Bias, fairness and discrimination-free insurance pricing

Katrien Antonio
LRisk - KU Leuven and ASE - University of Amsterdam

September 22, 2022

My personal website: https://katrienantonio.github.io

This workshop's mission is threefold

We aim to:

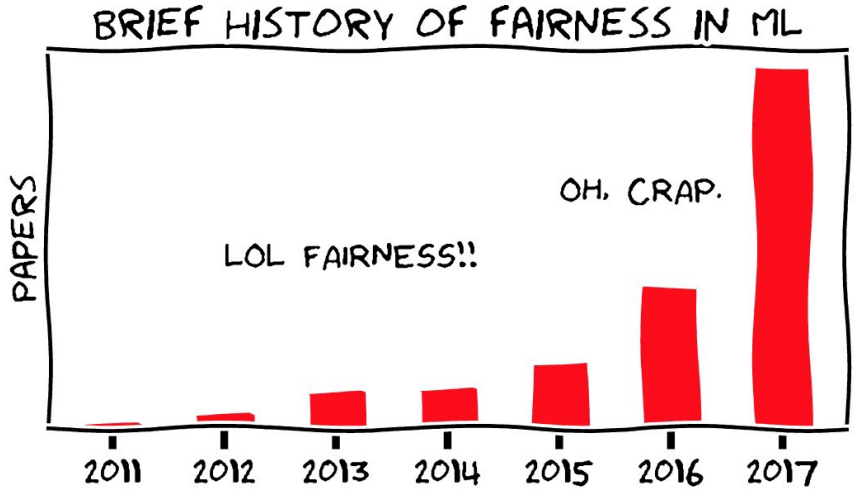
(1) demystify the (exploding) literature on fair machine learning

(2) explore in- and post-processing methods to establish fairness or to remove discrimination by proxy in insurance pricing

(3) with GLM ($\sim$ statistical learning) and GBM ($\sim$ machine learning) based pricing methods.

A widely expanding literature

In machine learning . . .



A widely expanding literature

In machine learning . . .

(Picture taken from A. Charpentier, presentation on Sept 9, see here.)

A widely expanding literature

. . . and in actuarial science!

> "Insurance is particularly interesting because the **entire industry is based on discrimination**. Here, we use the word discrimination in an **entirely neutral way**, taking it to mean the act of treating different groups differently. (. . .) The modern-day insurance industry is founded on the ability **to differentiate, or discriminate, among risks**, known as risk classification." (Frees & Huang, 2021, NAAJ)

In fact, insurers discriminate among customers via a diverse set of actions, e.g.:

- the decision to insure
- the coverage offered
- by charing different prices.

A widely expanding literature

. . . and in actuarial science!

- Frees & Huang (2021, NAAJ) argue how the meaning of **actuarial fairness** or **fair insurance systems**, may depend on:
  - historical context, with a shift of responsibility from individual to a pool
  - the nature of the pool, e.g. a mutual company vs a stock insurance company
  - whether the insurance product can be viewed as a social or type of public good, or not.
- Today, I will interpret actuarial fairness as a pricing system where each customer should pay a **premium proportional to their own risk**.

A selection of references

Sources I used to put this workshop together:

- Mehrabi et al. (2022), A survey on bias and fairness in machine learning, ACM Computing Surveys
- Caton & Haas (2021), Fairness in machine learning: a survey
- E. Frees & F. Huang (2021), The discriminating (pricing) actuary, North American Actuarial Journal
- M. Lindholm, R. Richman, A. Tsanakas & M. Wüthrich (2022), Discrimination-free insurance pricing, ASTIN Bulletin
- A. Charpentier (2022), Insurance: discrimination, biases and fairness, Institut Louis Bachelier

> "In the context of decision-making **fairness** is the **absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics**. Thus, an unfair algorithm is one whose decisions are skewed toward a particular group of people." (Mehrabi et al., 2019)

Mehrabi et al. (2019) identify two potential sources of unfairness:

- from biases in the data
- from the learning algorithms.

> "Bias can exist in many shapes and forms, some of which can lead to unfairness in different downstream learning tasks." (Mehrabi et al., 2019)

Most important **sources of bias** discussed in the literature: from

- data to algorithm, e.g. *measurement bias* or *omitted variable bias*
- algorithm to user, e.g. *algorithmic bias* where bias is not present in the data but purely added by the algorithm
- user to data, e.g. when data are user-generated, any inherent biases in users might be reflected in the data they generate.

Discrimination

Different notions

- **Explainable** discrimination
  - when differences in treatment and outcomes amongst different groups can be justified and explained via some (acceptable, non-protected) attributes.
- **Direct** discrimination
  - when protected attributes of individuals **explicitly** result in non-favorable outcomes toward them.
- **Indirect** discrimination
  - individuals appear to be treated based on seemingly neutral and non-protected attributes
  - however, protected groups, or individuals still get to be treated unjustly as a result of **implicit effects from their protected attributes**.

Discrimination

Indirect discrimination: by proxy and disparate impact

- **Proxy discrimination**:
  - arises from correlation between protected and unprotected characteristics
  - the implicit ability to infer protected characteristics from other (legitimately used) policyholder features
  - e.g. geographic area serves as a substitute for a protected variable such as race, or proxy produced by an AI that summarizes the effects of many variables.
- **Disparate impact**:
  - a systematic disadvantage resulting for a group that is protected by a nondiscrimination provision.

Discrimination

Protected features?

> "Grouping, or classifying, insureds into homogeneous categories for the purposes of risk sharing is at the heart of the insurance function. Many variables that insurers use are seemingly innocuous (e.g., blindness for auto insurance), yet others can be viewed as *wrong* (e.g., religious affiliation), *unfair* (e.g., onset of cancer for health insurance), *sensitive* (e.g., marital status), or *mysterious* (e.g., Artificial Intelligence produced)." (Frees & Huang, 2021, NAAJ)

Protected variables:

- not permitted in risk classification
- their choice is a normative one, after societal debate with many actors.

Discrimination

Protected or sensitive features?

	CA	HI	GA	NC	NY	MA	PA	FL	TX	AL	ON	NB	NL	QC
Gender	x	x	•	x	•	x	x	•	•	•	•	x	x	•
Age	x	x	•	x*	•	x	•	•	•	•*	•	x	x	•
Driving experience	•	x	•	•	•	•	•	•	•	•	•	•	•	•
Credit history	x	x	•	•	•	x	•*	•	•	x*	x	•*	x	•
Education	x	x	x	x	x	x	•	•	•	•	•	•	•	•
Profession	x	x	x	•	x	x	•	•	•	•	•	•	•	•
Employment	x	x	x	•	x	x	•	•	•	•	•	•	•	•
Family	•	x	•	•	•	x	•	•	•	•	•	•	•	•
Housing	x	x	•	•	•	x	•	•	•	x	x	•	•	•
Address/ZIP code	•	•	•	•	•	•	•	•	•	x	x	•	•	•

Table 2.1: A factor is considered "permitted" (•) when there are no laws or regulatory policies in the state or province that prohibit insurers from using that factor. Otherwise, it will be "prohibited" (x). In North Carolina, age is only allowed when giving a discount to drivers 55 years of age and older. In Pensylvania, credit score can be used for new business and to reduce rates at renewal, but not to increase rates at renewal. In Alberta, credit score and driver's license seniority cannot be used for mandatory coverage (but can be used on optional coverage). In Labrador, age cannot be used before 55, and beyond that, it must be a discount (as in North Carolina).

(Picture taken from A. Charpentier, report on *Insurance: discrimination, biases and fairness*, see here.)

Discrimination

Sensitive features?

Structure to identify whether or not a variable contains sensitive information, from Frees & Huang (2021), inspired by Avraham (2018) and Prince and Schwarcz (2020):

Property	Explanation
Control	If a policyholder has control over a certain characteristic, e.g. `Type of car`, it is deemed appropriate to use for risk classification.
Mutability	A variable for which the value changes over time is considered to be fair, as individuals have the chance to be on the winning and losing side during their lifetime. An example is the variable `Age`.
Causality	When there is a proven causal connection between the variable and the insured event, it is fair to use the variable for risk classification. Note that a lot of research needs to be done to conform causality.
Statistical discrimination	If a variable makes no significant contribution to the predictive accuracy it is best not to use it.
Limiting or reversing the effects of past discrimination	A variable used to preserve a historical negative stereotype is deemed unfair.
Inhibiting socially valuable behaviour	A variable that, when used to classify risks, will hinder socially desirable behaviour is considered as an unfair variable.

Fairness criteria and metrics

A multitude of fairness criteria and corresponding metrics has been proposed in the ML literature, almost exclusively linked to classification problems (with target Y being 0 or 1).

A selection: (with Y the target, $\hat{Y}$ the prediction and D the protected feature)

- **demographic or statistical parity**:

  a predictor $\hat{Y}$ *satisfies demographic parity if* $\mathbb{P}(\hat{Y} \mid D = 0) = \mathbb{P}(\hat{Y} \mid D = 1)$

- **equal opportunity**:

  a predictor $\hat{Y}$ *satisfies equal opportunity with respect to protected attribute* D *and outcome* Y*, if* $\mathbb{P}(\hat{Y} = 1 \mid D = 0,\ Y = 1) = \mathbb{P}(\hat{Y} = 1 \mid D = 1,\ Y = 1)$*. Thus, equal true positive rates for protected and unprotected groups.*

See e.g. http://research.google.com/bigpicture/attacking-discrimination-in-ml/ for visuals and https://developers.google.com/machine-learning/glossary/fairness for detailed illustrations and definitions.

A multitude of fairness criteria and corresponding metrics has been proposed in the ML literature, almost exclusively linked to classification problems (with target Y being 0 or 1).

A selection:

- **fairness through awareness**:

  an algorithm is fair if it gives similar predictions to similar individuals

- **fairness through unawareness**:

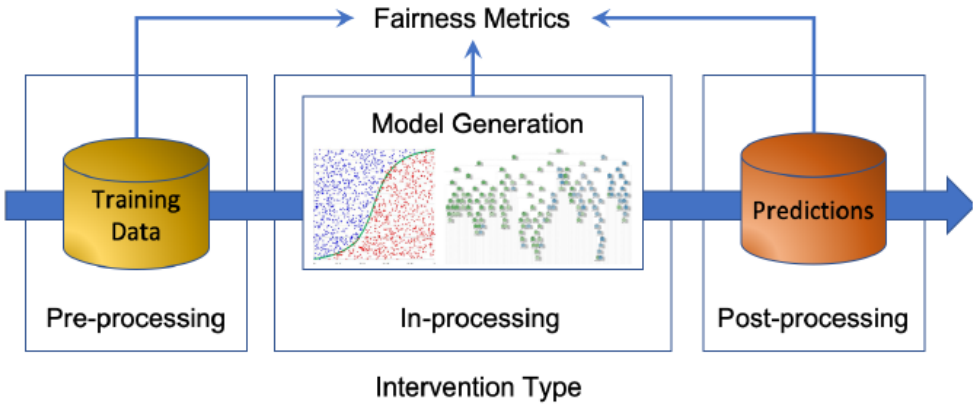
  an algorithm is fair as long as any protected attributes D are not explicitly used in the decision-making process.

Some reflections:

- very much focused on classification problems!
- impossible to satisfy some of these fairness conditions simultaneously, except in some very special cases.
- users must decide on where to place emphasis, but be mindful of the trade off between any fairness measure and model accuracy.

Mitigation strategies

Methods for fair machine learning



(Picture taken from Caton & Haas, see here.)

Mitigation strategies

Methods for fair machine learning

With **pre-processing**:

- transform the data with the aim to remove bias/discrimination from the training data
- then fit a learning model on the *repaired* data

With **in-processing**: (∼ Marchi, Antonio, et al., 2022, ongoing)

- try to find a balance between multiple model objectives, e.g. accuracy and fairness

With **post-processing**: (∼ Lindholm et al., 2022, ASTIN)

- apply transformations to model output to improve prediction fairness
- only needs access to the predictions and sensitive attribute information.

In ongoing research with Marchi (KU Leuven), Avanzi and Zhou (Uni of Melbourne), we aim to establish fair insurance pricing **via regularization** ($\sim$ Lasso)

$$\min_{w_h} \left\{ \sum_{i=1}^{n} \mathcal{L}(h_{w_h}(x_i), y_i) + \lambda \cdot \Phi(h_{w_h}, \{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n, \{d_i\}_{i=1}^n) \right\}. \tag{1}$$

where the data are $(x_i, y_i, d_i) \in \mathcal{X}^n \times \mathcal{Y}^n \times \mathcal{D}^n$, a generic loss function $\mathcal{L} : \mathbb{R}^2 \to \mathbb{R}$, a predictive model $h_{w_h} \in \mathcal{F}$ and a fairness criterion $\Phi : \mathcal{F} \times \mathcal{X}^n \times \mathcal{Y}^n \times \mathcal{D}^n \to \mathbb{R}$.

Hence, the goal is to balance accuracy (via $\mathcal{L}$) and fairness (via Φ).

Let's dive into:

Discrimination-free insurance pricing, by M. Lindholm, R. Richman, A. Tsanakas & M. Wüthrich (2022, ASTIN Bulletin)

with reproducible example in Python prepared by Lorenzo Marchi and Katrien Antonio, accessible via a Google Colab

further explored in the MSc thesis of Elien Baeten (KU Leuven) and Remco Bruinsma (UvA).

Lindholm et al. (2022):

- propose a **post-processing** pricing adjustment formula that explicitly addresses discrimination by proxy
- with the goal **to remove indirect discrimination** – if it happens to exist – from insurance pricing models
- purely reason from an **actuarial** rather than a legal perspective
- assume **knowledge of a policyholder's discriminatory features**.

No connection with fairness, or a metric to express fairness. In fact, open discussion going on the connection between insurance pricing and notions of fairness.

Motivating example

From Baeten (2021), inspired by Lindholm et al. (2022)

$n_{i,j}$	Small	Large	**Row total**
Female	83	14	97
Male	47	60	107
Column total	130	74	204

Table: Fictional claim counts.

$e_{i,j}$	Small	Large	**Row total**
Female	567	83	650
Male	269	253	522
Column total	836	336	1172

Table: Fictional exposures.

Here:

- $i \in \{0,1\}$ refers to `Gender` with $i = 1$ a *male* driver
- $j \in \{0,1\}$ refers to `Type of car` with $j = 1$ a *large* car.

Motivating example

From Baeten (2021), inspired by Lindholm et al. (2022)

Let us focus on the following set of questions:

- how to estimate the overall expected claim frequency, say $\hat{\lambda}$, in this portfolio?
- what is $\hat{\lambda}_{ij}$ when risks are classified using both `Gender` and `Type of car`?
- what if prices can not discriminate based on `Gender`, hence: `Gender` is a protected feature?

Motivating example

From Baeten (2021), inspired by Lindholm et al. (2022)

Overall expected claim frequency:

$$\hat{\lambda} = \frac{\sum_{i,j} n_{i,j}}{\sum_{i,j} e_{i,j}} = \frac{n_\bullet}{e_\bullet} = \frac{204}{1\ 172} = 0.174.$$

Price discrimination, or risk classification, using both `Gender` and `Type of car` leads to **best-estimate prices**:

$\hat{\lambda}_{i,j}$	Small	Large
Female	0.146	0.169
Male	0.175	0.237

where, for instance, $\hat{\lambda}_{0,0} = \frac{83}{567} = 0.146$. Try to summarize some findings here!

Motivating example

From Baeten (2021), inspired by Lindholm et al. (2022)

Some initial observations:

- claim frequencies higher for men than for women
- male drivers with a large car are high risk
- claim frequencies higher for large car drivers compared to small cars.

Motivating example

From Baeten (2021), inspired by Lindholm et al. (2022)

Now we treat `Gender` as a protected feature, and proceed by **simply ignoring** the variable in the pricing:

$$\hat{\lambda}_{\bullet,j} = \frac{n_{\bullet,j}}{e_{\bullet,j}} = \frac{n_{0,j} + n_{1,j}}{e_{0,j} + e_{1,j}} \text{ with } j \in \{0, 1\}$$

which results in

$$\begin{aligned} \hat{\lambda}_{\bullet,0} &= \frac{130}{836} = 0.156 \text{ for a } \textit{Small} \text{ car} \\ \hat{\lambda}_{\bullet,1} &= \frac{74}{336} = 0.220 \text{ for a } \textit{Large} \text{ car.} \end{aligned}$$

These prices adhere to the **unawareness** principle and are called **unawareness prices**.

Are these prices truly not discriminating based on `Gender`?

Motivating example

From Baeten (2021), inspired by Lindholm et al. (2022)

`Type of car` is very informative for `Gender` in our example.

Indeed,

$$\begin{aligned}\hat{\mathbb{P}}(\text{Male} \mid \text{Large}) &= \frac{e_{1,1}}{e_{0,1}+e_{1,1}} = \frac{253}{336} = 0.75 \\ \hat{\mathbb{P}}(\text{Male} \mid \text{Small}) &= \frac{e_{1,0}}{e_{0,0}+e_{1,0}} = \frac{269}{836} = 0.32.\end{aligned}$$

Thus, higher propensity of drivers with a *Large* car to be *Male* drivers.

At portfolio level, $\hat{\mathbb{P}}(\text{Male}) = \frac{522}{1\,172} = 0.45$ and $\hat{\mathbb{P}}(\text{Female}) = \frac{650}{1\,172} = 0.55$.

Indirect discrimination

Motivating example, from Baeten (2021), inspired by Lindholm et al. (2022)

In fact, the unawareness price $\hat{\lambda}_{\bullet,1}$ for a driver of a *Large* car can be rewritten as:

$$\begin{aligned}
\hat{\lambda}_{\bullet,1} &= \frac{n_{0,1}+n_{1,1}}{e_{0,1}+e_{1,1}} = \frac{n_{0,1}}{e_{0,1}} \cdot \frac{e_{0,1}}{e_{0,1}+e_{1,1}} + \frac{n_{1,1}}{e_{1,1}} \cdot \frac{e_{1,1}}{e_{0,1}+e_{1,1}} \\
&= \hat{\lambda}_{0,1} \cdot \frac{e_{0,1}}{e_{0,1}+e_{1,1}} + \hat{\lambda}_{1,1} \cdot \frac{e_{1,1}}{e_{0,1}+e_{1,1}} \\
&= \hat{\lambda}_{0,1} \cdot \hat{\mathbb{P}}(\text{Female} \mid \text{Large}) + \hat{\lambda}_{1,1} \cdot \hat{\mathbb{P}}(\text{Male} \mid \text{Large}) \\
&= 0.169 \cdot 0.25 + 0.237 \cdot 0.75 \\
&= 0.22.
\end{aligned}$$

Not only information about the influence of `Type of car` on producing a claim is used, but also about the propensity of drivers with a specific `Type` to be *male* or *female*.

The correlation between `Type of car` and `Gender` is exploited.

Motivating example

From Baeten (2021), inspired by Lindholm et al. (2022)

Similarly, the unawareness price $\hat{\lambda}_{\bullet,0}$ for a driver of a *Small* car becomes:

$$\begin{aligned}
\hat{\lambda}_{\bullet,0} &= \frac{n_{0,0}+n_{1,0}}{e_{0,0}+e_{1,0}} = \frac{n_{0,0}}{e_{0,0}} \cdot \frac{e_{0,0}}{e_{0,0}+e_{1,0}} + \frac{n_{1,0}}{e_{1,0}} \cdot \frac{e_{1,0}}{e_{0,0}+e_{1,0}} \\
&= \hat{\lambda}_{0,0} \cdot \frac{e_{0,0}}{e_{0,0}+e_{1,0}} + \hat{\lambda}_{1,0} \cdot \frac{e_{1,0}}{e_{0,0}+e_{1,0}} \\
&= \hat{\lambda}_{0,0} \cdot \hat{\mathbb{P}}(\text{Female} \mid \text{Small}) + \hat{\lambda}_{1,0} \cdot \hat{\mathbb{P}}(\text{Male} \mid \text{Small}) \\
&= 0.146 \cdot 0.678 + 0.175 \cdot 0.322 \\
&= 0.156.
\end{aligned}$$

Motivating example

From Baeten (2021), inspired by Lindholm et al. (2022)

We see the **potential for indirect discrimination** reflected in the **unawareness price**.

Indeed, in our example:

- men cause on average more claims than women
- in the sub-population of *Large* car drivers men are more prevalent, compared to the `Gender` composition at portfolio level
- hence, the unawareness price for *Large* is leveraged, and vice versa for *Small* car drivers.

Motivating example

From Baeten (2021), inspired by Lindholm et al. (2022)

Lindholm et al. (2022) replace the conditional probabilities by **unconditional probabilities** to obtain a **discrimination-free** price:

$$\hat{\lambda}^{\mathsf{DF}}_{\bullet,j} \quad = \quad \hat{\lambda}_{0,j} \cdot \hat{\mathbb{P}}(\mathsf{Female}) + \hat{\lambda}_{1,j} \cdot \hat{\mathbb{P}}(\mathsf{Male}).$$

The discrimination-free expected claim frequencies for *Small* and *Large* car drives then becomes:

$$\begin{aligned} \hat{\lambda}^{\mathsf{DF}}_{\bullet,0} &= 0.146 \cdot 0.55 + 0.175 \cdot 0.45 = 0.168 > \hat{\lambda}_{\bullet,0} = 0.156 \\ \hat{\lambda}^{\mathsf{DF}}_{\bullet,1} &= 0.169 \cdot 0.55 + 0.237 \cdot 0.45 = 0.199 < \hat{\lambda}_{\bullet,1} = 0.22. \end{aligned}$$

Note: the resulting price list still discriminates on the basis of `Type of car` (as it should do).

Motivating example

From Baeten (2021), inspired by Lindholm et al. (2022)

The **best estimate** and **unawareness** prices in this example are **unbiased**.

Indeed, assume claim cost is 1 EUR, then at portfolio level:

$$\sum_{i,j} \hat{\lambda}_{i,j} \cdot e_{i,j} \quad = \quad 567 \cdot 0.146 + 83 \cdot 0.169 + 269 \cdot 0.175 + 253 \cdot 0.237 = 204.$$

and

$$\sum_{j} \hat{\lambda}_{\bullet,j} \cdot e_{\bullet,j} \quad = \quad 836 \cdot 0.156 + 336 \cdot 0.22 = 204.$$

Thus, at portfolio level, the total premium volume (here: expected claim counts) equals the observed total loss (here: observed claim counts).

Motivating example

From Baeten (2021), inspired by Lindholm et al. (2022)

However, the **discrimination-free** prices **imply a bias** at portfolio level. Indeed,

$$\hat{\lambda}^{\mathsf{DF}}_{\bullet,0} \cdot (e_{0,0} + e_{1,0}) + \hat{\lambda}^{\mathsf{DF}}_{\bullet,1} \cdot (e_{0,1} + e_{1,1}) \quad = \quad 0.168 \cdot 836 + 0.199 \cdot 336 = 207.3 > 204.$$

To de-bias, one option is to adjust $\hat{\mathbb{P}}(\text{Female})$ and $\hat{\mathbb{P}}(\text{Male})$ so that portfolio bias is removed.

In our example, choose $\hat{\mathbb{P}}^{\star}(\text{Female}) = 0.65$ and $\hat{\mathbb{P}}^{\star}(\text{Male}) = 0.35$, then

$$\begin{aligned} \hat{\lambda}^{\mathsf{DF}\star}_{\bullet,0} &= 0.162 \cdot 0.65 + 0.175 \cdot 0.35 = 0.167 \\ \hat{\lambda}^{\mathsf{DF}\star}_{\bullet,1} &= 0.169 \cdot 0.65 + 0.237 \cdot 0.35 = 0.193, \end{aligned}$$

and at portfolio level

$$\hat{\lambda}^{\mathsf{DF}\star}_{\bullet,0} \cdot (e_{0,0} + e_{1,0}) + \hat{\lambda}^{\mathsf{DF}\star}_{\bullet,1} \cdot (e_{0,1} + e_{1,1}) \quad = \quad 0.167 \cdot 836 + 0.193 \approx 204.$$

Pricing formulas

Lindholm et al., 2022

Let $\boldsymbol{D}$ correspond to some protected (or: sensitive) features and let the non-protected features be denoted by $\boldsymbol{X}$.

Let Y denote the target variable of interest, e.g. number of claims or claim sizes.

Lindholm et al. (2022) then consider three pricing formulas:

- the **best-estimate** price, where both $\boldsymbol{D}$ and $\boldsymbol{X}$ are used
- the **unawareness** price, simply ignoring $\boldsymbol{D}$
- the **discrimination-free** price, where best-estimate prices are averaged over discriminatory covariates.

Pricing formulas

Lindholm et al., 2022 - best-estimate price

The **best-estimate** price for Y wrt $(\boldsymbol{X}, \boldsymbol{D})$ is:

$$\mu(\boldsymbol{X}, \boldsymbol{D}) := \mathbb{E}[Y \mid \boldsymbol{X}, \boldsymbol{D}].$$

Hereby, the price list is

- in general not discrimination-free, unless $\mu(\boldsymbol{X}, \boldsymbol{D})$ reduces to $\mu(\boldsymbol{X})$ because of independence between $\boldsymbol{X}$ and $\boldsymbol{D}$
- obtained from some predictive model that **can use both** $\boldsymbol{X}$ and $\boldsymbol{D}$.

Pricing formulas

Lindholm et al., 2022 - unawareness price

The **unawareness** price for Y wrt $(\boldsymbol{X})$ is:

$$\mu(\boldsymbol{X}) := \mathbb{E}[Y \mid \boldsymbol{X}].$$

Hereby, the price list

- attempts at avoiding discrimination by simply ignoring the protected features in $\boldsymbol{D}$
- may produce indirect discrimination via ($\boldsymbol{D}$ proxied by $\boldsymbol{X}$)

$$\mu(\boldsymbol{X}) \quad = \quad \int_{\boldsymbol{d}} \mu(\boldsymbol{X}, \boldsymbol{d})\, d\mathbb{P}(\boldsymbol{D} = \boldsymbol{d} \mid \boldsymbol{X}),$$

where the conditional probability enables inference of protected features $\boldsymbol{D}$ via unprotected features $\boldsymbol{X}$.

Pricing formulas

Lindholm et al., 2022 - discrimination-free price

A discrimination-free price for Y wrt $(\boldsymbol{X})$ is:

$$h^*(\boldsymbol{X}) := \int_{\boldsymbol{d}} \mu(\boldsymbol{X}, \boldsymbol{d})\, d\mathbb{P}^*(\boldsymbol{D} = \boldsymbol{d}).$$

Hereby, the price list

- averages best-estimate prices over discriminatory covariates $\sim \mathbb{P}^*(\boldsymbol{D} = \boldsymbol{d})$
- is free of direct discrimination, since $h^*(\boldsymbol{X})$ does not explicitly use $\boldsymbol{D}$
- is free of indirect discrimination, since the possible explanatory power that $\boldsymbol{X}$ may have for $\boldsymbol{D}$ is removed
- is in general not unbiased.

In fact, the problem that arises with the unawareness prices is referred to in econometrics as **omitted variable bias**.

When features $\boldsymbol{X}$ included in a linear regression model are correlated with omitted features $\boldsymbol{Z}$, the $\boldsymbol{X}$ will (partially) proxy for the $\boldsymbol{Z}$ and the estimated regression parameters β will be biased.

For more details:

- Pope & Sydnor (2011) on *Implementing Anti-Discrimination Policies in Statistical Profiling Models* in the American Economic Journal: Economic Policy
- our Colab.

We now demonstrate technical insurance pricing with:

- Generalized Linear Models (GLMs)
- Gradient Boosting Machines (GBMs).

We use the set-up from Lindholm et al. (2022) and consider the best-estimate, the unawareness and discrimination-free pricing formulas.

Stylized example

Set up

Consider three covariates, X_1, X_2 (unprotected) and D (protected), with

- $D \in \{\text{female}, \text{male}\}$
- $X_1 \in \{15, ..., 80\}$, the age of the policyholder
- $X_2 \in \{\text{non-smoker}, \text{smoker}\}$.

For the $\mathbb{P}$ distribution of $(\boldsymbol{X}, D)$ we assume:

$$\begin{aligned}
&\mathbb{P}(D = \text{female}) = 0.45 \\
&\mathbb{P}(X_2 = \text{smoker}) = 0.3 \\
&\mathbb{P}(D = \text{female} \mid X_2 = \text{smoker}) = 0.8.
\end{aligned}$$

Stylized example

Set up

The example assumes **three types of health related claims**:

(1) type 1 (related to giving birth)

$$\lambda_1(\boldsymbol{X}, D) \quad := \exp(\alpha_0 + \alpha_1 1_{\{X_1 \in [20,40]\}} 1_{\{D=\text{female}\}})$$

(2) type 2 (cancer related)

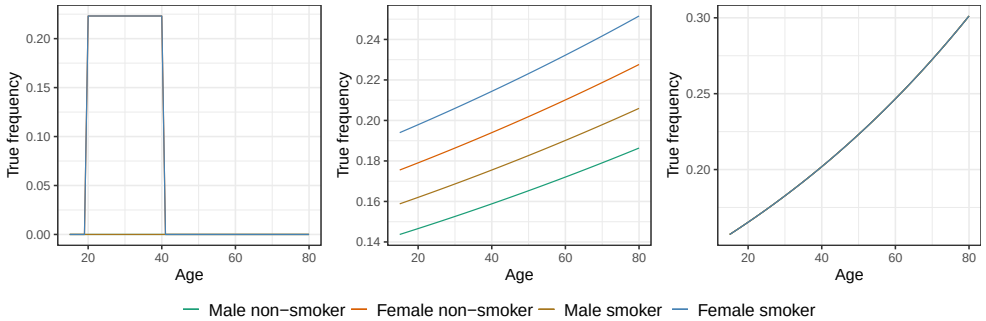
$$\lambda_2(\boldsymbol{X}, D) \quad := \exp(\beta_0 + \beta_1 X_1 + \beta_2 1_{\{X_2=\text{smoker}\}} + \beta_3 1_{\{D=\text{female}\}})$$

(3) type 3 (other diseases)

$$\lambda_3(\boldsymbol{X}, D) \quad := \exp(\gamma_0 + \gamma_1 X_1).$$

Stylized example

Set up



Assumptions on previous sheet with $(\alpha_0, \alpha_1) = (-40, 38.5)$, $(\beta_0, \beta_1, \beta_2, \beta_3) = (-2, 0.004, 0.1, 0.2)$ and $(\gamma_0, \gamma_1) = (-2, 0.01)$.

Technical prices

Let us assume deterministic claim amounts, namely $(c_1, c_2, c_3) = (0.5, 0.9, 0.1)$.

Using these assumptions, the pricing formulas become:

- the best-estimate price

$$\mu(\boldsymbol{X}, D) = c_1\lambda_1(\boldsymbol{X}, D) + c_2\lambda_2(\boldsymbol{X}, D) + c_3\lambda_3(\boldsymbol{X}, D)$$

- the unawareness price

$$\mu(\boldsymbol{X}) = \sum_{d\in\{\text{female,male}\}} (c_1\lambda_1(\boldsymbol{X}, d) + c_2\lambda_2(\boldsymbol{X}, d) + c_3\lambda_3(\boldsymbol{X}, d))\mathbb{P}(D = d \mid \boldsymbol{X})$$

- a discrimination-free price

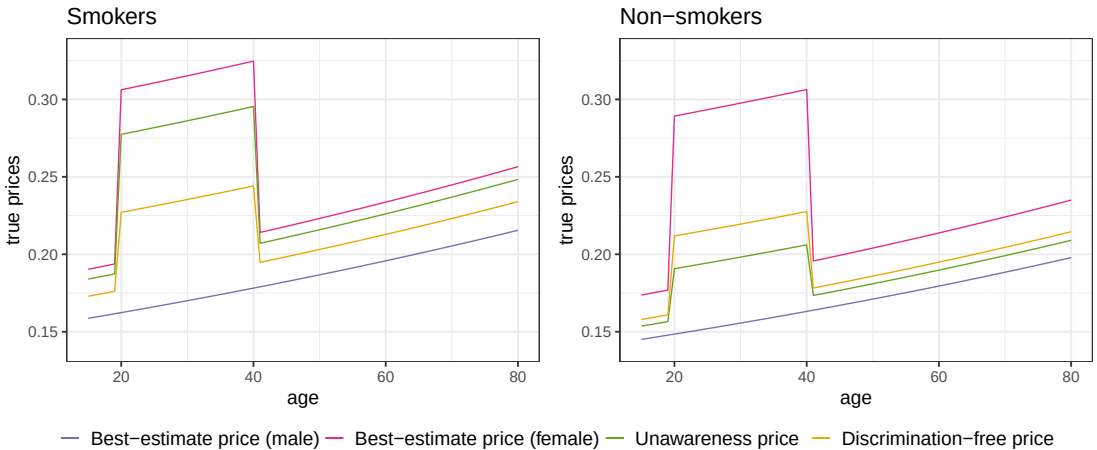
$$h^*(\boldsymbol{X}) = \textstyle\sum_{d\in\{\text{female,male}\}} (c_1\lambda_1(\boldsymbol{X}, d) + c_2\lambda_2(\boldsymbol{X}, d) + c_3\lambda_3(\boldsymbol{X}, d))\mathbb{P}(D = d).$$

To calculate the unawareness prices, we use:

$$\begin{aligned}\mathbb{P}(D = d \mid \boldsymbol{X}) &= \frac{\mathbb{P}(D = d, \boldsymbol{X} = \boldsymbol{x})}{\mathbb{P}(\boldsymbol{X} = \boldsymbol{x})} \\ &= \frac{\mathbb{P}(D = d, X_2 = x_2)}{\mathbb{P}(X_2 = x_2)},\end{aligned}$$

assuming X_1 is independent of (D, X_2).

Technical prices



Some reflections:

- smokers are charged a higher price than non-smokers, due to a higher expected number of type 2 claims
- for smokers: unawareness price $>$ discrimination-free price, and vice versa for non-smokers
- if a policyholder indicates to be a smoker, the unawareness price implicitly incorporates the higher price for women (on type 1 and type 2 claims) because - in this portfolio - smokers are more likely women
- explore the impact on the unawareness prices when - say - $\mathbb{P}(D = \text{female} \mid X_2 = \text{smoker}) = 0.2$ instead of the assumed 0.8.

We obtain the following `Gender-Smoking habits` probabilities

$$\begin{aligned}
\mathbb{P}(D = \text{female} \cap X_2 = \text{smoker}) &= \mathbb{P}(D = \text{female} \mid X_2 = \text{smoker}) \cdot \mathbb{P}(X_2 = \text{smoker}) \\
&= 0.8 \cdot 0.3 = 0.24 \\
\mathbb{P}(D = \text{female} \cap X_2 = \text{non-smoker}) &= 0.21 \\
\mathbb{P}(D = \text{male} \cap X_2 = \text{smoker}) &= 0.06 \\
\mathbb{P}(D = \text{male} \cap X_2 = \text{non-smoker}) &= 0.49,
\end{aligned}$$

using Bayes rule.

We generate 100 000 policyholder records, and distribute their `Gender-Smoking habits` risk profiles along this probability distribution.

Generating a data set

Age of the policyholder is simulated from a specified probability distribution.

Number of claims per type is sampled from a POI distribution with mean λ_j (for $j = 1, 2, 3$).

Excerpt from this data set:

#	Gender	Smoker	Age	Type_1	Type_2	Type_3
1	0	0	61	0	0	0
2	1	0	46	0	0	0
3	0	0	68	0	0	0
4	1	0	46	0	0	0
5	0	1	39	0	1	0
6	0	0	70	0	0	1

For the **best-estimate frequencies**, we calibrate three POI GLMs: (for $j = 1, 2, 3$)

$$\begin{aligned} N_j &\sim \text{POI}(\lambda_j) \\ \lambda_j &= \exp\left(\beta_0^j + \beta_1^j X_1 + \beta_2^j X_2 + \beta_3^j D\right). \end{aligned}$$

The **best-estimate prices** then follow from:

$$\hat{\mu}(\boldsymbol{X}, D) = c_1 \hat{\lambda}_1(\boldsymbol{X}, D) + c_2 \hat{\lambda}_2(\boldsymbol{X}, D) + c_3 \hat{\lambda}_3(\boldsymbol{X}, D),$$

with the c's denoting the (fixed, type-specific) claim amounts.

For the **unawareness frequencies**, we drop the D variable from the linear predictor:

$$\begin{aligned} N_j &\sim \text{POI}(\lambda_j) \\ \lambda_j &= \exp\left(\beta_0^j + \beta_1^j X_1 + \beta_2^j X_2\right). \end{aligned}$$

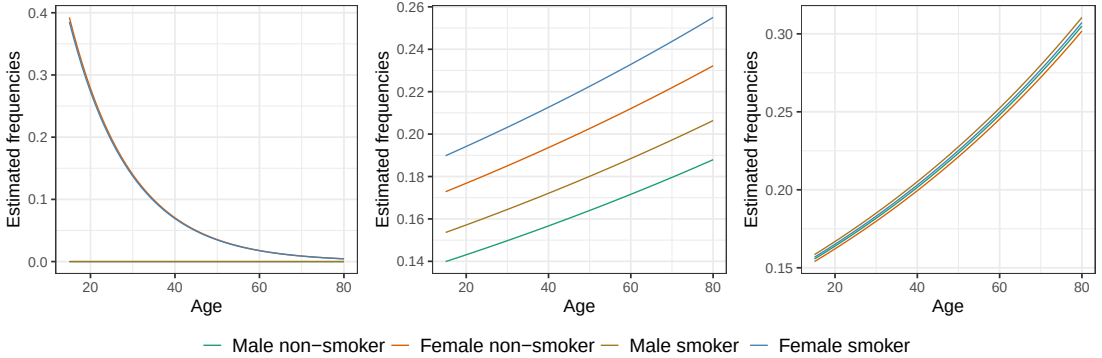
For the **discrimination-free prices**, we average the best-estimate prices over the values of D, using the empirical probabilities:

$$\hat{h}^*(\boldsymbol{x}) = \sum_d \hat{\mu}(\boldsymbol{x}, d)\frac{n_d}{n},$$

with $n = 100\ 000$ the sample size.

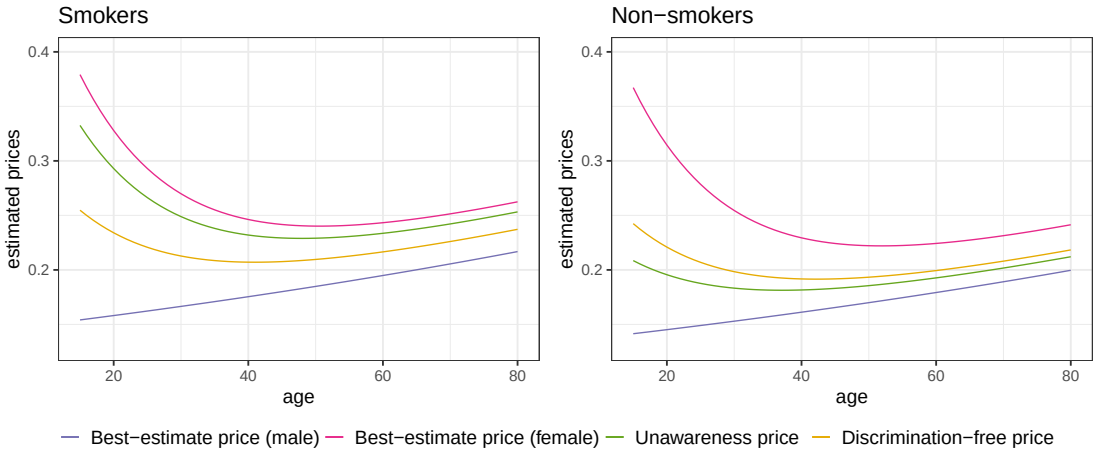
Estimated prices

GLMs



Estimated prices

GLMs



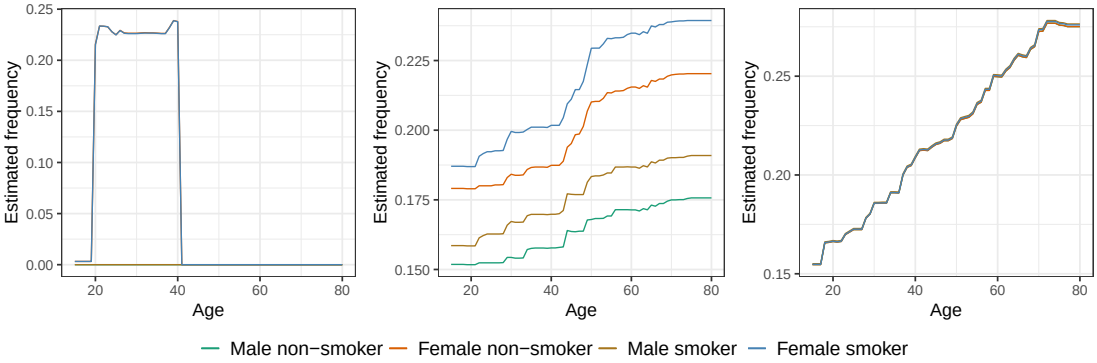
Lindholm et al. (2022) analyze the simulated data set with neural network regression.

We opt for a **Gradient Boosting Machine (GBM)** (see Henckaerts et al., 2021):

- iteratively combines weak learners into a powerful predictor
- at each iterative step a new tree is fit using information from previously grown trees
- tuning parameters: shrinkage, interaction depth, number of trees, minimum number of observations in a node and the bag fraction.

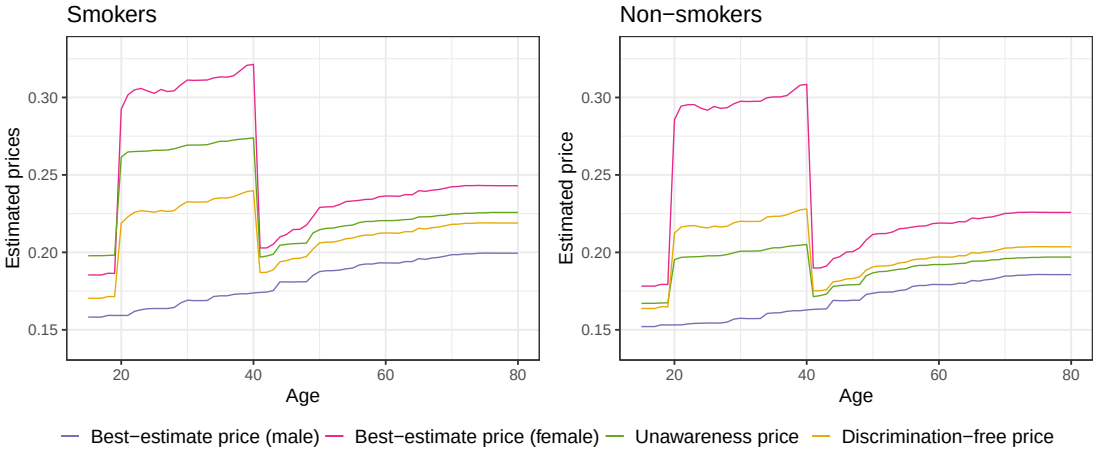
Estimated prices

GBMs - estimated frequencies



Estimated prices

GBMs - estimated prices



Estimated prices

GLMs & GBMs - some findings

Comparing the assumed technical prices and the calibrated ones:

- GLM prices are a poor approximation to the true prices, because of difficulties capturing the highly nonlinear birthing-related effects
- GBMs perform better
- unawareness price discriminates indirectly by learning the gender D from smoking habits X_2.

Throughout the workshop we

- took a quick tour of the literature on bias and fairness in machine learning
- explored a recently proposed strategy to adjust insurance prices for discrimination by proxy, when information on protected features is available.