MILLIMAN WHITE PAPER

Trustworthy artificial intelligence in insurance: Navigating fairness and performance in predictive modeling

A technical guide for group fairness metrics and an illustration of fairness remediation results on real-world datasets

François Hu
Alexandre Boumezoued

Milliman

Predictive models rely heavily on often biased data shaped by human decisions, societal inequities, and, predominantly, representation bias. While no universal standard for fairness exists, it is widely acknowledged that simply removing sensitive attributes does not eliminate bias.

Therefore, predictive modeling requires ensuring algorithmic fairness, whether using traditional generalized linear models (GLM), Generalized additive model (GAMs), or other more complex predictive models. This paper discusses the legal and business imperatives for fairness in actuarial practices and introduces key concepts of group fairness. It also presents fairness mitigation methodologies designed to prevent legal issues while preserving predictive performance as much as possible with practical applications in insurance.

Insurance companies collect and store vast amounts of structured and unstructured data daily, including claim reports, claim characteristics, emails, and customer reviews. By leveraging information technology resources such as cloud storage, computing power, and the Internet of Things (IoT), these companies have significantly increased their use of predictive models driven by machine learning (ML) and artificial intelligence (AI). These technologies are key drivers in the insurance industry, offering numerous benefits such as improved risk segmentation, fraud detection, claims prediction, automated processes, and efficient decision-making. However, while these technologies greatly improve predictive performance, actuaries and data scientists are increasingly aware of their limitations in providing unbiased, robust, and transparent estimates. These methods often struggle to prevent unjustified discrimination against groups protected due to sensitive attributes such as religion, race, or gender, which can lead to ethical concerns and pose substantial reputational and legal risks for organizations. More specifically, according to "Google's Machine Learning Glossary,"[1] a sensitive attribute can be defined as "a human attribute that may be given special consideration for legal, ethical, social, or personal reasons."

Reputational and legal risks of unjustified discrimination

As AI technologies are increasingly deployed and used, many decision makers base their analyses and decisions on scores or predictions generated by predictive models. However, the reliance on these algorithm-driven decisions has brought to light numerous reputational risks worldwide, significantly impacting the industry. Recent investigations have highlighted biases in algorithms used by public and private organizations. For example, a study published in Science found that a commercial health score prediction algorithm used in hospitals was biased against Black patients, who were sicker than white patients at the same risk score[2]. Another study analyzed racial discrimination in home insurance claims resolution, with claims from predominantly Black regions less likely to be compensated.[3]

Alongside these rising ethical and reputational risks, there is an increasing focus on revising regulatory requirements so that the use of predictive models and their underlying data is carefully overseen. Actuaries and data scientists are now shifting towards providing oversight for complex model analysis, ensuring that predictive models are explainable, ethical, and compliant with regulatory standards. Specifically, the regulatory landscape has evolved, with variations depending on the country or region where the AI is developed and deployed. Without providing exhaustive coverage, we highlight some recent regulatory developments in the European Union (EU) and United States (U.S.) below.

THE EU

For EU insurers, the fairness principle is an obligation. A notable example is the 2004 EU Goods and services directive by the Council of the European Union[4], which mandates equal treatment between men and women (known as "unisex") in accessing and providing goods and services, including insurance services. However, some exceptions in insurance premiums remained in place until December 2012, when the unisex pricing directive in the insurance industry began to be fully enforced by the European Court of Justice. Specifically, this means men and women with the same characteristics should pay the same price for the same product. Recently in data processing and predictive modeling, regulatory advancements in the EU regarding the ethical development include:

1) **General data protection regulation (GDPR):** With the democratization of data-based algorithms and procedures, the GDPR came into force across the EU in May 2018. Specifically, this regulation aims to protect the privacy of personal data, including sensitive attributes such as ethnicity and health conditions. According to Article 5(1) of the GDPR[5], personal data shall be "processed lawfully, fairly and in a transparent manner in relation to the data subject."

2) **European AI act (EU AI Act)**: With the rise of AI systems and generative AI models such as OpenAI's ChatGPT, the EU AI Act has been introduced to regulate them based on their level of risk, leading to different obligations for the companies. Unacceptable risks, such as social scoring, are prohibited. High-risk AI applications, such as those used in healthcare, require strict regulation to ensure transparency and mitigate algorithmic biases. Limited-risk applications, such as chatbots, must be documented and have some lighter transparency obligations, while other safe applications have minimal requirements. Specifically, in the insurance sector, health and life insurance are heavily affected as ML-based scoring is considered high-risk when using AI systems. Furthermore, according to Recital 44, "to protect the right of others from the discrimination that might result from the bias in AI systems, the providers should be able to process special categories of personal data […] to ensure the bias monitoring, detection and correction in relation to high-risk AI systems." We observe that general-purpose AI such as ChatGPT has a category of its own and we refer to the Regulation (EU) 2024/1689[6] of the European Parliament, for more information.

Furthermore, we note that insurance is already a highly regulated domain with notable overlap between the above legal requirements and existing insurance regulations for AI/data. For example, the fairness principle is also recognized in Article 17(1) of the Insurance distribution directive[7] where it states that insurance distributors shall "always act honestly, fairly and professionally in accordance with the best interests of their customers." Similarly, the Solvency II framework, particularly Delegated regulation 2015/35, includes detailed provisions on data quality, model validation, model calibration, documentation, and record-keeping.

THE U.S.

For insurers in the U.S., fairness requirements can vary depending on the state and the insurance domain (e.g., property and casualty [P&C] vs. life insurance). While there is no universal consensus on fairness, discrimination based on race, ethnicity, and religion (sometimes called the "big three") is generally prohibited. Beyond these sensitive attributes, scholars and actuaries have observed that certain attributes may be prohibited in some states or specific insurance domains but permitted in others. For instance, using genetic information is typically permitted in most states within P&C practices but is strictly prohibited in health insurance. Figure 1, adapted from Avraham (2013) and Charpentier (2024) in the book Insurance, biases, discrimination and fairness, highlights the attributes considered sensitive across different lines of insurance in the U.S.[8]

FIGURE 1: U.S. INSURANCE ANTIDISCRIMINATION LAWS

	AUTO	P&C	DISABILITY	HEALTH	LIFE
Race (or origin)	x	x	x	x	x
Religion	x	x	x	x	x
Gender	o	●	o	o	●
Sexual orientation	●	●	o	o	o
Age	o	●	o	o	●
Credit score	o	o	●	●	●
Zip code	o	o	●	●	●
Genetics	●	●	o	x	o

A characteristic is classified as ● if it is permitted without specific regulation, x if it is prohibited, and o if its status varies across states.

The recent enforcement of fairness in insurance is particularly evident in the U.S. We outline some key fairness principles in insurance, particularly in Colorado and New York.

1. **Colorado:** The Division of Insurance (DOI) issued a regulation effective November 14, 2023[9] to prevent race-based discrimination in life insurers' use of AI models. This regulation aims to govern life insurers that use external consumer data and information sources (ECDIS) that could lead to discriminatory models. ECDIS includes data sources such as credit scores, social media habits, locations, purchasing habits, home ownership, educational attainment, occupation, licensures, civil judgments, and court records. The regulation required all Colorado licensed life insurers to submit a compliance progress report by June 1, 2024, and an annual compliance attestation starting December 1, 2024. Notably, the DOI does not require insurers to collect data on the race or ethnicity of their policyholders. Instead, the DOI prescribes using Bayesian improved first name surname geocoding to infer policyholders' race. This statistical method uses geodemographic data from the 2010 U.S. census and lists of common first names and surnames associated with various races and ethnicities to provide probabilistic predictions. The DOI has determined this method to be adequate for testing insurance models for compliance with the law.
2. **New York:** On July 11, 2024, the New York State Department of Financial Services issued the AI Circular Letter (Circular Letter No. 7)[10], regulating the use of ECDIS and AI systems in insurance underwriting and pricing. New York licensed insurers must evaluate whether ECDIS correlates with protected class status, potentially causing unfair discrimination. If correlations are found, insurers must (1) justify using ECDIS as a legitimate business necessity and (2) annually search for less discriminatory alternative variables or methodologies that would meet their business needs. Insurers are required to maintain comprehensive documentation for their use of all AI systems, including the ECDIS that were used. This documentation should include descriptions of testing conducted at least annually to assess the output of AI models, including any drift resulting from ML or other automated updates.

Despite numerous regulatory developments worldwide, there is no clear consensus on a gold standard for evaluating unwanted discrimination in predictive models. Various fairness notions exist for specific applications but they are often incompatible with each other. In the field of algorithmic fairness, ongoing developments focus on local (individual-level) and global (group-level) fairness approaches. These are commonly known as individual fairness and group fairness, respectively.

Individual fairness assesses whether a model would make the same prediction if the same individual were part of the privileged group. It ensures that similar individuals receive similar outcomes, regardless of their group membership. This article instead focuses on group fairness, which examines whether individuals in privileged and unprivileged groups are treated differently. Group fairness aims to ensure that groups defined by protected attributes, such as race or gender, receive equitable treatment overall.

Main notions of (group) fairness and bias diagnosis

Before detailing the main approaches to mitigate unfairness, this section introduces the main notions of group fairness. This provides a foundation for understanding how to evaluate and address biases in predictive models.

For this purpose, we use the following notations: Let $\boldsymbol{X}$ represent the non-sensitive features, $\boldsymbol{S}$ the sensitive attribute we want to ensure fairness for, $\boldsymbol{Y}$ the true response and $\widehat{\boldsymbol{Y}}$ the predictive response of the model. Within the technical developments that follow, we assume that the sensitive attribute is available and can be used for testing fairness and creating a predictive model, denoted as $\hat{\boldsymbol{f}}$.

For simplicity, we consider a binary classification framework where $\widehat{\boldsymbol{Y}}, \boldsymbol{Y} \in \{\boldsymbol{0}, \boldsymbol{1}\}$ with binary sensitive attribute $\boldsymbol{S} \in \{-\boldsymbol{1}, \boldsymbol{1}\}$. Without loss of generality, we can assume $\boldsymbol{S}$ represents gender. Additionally, let $\widehat{\boldsymbol{Z}} \coloneqq \hat{\boldsymbol{f}}(\boldsymbol{X}, \boldsymbol{S})$ be the estimated probability score of the event $\{\widehat{\boldsymbol{Y}} = \boldsymbol{1}\}$ before applying any threshold.

We categorize fairness notions into three main sets: independence, separation, and sufficiency. Each is detailed below with examples in the insurance context.

INDEPENDENCE

The independence notion of fairness, or demographic parity (DP), is likely the most intuitive and widely used. It compares whether the model predictions (or scores) are similar between different populations of the sensitive attribute. Formally, this is expressed as:

$$\hat{Y} \perp S.$$

This means that predicted outcomes should be independent of the sensitive attribute, ensuring equitable treatment of individuals from different groups by the model. More specifically, in the binary classification assumption, the DP criterion states:

$$\mathbb{P}\big(\hat{Y} = 1 \big| S = 1\big) = \mathbb{P}\big(\hat{Y} = 1 \big| S = -1\big).$$

As an example, in the case of a fraud scoring or default risk (such as a loan) algorithm, if the average score for women is 2%, and the average score for men is also 2%, the algorithm is considered "gender-fair" in terms of DP.

In the algorithmic fairness literature, this notion of fairness is known as a weak DP. If we denote the conditional cumulative distribution function (CDF) $F_s(t) = \mathbb{P}\big(\hat{Z} \leq t \big| S = s\big)$, expressed with estimated scores, the stronger counterpart of the previous DP version is:

$$F_1(t) - F_{-1}(t) = 0 \text{ for all } t \in [0, 1].$$

We define the following (strong) DP unfairness measure:

$$U_{DP}\big(\hat{f}\big) \coloneqq \max_{t \in [0,1]} \left|F_1(t) - F_{-1}(t)\right|,$$

where the predictive model $\hat{f}$ is (strongly) DP fair if and only if $U_{DP}\big(\hat{f}\big) = 0$. The empirical version of this quantity is well-established in statistics as it relates to the nonparametric Kolmogorov–Smirnov (KS) test, which compares two samples to assess the similarity between their empirical CDFs, denoted $\hat{F}_s$. In this context, $\widehat{U}_{DP}$, the empirical version of U_{DP}, corresponds to the D statistic of the KS test.

This U_{DP} metric is suitable when the response is expected to be independent of the sensitive variable or when observations are biased against the unprotected population. Such bias may arise from historical or societal inequalities reflected in the data.

However, this metric does not account for the actual risk levels observed in different groups and is therefore unsuitable if the risk factors are dependent on the sensitive attribute. More generally, if the true response Y is dependent on the sensitive attribute S, the notion of independence $\hat{Y} \perp S$ may be inappropriate. Consequently, we can consider an alternative to the independence criterion through the so-called "separation" criterion, which incorporates the true response.

SEPARATION

The separation notion of fairness evaluates whether the model's predictions (or scores) are consistent across different populations defined by the sensitive attribute, conditional on the true response. This is expressed as:

$$\left(\hat{Y} \perp S\right) \mid Y.$$

This means that the model's predictive performance should be consistent across different values of the sensitive attribute. In binary classification, the separation criterion leads to subnotions of fairness such as “equal opportunity” and “equalized odds” (the latter will be denoted by “EO” in this paper). Equal opportunity requires that the true positive rates (TPR) be equal across groups defined by the sensitive attribute. More formally:

$$\mathbb{P}\left(\hat{Y} = 1 \middle| S = 1, Y = 1\right) = \mathbb{P}\left(\hat{Y} = 1 \middle| S = -1, Y = 1\right) \text{ (EqTPR).}$$

For example, in a car insurance model estimating the probability of a serious claim, if the TPR is 87% for women and men, the algorithm is considered gender-fair in terms of equal opportunity.

Instead of only ensuring equal chances of correct classification, one can also ensure that false positive rates (FPRs) are equal across different values of the sensitive attribute, ensuring equal chances of incorrect classification:

$$\mathbb{P}\left(\hat{Y} = 1 \middle| S = 1, Y = 0\right) = \mathbb{P}\left(\hat{Y} = 1 \middle| S = -1, Y = 0\right) \text{ (EqFPR).}$$

If both criteria (EqTPR) and (EqFPR) are satisfied simultaneously, the model achieves EO fairness, providing a more nuanced approach by balancing both types of errors. Let us denote the conditional TPR $TPR_s\left(\hat{f}\right) = \mathbb{P}\left(\hat{Y} = 1 \middle| S = s, Y = 1\right)$ and the conditional FPR $FPR_s\left(\hat{f}\right) = \mathbb{P}\left(\hat{Y} = 1 \middle| S = s, Y = 0\right)$. Then the (weak) EO unfairness measure is defined by:

$$U_{EO}\left(\hat{f}\right) \coloneqq max\left(\left|TPR_1\left(\hat{f}\right) - TPR_{-1}\left(\hat{f}\right)\right| , \left|FPR_1\left(\hat{f}\right) - FPR_{-1}\left(\hat{f}\right)\right| \right)$$

Like DP, these criteria are considered a weaker form of fairness. The stronger form involves evaluating fairness based on the estimated score $\hat{Z}$, which includes assessing the equality of receiver operating characteristic (ROC) curves or equality of area under the ROC curve (AUC-ROC or AUROC) for different values of the protected attribute. This aspect is beyond the scope of this paper.

An example of EO is the following: If the rate of loan acceptance, from a credit lending algorithm, among those who would have repaid and those who would not have repaid is similar for men and women, the algorithm is considered gender-fair in terms of EO.

This notion of fairness is appropriate when the response is linked to the sensitive attribute and when observations are not biased with regard to the sensitive attribute (e.g., no historical or societal bias). EO ensures that errors made by the model do not disproportionately impact any one group. For instance, such errors may arise from sampling bias (e.g., when certain groups are underrepresented in the data) or representation bias (e.g., when features related to a sensitive attribute are unevenly distributed) contained in the data.

SUFFICIENCY

Another commonly used fairness criterion is sufficiency. This notion requires true outcome parity among individuals who receive the same decision from the model, regardless of the value of the sensitive attribute. This is expressed as:

$$\left(Y \perp S\right) \mid \hat{Y}.$$

While separation deals with error rates in terms of the fraction of errors over the ground truth, such as the number of individuals whose loan requests are likely to be denied by the model among those who would have repaid, sufficiency considers the decision maker’s perspective. Specifically, it looks at the number of individuals who will not repay the loan among those who are given the loan. In the binary classification framework, the most frequently used fairness subnotion of sufficiency is predictive parity (PP), also known as outcome test, defined as:

$$\mathbb{P}\left(Y = 1 \middle| S = 1, \hat{Y} = 1\right) = \mathbb{P}\left(Y = 1 \middle| S = -1, \hat{Y} = 1\right),$$

or equivalently,

$$TDR_1\left(\hat{f}\right) = TDR_{-1}\left(\hat{f}\right),$$

where $TDR_s(\hat{f}) \coloneqq \mathbb{P}(Y = 1 | S = s, \hat{Y} = 1)$ denotes the true discovery rate (TDR). This metric measures the proportion of true positives among all positive predictions made by the model for group $\{S = s\}$. Conversely, the false discovery rate (FDR) for group $\{S = s\}$, defined as $FDR_s(\hat{f}) \coloneqq \mathbb{P}(Y = 1 | S = s, \hat{Y} = 0)$ looks at the proportion of false positives among all positive predictions made by the model. The above PP notion of fairness is later extended to consider the FDR as well:

$$\mathbb{P}(Y = 1 | S = 1, \hat{Y} = y) = \mathbb{P}(Y = 1 | S = -1, \hat{Y} = y), y \in \{0, 1\}$$

Using the quantities we defined, the proposed PP unfairness measure is written as

$$U_{PP}(\hat{f}) \coloneqq max(\left|(TDR_1 - TDR_{-1})(\hat{f})\right|, \left|(FDR_1 - FDR_{-1})(\hat{f})\right|).$$

A stronger version of PP is calibration parity (CP), which requires conditioning directly on the estimated score $\hat{Z}$ rather than the prediction $\hat{Y}$. It is formally expressed as:

$$\mathbb{P}(Y = 1 | S = 1, \hat{Z} = z) = \mathbb{P}(Y = 1 | S = -1, \hat{Z} = z), z \in [0,1].$$

EXAMPLE OF CP

Let us recall the 2019 Science article cited above that highlighted a health score prediction algorithm used in hospitals. This study, relevant to the life and health insurance domains, found that Black patients were sicker than white patients despite having the same risk score. This is an example where CP is not verified as the true outcomes differed significantly between racial groups despite having the same model outcome.

In Figure 2, we outline the advantages and disadvantages of each fairness metric.

FIGURE 2: STRENGTHS AND LIMITATIONS OF FAIRNESS METRICS

	OBJECTIVE	LIMITATIONS	FORMULA
Independence (Demographic parity)	Ensures equal outcomes across sensitive groups, ignoring all other factors	Does not make use of the true target	$\hat{Y} \perp S$
Separation (Equalized odds)	Ensures the same performance and/or the same error rates across sensitive groups (focuses on those affected by decisions)	The true outcome should not be biased	$(\hat{Y} \perp S) \mid Y$
Sufficiency (Predictive parity)	Requires equal treatment for individuals receiving the same model prediction, irrespective of sensitive attributes (focuses on the decision maker's perspective)	Systemic biases in the data (e.g., historical discrimination) can still result in unfair outcomes	$(Y \perp S) \mid \hat{Y}$

IMPOSSIBILITY THEOREM IN GROUP FAIRNESS

In a binary classification framework, except in rare cases, any two of the three fairness criteria (independence, separation, and sufficiency) are mutually exclusive. More specifically,

- If S and $\hat{Y}$ are dependent, the exact **DP and EO** are incompatible, i.e., $\left(U_{DP}(\hat{f}), U_{EO}(\hat{f})\right) \neq (0, 0)$.
- If S and Y are dependent, the exact **DP and PP** are incompatible, i.e., $\left(U_{DP}(\hat{f}), U_{PP}(\hat{f})\right) \neq (0, 0)$.
- If S and Y are dependent, the exact **EO and PP** are incompatible, i.e., $\left(U_{EO}(\hat{f}), U_{PP}(\hat{f})\right) \neq (0, 0)$.

Nevertheless, rather than attempting to meet multiple exact fairness criteria, such as $\left(U_{DP}(\hat{f}), U_{EO}(\hat{f}), U_{PP}(\hat{f})\right) = (0, 0, 0)$ or closely approaching accepted thresholds (e.g., 0.05) with $U_{DP}(\hat{f}) \leq 0.05, U_{EO}(\hat{f}) \leq 0.05$ and $U_{PP}(\hat{f}) \leq 0.05$, which may be incompatible, instead, it is more appropriate to:

1) select the most relevant fairness criterion for the specific application, with the target of $\mathrm{U_{EO}}(\hat{\mathrm{f}}) \leq 0.05$;
2) relax these criteria and focus on effectively achieving several approximate fairness measures such as $\left(\mathrm{U_{DP}}(\hat{\mathrm{f}}), \mathrm{U_{EO}}(\hat{\mathrm{f}}), \mathrm{U_{PP}}(\hat{\mathrm{f}})\right) = (0.1, 0.1, 0.1)$.

Main techniques of bias remediation

There are three primary approaches to achieving fair predictions: Pre-processing methods address bias in the data before applying predictive models, in-processing methods reduce bias during the model calibration, and post-processing methods enforce fairness by adjusting model outcomes in a final step.

PRE-PROCESSING APPROACHES

In this section, we present the pre-processing approaches based on the reweighting scheme proposed by Kamiran and Calders in 2012[11].

Focus on the reweighting instances methodology

Kamiran and Calders (2012) propose a method to reduce DP unfairness by adjusting the weights of training instances based on group membership, promoting fairness in the model's predictions across different sensitive groups. Instead of assigning equal weights to all observations, they generate a weight $w(s)$ for each value s of the sensitive attribute to satisfy the independence criterion. These weights are designed to ensure

$$\mathbb{P}(Y = 1, S = s) = w(s)\mathbb{P}(Y = 1)\,\mathbb{P}(S = s),$$

thereby enforcing DP fairness. Empirically, given n observations $\{(x_i, s_i, y_i)\}_{i=1,\dots,n}$, the weights are approximated by the plug-in formula:

$$\hat{w}(s) = \frac{1}{n} \times \frac{\sum_i 1_{y_i=1} \times \sum_i 1_{s_i=s}}{\sum_i 1_{y_i=1, s_i=s}}.$$

During the training of any predictive model, the weight $\hat{w}(s)$ is applied to each instance (x, s).

IN-PROCESSING APPROACHES

Classical predictive models aim to minimize predictive errors or maximize accuracy. To incorporate fairness metrics, two main approaches are commonly used: model-specific and model-agnostic approaches. A model-specific approach leverages the architecture of the predictive model, such as adversarial debiasing introduced by Zhang et al. (2018)[12]. This method enforces DP and EO fairness in neural network architectures using adversarial training, involving a primary model trained for a specific task and a bias correction model trained to reduce bias in the primary model's predictions. Conversely, a model-agnostic approach can be applied to a wide variety of predictive models. This article focuses on the latter, specifically the method proposed by Agarwal in 2018[13] (and later in 2019[14] for regression methods), which adds a penalty term to the objective function to achieve a trade-off between accuracy and fairness.

Focus on the exponentiated gradient reduction (EGR) methodology

Let us focus on EO fairness. Formally, the original optimization problem is expressed as:

$$\hat{f} \in \operatorname{argmin}_f R(f) \text{ such that } f \text{ is EO-fair},$$

with R representing the risk function, typically a quadratic risk for regression tasks and misclassification risk for classification tasks. The authors propose EGR, a reduction-based algorithm that incorporates either DP or EO fairness constraints into a sequence of standard cost-sensitive classification tasks. In each iteration, the algorithm reweights biased data points using Lagrange multipliers, which impose higher penalties on errors that violate fairness constraints the most. A new classifier is then trained on the adjusted dataset, therefore progressively reducing bias. Note that such an approach could be extended to other fairness constraints like PP fairness.

POST-PROCESSING APPROACHES

Post-processing techniques adjust a model's predictions to ensure fairness, without modifying the training data or the model itself. Recent methods apply optimal transport theory for DP-fairness in both regression and classification, as demonstrated by Chzen et al. (2020)[15] and Hu et al. (2024)[16]. Additionally, constraint-based methodologies, such as those developed by Denis et al.(2024)[17] and Alghambi (2022)[18], are used to ensure DP or EO fairness. Finally, methodologies such as those developed by Zeng et al.(2022)[19] introduce groupwise thresholding rules (GWTR) to ensure fairness in final binary outcomes, even when the underlying scores remain biased.

Focus on GWTR in Zeng et al. (2022)

The authors demonstrate that if differences in group performance are moderate, as defined by the sufficient condition:

$$\min_s \mathbb{P}(Y = 1 | \hat{Z} \geq c, S = s) \geq \max_s \mathbb{P}(Y = 1 | S = s),$$

where $\hat{Z}$ represents the score for the individual x belonging to group s, and c is a cost-related parameter, then all optimal classifiers satisfying PP fairness follow GWTR. In other words, the decision threshold for determining the final outcome is entirely dependent on the individual's sensitive group, ensuring that fairness constraints are met at the classification level.

For further details on optimization, we refer to Zeng et al.'s original paper.

COMPARING PRE-, IN- AND POST-PROCESSING APPROACHES

Pre-processing methods promote group fairness early in the ML lifecycle by addressing bias at the data level, offering simplicity and model-agnostic application; however, they often underperform compared to in- and post-processing techniques, especially given that they do not provide control on potential bias further introduced by the modeling process itself.

In-processing methods directly optimize fairness constraints during training and preserve the original model structure. For example, if the modeler relies on a generalized linear model (GLM), the fair model after in-processing is still a GLM model, which preserves some form of transparency with access to coefficients. However, in-processing methods can be computationally demanding and require access to the training process.

Post-processing methods are the fastest and most flexible as they can be applied to any pre-trained classifier without altering the training phase. However, they trade off model structure preservation—i.e., a GLM is no longer a GLM after post-processing—potentially impacting model transparency and interpretability.

In Figure 3, we outline the advantages and disadvantages of each approach.

FIGURE 3: STRENGTHS AND LIMITATIONS OF PRE-, IN-, AND POST-PROCESSING APPROACHES

	OBJECTIVE	PROS	CONS
Pre-processing	Generate a corrected dataset by modifying existing data or adding synthetic data	Model-agnostic Mitigates biases at the beginning	No direct control over biases in the model's output
In-processing	Incorporate fairness constraints into the objective function used for model calibration	Remains within the same model class: A GLM is still a GLM after fairness mitigation	Model-specific and requires greater time complexity
Post-processing	Modify model outputs after training to reduce bias, typically by adjusting decision thresholds or reweighting scores	Direct control over biases Model-agnostic	The final predictions do not remain within the original model class

Case studies

This section examines fairness diagnosis and remediation across various actuarial use cases, using datasets from bank data, car insurance (French motor personal line [freMPL]), and mortality data (global open source severity of illness score [GOSSIS]). For each dataset, we explore the principles of DP, EO, and PP fairness, using GLM, GBM, and "black-box" scoring as the respective predictive models.

DP USING PRE-PROCESSING

The bank dataset contains information on 30,000 credit card customers, including demographic details such as education level, marital status, and gender. Financial data such as credit card limits, repayment status, spending habits, transaction frequency, and median purchases are also included. We focus on evaluating the DP unfairness measure using $\widehat{\mathrm{U}}_{\mathrm{DP}}$ for the base and fair scores on this dataset, where the goal is to produce gender-fair scores for loan defaults. Specifically, we apply a pre-processing approach based on the reweighting strategy described earlier to generate fair probability scores and compare them to the base probability scores from the logistic regression model.

Figure 4 uses kernel density estimation (KDE) graphs to compare model score distributions by gender. The left graph represents the base model without fairness mitigation while the right graph shows the fair model after mitigation. KDE smooths data points into continuous curves, making visualizing differences in score distributions easier. Group 1 represents men, and Group 2 represents women. In this framework, we observe in the left graph that DP fairness is not achieved in the reference model, which predicts a higher probability of loan default for men than women. Using the reweighting approach introduced earlier, as shown in the right graph, we achieve our main goal by significantly reducing DP unfairness by approximately seven times (from 0.14 to 0.02) with minimal impact on predictive performance and maintaining an AUC of 0.72. Overall, the right graph also shows that all fairness metrics (DP, EO, PP) are close to or below the 5% threshold, except for PP-fairness.

FIGURE 4: STRONG DP ANALYSIS: MODEL OUTPUT COMPARISON

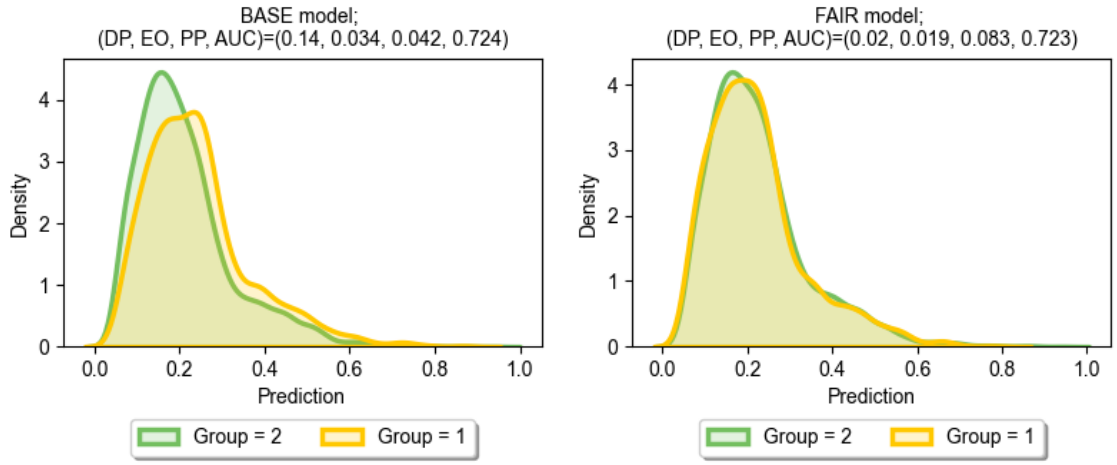

Figure description: Comparison of model distributions for Group 1 (men) and Group 2 (women).

Figure 5 shows a $(\mathrm{U}_{\mathrm{DP}}, \mathrm{AUC})$ diagram where each point represents the mean and one-third of the standard deviation across 30 bootstrap iterations, varying the complexity by selecting 3, 4, or 10 features based on the top variable importances of the model (e.g., for GLM, this would be the absolute value of the coefficients with the features being normalized). In the diagram, orange points represent the fair predictions, and blue points represent the baseline predictions.

We observe that as complexity increases, defined by the number of features, the methodology consistently maintains good DP fairness. However, this is accompanied by a slight overall average decrease in predictive performance. In other words, after mitigation, the model predicts similar probabilities of loan default for both men and women with only a slight reduction in predictive performance.

Note that focusing the reweighing approach on scores rather than class predictions (strong DP vs. weak DP) allows for a more in-depth analysis. This approach enables us to conduct additional analyses, such as examining score distributions, as shown in Figures 5 and 6. More specifically, Figure 6 shows the graphical representation of observed and predicted default rates, broken down by individual categorical variables. All other factors being equal, the base model (in blue, see the last panel) introduces a more significant bias based on gender than the "fair" model (in orange, see the last panel), adjusted for DP, which ensures consistent default rates (labeled as “average positive rate” in the graph) between the two groups.

FIGURE 5: (DP) UNFAIRNESS AND PERFORMANCE ACROSS DIFFERENT MODEL COMPLEXITIES

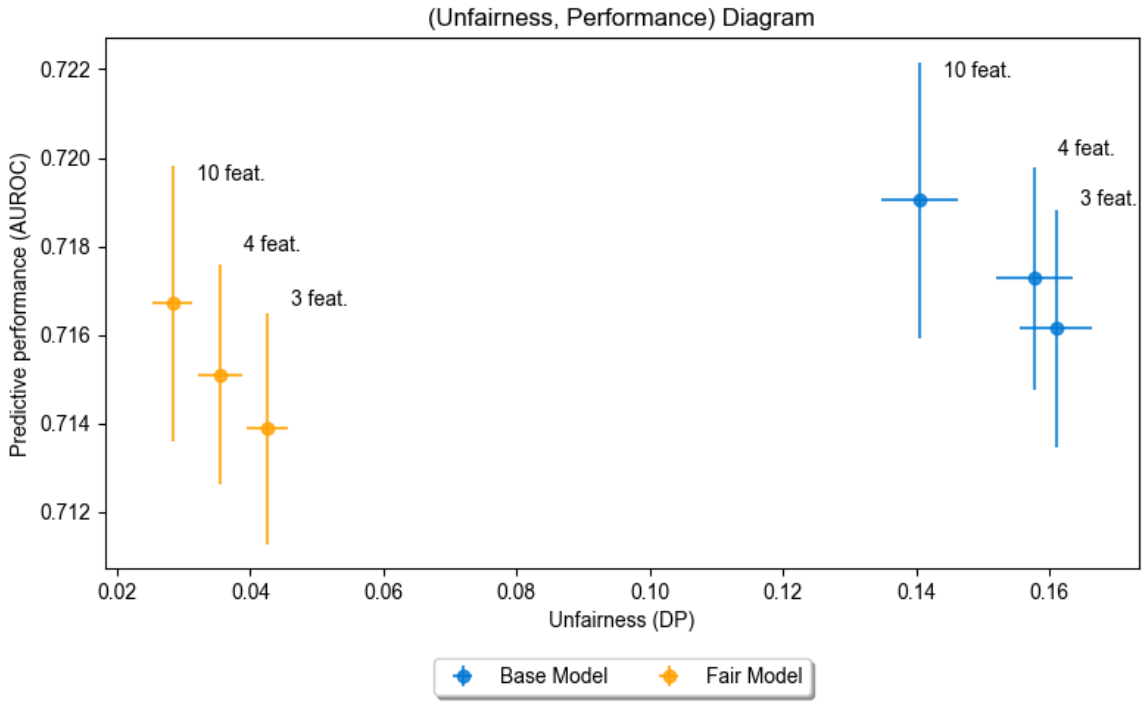

Figure description: (Unfairness, predictive performance) diagram for base and fair predictions across 30 simulations with varying model complexity (number of features used for training the predictive model). Means (o) and standard deviations (+) are shown.

FIGURE 6: COMPARISON OF ACTUAL AND PREDICTED RATES (FAIR AND BASE) BY CATEGORICAL FEATURE

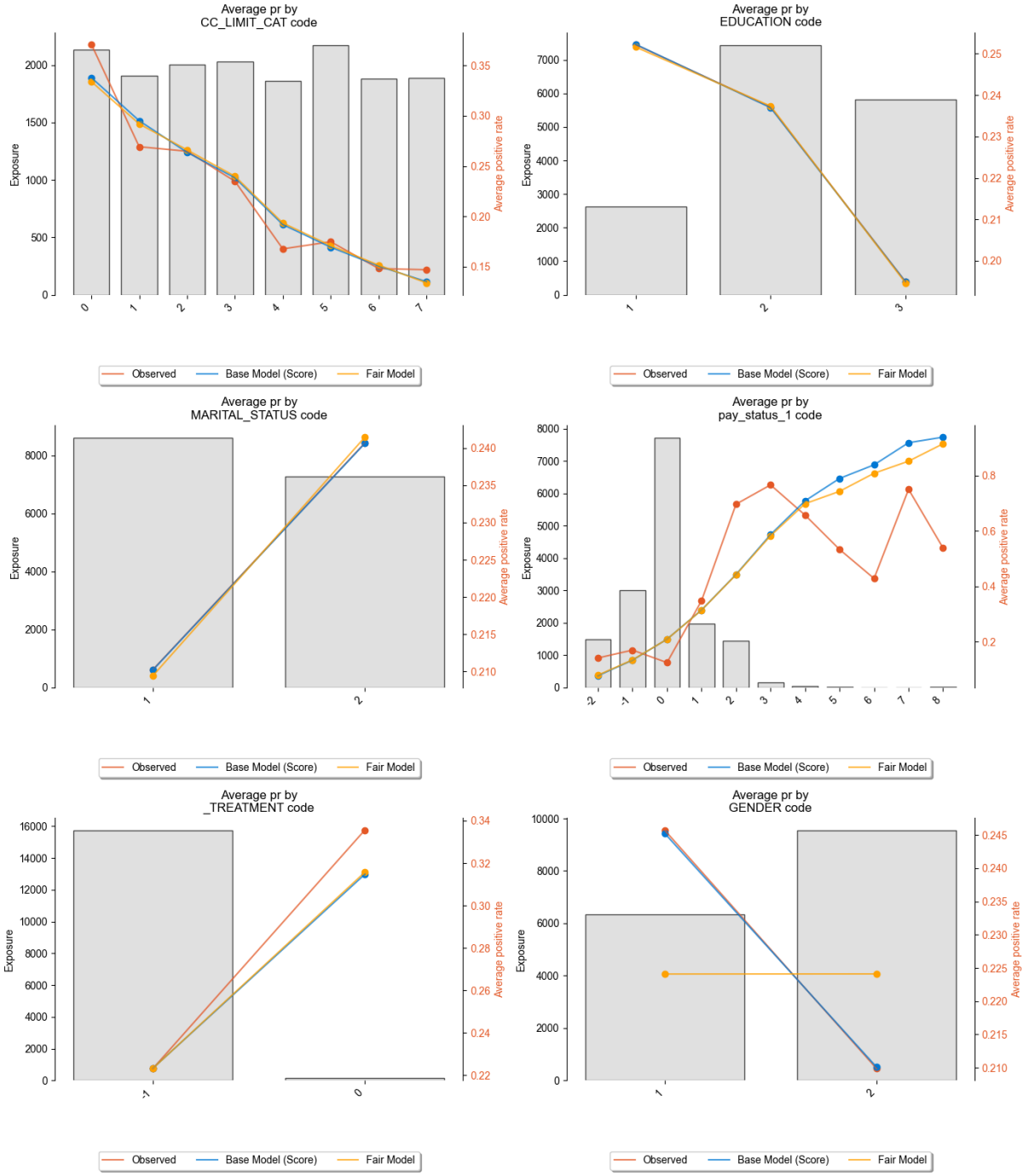

Figure description: In each panel, bar plots display the frequency of each feature value, while overlaid line plots show the observed and predicted default rates.

EO USING IN-PROCESSING

In this section, we focus on evaluating the EO unfairness measure using $\widehat{\mathrm{U}}_{\mathrm{EO}}$ for the base and fair predictions on the motor insurance dataset freMPL. The goal is to produce gender-fair predictions for classifying claim occurrence. This dataset is provided by a private French motor insurer and includes various demographic and vehicle-related features as well as risk factors, claim amounts, and claim histories for approximately 30,000 policies. We apply the EGR method described above to generate fair predictions and compare them to the base predictions from the GLM model.

Agarwal et al. (2018)'s methodology examines weak EO by directly predicting binary outcomes—claim or no claim—rather than relying on model scores. Since model scores are often poorly calibrated and may not accurately reflect probabilities, score density plots provide limited insights. Therefore, we present binary outcome rates instead.

Figure 7 illustrates this using bar plots, which show the rates of predicted claim and no claim outcomes within the "true claim" cases, stratified by different values of the sensitive group. Notably, this figure shows that we achieve improvements in EO fairness (and strong DP fairness) with only a minor decrease of approximately 0.02 in predictive performance (AUC).

FIGURE 7: EQUAL OPPORTUNITY ANALYSIS FOR BASE AND FAIR MODELS

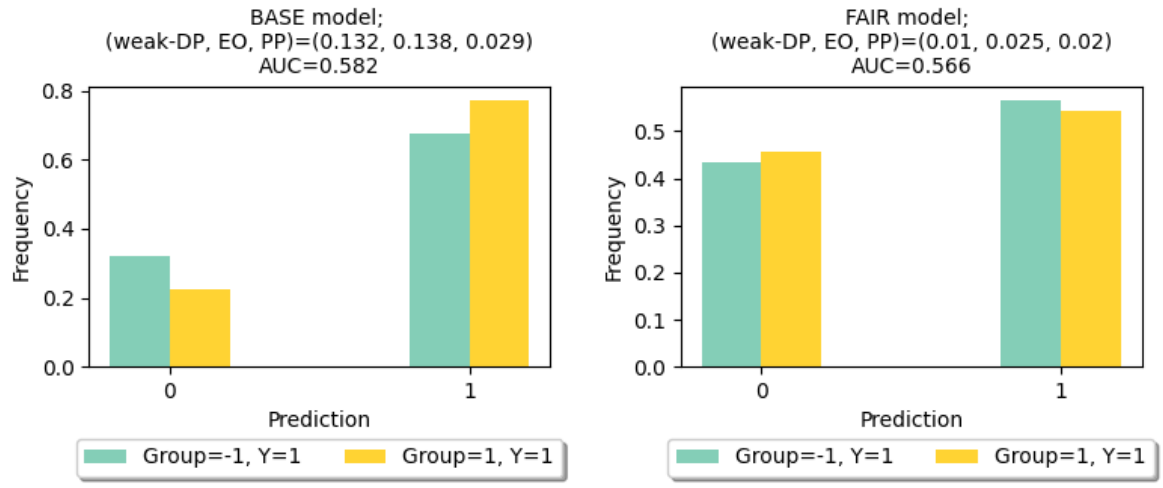

Figure description: Bar plots illustrate the rates of predictive positive and negative outcomes within the (true) positive outcomes, stratified by different values of the sensitive group.

Furthermore, Figure 8 shows that across 30 bootstrap iterations, the proposed methodology (orange) consistently enforces EO fairness across all ranges of explanatory features compared to the base model (blue), with only a slight drop in predictive performance.

FIGURE 8: (EO) UNFAIRNESS AND PERFORMANCE ACROSS DIFFERENT MODEL COMPLEXITIES

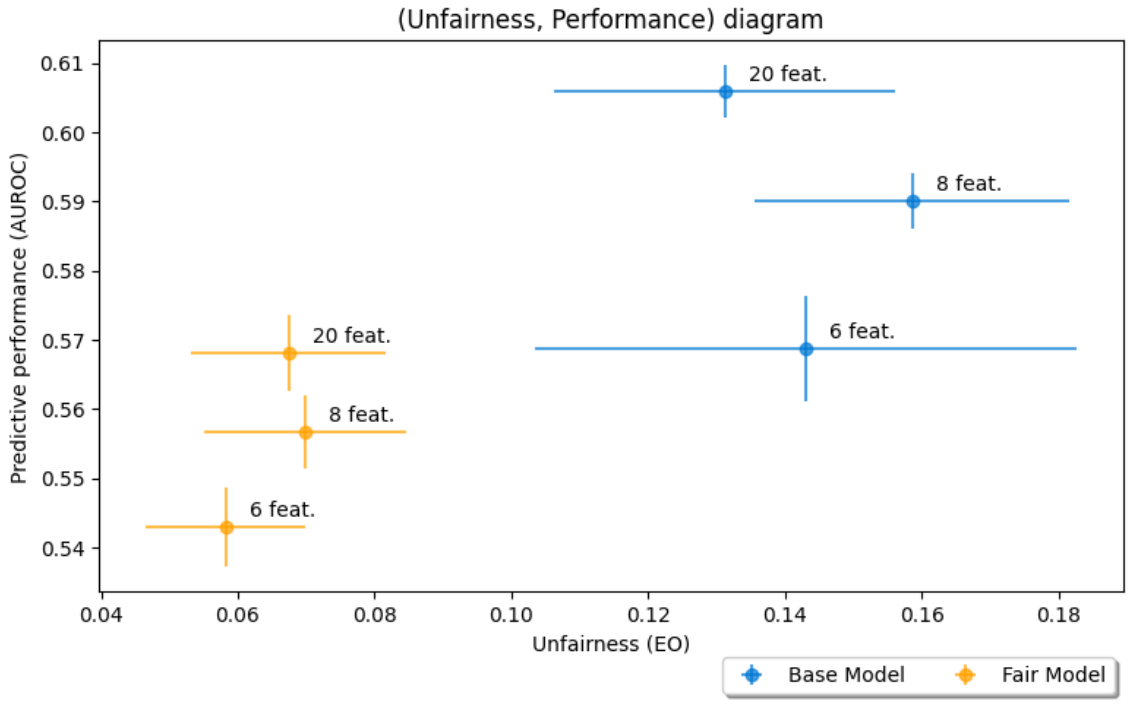

Figure description: (Unfairness, Predictive performance) diagram for base and fair predictions across 30 simulations with varying model complexities (number of features used for training the predictive model). Means (o) and standard deviations (+) are shown.

PP USING POST-PROCESSING

We focus on evaluating the PP unfairness measure using $\widehat{\mathrm{U}}_{\mathrm{PP}}$ for the base and fair predictions on the GOSSIS dataset, where the goal is to produce ethnicity-fair predictions for classifying mortality. This dataset includes over 130,000 intensive care unit visits from patients across a one-year period and covers Argentina, Australia, Brazil, India, New Zealand, Sri Lanka, and more than 200 hospitals in the U.S. Moreover, this dataset provides a black-box score corresponding with the Acute Physiology and Chronic Health Evaluation (APACHE) IV probabilistic prediction of in-hospital mortality for the patient. This score is derived from the APACHE III score and other covariates, including diagnosis.

Figure 9 illustrates the evolution of the mean hospital death count with respect to the mortality risk score percentile, conditional on the sensitive group. The given black-box score appears to be PP-fair. This observation is further supported by Figure 10, which displays the distribution of true outcomes among positive predictions, again broken down by sensitive group.

FIGURE 9: CALIBRATION PARITY ANALYSIS FOR BLACK-BOX SCORES

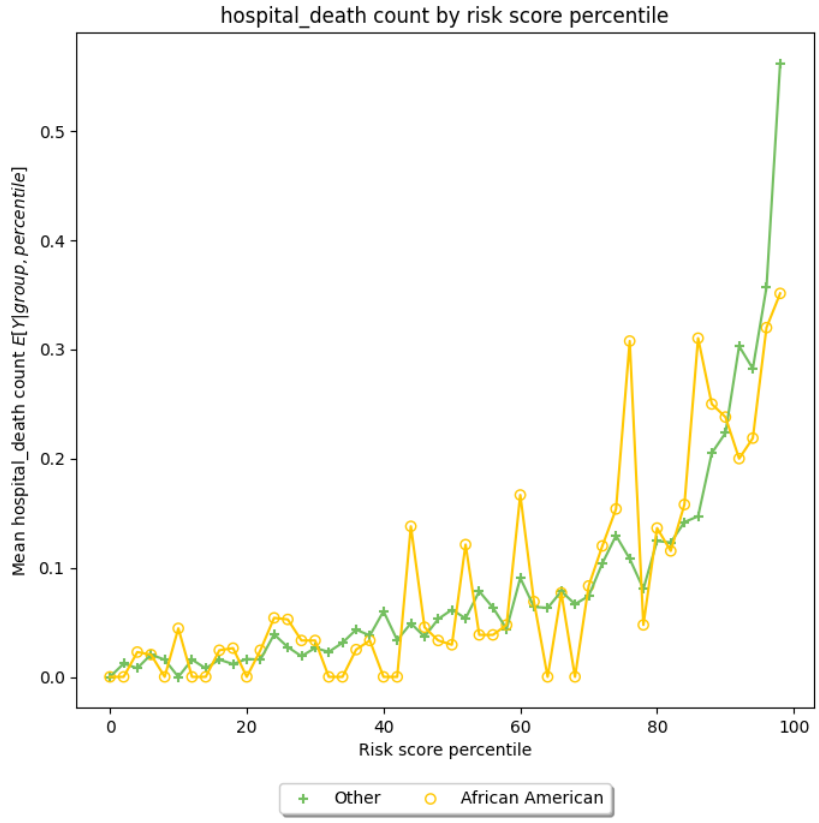

Figure description: Line plots illustrate the evolution of the mean hospital death count by mortality risk score percentile, stratified by different values of the sensitive group.

FIGURE 10: PP ANALYSIS FOR BLACK-BOX SCORES

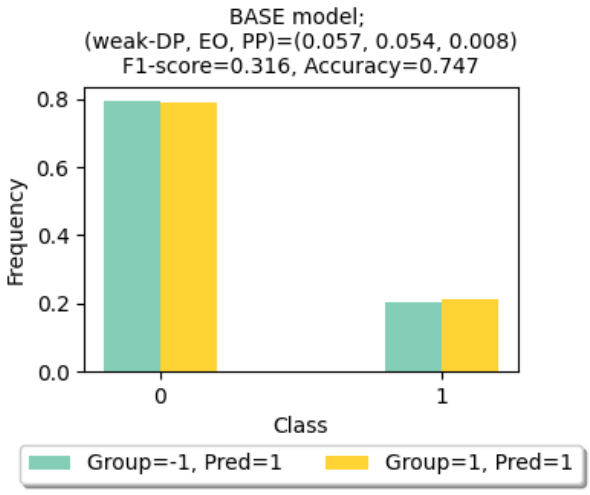

Figure description: Bar plots illustrate the rates of (true) positive and negative classes within the positive predicted outcomes, stratified by different values of the sensitive group.

To explore the impact of bias, we introduce an artificial shift in the black-box score by adding a constant value of 0.1 to all scores from one sensitive group while subtracting 0.1 from those of the other sensitive group. Based on this artificially biased dataset, we apply the previously described Lagrange-based method to generate fair predictions. We then compare these adjusted predictions to the given (biased) black-box predictions, with the results presented in Figure 11. This shows the rates of predicted death and no death outcomes among actual mortality cases, stratified by different values of the sensitive group. The (artificially biased) black-box score is PP-unfair, with a PP unfairness measure of 0.199 (well above 0.05). This indicates that within actual mortality cases, mortality predictions significantly differ between the group "African American" and "Other." However, as shown in Figure 11, fairness can be further enhanced (from 0.199 to 0.017 of PP unfairness), leading to overall improvements in the predictive model.

FIGURE 11: PP ANALYSIS FOR BASE AND FAIR MODELS

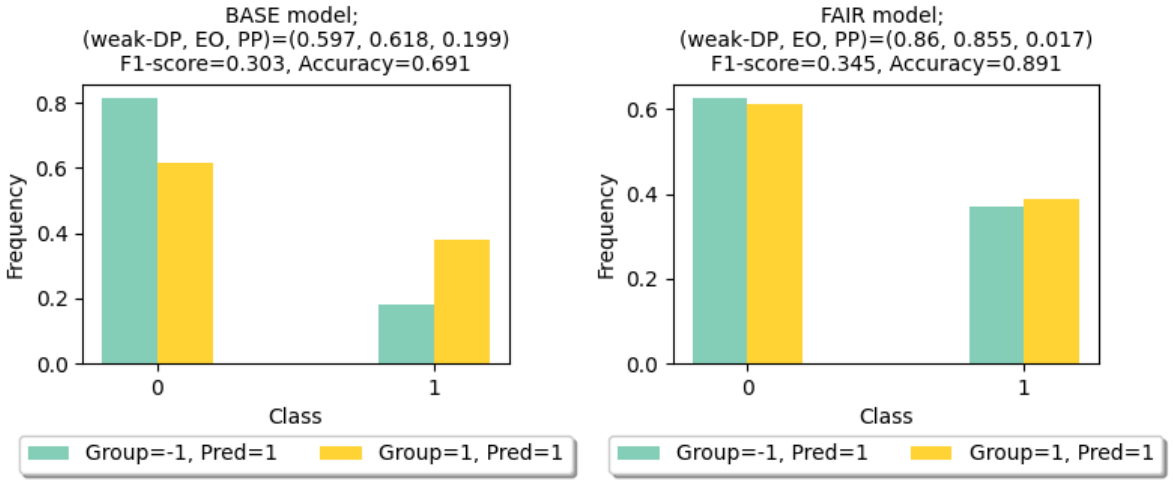

Figure description: Bar plots illustrate the rates of positive and negative classes within the positive predicted outcomes, across different values of the sensitive group.

Summary

Driven by ethical considerations or imposed regulations, various approaches can be employed to diagnose and mitigate biases and unfairness. As highlighted, biases often stem from issues such as a lack of data representation, poor data quality, or societal and historical human biases embedded in decisions. This article demonstrates that these biases can be defined with the predictive modeling objective in mind and, more importantly, addressed at every stage of the ML cycle—pre-processing, in-processing, and post-processing. This paper outlined several established fairness considerations and methodologies from an actuarial perspective, including ensuring gender-fair performance in non-life insurance for P&C practices, preventing unfair treatment in fraud or loan default detection, and promoting ethnicity-fair procedures in life insurance mortality scores.

CONTACT

François Hu
francois.hu@milliman.com

Alexandre Boumezoued
alexandre.boumezoued@milliman.com

ENDNOTES

1. Machine learning glossary: Responsible AI. (n.d.). Google Developers. Retrieved Month 04, 2025 from https://developers.google.com/machine-learning/glossary/responsible-ai#a
2. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science. 366(6464), 447–453.
3. Lin, X., Browne, M. J., and Hofmann, A. (2022). Race discrimination in the adjudication of claims: Evidence from earthquake insurance. Journal of Risk and Insurance, 89(3), 553–580.
4. Council directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services. (2004). Official Journal of the European Union, 373, 37–43.
5. See Regulation (EU) 2016/679 of the European Parliament at https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679
6. See Regulation (EU) 2024/1689 of the European Parliament at https://eur-lex.europa.eu/eli/reg/2024/1689/oj.
7. Directive (EU) 2016/97 of the European Parliament and the Council of 20 January 2016 on insurance distribution(recast)Text with EEA relevance. (n.d.). EUR-Lex. Retrieved Month 01, 2025 from https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:32016L0097#document1
8. Charpentier, A. (2024). Insurance, Biases, Discrimination, and Fairness. Berlin: Springer.
9. Regulation 10-1-1 Governance and Risk Management Framework Requirements for Life Insurers' Use of External Consumer Data and Information Sources, Algorithms, and Predictive Models. For more details, refer to https://doi.colorado.gov/announcements/notice-of-adoption-new-regulation-10-1-1-governance-and-risk-management-framework
10. For more detail, refer to https://www.dfs.ny.gov/industry-guidance/circular-letters/cl2024-07
11. Kamiran F. & Calders T. (2012). Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems, 33(1), 1–33.
12. Zhang, B. H., Lemoine, B., & Mitchell, M. (2018, December) Mitigating Unwanted Biases with Adversarial Learning. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 335–340).
13. Agarwal, A. et al. (2018, July). A reductions approach to fair classification. In International Conference on Machine Learning (pp. 60–69).
14. Agarwal, A., Dudík, M., & Wu, Z. S. (2019, May) Fair regression: Quantitative definitions and reduction-based algorithms. In International Conference on Machine Learning (pp. 120–129). PMLR.
15. Chzhen, E. et al. (2020) Fair regression with Wasserstein barycenters. Advances in Neural Information Processing Systems, 33, 7321–7331.
16. Hu, F., Ratz, P., Charpentier, A. (2024, March) A sequentially fair mechanism for multiple sensitive attributes. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 11, pp12502–12510).
17. Denis, C. et al. (2024) Fairness guarantees in multi-class classification with demographic parity. Journal of Machine Learning Research, 25(130), 1–46.
18. Alghamdi, W. et al. (2022). Beyond adult and COMPAS: Fairness in multi-class prediction. arXiv preprint arXiv:2206.07801.
19. Zeng, X., Dobriban, E., & Cheng, G. (2022). Fair Bayes-optimal classifiers under predictive parity. Advances in Neural Information Processing Systems, 35, 27692–27705.