 10% missing values.

Demographic features. The full list of six demographic features we use is:

- AGE
- GENDER
- RACE
- EDUCATION
- MARITAL_STATUS
- CITIZENSHIP_STATUS

Target variable. The target variable is constructed based on the feature thcyincpov in Wave 2, which reflects the household income-to-poverty ratio in the 2019 calendar year, excluding Type 2 individuals. Type 2 individuals are individuals that lives in the household for some month but no longer reside there.

We threshold thcyincpov at 3 so that all instances with thcyincpov strictly greater than 3 are labeled positive (1) and all others are labeled negative (0). This leads to 51.12% positive instances. Hence an all ones predictor has accuracy 48.88%.

Full set of figures. Figure 8 shows all results for the zero-one loss, Figure 9 for the squared loss, and Figure 10 for ROC curves.

C.3 ProPublica COMPAS Recidivism Scores

Data sources and use conditions. We use the COMPAS score dataset collected and made available by Problica [ALMK16], which is widely used througout the algorithmic fairness literature. The Propublica COMPAS score dataset is available online: `https://github.com/propublica/compas-analysis` The data repository does not specify a license or data use agreement.

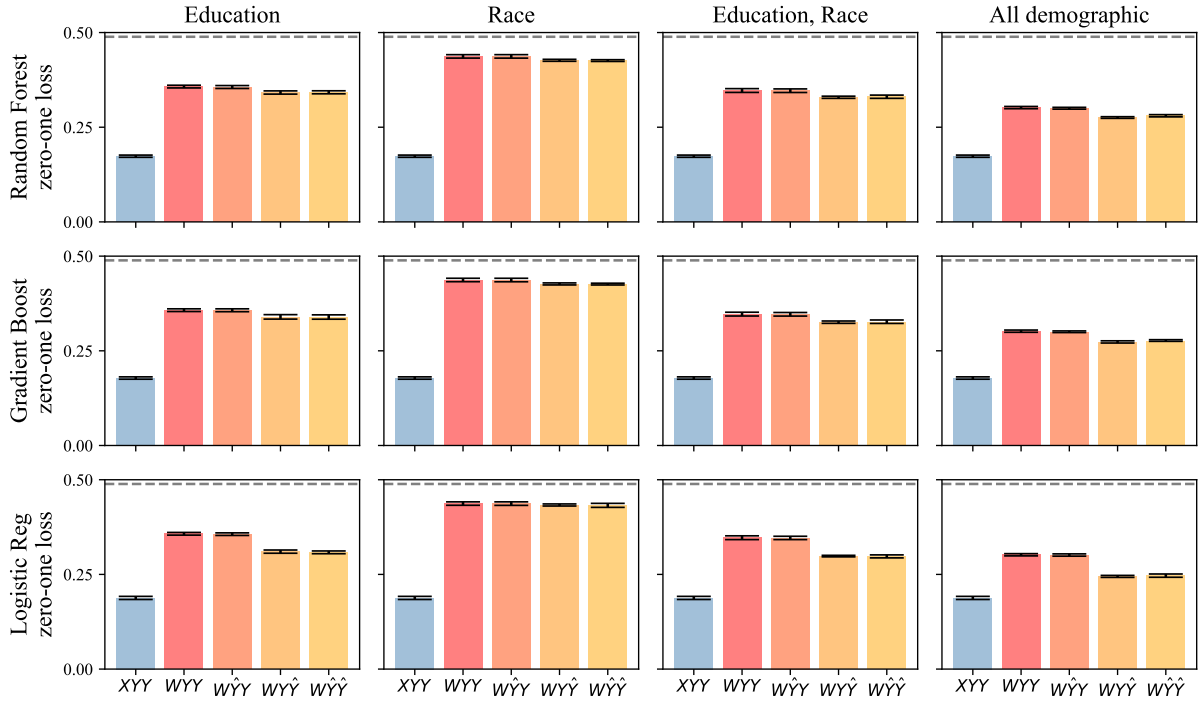


Figure 8: Baselines on SIPP for varying features and classifiers (zero-one loss)

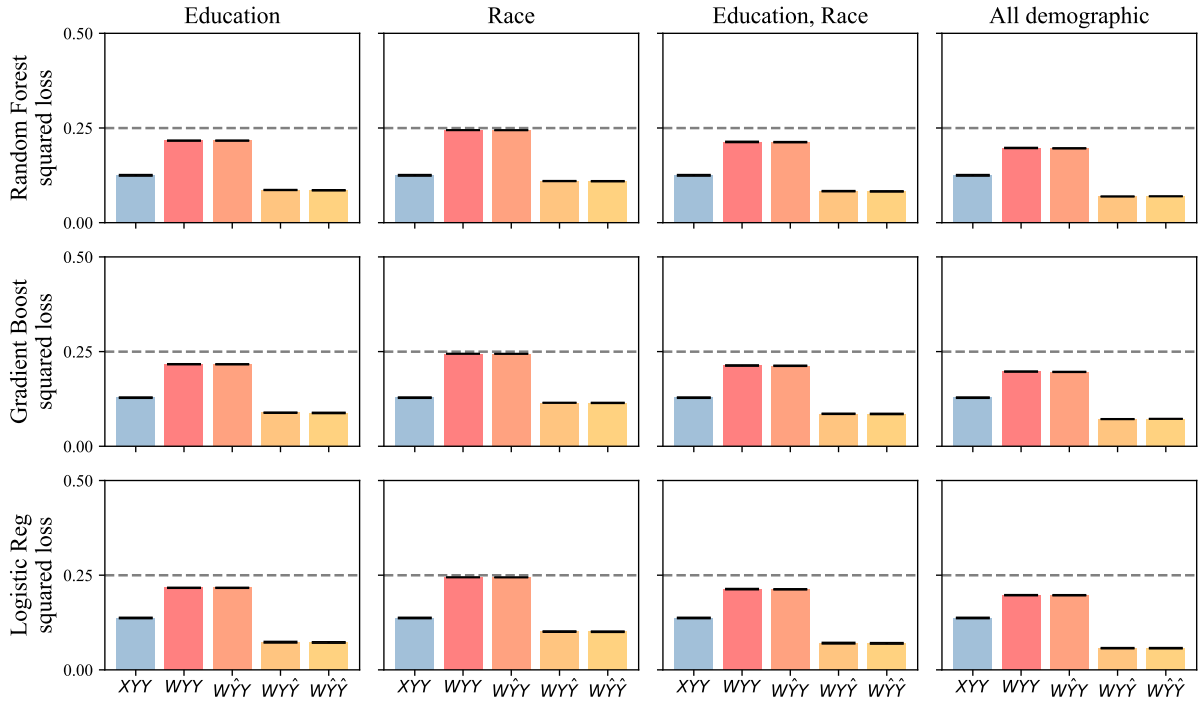


Figure 9: Baselines on SIPP for varying features and classifiers (squared loss)

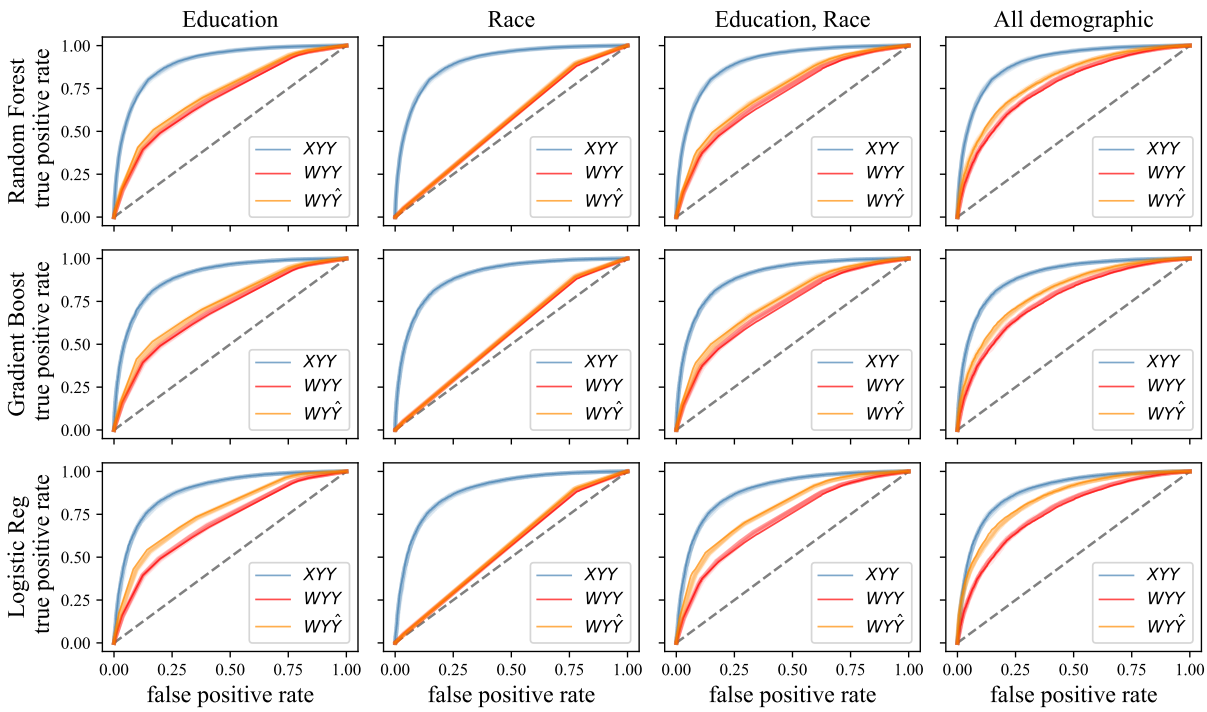


Figure 10: Baselines on SIPP for varying features and classifiers (ROC curves)

Demographic features. We use the following demographic features avilable in the dataset:

- 'race',
- 'age',
- 'juv_fel_count', 'juv_misd_count', 'juv_other_count' : juvenile priors
- 'prior_count'

Target variable. We use *two-year recidivism* ('two_year_recid') as the target variable.

Predictor. Since we lack training data, we instead audit COMPAS scores as a black-box. The column in the data corresponding to COMPAS scores is called 'decile_score' and provides score deciles. To obtain a predictor we fit a single-variable model to predict the target variable from the score deciles. This amounts to a recalibration of the score values to the target variable, ensuring that we obtain the best possible predictor we can from the score deciles.

Full set of figures. Figure 11 shows all results we report on the COMPAS dataset.

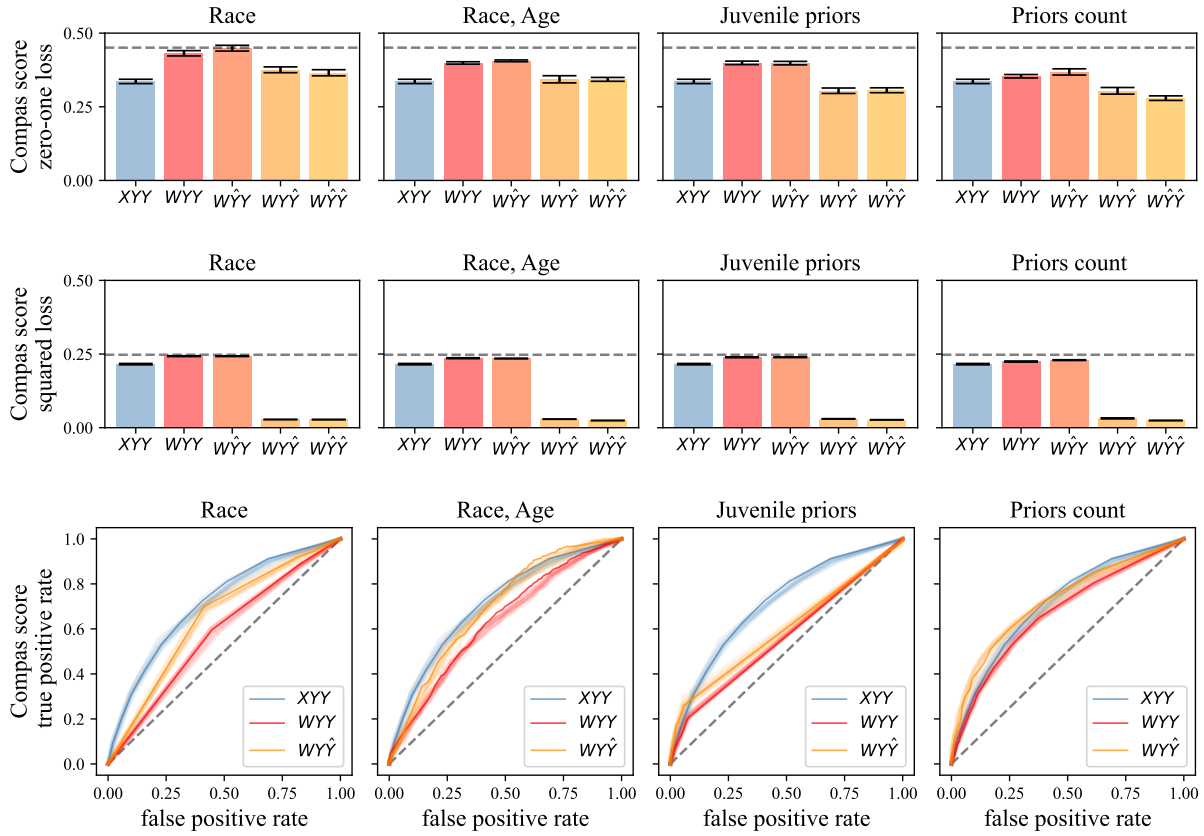


Figure 11: Baselines on COMPAS for varying features and metrics

D Reference implementation of backward baselines

```
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.metrics import accuracy_score
from sklearn.model_selection import train_test_split


def backward_baselines(X, y, features, model):
    """Compute backward baselines.

    Parameters
    ----------
    X : numpy.ndarray
        data matrix (n, d)
    y : numpy.ndarray
        target variable (n,)
    features : list
        list of column names
    model : object
        model supporting fit and predict

    Returns
    -------
    dict
        Scores of all backward baselines.
    """

    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33)

    scores = {}
    # XYY
    model.fit(X_train, y_train)
    scores['XYY'] = accuracy_score(model.predict(X_test), y_test)

    # WYY
    baseline = GradientBoostingClassifier()
    baseline.fit(X_train[features], y_train)
    scores['WYY'] = accuracy_score(baseline.predict(X_test[features]), y_test)

    # WY^Y
    baseline.fit(X_test[features], model.predict(X_test))
    scores['WY^Y'] = accuracy_score(baseline.predict(X_test[features]), y_test)

    # WYY^
    baseline.fit(X_test[features], y_test)
    scores['WYY^'] = accuracy_score(baseline.predict(X_test[features]),
                                    model.predict(X_test))

    # WY^Y^ requires new train/test split
    X_testA, X_testB, y_testA, y_testB = train_test_split(X_test[features],
                                           model.predict(X_test), test_size=0.5)
    baseline.fit(X_testA, y_testA)
    scores['WY^Y^'] = accuracy_score(baseline.predict(X_testB), y_testB)

    return scores
```