IN-PROCESSING OF ACTUARIAL AND EQUITY FAIRNESS CONSTRAINTS FOR NEURAL NETWORKS

Donatien Hainaut

LIDAM Discussion Paper ISBA
2025 / 11

UCLouvain

ISBA | LIDAM
Louvain Institute of Data Analysis and Modeling in economics and statistics

ISBA

Voie du Roman Pays 20 - L1.04.01

B-1348 Louvain-la-Neuve

Email : lidam-library@uclouvain.be

https://uclouvain.be/en/research-institutes/lidam/isba/publication

In-processing of actuarial and equity fairness constraints for Neural networks

Donatien Hainaut*
UCLouvain, ISBA-LIDAM

May 12, 2025

This article introduces a novel in-processing method for integrating actuarial and equity fairness into neural networks used for actuarial valuation. We consider one primary network penalized during training to ensure balanced predictions (actuarial fairness) and independence from sensitive features (equity fairness). Global and local actuarial equilibrium is obtained by aligning the inter-quantile averages of predicted and observed responses. Meanwhile, a second auxiliary network penalizes the primary network for discriminatory predictions. The combined training algorithm effectively preserves predictive accuracy while mitigating discrimination. Numerical illustrations on real-world datasets demonstrate the method's efficacy in achieving fair and reliable insurance pricing models.

KEYWORDS : neural network, equity fairness, actuarial fairness, non-life pricing

1 Introduction

In recent years, the integration of machine learning techniques, particularly neural networks, into actuarial science has gained significant attention due to their potential to enhance predictive accuracy and efficiency. We refer the reader to Denuit et al. (2019b, 2020) for an introduction. Traditional actuarial models, such as Generalized Linear Models (GLMs), have long been the cornerstone of insurance pricing and risk assessment. However, neural networks offers a more flexible and powerful alternative, capable of capturing complex, non-linear relationships within data, as underlined by Richman and Wüthrich (2024).

Despite their advantages, neural networks pose unique challenges, discussed in Aas et al. (2024), particularly concerning fairness and transparency. Actuarial fairness ensures that predictions are unbiased and balanced across individuals with similar risk profiles, while equity fairness guarantees that predictions do not discriminate based on protected characteristics such as race, gender, or age. Addressing these fairness concerns is crucial, given the regulatory and ethical implications in the insurance industry.

This article introduces a novel in-processing method for incorporating both actuarial and equity fairness constraints into neural networks. Unlike pre-processing or post-processing approaches, in-processing methods modify the learning algorithm itself to enforce fairness during

*Corresponding author. Postal address: Voie du Roman Pays 20, 1348 Louvain-la-Neuve (Belgium). E-mail to: donatien.hainaut(at)uclouvain.be

model training. Padala and Gujar (2020) propose an in-processing method that includes fairness constraints in the loss function using Lagrangian multipliers. Mao et al. (2023) fine-tune the last layers of the network to enforce fairness. Our approach instead leverages two penalties introduced in the learning phase of a primary network. The first penalty aligns predicted and observed inter-quantile averages to ensure global and local actuarial fairness. The second penalty is computed by an auxiliary neural network, trained simultaneously, which enforces equity fairness by ensuring predictions are independent of sensitive features.

The proposed method builds on existing literature that highlights the importance of fairness in machine learning. For instance, Kamishima et al. (2012) discuss fairness-aware classifiers that incorporate penalties to counteract unfair decisions. Similarly, Richman and Wüthrich (2024) introduce ICEnet, a method for enforcing smoothness and monotonicity constraints in neural networks. Our approach extends these concepts by integrating fairness constraints directly into the training process, thereby enhancing both predictive accuracy and fairness.

Numerical illustrations using real-world datasets, such as a bank churn dataset and a motorcycle insurance dataset, demonstrate the effectiveness of our method. These case studies show that our approach not only mitigates discrimination but also maintains or improves predictive performance compared to traditional models and non-constrained neural networks.

In summary, this article presents a new application of neural networks to actuarial science, addressing critical fairness concerns while leveraging the predictive power of modern machine learning techniques. This dual fairness approach represents a robust framework for developing fair and reliable insurance pricing models.

2 Actuarial neural network regression

In actuarial regressions, we aim to calculate expected claim frequency, amount or probability of an event for a given exposure to risk and given some available information. The explanatory factors are stored in a vector $\boldsymbol{z} = (\boldsymbol{x}, s) \in \mathcal{Z} \subset \mathbb{R}^{m+1}$. The vector $\boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^m$ contains non-discriminatory features $x_1, x_2, \ldots, x_m$ whereas s is a sensitive attribute with q levels. The discriminatory feature s may not used as covariate. Nevertheless, ignoring the discriminatory factor is not sufficient to guarantee the fairness of a regression model. We will come back on this point in Section 3.1 and in a first time, simply exclude it from the set of explanatory variables. In the remainder of this section, we briefly review the method of neural network regressions.

The quantity of interest is a random continuous or discrete random variable R, with an exposure denoted by ω. For instance, R is the total amount of claims, the total number of claims or of defaults for a portfolio of ω insurance policies with common features $\boldsymbol{z}$. The key ratio or response is the quantity of interest scaled by the exposure: $Y = \frac{R}{\omega}$. The domain of Y is $\mathcal{Y} \subset \mathbb{R}$. If R and Y are continuous random variables their pdf's denoted by $p_R(r \,|\, \boldsymbol{x})$ and $p(y \,|\, \boldsymbol{x})$ are related by:

$$p(y \,|\, \boldsymbol{x}) = \omega p_R(\omega y \,|\, \boldsymbol{x}) \,.$$

If R and Y are discrete random variables, their pmf's denoted by $p_R(r \,|\, \boldsymbol{x})$ and $p(y \,|\, \boldsymbol{x})$ satisfy the following equality

$$p(y \,|\, \boldsymbol{x}) = p_R(\omega y \,|\, \boldsymbol{x}) \,.$$

In this article, we assume that Y belongs to the family of exponential dispersed (ED) distributions. The generic probability density function admits in this case the general representation:

$$p\left(y \,|\, \boldsymbol{x}\right) = \exp\left\{\frac{y\,\theta(\boldsymbol{x}) - \gamma\left(\theta(\boldsymbol{x})\right)}{\phi/\omega} + c(y, \phi, \omega)\right\} \tag{1}$$

where $\theta(\boldsymbol{x}) : \mathcal{X} \to \mathbb{R}$ and $\phi \in \mathbb{R}^+$ is a dispersion parameter. The $\gamma(.)$ function is $\mathcal{C}^2$ and admits an inverse second order derivative whereas the function $c(.)$ is independent from $\theta(\boldsymbol{x})$. Table 1 presents the most common and useful distributions of responses for actuarial purposes. The corresponding functions $\theta(\boldsymbol{x})$, $\gamma(.)$ and parameter ϕ are provided in Table 2.

R	Key ratio, $Y = R/\omega$	$p(y \mid \boldsymbol{x})$
Sum of ω Normal claims, $\mathcal{N}(\omega\mu(\boldsymbol{x}), \omega\sigma^2)$	Y is Normal, $\mathcal{N}(\mu(\boldsymbol{x}), \frac{\sigma^2}{\omega})$	$\frac{\sqrt{\omega}}{\sigma\sqrt{2\pi}} e^{-\frac{\omega}{2}\left(\frac{y-\mu(\boldsymbol{x})}{\sigma}\right)^2}$.
Sum of ω Gamma claims, $\mathcal{G}(\omega\alpha, \beta(\boldsymbol{x}))$	Y is Gamma, $\mathcal{G}(\omega\alpha, \omega\beta(\boldsymbol{x}))$	$\frac{(\omega\beta(\boldsymbol{x}))^{\omega\alpha}}{\Gamma(\omega\alpha)} y^{\omega\alpha-1} e^{-\omega\beta(\boldsymbol{x})y}$
Sum of ω Inverse Gaussian claims $\mathcal{IG}(\omega\mu(\boldsymbol{x}), \omega^2\alpha)$	Y is Inverse Gaussian, $\mathcal{IG}(\mu(\boldsymbol{x}), \omega\alpha)$	$\sqrt{\frac{\alpha\omega}{2\pi y^3}} e^{-\frac{\alpha\omega(y-\mu(\boldsymbol{x}))^2}{2y\mu(\boldsymbol{x})^2}}$
Sum of ω Poisson number of claims, $\mathcal{P}(\lambda(\boldsymbol{x})\,\omega)$	ωY is Poisson, $\mathcal{P}(\lambda(\boldsymbol{x})\,\omega)$	$e^{-\omega\lambda(\boldsymbol{x})} \frac{(\omega\lambda(\boldsymbol{x}))^{\omega y}}{(\omega y)!}$
Sum of ω Bernoulli defaults $\mathcal{B}(\omega, p(x))$	ωY is Binomial, $\mathcal{B}(\omega, p(x))$	$\binom{\omega}{\omega y} p(\boldsymbol{x})^{\omega y} (1 - p(\boldsymbol{x}))^{\omega(1-y)}$

Table 1: Most common statistical distributions used for actuarial analysis.

Key ratio, $Y = R/\omega$	$\theta(\boldsymbol{x})$	ϕ	$\gamma(z)$
$\mathcal{N}(\mu(\boldsymbol{x}), \frac{\sigma^2}{\omega})$	$\mu(\boldsymbol{x})$	σ^2	$z^2/2$
$\mathcal{G}(\omega\alpha, \omega\beta(\boldsymbol{x}))$	$-\frac{\beta(\boldsymbol{x})}{\alpha}$	$\frac{1}{\alpha}$	$-\ln(-z)$
$\mathcal{IG}(\mu(\boldsymbol{x}), \omega\alpha)$	$-\frac{1}{2\mu(\boldsymbol{x})^2}$	$\frac{1}{\alpha}$	$-\sqrt{-2z}$
$\mathcal{P}(\lambda(\boldsymbol{x})\,\omega)$	$\ln\lambda(\boldsymbol{x})$	1	e^z
$\mathcal{B}(\omega, p(x))$	$\ln\frac{p(\boldsymbol{x})}{1-p(\boldsymbol{x})}$	1	$\ln(1+e^z)$

Table 2: Reformulation of distributions in Table 1 as exponential dispersed (ED) distributions.

Using the canonical representation of ED distributions allows us to standardize future developments. Within this framework, the cumulant generating function is

$$E(e^{tY}) = \exp\left(\frac{\gamma\left(\theta(\boldsymbol{x}) + t\frac{\phi}{\omega}\right) - \gamma(\theta(\boldsymbol{x}))}{\phi/\omega}\right). \tag{2}$$

Deriving this expression allows us to infer the expectation and variance of responses:

$$\begin{cases} \mu(\boldsymbol{x}) = \mathbb{E}(Y) = \gamma'(\theta(\boldsymbol{x})), \\ \sigma^2(\boldsymbol{x}) = \mathbb{V}(Y) = \frac{\phi}{\omega}\gamma''(\theta(\boldsymbol{x})). \end{cases}$$

We deduce that $\theta(\boldsymbol{x})$ is related to $\mu(\boldsymbol{x})$ by the inverse of the derivative of $\gamma(.)$:

$$\theta(\boldsymbol{x}) \;=\; \gamma'^{-1}(\mu(\boldsymbol{x})).$$

Whereas $\sigma^2(\boldsymbol{x})$ depends on $\mu(\boldsymbol{x})$ through the variance function $v(.)$, that is defined by:

$$\sigma^2(\boldsymbol{x}) = \frac{\phi}{\omega} v(\mu(\boldsymbol{x})) \text{ where } v(.) = \gamma''\left(\gamma'^{-1}(.)\right).$$

The variance functions of main exponential dispersed distributions are reported in Table 3 for information.

Key ratio, $Y = R/\omega$	$\mathbb{E}(Y)$	$\mathbb{V}(Y)$	$v(z)$
$\mathcal{N}(\mu(\boldsymbol{x}), \frac{\sigma^2}{\omega})$	$\mu(\boldsymbol{x})$	$\frac{\sigma^2}{\omega}$	1
$\mathcal{G}(\omega\alpha, \omega\beta(\boldsymbol{x}))$	$\frac{\alpha}{\beta(\boldsymbol{x})}$	$\frac{\alpha}{\omega\beta(\boldsymbol{x})^2}$	z^2
$\mathcal{IG}(\mu(\boldsymbol{x}), \omega\alpha)$	$\mu(\boldsymbol{x})$	$\frac{\mu(\boldsymbol{x})^3}{\omega\alpha}$	z^3
$\mathcal{P}(\lambda(\boldsymbol{x})\,\omega)$	$\lambda(\boldsymbol{x})$	$\frac{\lambda(\boldsymbol{x})}{\omega}$	z
$\mathcal{B}(\omega, p(x))$	$p(\boldsymbol{x})$	$\frac{p(\boldsymbol{x})(1-p(\boldsymbol{x}))}{\omega}$	$z(1-z)$

Table 3: Expectations, variances and variance functions of exponential dispersed (ED) distributions.

In numerical illustrations, we use the generalized linear model (GLM) as a benchmark, which is one of the most common regression tools for actuarial pricing. In this framework, the expectation is related to a linear combination of features through a link function, $g(.)$ as follows:

$$g(\mu(\boldsymbol{x})) = \boldsymbol{\beta}^\top \boldsymbol{x}\,, \tag{3}$$

where $\boldsymbol{\beta} \in \mathbb{R}^m$ is a vector of regression coefficients to estimate. If $g(\mu(\boldsymbol{x})) = \theta(\boldsymbol{x})$, the link function is called canonical and equal to $g(.) = \gamma'^{-1}(.)$. Table 4 reports the canonical link functions of distributions in Table 1. In certain circumstances, using a canonical link function simplifies developments. For further details on ED distributions and GLM, we refer to the first chapters of Denuit et al. (2019 a).

Link function	$g(\mu) = \gamma'^{-1}(\mu)$	Canonical link for
Identity	μ	Normal
Power	$-\mu^{-1}$	Gamma
	$-\frac{1}{2}\mu^{-2}$	Inv. Gaussian
Log	$\ln \mu$	Poisson
Logit	$\ln \frac{\mu}{1-\mu}$	Binomial

Table 4: Canonical link functions.

A neural network regression is a non-linear extension of a GLM in which the expectation is related to explanatory features by a neural network, $F(\boldsymbol{x}|\Omega)$:

$$\mu(\boldsymbol{x}) = F(\boldsymbol{x}|\Omega)\,, \tag{4}$$

where Ω is a set of weights to estimate. We consider feed-forward networks where information flows forward through several neural layers. We recall the definition of this type of neural network.

Definition Let l, n_0, n_1, ..., $n_l \in \mathbb{N}$ be respectively the number of layers and neurons in each layer. n_0 is also the size of input vector. The activation function of layer $k = 1, 2, ..., l$ is noted $\phi_k(.) : \mathbb{R} \to \mathbb{R}$. Let $C_k \in \mathbb{R}^{n_k} \times \mathbb{R}^{n_{k-1}}$ and $\boldsymbol{c}_k \in \mathbb{R}^{n_k}$ for $k = 1, ..., l$, be neural weights defining the input, intermediate and output layers. We define the following functions

$$\psi_k(\boldsymbol{x}) = \phi_k(C_k\boldsymbol{x} + \boldsymbol{c}_k)\,, k = 1, ..., l$$

where the activation functions $\phi_k(.)$ are applied component wise. The neural network is a function $F_\Omega : \mathbb{R}^{n_0} \to \mathbb{R}^{n_l}$ defined by

$$F(\boldsymbol{x}|\Omega) \quad = \quad \psi_l \circ \psi_{l-1} \circ ... \circ \psi_1(\boldsymbol{x})\,.$$

	Unscaled deviance, $D\left(y,\hat{y}\right)$
Normal	$\omega\left(y-\widehat{y}\right)^2$
Gamma	$\begin{cases} 2\omega\left(\frac{y}{\widehat{y}}-1-\ln\left(\frac{y}{\widehat{y}}\right)\right) & y>0 \\ 0 & y=0 \end{cases}$
Inverse Gaussian	$\omega\frac{(y-\widehat{y})^2}{y\widehat{y}^2} \quad y>0$
Poisson	$\begin{cases} 2\omega\left(y\ln y-y\ln\widehat{y}-y+\widehat{y}\right) & y>0 \\ 2\omega\widehat{y} & y=0 \end{cases}$
Binomial	$\begin{cases} 2\omega\left(y\ln\left(\frac{y}{\widehat{y}}\right)+(1-y)\ln\left(\frac{1-y}{1-\widehat{y}}\right)\right) & y\in(0,1) \\ -2\omega\ln\left(1-\widehat{y}\right) & y=0 \\ -2\omega\ln\left(\widehat{y}\right) & y=1 \end{cases}$

Table 5: Deviance statistics for Normal, Gamma, inverse Gaussian, Poisson and Binomial distributions.

In our context, $F(\boldsymbol{x}|\Omega)$ is a non-linear function from $\mathcal{X}\subset\mathbb{R}^m$ to $\mathcal{Y}\subset\mathbb{R}$. The activation function of the last layer, $\phi_l(.)$, is set to the inverse link function, $g^{-1}(.)$ of the equivalent GLM formulation. This ensures that network outputs are in the domain of responses. After choosing the network architecture, the model is trained by minimizing the deviance between the observation y and the response, $\widehat{y}(\boldsymbol{x})=F(\boldsymbol{x}|\Omega)$. We briefly recall this concept. Let $LL\left(\widehat{y}(\boldsymbol{x})\right)$ be the log-likelihood of $\widehat{y}(\boldsymbol{x})$. The highest log-likelihood is achieved with the saturated model in which the predictor is equal to the observation, $\widehat{y}=y$. The scaled deviance D^* is defined as the likelihood ratio test of the model under consideration against the saturated model:

$$D^*\left(y,\widehat{y}(\boldsymbol{x})\right)=2\left(LL(y)-LL(\widehat{y})\right) .$$

From the definition of exponential dispersed distributions and using $\theta(\boldsymbol{x})=\gamma'^{-1}\left(\hat{y}(\boldsymbol{x})\right)$, the scaled deviance may be rewritten as:

$$D^*\left(y,\hat{y}(\boldsymbol{x})\right)=2\frac{\omega}{\phi}\left(y\gamma'^{-1}\left(y\right)-\gamma\left(\gamma'^{-1}\left(y\right)\right)-y\gamma'^{-1}\left(\hat{y}(\boldsymbol{x})\right)+\gamma\left(\gamma'^{-1}\left(\hat{y}(\boldsymbol{x})\right)\right)\right) .$$

The unscaled deviance, defined as $D\left(y,\hat{y}(\boldsymbol{x})\right)=\phi D^*\left(y,\hat{y}(\boldsymbol{x})\right)$, measures the goodness of fit for estimating neural weights. Table 5 presents the deviance functions of Normal, Gamma, Inverse Gaussian, Poisson and Binomial distributions. For a data set $\mathcal{D}$ made of n policies $\left(\boldsymbol{x}_i,z_i,y_i,\omega_i\right)_{i=1,\ldots,n}$, optimal neural weights Ω^* are obtained by minimizing the total deviance, $\mathcal{L}_\Omega(\mathcal{D})$:

$$\begin{aligned}\Omega^* &= \arg\min_{\Omega}\sum_{\boldsymbol{x}_i\in\mathcal{D}}D\left(y_i,\widehat{y}(\boldsymbol{x}_i)\right) \\ &= \arg\min_{\Omega}\mathcal{L}(\mathcal{D},\Omega) .\end{aligned} \tag{5}$$

The minimization is performed by a gradient descent algorithm (or one of its variants) on a random sub-sample (a batch) of $\mathcal{D}$. We denote the neural weights after k training epochs by $\Omega^{(k)}$, the batch by $\mathcal{D}^{(k)}$ and the gradient of the total deviance by $\nabla_{\Omega^{(k)}}\mathcal{L}\left(\mathcal{D}^{(k)},\Omega^{(k)}\right)$. At epoch $k+1$, neural weights are updated as follows

$$\Omega^{(k+1)} \;=\; \Omega^{(k)}-\rho_k\nabla_{\Omega^{(k)}}\mathcal{L}\left(\mathcal{D}^{(k)},\Omega^{(k)}\right) , \tag{6}$$

where ρ_k is a decreasing learning rate. In the numerical illustration, the network is coded in TensorFlow and the gradient is computed with the command GradientTape(). The weights are updated with the Adaptive Moment (ADAM) algorithm. We refer the reader to Chapter 3 of

Denuit et al. (2019 b) for details about it. Despite neural networks being powerful regression tools, they suffer from two issues. First, there is no guarantee that predictions are unbiased and globally balanced compared to observations. Second, removing the sensitive feature from covariates does not prevent discriminatory predictions. In the next two sections, we will adapt the training procedure of the neural network regression to enforce actuarial and equity fairness.

3 Managing the fairness constraints

We distinguish between two types of fairness constraints. Actuarial fairness ensures that predictions for individuals with the same risk level are identical and unbiased. With this constraint, the model is balanced in the sense that the average of predictions (claim frequencies or amounts) does not deviate from observed ones. Nevertheless, neural networks may violate this balance condition. As pointed out by Wüthrich (2021), machine learning algorithms provide an accurate fit at the individual policy level, but the global price level may be completely wrong. It is therefore crucial to rebalance predictions to restore the global equilibrium of the tariff. We will propose in sub-section 3.2, an approach in which the balance property is preserved by constraining the learning phase.

On another side, the rating factors used in insurance segmentation often statistically correlate with characteristics protected under anti-discrimination laws (including race, color, national origin, gender, religion, or age). Equity fairness guarantees that insureds are equally treated, without discriminating based on these legally prohibited characteristics. There is no consensus about the best definition of fairness in the economic and actuarial literature. In the next section, we develop a training method based on an adversarial network checking the independence between predictions and the sensitive features. The neural network trained in this way achieves demographic parity, which means that different demographic groups are treated equitably.

3.1 Equity fairness

There are several approaches for mitigating fairness which are organized into three types of intervention (Kamishima et al., 2012). "pre-processing" is about modifying the original features to make them independent of the prohibited attribute (Charpentier et al., 2023; Frees & Huang, 2023; Komiyama & Shimao, 2017; Lindholm et al. 2024). "In-processing" is about modifying the learning algorithm itself by introducing some penalties in the objective function to counteract unfair decisions (Bechavod & Ligett, 2017; Grari et al., 2022; Kamishima et al., 2012). Finally, "Post-processing" is about modifying the model outcomes to align with some fairness criteria (see e.g. Frees & Huang (2023) or Charpentier (2023) for a comprehensive overview). In the absence of anti-discrimination regulations, the actuary would consider "all" available covariates as rating factors $\boldsymbol{z} = (\boldsymbol{x}, s)$ to estimate the response:

$$y(z) = \mathbb{E}\left(Y|z\right) = \mathbb{E}\left(Y|\boldsymbol{x}, s\right) .$$

Under anti-discrimination laws, a naive approach is to not explicitly use s as a rating factor. We denote by $\widehat{y}(\boldsymbol{X})$ the prediction of the neural network $F(\boldsymbol{X}|\Omega)$ where $\boldsymbol{X}$ is the vector of random features and S the associated random discriminatory variable. As $F(\boldsymbol{X}|\Omega)$ only depends on $\boldsymbol{X}$, it satisfies the fairness through unawareness criterion (Dwork et al. 2012) with respect to the sensitive attribute. But removing discriminatory covariates does not guarantee discrimination-free insurance prices because other rating factors $\boldsymbol{X}$ are usually not independent from S. In the literature, most of the definitions of fairness fall inside or try to reflect one of the following three principles: $\widehat{y}(\boldsymbol{X}) \perp\!\!\!\perp S$ (independence), $\widehat{y}(\boldsymbol{X}) \perp\!\!\!\perp S \,|\, Y$ (separation), $\widehat{y}(\boldsymbol{X}) \perp\!\!\!\perp S \,|\, \widehat{y}(\boldsymbol{X})$ (sufficiency). In this article, we adapt the network during the training process in order to obtain independent predictions from the sensitive feature (principle 1). We recall that the independence

implies a first weak criterion of fairness, the weak demographic parity (WDP). A model satisfies this criterion if

$$\mathbb{E}\left(\widehat{y}(\boldsymbol{X}) \,|\, S=i\right) \quad = \quad \mathbb{E}\left(\widehat{y}(\boldsymbol{X}) \,|\, S=j\right) \; i,j \in \{1,...,q\}. \tag{7}$$

Independence also induces a stronger criterion of fairness called strong demographic parity (SDP). A model fulfills the SDP criterion when

$$\mathbb{P}\left(\widehat{y}(\boldsymbol{X}) \in \mathcal{A} \,|\, S=i\right) \quad = \quad \mathbb{P}\left(\widehat{y}(\boldsymbol{X}) \in \mathcal{A} \,|\, S=j\right) \quad \forall \mathcal{A} \subset \mathcal{Y}\,. \tag{8}$$

To achieve independence, we consider an auxiliary neural network, $F_D(.\,|\,\Omega_D)$, with weights Ω_D, that classifies policies between the q-discriminated categories of S. $F_D : \mathbb{R}^{n_d} \to [0,1]^q$, takes as inputs, the outputs of the d^{th} intermediate layer of $F(\boldsymbol{x}|\Omega)$:

$$\varphi(\boldsymbol{x}) \quad = \quad \psi_d \circ \psi_{d-1} \circ ... \circ \psi_1(\boldsymbol{x})\,. \tag{9}$$

The input layer for $F_D(.)$ is a-priori chosen (typically, we select a mid-network layer). $F_D(\varphi(\boldsymbol{x})\,|\,\Omega_D)$ returns a $q-$vector of probabilities, $\widehat{\boldsymbol{s}}(\boldsymbol{x}) = \left(\widehat{s}_j(\boldsymbol{x})\right)_{j=1,...,q}$ where

$$\widehat{s}_j(\boldsymbol{x}) = \widehat{\mathbb{P}}(S=j|\boldsymbol{X}=\boldsymbol{x}) \quad \text{for } j=1,...,q.$$

To ensure that outputs of $F_D(.)$ are probabilities, The activation function of its last layer is the softmax function:

$$\left(\phi(\boldsymbol{u})\right)_{j=1,...,q} = \frac{\mathrm{e}^{u_j}}{\sum_{i=1}^{q} \mathrm{e}^{u_i}}\,.$$

The predictive power of $F_D(.)$ is measured by the deviance of a multinomial distribution, that is equal to twice the cross-entropy:

$$D_M\left(s, \hat{\boldsymbol{s}}(x)\right) = -2 \sum_{j=1}^{q} \log\left(\widehat{s}_j(\boldsymbol{x})\right) \mathbf{1}_{\{s=j\}}\,.$$

We simultaneously train the primary and auxiliary networks. Our goal is to wipe out during the training, any dependence between the sensitive feature and the output $\varphi(\boldsymbol{x})$, of the d^{th} intermediate layer of $F(\boldsymbol{x}|\Omega)$. The auxiliary network $F_D(.)$ only serves in the training phase to avoid discriminatory predictions of $F(.)$.

The weights and the batch after k training epochs are denoted by $\Omega_D^{(k)}$ and $\mathcal{D}^{(k)}$ while $\boldsymbol{\varphi}^{(k)}(\boldsymbol{x})$ is the n_d-vector of outputs of the d^{th}-layer of $F(\boldsymbol{x}|\Omega^{(k)})$. At epoch k, the predicted probabilities are denoted by $\widehat{\boldsymbol{s}}_i^{(k)} = F_D(\varphi^{(k)}(\boldsymbol{x}_i)\,|\,\Omega_D^{(k)})$ for $\boldsymbol{x}_i \in \mathcal{D}$. The loss function for the training of $F_D(.)$ is the total multinomial deviance:

$$\mathcal{L}_D(\mathcal{D}^{(k)}, \Omega_D^{(k)}) \quad = \quad \sum_{\boldsymbol{x}_i \in \mathcal{D}^{(k)}} D_M\left(s_i, \widehat{\boldsymbol{s}}_i^{(k)}\right)\,. \tag{10}$$

The gradient of this function with respect to neural weights, is$\nabla_{\Omega_D^{(k)}} \mathcal{L}_D\left(\mathcal{D}^{(k)}, \Omega_D^{(k)}\right)$. At epoch $k+1$, neural weights of $F_D(.)$ are updated in order to reduce the deviance (10):

$$\Omega_D^{(k+1)} \quad = \quad \Omega_D^{(k)} - \rho_k \nabla_{\Omega_D^{(k)}} \mathcal{L}_D\left(\mathcal{D}^{(k)}, \Omega_D^{(k)}\right)\,, \tag{11}$$

and the probability estimates becomes $\widehat{\boldsymbol{s}}_i^{(k+1)} = F_D\left(\varphi^{(k)}(\boldsymbol{x}_i)|\Omega_{BC}^{(k+1)}\right)\,.$

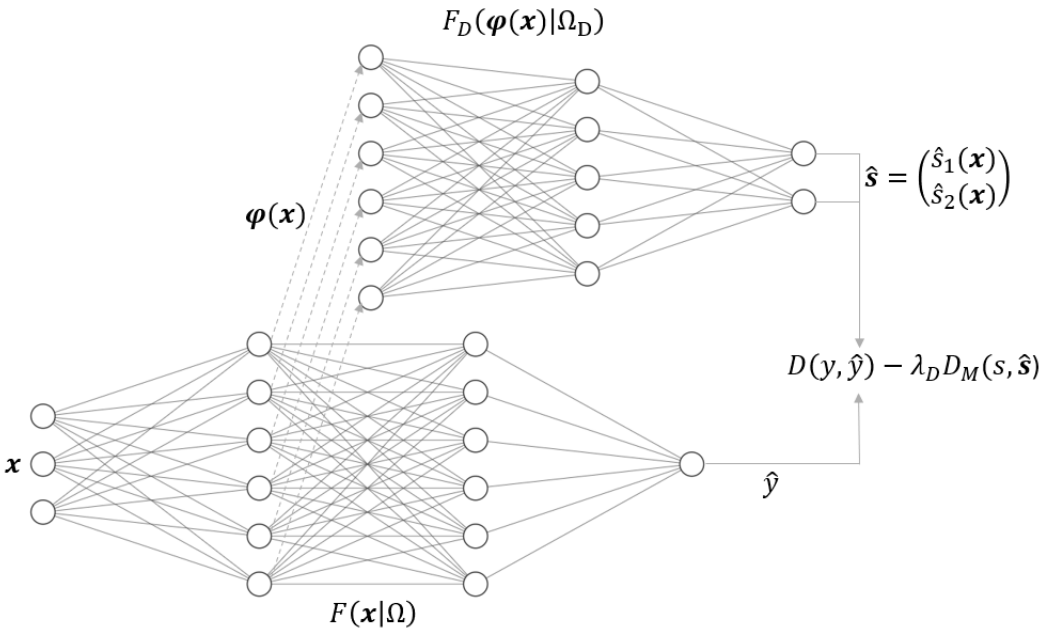


Figure 1: Plot of primary and auxiliary neural networks for equity fairness, when $q = 2$.

As illustrated in Figure 1, the role of $F_D(.)$ is to monitor the information about the sensitive feature, contained in the outputs of the d^{th} layer of the primary network. During the training of $F(.)$, we aim to decrease as much as possible the predictive power of $F_D(.)$ by tuning neural weights involved in the calculation of $\boldsymbol{\varphi}^{(k)}(\boldsymbol{x})$. This is equivalent to adapting the primary network in order to maximize the deviance $\mathcal{L}_D(\mathcal{D}^{(k)}, \Omega_D^{(k+1)})$ of the auxiliary network. Based on this, we define the equity fair loss function as a penalized version of Equation (5):

$$\mathcal{L}_{EF}(\mathcal{D}^{(k)}, \Omega^{(k)}) \quad = \quad \sum_{\boldsymbol{x}_i \in \mathcal{D}^{(k)}} D\left(y_i, \widehat{y}_i^{(k)}\right) - \lambda_D \mathcal{L}_D(\mathcal{D}^{(k)}, \Omega_D^{(k)}) \,,$$

where $\lambda_D \in \mathbb{R}^+$ is a penalty weight for equity fairness. The weights of the primary neural network $F(.)$ are again updated by gradient descent (or one of its variants):

$$\Omega^{(k+1)} \quad = \quad \Omega^{(k)} - \rho_k \nabla_{\Omega^{(k)}} \mathcal{L}_{EF}\left(\mathcal{D}^{(k)}, \Omega^{(k)}\right) \,, \tag{12}$$

and predictions are recomputed with these new parameters, $\widehat{y}_i^{(k+1)} = F\left(\boldsymbol{x}_i|\Omega^{(k+1)}\right)$. $F_D(.)$ may be seen as an adversarial network for $F(.)$ which is trained to annihilate the classification power of $F_D(.)$. After convergence, the last $(l-d)$ layers of $F(.)$ are fed by input signals without (or nearly without) informative content about the sensitive feature. This ensures that predictions are not discriminating. We can draw a parallel between our approach and the pre-processing methods which a priori modify the original features. We internalize this pre-processing during the training of the neural network.

In numerical illustrations, we measure the degree of independence between the prohibited feature and the model output by comparing divergences and distances between distributions of predictions for each discriminated group. The Jenssen-Shannon divergence, the $p-$Wasserstein and Cramèr distance are reminded in Appendix B.

3.2 Actuarial fairness

A model is globally balanced when the average of predictions, weighted by exposures, does not depart from the empirical equivalent average:

$$\sum_{\boldsymbol{x}_i \in \mathcal{D}} \omega_i \widehat{y}(\boldsymbol{x}_i) \quad = \quad \sum_{\boldsymbol{x}_i \in \mathcal{D}} \omega_i y(\boldsymbol{x}_i)\,. \tag{13}$$

This global equilibrium is naturally achieved with a GLM and the canonical link function as recalled in Appendix A. The global balance condition BC (13) is not necessarily fulfilled by predictions of a neural network. A natural remedy consists in restoring global balance between average fitted and observed values, at each step of the iterative procedure used to optimize the loss function. This is easily implemented by constraining the optimization so that the observed total matches its fitted counterpart. An alternative proposed in Wüthrich (2023, Chapter 7) is to use the canonical link function as the activation function of the last neuron of F. The last layer has in this case the same specification as a GLM. After optimizing the weights of the last layer, it shares all the properties of a GLM and fulfills the balance condition (13). These solutions restore the global balance but does not ensure that financial equilibrium also holds in meaningful sub-groups. Denuit et al. (2021) solve this issue by auto-calibration with local GLM regressions.

Instead, we propose to enforce the local balance condition by adding a constraint to the loss function of the primary network. Contrary to existing approaches, the balance condition is no longer post-processed. The neural weights are tuned during training to align inter-quantile averages of predicted and observed responses. Our approach is also designed to be compatible with the one in Section 3.1, for restoring equity fairness. Nevertheless, we delay the presentation of the combined methods to the next section for clarity.

To achieve global and local actuarial fairness, we modify the loss function of the primary neural network, $F(\boldsymbol{x}|\Omega)$, with neural weights, Ω. This network takes as input the non-prohibited features, $\boldsymbol{x}$, and outputs $\widehat{y}(\boldsymbol{x})$. We adjust Ω to align inter-quantile averages of predicted and observed responses. This method ensures that $\widehat{y}(\boldsymbol{x})$ fulfills the balance condition (13). This also guarantees local equilibrium between insureds with a similar risk profile. The weights and the batch after k training epochs are noted $\Omega^{(k)}$, $\mathcal{D}^{(k)}$. At epoch k, the predictions are $\widehat{y}_i^{(k)}$ for $\boldsymbol{x}_i \in \mathcal{D}^{(k)}$. We first rank $\widehat{y}_i^{(k)}$ by ascending order and denote by $\widehat{y}_{(i)}^{(k)}$, the ordered predictions. The corresponding responses and exposures are $y_{(i)}^{(k)}$ and $\nu_{(i)}^{(k)}$. If $u(s) = s\left\lfloor \frac{|\mathcal{D}^{(k)}|}{n_q} \right\rfloor$, we define the n_q-inter-quantile averages of observed and predicted responses as follows:

$$\begin{cases} \widehat{q}_{(s)}^{(k)} & = \sum_{j=u(s-1)}^{u(s)} \nu_{(j)}^{(k)} \widehat{y}_{BC,(j)}^{(k)} \\ q_{(s)}^{(k)} & = \sum_{j=u(s-1)}^{u(s)} \nu_{(j)}^{(k)} y_{(j)}^{(k)} \end{cases}, \; s = 1, \ldots, n_q\,. \tag{14}$$

he penalty for preserving the balance condition is the sum of the roots of the squared spreads between the observed and predicted inter-quantile averages:

$$\mathcal{L}_{BC}(\mathcal{D}^{(k)}, \Omega^{(k)}) \quad = \quad \sqrt{\sum_{s=1}^{n_q} \left(q_{(s)}^{(k)} - \widehat{q}_{(s)}^{(k)} \right)^2}\,. \tag{15}$$

We define the actuarial fair loss function of the primary network as follows,

$$\mathcal{L}_{AF}(\mathcal{D}^{(k)}, \Omega^{(k)}) = \sum_{\boldsymbol{x}_i \in \mathcal{D}^{(k)}} D\left(y_i, \widehat{y}_i^{(k)}\right) + \lambda_{BC} \mathcal{L}_{BC}(\mathcal{D}^{(k)}, \Omega^{(k)})\,,$$

where $\lambda_{BC} \in \mathbb{R}^+$ is a penalty weight. In practice, the number of quantiles, n_q, is proportional to the size of the data set and batches. For a small data sample, n_q must be small enough to yield statistically efficient estimates of $q_{(s)}^{(k)}$. The gradient of the loss function (15) is denoted by $\nabla_{\Omega^{(k)}} \mathcal{L}_{AF}\left(\mathcal{D}^{(k)}, \Omega^{(k)}\right)$. At epoch $k+1$, weights $\Omega^{(k)}$ of $F(.)$ are updated with a gradient descent:

$$\Omega^{(k+1)} \quad = \quad \Omega^{(k)} - \rho_k \nabla_{\Omega^{(k)}} \mathcal{L}_{AF}\left(\mathcal{D}^{(k)}, \Omega^{(k)}\right), \tag{16}$$

while balanced-corrected estimates are recomputed with these new parameters.

In a similar manner to Denuit et al. (2021), we could instead use a local GLM to a posteriori tune the outputs of $F(.)$. This post-processing involves two estimation steps and any modification of the network requires refitting the GLM. From an operational point of view, managing two separate models increases the risk of misuse or manipulation. It is often preferable to have a unique model which directly provides balanced forecasts. Furthermore, the a posteriori adjustments may deteriorate the deviance and lead to a sub-optimal model from this point of view. This motivates us to instead include a balance penalization during the training of $F(.)$.

In numerical illustration, we measure the model actuarial fairness by the root mean squared errors between inter-quantile averages of observed and predicted responses. This point is detailed in the first case study.

Algorithm 1 Training of an actuarial and equity fair network.

Initialization:

Initial random weights $\Omega_D^{(0)}$ and $\Omega^{(0)}$.

Loop on k :

Sampling of $\mathcal{D}^{(k)}$ and computations of equity and actuarial penalties:

$$\begin{cases} \mathcal{L}_D(\mathcal{D}^{(k)}, \Omega_D^{(k)}) & = \sum_{\boldsymbol{x}_i \in \mathcal{D}^{(k)}} D_M\left(s_i, \widehat{\boldsymbol{s}}_i^{(k)}\right), \\ \mathcal{L}_{BC}(\mathcal{D}^{(k)}, \Omega_{BC}^{(k)}) & = \sqrt{\sum_{s=1}^{n_q}\left(q_{(s)}^{(k)} - \widehat{q}_{(s)}^{(k)}\right)^2}. \end{cases}$$

Update weights of the auxiliary network $F_D(.)$:

$$\Omega_D^{(k+1)} = \Omega_D^{(k)} - \rho_k \nabla_{\Omega_D^{(k)}} \mathcal{L}_D\left(\mathcal{D}^{(k)}, \Omega_D^{(k)}\right),$$

Refresh the output of the auxiliary network:

$$\widehat{\boldsymbol{s}}_i^{(k+1)} = F_D\left(\varphi^{(k)}(\boldsymbol{x}_i) | \Omega_D^{(k+1)}\right),$$

Evaluate the penalized loss function of the primary network:

$$\begin{aligned} \mathcal{L}_{EAF}(\mathcal{D}^{(k)}, \Omega^{(k)}) = & \sum_{\boldsymbol{x}_i \in \mathcal{D}^{(k)}} D\left(y_i, \widehat{y}_i^{(k)}\right) \\ & + \lambda_{BC} \mathcal{L}_{BC}(\mathcal{D}^{(k)}, \Omega^{(k)}) - \lambda_D \mathcal{L}_D(\mathcal{D}^{(k)}, \Omega_D^{(k+1)}). \end{aligned}$$

Update neural weights of $F(.)$ and predictions:

$$\begin{cases} \Omega^{(k+1)} & = \Omega^{(k)} - \rho_k \nabla_{\Omega^{(k)}} \mathcal{L}_{EAF}\left(\mathcal{D}^{(k)}, \Omega^{(k)}\right), \\ \widehat{y}_i^{(k+1)} & = F\left(\boldsymbol{x}_i | \Omega^{(k+1)}\right). \end{cases}$$

End loop

Features	Description	Type
Customer_Age	Customer's Age in Years	num
Gender	M=Male, F=Female	cat
Dependent_count	Number of dependents	num
Education_Level	high school, college graduate, etc.	cat
Marital_Status	Married, Single, Divorced, Unknown	cat
Income_Category	Annual Income Category	cat
Card_Category	Blue, Silver, Gold, Platinum	cat
Contacts_Count_12_mon	No. of Contacts in the last 12 months	num
Total_Relationship_Count	Total no. of products held by the customer	num
Credit_Limit	Credit Limit on the Credit Card	num
Total_Revolving_Bal	Total Revolving Balance on the Credit Card	num
Total_Trans_Amt	Total Transaction Amount (last 12M)	num
Total_Trans_Ct	Total Transaction Count (Last 12M)	num
Avg_Utilization_Ratio	Average Card Utilization Ratio	num

Table 6: Features used for predicting the attrition probability and they type (cat: categorical, num: numerical).

3.3 Actuarial and Equity fairness

The methods of Sections 3.1 and 3.2 for restoring equity and actuarial fairness may be combined. In this setting, the primary network is doubly penalized for preserving the actuarial balance. The penalty for discrimination is computed by the auxiliary network. Algorithm 1 summarizes the full procedure. In the next sections, we test this algorithm on two real-world data sets and focus on gender discrimination.

4 Numerical illustrations

4.1 Churn data set

We analyze the bank churn dataset[1], which contains 10 207 records about 5358 female and 4769 male credit-card holders. We aim to estimate a neural network for predicting the customer's attrition probability, under equity and actuarial fairness constraints. Table 6 presents the available features. The gender is the prohibited attribute, s. Categorical variables are dummified and numerical variables are rescaled with a standard normalization. The average churn rates by gender are 85.38% for men and 82.64% for women. To test the capacity of our approach for mitigating the gender discrimination, we artificially reduce the female average churn rate to 66.01%. To do this, we randomly draw the attrition flag from a Bernoulli distribution with a probability of success of 80%, for women who have closed their bank account.

In this data set, gender is highly correlated with other explanatory variables. This is confirmed by a logistic regression of gender on other features. We can predict gender with an accuracy of 90.02% from other explanatory variables. As we will see, withdrawing the gender from the set of covariates is therefore not sufficient to avoid discrimination.

We sample the training and validation sets with 80% and 20% of the records, respectively. We estimate five models: two non-constrained neural networks (NN 16, NN 32), two fair neural networks (fair-NN 16 and fair-NN 32) with 16 or 32 neurons in their intermediate layers, and a GLM. The architecture of the networks is detailed in Table 7. The auxiliary networks take as input the outputs of the second layer of the primary network.

The penalty factors are set to $\lambda_D = 1$ and $\lambda_{BC} = 1$. We run 1500 epochs for training the networks with the ADAM optimizer and a learning rate of 0.005. We consider relatively small

[1] https://www.kaggle.com/sakshigoyal7/credit-card-customers

networks to avoid the recourse to, e.g. dropping out layers or Lasso penalties to limit over-fitting. This makes the effects of fairness constraints more visible. The training set being relatively small (8101 records) we do not use batches and only consider $n_q = 20$ quantiles of predicted responses to calculate the loss $\mathcal{L}_{BC}(\mathcal{D}^{(k)}, \Omega_{BC}^{(k)})$, as defined in Equation (15).

	Layers	Shape	Activation
NN, Fair-NN, $F(.)$	Input	26	
Number of weights with	Layer 1	16, 32	sigmoid
16 or 32 neurons : 993, 3009	Layer 2	16, 32	sigmoid
	Layer 3	16, 32	sigmoid
	Ouput	1	sigmoid
Auxiliary NN, $F_D(.)$	Input (Layer 2 of $F(.)$)	16, 32	
Number of weights with	Layer 1	16, 32	sigmoid
16 or 32 neurons : 289, 1089	Output	1	sigmoid

Table 7: Architecture of primary and secondary neural networks.

After the training of fair-NN 16 / fair-NN 32, the accuracies for predicting the female gender with the network $F_D(.)$, are 47.31% / 47.62% on the training set and 47.10% / 46.72% on the validation sample. This confirms that the outputs of the second layer of the primary network no longer contain information about the gender of the customer.

	Number of neurons	Epochs	Training	Validation
GLM			7310.02	1803.69
Fair-NN	16	1500	5632.74	1961.51
NN		1500	4721.66	2478.23
Fair-NN	32	1500	3860.61	3091.47
NN		1500	2237.87	4172.20

Table 8: Deviances on the training and validation data sets.

Table 8 reports the deviances of the fair and non-constrained neural networks. The comparison of these statistics for the validation sample reveals that the NN 16, fair-NN 32, and NN 32 overfit the data. In terms of deviance, the fair-NN 16 outperforms the GLM for the training set and offers a reasonable goodness of fit when evaluated on the validation set.

Table 9 presents the average predicted attrition rates by gender and model. Even though the gender is not a covariate, we observe that NN 16, NN 32 and GLM yield, on average, discriminatory attrition probabilities. In comparison, the average predictions of fair-NN 16 and fair-NN 32 are less discriminatory. Nevertheless, we need to investigate other measures to assess the degree of equity fairness of models.

Table 10 reports the Jenssen-Shannon divergences, Wasserstein, and Cramér distances of distributions of $\hat{y}$ per gender. To compute these statistics, we estimate the density functions of attrition rates using a kernel method applied to log-probabilities (bandwidth of 0.5), computed for all records (training and validation sets). Regardless of the metric, the fair neural networks drastically reduce discrimination between genders. The least discriminatory network is the NN 16. This is confirmed by Figure 2, which shows kernel estimate densities of attrition rates by gender, with the fair-NN 16, the NN 16, and the GLM. The densities by gender are much closer to each other with the fair-NN 16 than with the NN 16 or the GLM. The same plots are provided in Figure 9 of Appendix C for the NN 32 and fair-NN 32. These graphs clearly reveal the trend

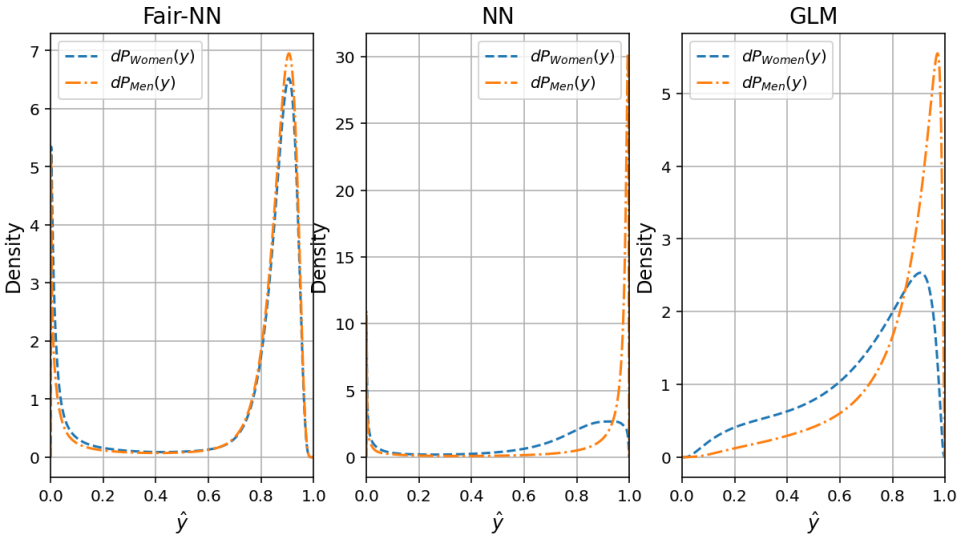
of the NN to overfit data. We also see that female and male densities of attrition probabilities are close when computed with the fair-NN 32.

	Number of neurons	Training		Validation	
		$\hat{y}\|W$	$\hat{y}\|M$	$\hat{y}\|W$	$\hat{y}\|M$
Fair-NN	16	0.7304	0.7736	0.7389	0.7765
NN		0.6791	0.8289	0.6809	0.8232
Fair-NN	32	0.7057	0.8014	0.7065	0.7891
NN		0.6709	0.8420	0.6746	0.8336
GLM		0.6911	0.8160	0.6872	0.8173

Table 9: Average predicted attrition probabilities by gender and model.

	Number	**Equity fairness**		
	of neurons	Fair-NN	NN	GLM
$\left\|\bar{\hat{y}}\|F - \bar{\hat{y}}\|M\right\|$	16	0.0421	0.1483	0.1260
Jenssen-Shannon		0.0119	0.8235	0.2620
Wasserstein, $p = 2$		0.1072	0.1863	0.1385
Cramér, $p = 2$		0.2223	2.9936	0.9463
$\left\|\bar{\hat{y}}\|F - \bar{\hat{y}}\|M\right\|$	32	0.0932	0.1705	0.1260
Jenssen-Shannon		0.0539	0.2632	0.2620
Wasserstein, $p = 2$		0.1468	0.1675	0.1385
Cramér, $p = 2$		0.5175	2.3374	0.9463
		Actuarial fairness		
RMSE (16)	16	0.02350	0.02811	0.03481
RMSE (32)	32	0.03602	0.04253	0.03480

Table 10: Measures of equity and actuarial fairness (training and test sets).



to 1

Figure 2: From the left to the right: plots of kernel estimate densities of predictions for women and men with the fair-NN 16, the NN 16 and the GLM.

To quantify the level of actuarial fairness, we sort the model predictions for the entire data set, by ascending order. The ordered sample is denoted by $\{\widehat{y}_{(1)}, ..., \widehat{y}_{(n)}\}$ where $\widehat{y}_{(i)} = \widehat{y}\left(\boldsymbol{x}_{(i)}\right)$. The

corresponding responses and exposures are $y_{(i)}$ and $\nu_{(i)}$. We define the n_p inter-quantile averages of observed and predicted responses, per unit of exposure, in a similar way to Equation (14):

$$\begin{cases} \widehat{p}_{(s)} &= \frac{1}{\sum_{j=u(s-1)}^{u(s)} \nu_{(j)}} \sum_{j=u(s-1)}^{u(s)} \nu_{(j)} \hat{y}_{BC,(j)} \\ p_{(s)} &= \frac{1}{\sum_{j=u(s-1)}^{u(s)} \nu_{(j)}} \sum_{j=u(s-1)}^{u(s)} \nu_{(j)} y_{(j)} \end{cases} , \, s = 1, \ldots, n_p \,, \tag{17}$$

where $u(s) = s \left\lfloor \frac{|\mathcal{D}|}{n_p} \right\rfloor$ and $\mathcal{D}$ is the union of training and validation sets. A model is actuarially fair if the $\left(\widehat{p}_{(s)}\right)_{s=1,\ldots,n_p-1}$ are close to their empirical equivalents, $\left(p_{(s)}\right)_{s=1,\ldots,n_p-1}$. Figure 3 shows the observed $p_{(s)}$ and predicted $\widehat{p}_{(s)}$ with respect to the probability s/n_q, for the fair-NN 16, NN 16, and GLM, with $n_p = 20$. We see that the inter-quantile averages of the fair-NN 16, match their empirical equivalents, even after removing the gender influence. Figure 9 in Appendix C presents the same plots for the NN 32 and fair-NN 32. We again check that the Fair-NN 32 yields locally balanced estimates of the attrition probabilities. The GLM and the NN also appear actuarially fair. To measure the degree of actuarial imbalance, we calculate the root mean square errors between inter-quantile averages:

$$RMSE \;=\; \sqrt{\frac{1}{l-1} \sum_{s=1}^{n_p-1} \left(\widehat{p}_{(s)} - p_{(s)}\right)^2} \,. \tag{18}$$

These RMSEs are reported in Table 10. The most actuarially fair model is the network with 16 neurons in its intermediate layers. As discussed in Section 3.2, the unconstrained networks and GLM are based on the canonical link function and therefore naturally fulfill the global balance condition 13. Nevertheless, we observe some local bias, even in the GLM. With our approach, we reduce this by imposing local balance conditions based on the proximity between predicted and observed n_q inter-quantile averages.

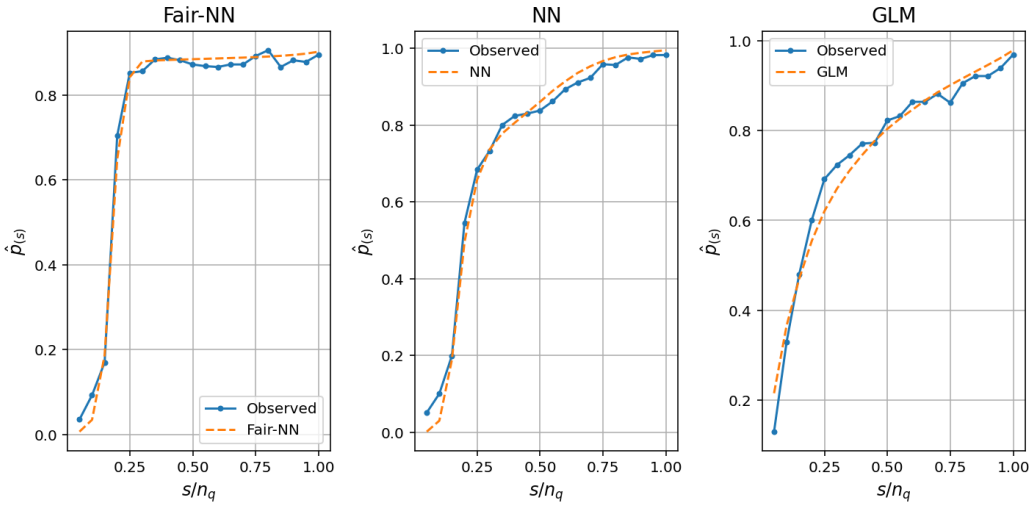


Figure 3: From the left to the right: QQ plots of empirical versus predicted inter-quantile averages of attrition probabilities, with the fair-NN 16, NN 16 and GLM.

To get a better insight into the impact of the local balance condition (15) on predictions, we fit a fair-NN 16 model without imposing a constraint of actuarial fairness ($\lambda_{BC} = 0$). Figure 4 compares the inter-quantile averages. The left graph is a QQ plot of empirical versus predicted inter-quantile averages ($n_q = 20$), with the fair-NN 16 with and without balance conditions. Most of these averages for both models are grouped in the interval 80%-90%. The right plot shows the inter-quantile averages for the fair-NN 16 with and without BC, ranked by probability.

The impact of the BC is clearly visible here. For the 15%-20%, 20%-25%, 25%-30%, 30%-35% quantile intervals, the averages are higher with the local balance condition than without it. For quantiles above 60%, the inter-quantile averages with the BC are slightly lower than wihout this constraint. The RMSEs (18) with and without BC are equal to 2.3497% and 2.4222%, respectively.

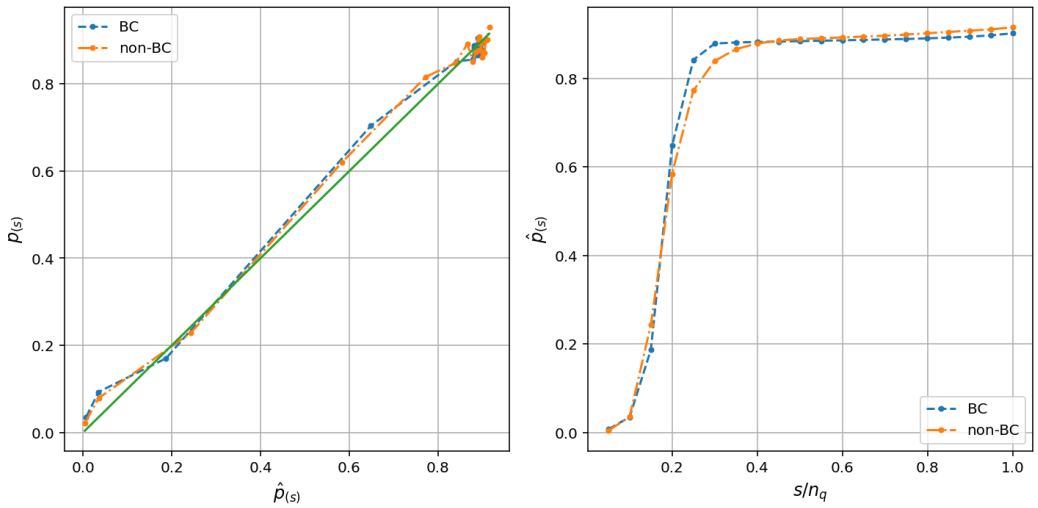


Figure 4: Left: QQ plots of empirical versus predicted inter-quantile averages, with the fair-NN 16 with and without BC. Right: Inter-quantile averages for the fair-NN 16 with and without BC.

4.2 MCC data set

In this section, we fit a model for predicting accident rates, to a data set from the Swedish insurance company *Wasa*. The data set is provided with the book by Ohlsson and Johansson (2010) and contains information about motorcycle insurance from 1994 to 1998. Each policy is described by quantitative and categorical features, detailed in Table 11.

Features	**Description**	**Type**
OwnersAge	Customer's Age in Years	num
VehicleAge	Age of the motorcycle	num
Gender	M=Male, K=Female	cat
Zone	7 categories (Urban, countryside,...)	cat
Class	7 vehicle class (1: low power 7:high power)	cat
BonusClass	Bonus malus, 7 levels	cat

Table 11: Rating factors of motorcycle insurances.

The database counts n =62436 policies after removing contracts with a null duration. The average accident rates by gender are 1.09% for men and 0.86% for women. We again consider gender as the discriminatory attribute. The portfolio in unbalanced in terms of gender: 9482 female versus 52954 male insureds. Gender is highly correlated with other features. For instance, we can predict it with an accuracy of 84.81% from other features using a logistic regression. As with the churn data set, removing gender from covariates will not prevent discrimination.

The training and validation sets contain, respectively, 80% and 20% of records. We estimate five models: two non-constrained neural networks (NN 16, NN 32), two fair neural networks (fair-NN 16 and fair-NN 32) with 16 or 32 neurons in their intermediate layers, and a GLM. The architecture of the networks is detailed in Table 12. The penalty factors are set to $\lambda_D = 1$ and $\lambda_{BC} = 1$. We run 1500 epochs for training the networks with respectively 16 and 32 neurons

in their intermediate layers. We consider $n_q = 20$ inter-quantile averages for defining the loss function (15) of the second auxiliary network. Notice that the activation function of the last layer is a sigmoid, which is not the inverse of the canonical link of the Poisson distribution. We have made this choice to move away from the GLM specification for the last layer.

	Layers	Shape	Activation
NN, Fair NN, $F(.)$	Input	20	
Number of weights with	Layer 1	16, 32	sigmoid
16 or 32 neurons : 897, 2817	Layer 2	16, 32	sigmoid
	Layer 3	16, 32	sigmoid
	Ouput	1	sigmoid
Auxiliary NN, $F_D(.)$	Input (Layer 2 of $F(.)$)	16, 32	
Number of weights with	Layer 1	16, 32	sigmoid
16 or 32 neurons : 289, 1089	Output	1	sigmoid

Table 12: Architecture of networks.

The accuracies for predicting the male gender are 84.80% / 84.60% on the training set and 84.76% / 84.61% on the validation sets, for networks with 16 / 32 neurons in intermediate layers. These percentages are close to the proportion of men in the portfolio because the adverse network classifies all insureds as male and is therefore unable to identify the gender from $\varphi(\boldsymbol{x})$.

	Number of neurons	Epochs	Training	Validation
GLM			4612.88	1214.94
Fair-NN	16	1000	4445.10	1172.69
NN		1000	4350.03	1186.75
Fair-NN	32	1000	4426.57	1214.93
NN		1000	3872.87	1285.70

Table 13: Deviances on the training and validation data sets.

Table 13 presents the deviances for the training and validation sets. The non-constrained network with 32 neurons in each intermediate layer achieves the lowest deviance on the training set and a goodness of fit comparable to the GLM on the validation set. Table 14 reports the average predicted accident rates by gender. The lowest spreads between them are obtained with the GLM on both the training and validation sets. The spreads between average male and female accident rates are not large for other models. It is therefore difficult to evaluate the level of discrimination induced by the models, based only these gaps.

	Number of neurons	Training		Validation	
		$\hat{y}\|W$	$\hat{y}\|M$	$\hat{y}\|W$	$\hat{y}\|M$
Fair NN	16	0.0136	0.0140	0.0134	0.0142
NN		0.0130	0.0134	0.0132	0.0133
Fair NN	32	0.0138	0.0142	0.0136	0.0141
NN		0.0131	0.0140	0.0135	0.0138
GLM		0.0132	0.0134	0.0131	0.0136

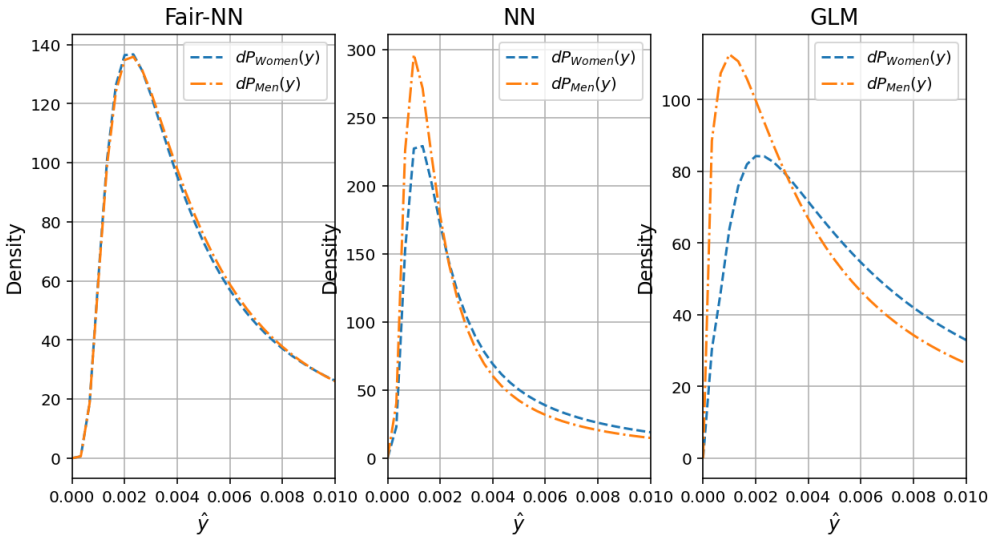
Table 14: Average predicted accident rates by gender and model.

Table 15 reports the divergences and distances of distributions by gender, computed using kernel estimates of density functions. Based on these measures, the fair neural networks clearly

reduce discrimination between predicted accident rates by gender compared to non-constrained networks or GLMs. The least discriminating network contains 32 neurons in its intermediate layers. This is confirmed by the left plot of Figure 5, which shows kernel estimates of density functions by gender for fair-NN 32 predictions. We see that the male and female densities of estimated accident rates are much closer with the fair-NN 32 than with other models. Figure 10, provided in Appendix C, presents the same plots for the models with 16 neurons in their intermediate layers. It confirms that the fair-NN 16 yields non-discriminatory estimates of accident rates.

	Number	**Equity fairness**		
	of neurons	Fair NN	standard NN	GLM
$\lvert\bar{\hat{y}}\lvert F - \bar{\hat{y}}\lvert M\rvert$	16	0.0005	0.0004	0.0002
Jenssen-Shannon		0.0008	0.0099	0.0376
Wasserstein, $p = 2$		0.0023	0.0027	0.0031
Cramér, $p = 2$		0.2015	1.1446	2.0992
$\lvert\hat{\hat{y}}\lvert F - \hat{\hat{y}}\lvert M\rvert$	32	0.0004	0.0007	0.0002
Jenssen-Shannon		0.0005	0.0191	0.0376
Wasserstein, $p = 2$		0.0017	0.0044	0.0031
Cramér, $p = 2$		0.1649	2.1499	2.0992
		Actuarial fairness		
RMSE (16)	16	0.00149	0.00125	0.00226
RMSE (32)	32	0.00148	0.00101	0.00226

Table 15: Measures of equity and actuarial fairness (training and test sets).



to l

Figure 5: From the left to the right graph : plot of kernel estimate densities of predictions for women and men with the fair-NN 32, the NN 32 and the GLM.

The RMSEs between the 20 inter-quantile averages of observed and predicted responses per unit of exposure, are reported in Table 15. Figure 6 shows the plots of these quantities. The graphs clearly emphasize that the fair-NN 32 yields locally balanced accident rates, even if their statistical distributions by gender are different from those obtained with the GLM. We observe that the NN 32 model is also actuarially fair, even though no particular constraint is imposed and the last layer is not a canonical GLM. We also notice that the GLM is less actuarially fair. Indeed, a GLM satisfies the global balance condition by construction, but there is no guarantee

that estimates are locally balanced. In Appendix C, Figure 11 reports the same information for the fair-NN 16 and NN 16. The plots confirm that the fair-NN16 estimates are actuarially fair.

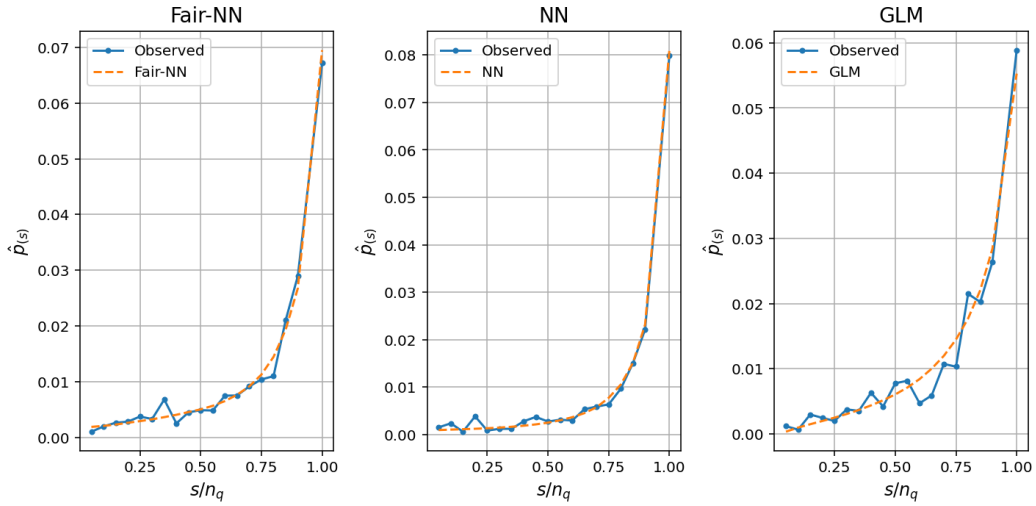


Figure 6: From the left to the right: QQ plots of empirical versus predicted inter-quantile averages of attrition probabilities, with the fair-NN 32, NN 32 and GLM.

To conclude, we fit a fair-NN 32 model without imposing actuarial fairness ($\lambda_{BC} = 0$). The left graph of Figure 7 is a QQ plot of empirical versus predicted inter-quantile averages ($n_q = 20$), with the fair-NN 16 with and without balance conditions. Most of the averages for both models, are located between 0% and 1%. The right plot shows the inter-quantile averages for the fair-NN 16 with and without BC, ranked by probability. The difference is not significant in this case. We observe small deviations between 85%-90%, 90%-95% and 95%-100% inter-quantile averages with and without BC. The RMSEs (18) with and without BC are equal to 0.1477% and 0.177%, respectively.

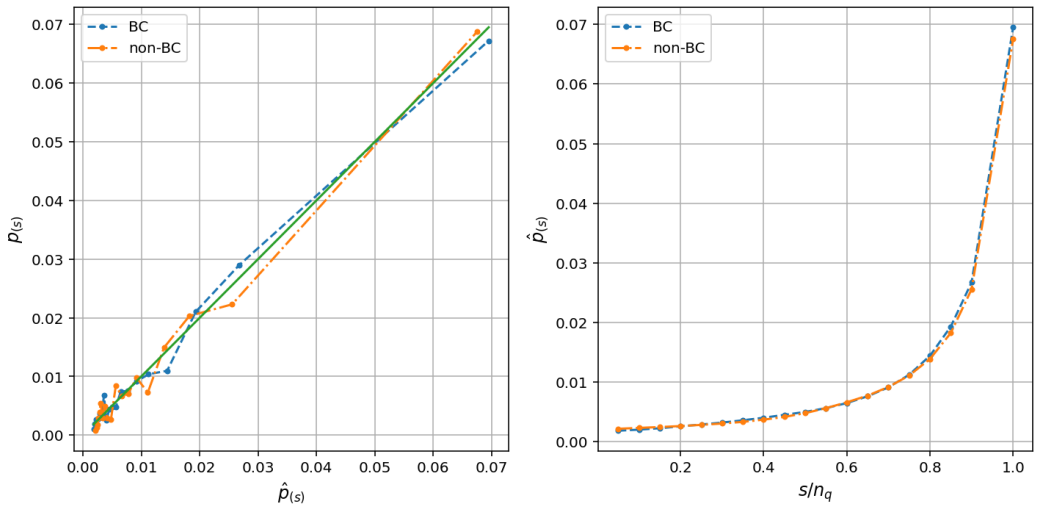


Figure 7: Left: QQ plots of empirical versus predicted inter-quantile averages, with the fair-NN 32 with and without BC. Right: Inter-quantile averages for the fair-NN 32 with and without BC.

5 Conclusions

The integration of actuarial and equity fairness constraints into neural networks represents is a serious challenge in the field of actuarial science. Traditional actuarial models, such as Generalized Linear Models (GLMs), have provided a solid foundation for insurance pricing and risk assessment. However, the flexibility and predictive power of neural networks offer a promising alternative, capable of capturing complex, non-linear relationships within data.

Despite their advantages, neural networks pose unique challenges, particularly concerning fairness and transparency. Ensuring actuarial fairness involves making unbiased and balanced predictions across individuals with similar risk profiles, while equity fairness guarantees that predictions do not discriminate based on protected characteristics such as race, gender, or age. Addressing these fairness concerns is crucial, given the regulatory and ethical implications in the insurance industry.

This article introduces a novel in-processing method for incorporating both actuarial and equity fairness constraints into neural networks. By leveraging an auxiliary neural network, we enforce equity fairness by ensuring predictions are independent of sensitive features. We achieve actuarial fairness by constraining the primary network with a penalty proportional to the spreads between predicted and observed inter-quantile averages. The proposed method effectively mitigates discrimination while preserving predictive accuracy.

Numerical illustrations using real-world datasets, such as a bank churn dataset and a motorcycle insurance dataset, demonstrate the effectiveness of the proposed method. These case studies show that the approach not only reduces gender discrimination but also maintains or improves predictive performance compared to traditional models and non-constrained neural networks. The results highlight the importance of integrating fairness constraints directly into the training process, rather than relying on pre-processing or post-processing methods.

In conclusion, the dual fairness approach presented in this article offers a robust framework for developing fair and reliable insurance pricing models. By addressing both actuarial and equity fairness, the method ensures that neural networks can be used responsibly and ethically in actuarial science. This advancement not only enhances the predictive capabilities of machine learning models but also aligns with the industry's commitment to fairness and transparency. As the field continues to evolve, the integration of fairness constraints into neural networks will play a crucial role in shaping the future of actuarial science and insurance pricing.

Appendix A, actuarial fairness of GLM

As underlined by Wüthrich (2020), a GLM with the canonical link fulfills the global balance condition (13). We briefly sketch the proof. If $g = \gamma'^{-1}(.) = h$ is the canonical link then $h\left(g^{-1}(u)\right) = u$ and the deviance of a data set of size n, simplifies as follows:

$$D = 2\sum_{i=1}^{n} \omega_i \left(y_i h\left(y_i\right) - \gamma\left(h\left(y_i\right)\right) - y_i\, \boldsymbol{\beta}^\top \boldsymbol{x}_i + \gamma\left(\boldsymbol{\beta}^\top \boldsymbol{x}_i\right)\right) .$$

The optimal regression coefficients cancel the partial derivatives of the deviance, which are proportional to

$$\frac{\partial D}{\partial \beta_j} \;\propto\; \sum_{i=1}^{n} \omega_i \left(\frac{\partial}{\partial \beta_j} \gamma\left(\boldsymbol{\beta}^\top \boldsymbol{x}_i\right) - y_i\, \beta_j x_{i,j} \right) \quad j = 1, ..., m\,.$$

Given that $\frac{\partial}{\partial \beta_j}\gamma\left(\boldsymbol{\beta}^\top \boldsymbol{x}_i\right) = g^{-1}\left(\boldsymbol{\beta}^\top \boldsymbol{x}_i\right)\beta_j x_{i,j}$,the optimal weights fulfill the equality:

$$\sum_{i=1}^{n} \omega_i g^{-1}\left(\boldsymbol{\beta}^{*\top}\boldsymbol{x}_i\right)\beta_j^* x_{i,j} \quad = \quad \sum_{i=1}^{n}\omega_i y_i \beta_j^* x_{i,j} \quad j=1,...,m\,.$$

Summing up over all j leads to the following relation

$$\sum_{i=1}^{n} \omega_i g^{-1}\left(\boldsymbol{\beta}^{*\top}\boldsymbol{x}_i\right)\left(\boldsymbol{\beta}^{*\top}\boldsymbol{x}_i\right) \quad = \quad \sum_{i=1}^{n}\omega_i y_i \left(\boldsymbol{\beta}^{*\top}\boldsymbol{x}_i\right)\,,$$

which implies that the balance condition (13) holds.

Appendix B, divergences and distances

The Kullback-Leibler divergence is a measure of entropy. For continuous distributions $\mathbb{P}_A(y)$ and $\mathbb{P}_B(y)$, this divergence is defined by :

$$D_{KL}\left(\mathbb{P}_A||\mathbb{P}_B\right) = \int_{y\in\mathbb{Y}} \log\left(\frac{d\mathbb{P}_A(y)}{d\mathbb{P}_B(y)}\right) d\mathbb{P}_A(y).$$

This measure is asymmetric : $D_{KL}\left(\mathbb{P}_A||\mathbb{P}_B\right) \neq D_{KL}\left(\mathbb{P}_B||\mathbb{P}_A\right)$. The issue of asymmetry is solved by the Jenssen-Shannon divergence:

$$D_{JS}\left(\mathbb{P}_A,\mathbb{P}_B\right) \quad = \quad \frac{1}{2}\left(D_{KL}\left(\mathbb{P}_A||\mathbb{P}_B\right) + D_{KL}\left(\mathbb{P}_B||\mathbb{P}_A\right)\right)\,. \tag{19}$$

The $p-$Wasserstein (or Earth mover's) distance between two distributions $\mathbb{P}_A(y)$ and $\mathbb{P}_B(y)$ is defined by

$$W_p(\mathbb{P}_A,\mathbb{P}_B) = \left(\inf_{\mathbb{P}\in\Pi(\mathbb{P}_A,\mathbb{P}_B)}\int_{\mathbb{R}\times\mathbb{R}} |x-y|^p\, d\mathbb{P}(x,y)\right)\,,$$

where $\Pi\left(\mathbb{P}_A,\mathbb{P}_B\right)$ is the set of bivariate distributions $\mathbb{P}(x,y)$ with marginals $\mathbb{P}_A$ and ,$\mathbb{P}_B$. $d\mathbb{P}(x,y)$ represents how mass is transported from x (under $\mathbb{P}_A$) to y (under $\mathbb{P}_B$). The infimum is taken over all possible couplings , ensuring we find the optimal transport plan that minimizes the total cost of transporting mass from x to y.The general quantile function of the cdf is defined as

$$\mathbb{P}_s^{-1}(y) \quad = \quad \inf\{u\,:\,\mathbb{P}_s(u)\geq y\}\,,\; s\in\{A,B\}.$$

For one-dimensional distributions $\mathbb{P}_A$ and $\mathbb{P}_B$, the p-Wasserstein distance can be written in terms of their quantile functions :

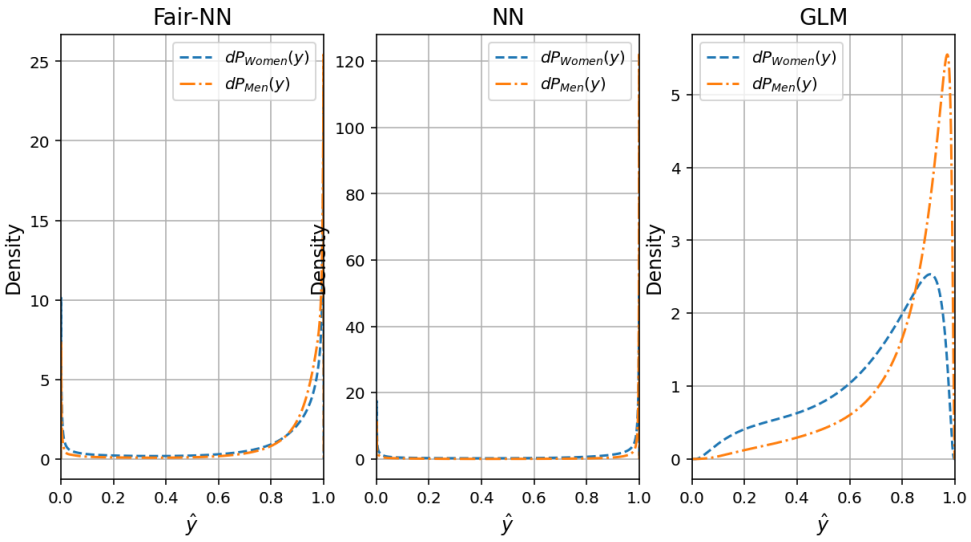
$$W_p\left(\mathbb{P}_A,\mathbb{P}_B\right) \quad = \quad \left(\int_0^1 \left|\mathbb{P}_A^{-1}(u)-\mathbb{P}_B^{-1}(u)\right|^p\, du\right)^{1/p}\,.$$

A last measure of similarity between two distributions is the Cramér distance. It is defined as the L_p distance between the cdfs of the two distributions:

$$l_p\left(\mathbb{P}_A,\mathbb{P}_B\right) = \left(\int_{-\infty}^{\infty} |\mathbb{P}_A(y)-\mathbb{P}_B(y)|^p\, dy\right)^{1/p}\,.$$

Appendix C Additional graphs

Churn data set



to l

Figure 8: From the left to the right: plots of kernel estimate densities of predictions for women and men with the fair-NN 32, the NN 32 and the GLM.

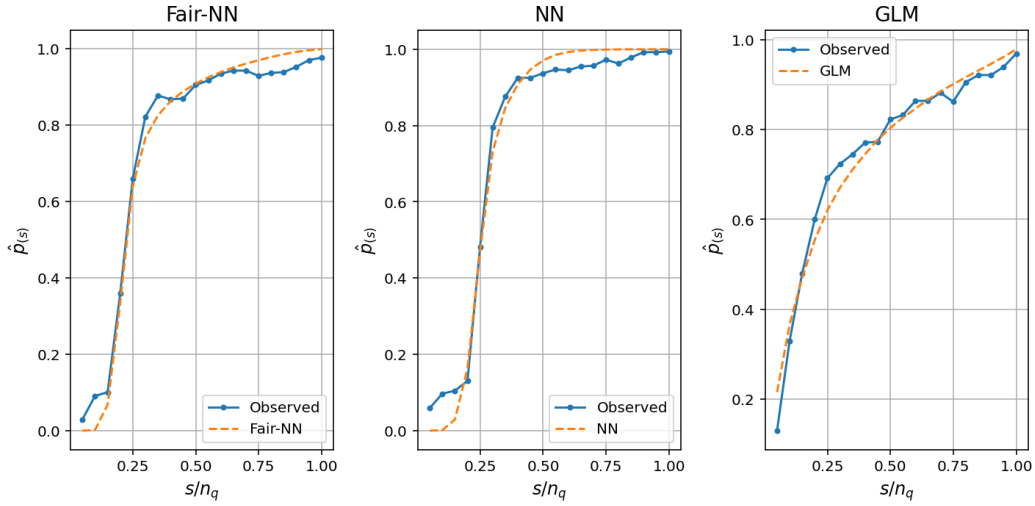
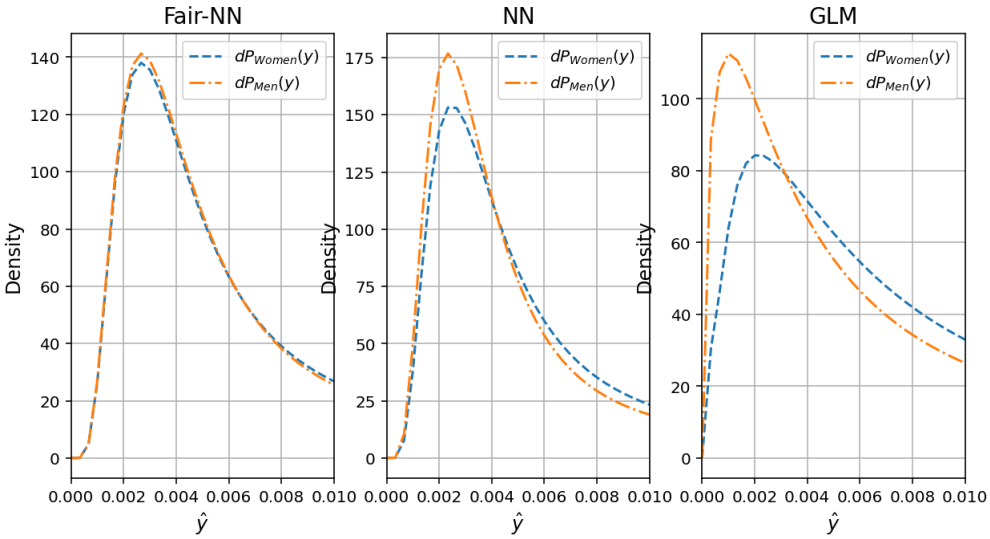


Figure 9: From the left to the right: QQ plots of empirical versus predicted inter-quantile averages of attrition probabilities, with the fair-NN 32, NN 32 and GLM.

MCC data set



to l

Figure 10: From the left to the right: plots of kernel estimate densities of predictions for women and men with the fair-NN 16, the NN 16 and the GLM.

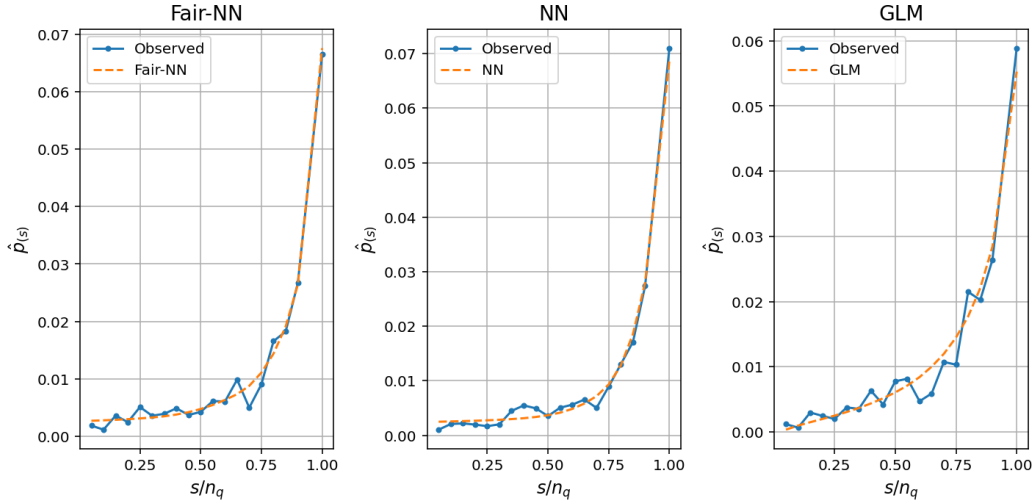


Figure 11: From the left to the right: QQ plots of empirical versus predicted inter-quantile averages of attrition probabilities, with the fair-NN 16, NN 16 and GLM.

Acknowledgement

I thank the FNRS (Fonds de la recherche scientifique) for the financial support through the EOS project Asterisk (research grant 40007517). I also thank Michel Denuit, for his relevant suggestions.

Disclosure statement

No potential conflict of interest was reported by authors.

References

[1] Aas K., Charpentier A. Huang F., Richman R. 2024. Insurance analytics: prediction, explainability, and fairness. Annals of Actuarial Sciences, 18 (3), 535-539.

[2] Bechavod, Y., Ligett, K., 2017. Penalizing unfairness in binary classification. arXiv preprint arXiv:1707.00044.

[3] Denuit M., Hainaut D., Trufin J. 2019 a. Effective statistical learning for actuaries I. GLMs and Extensions. Springer Actuarial Lecture Notes. Springer Nature Switzerland AG.

[4] Denuit M., Hainaut D., Trufin J. 2020. Effective statistical learning for actuaries II. Tree-Based Methods and Extensions. Springer Actuarial Lecture Notes. Springer Nature Switzerland AG.

[5] Denuit M., Hainaut D., Trufin J. 2019 b. Effective statistical learning for actuaries III. Neural Networks and Extensions. Springer Actuarial Lecture Notes. Springer Nature Switzerland AG.

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

[7] Grari V., Charpentier A., Detyniecki M., 2022. A fair pricing model via adversarial learning. arXiv preprint arXiv:2202.12008

[8] Kamishima, T., Akaho, S., Asoh, H., Sakuma J., 2012. Fairness-aware classifier with prejudice remover regularizer. Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23, 35–50.

[9] Charpentier A., Flachaire E., Gallic E., 2023. Optimal transport for counterfactual estimation: A method for causal inference. In Optimal transport statistics for economics and related topics (pp. 45–89). Springer.

[10] Frees E. W., Huang F., 2023. The discriminating (pricing) actuary. North American Actuarial Journal, 27(1), 2–24.

[11] Komiyama J., Shimao H., 2017. Two-stage algorithm for fairness-aware machine learning. arXiv preprint arXiv:1710.04924.

[12] Lindholm M., Richman R., Tsanakas A., Wuthrich M. V., 2024. What is fair? proxy discrimination vs. demographic disparities in insurance pricing. Proxy Discrimination vs. Demographic Disparities in Insurance Pricing (February 1, 2024).

[13] Mao Y., Deng Z., Yao H., Ye T., Kawaguchi K., Zou J., 2023. Last-Layer Fairness Fine-tuning is Simple and Effective for Neural Networks. arXiv preprint arXiv:2304.03935. Published at the ICML 2023 Workshop on Spurious Correlations, Invariance, and Stability.

[14] Padala M., Gujar S. 2020. FNNC: achieving fairness through neural networks. IJCAI'20: Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, Article No.: 315, Pages 2277 - 2283

[15] Ohlsson E., Johansson B., 2010. Non-Life Insurance Pricing with Generalized Linear Models. Springer-Verlag Berlin Heidelberg.

[16] Richman R., Wüthrich M.V. 2024. Smoothness and monotonicity constraints for neural networks using ICEnet. Annals of Actuarial Sciences, 18 (3), 712 - 739.

[17] Wüthrich M.V., Merz M., 2023. Statistical Foundations of Actuarial Learning and its Applications. Springer Nature Switzerland AG.

[18] Wüthrich, M.V., 2020. Bias regularization in neural network models for general insurance. European Actuarial Journal, 10, 179–202.

[19] Denuit M., Charpentier A., Trufin J. 2021. Autocalibration and Tweedie-dominance for insurance pricing with machine learning. Insurance: Mathematics and Economics. 101, 485-497.