Annals of Actuarial Science (2025), 1–19
doi:10.1017/S1748499525100092

ORIGINAL RESEARCH PAPER

DPTree and DPForest: tree-based methods fulfilling demographic parity

Pierre-Alexandre Simon[1], Michel Denuit[2] and Julien Trufin[1]

[1]Department of Mathematics, Université Libre de Bruxelles (ULB), Brussels, Belgium; and [2]Institute of Statistics, Biostatistics and Actuarial Science, UCLouvain, Louvain-la-Neuve, Belgium
Corresponding author: Julien Trufin; Email: julien.trufin@ulb.be

(Received 21 March 2024; revised 16 July 2025; accepted 17 July 2025)

Abstract

Tree-based methods are widely used in insurance pricing due to their simple and accurate splitting rules. However, there is no guarantee that the resulting premiums avoid indirect discrimination when features recorded in the database are correlated with the protected variable under consideration. This paper shows that splitting rules in regression trees and random forests can be adapted in order to avoid indirect discrimination related to a binary protected variable like gender. The new procedure is illustrated on motor third-party liability insurance claim data.

Keywords: Demographic parity; insurance risk classification; random forests; regression trees

1. Introduction

Discrimination is an important issue in the insurance industry because applicants and policyholders are subject to differentiation through risk classification at every stage of their customer relationship, from underwriting to possible cancelation. This paper is motivated by gender-based discrimination within the European Union (EU). In 2011, the European Court of Justice concluded that any gender-based insurance discrimination must be prohibited. From December 21, 2012, after the Test-Achats ruling, all insurers operating in the EU are required to offer unisex premiums and benefits. However, as part of the guidelines on the application of the 2011 ruling, the use of true risk factors that might be correlated with gender remains permitted. Precisely, if the candidate rating factor enters the model because of its predictive power in the presence of gender, then it can still be used by the insurers operating in the EU, in application of these guidelines. The latter only prohibits the use of proxies without impact on the premium in presence of the protected variable.

Let us justify this interpretation of Article 17 in the “Guidelines on the application of Council Directive 2004/113/EC to insurance, in the light of the judgment of the Court of Justice of the European Union in Case C-236/09 (Test-Achats),” together with its footnote (3). The first example given in these guidelines refers to power of the car in motor insurance, which is likely to be an example of true risk factor because actuarial analyses consistently report higher claim frequencies associated with more powerful cars in motor third-party liability insurance, corrected for the effect of gender. In such a case, the guidelines still allow insurers to use power as a rating factor even if it may capture part of the gender effect when this protected attribute is removed from the analysis (because men generally drive more powerful cars). The second example in these guidelines considers driver’s height and weight, which have never been considered as risk factors in motor insurance, as far as we know. Both height and weight are obviously correlated to gender,

https://doi.org/10.1017/S1748499525100092 Published online by Cambridge University Press

so the guidelines prohibit their use as rating factors because they would not qualify as risk factors once gender is included in the analysis. They only become statistically significant when gender is omitted, resulting in prohibited indirect discrimination.

Indirect discrimination resulting from proxy discrimination may thus still be present on the market in application of these guidelines. Proxy discrimination originates from the relationship between a protected variable and a set of one or more surrogate variables recorded in the database. This occurs when premiums vary according to a facially neutral risk factor correlated with the protected variable. For instance, amounts of motor insurance premiums often differ according to the power of the vehicle or the distance traveled. If these features are correlated to gender, e.g. because men tend to drive more powerful cars over longer distance, then the insurer's price list still indirectly discriminates according to gender.

The present paper demonstrates that demographic parity-adjusted prices are theoretically possible with tree-based techniques. The insurance premiums resulting from the proposed approach avoid indirect discrimination and thus strictly comply with the Gender EU Directive. This is achieved by adapting the splitting rules so that each risk class is gender-balanced. Therefore, it would theoretically be possible to enforce the prohibition of gender-based discrimination by adopting these techniques. In practice, some complications nevertheless remain, as explained in the final section of this paper.

The discrimination originating from insurance risk classification is an important topic in the literature. Let us briefly review some contributions related to the contents of the present paper. Abraham (1985) proposed five criteria to evaluate risk classification plans: separation, reliability, incentive value, homogeneity, and admissibility. The last one emphasizes possible indirect discrimination by the use of "characteristics that are admissible on their face but that have a disparate impact on the members of one race or sex." In a more general context, Avraham (2018) and Prince and Schwarcz (2020) proposed different properties that any rating factor should possess, irrespective of its discriminatory character. Frees and Huang (2021) discussed the predominant role of the actuary in the pricing process by choosing which information should be restricted or prohibited.

Discrimination has also been studied in relation to premium calculation algorithms. Lindholm *et al.* (2022) characterized direct and indirect discrimination and proposed discrimination-free premiums. Kusner *et al.* (2017), Kilbertus *et al.* (2017), Tschantz (2022), and Côté *et al.* (2024a) demonstrated that indirect discrimination issues can be formulated within the causal inference framework. Lindholm *et al.* (2022) and Araiza Iturria *et al.* (2024) illustrated the connection between discrimination-free price and the back-door adjustment formula of Pearl (2009) under Markovian Directed Acyclic Graph assumptions. Interestingly, Zhao and Hastie (2019) indicated that the definition of the Partial Dependence Plot (PDP-plot) introduced by Friedman (2001) is similar to the back-door adjustment of Pearl (2009). In this regard, the difference between the discrimination-free price and the unawareness price of Lindholm *et al.* (2022) is similar to the difference between the PDP-plot and the marginal plot (M-plot) of Cooks *et al.* (1997), used to identify a posteriori the relationship between the features and the predictions made by black-box models.

In the machine learning literature, several group-fairness criteria have been proposed in relation to discrimination. This paper considers independence, also referred to as demographic parity. Different approaches have been followed in order to reestablish model's fairness:

- Pre-processing approaches focus on removing discrimination in the input data by using appropriate transformations. See e.g. Kamiran and Calders (2012) and Calmon *et al.* (2017).
- In-processing approaches enforce a constraint with respect to a fairness measure in the learning process by introducing a penalization parameter in the objective function. See e.g. Kamishima *et al.* (2011), Dwork *et al.* (2012) and Grari *et al.* (2022).

https://doi.org/10.1017/S1748499525100092 Published online by Cambridge University Press

– Post-processing approaches consist of removing discrimination a posteriori from the model's predictions. See e.g. Hardt *et al.* (2016) and Petersen *et al.* (2021).

Recently, pre- and post-processing approaches have been increasingly developed using optimal transport theory, see Barrio *et al.* (2019), Silvia *et al.* (2020), Chzhen *et al.* (2020a), Charpentier *et al.* (2023), and Lindholm *et al.* (2024).

While most discussions focus on fair classification, Agarwal *et al.* (2019) proposed two definitions of fair regression fulfilling demographic parity or bounded group loss. The latter notion requires that the prediction error for each protected group remains below a defined level. By considering bounded predictions and responses lying in $[0, 1]$, they transformed the original regression problem into a classification one and obtained theoretical results on solutions. Chzhen *et al.* (2020b) decomposed the problem of finding an optimal predictor fulfilling demographic parity into two successive optimization problems. The first one is standard risk minimization, and the second one applies the demographic parity constraint by means of Lagrange multipliers on discretized bins of predictions. This estimation involved the conditional expectation of the non-discriminatory features given the discriminatory one on an unlabeled data set left aside. These methods, however, introduce a large number of additional parameters.

This paper adopts the in-processing fairness approach in the insurance pricing framework. Proxy discrimination arises from the fact that the gender proportion may vary between risk classes. Therefore, premiums bring some information about gender and thus implicitly discriminate on that basis. This drawback can be eliminated by constraining each risk class to be gender-balanced. This is the idea explored in this paper with tree-based methods.

In relation to this framework, Kamiran *et al.* (2010) and Raff *et al.* (2017) proposed integrating fairness into decision trees within a classification context. They evaluated discrimination during split creation by determining information gain concerning the protected variable, alongside the usual evaluation of accuracy through information gain related to the categorical response variable (class labels). Kamiran *et al.* (2010) replaced the standard information gain maximized for the best split selection with various alternatives, such as difference, ratio, or sum of information gain with respect to the class label and the protected class, respectively. They combined this approach with leaf relabeling as a second step to mitigate discrimination. In contrast, Raff *et al.* (2017) introduced a normalized Gini impurity measure and only considered the difference between both information gains, with the second acting directly as a fairness penalization. These methods control fairness with respect to demographic parity in a global and implicit manner by incorporating a penalty based on the information gain computed with respect to the protected attribute during split selection. Although these approaches highlight the fact that they do not add new parameters to tune during the learning process for considering fairness, this effectively amounts to considering a Lagrange multiplier with respect to the fairness constraint equal to one in the objective function. Indeed, although both information gains lie within [0,1], their typical magnitudes can differ substantially, potentially limiting the effectiveness and flexibility of these methods without careful tuning of the Lagrange multiplier.

The approach described in this paper focuses on fair regression considering strong demographic parity (in distribution) by enforcing fairness locally at each split through a simple and interpretable constraint on group balance, filling the gap left by classification-only methods and information-based approaches that do not guarantee distributional fairness. This not only simplifies implementation but also provides explicit fairness guarantees, making the method more transparent, auditable, and suitable for deployment in regulated or high-stakes decision-making contexts. By constraining the splitting procedure, it is shown that we can end up with all risk classes containing the same proportion of men and women policyholders. The knowledge of the premiums then brings no more information about gender and discrimination is entirely removed as long as the composition of the portfolio remains unchanged. We refer to this approach as demographic parity corrected regression tree (DPTree) to emphasize that trees are made such that they

https://doi.org/10.1017/S1748499525100092 Published online by Cambridge University Press

satisfy demographic parity with respect to a binary protected variable. An extension to ensemble models allows us to introduce demographic parity corrected random forest (DPForest) built from DPTrees.

The present paper does not innovate in terms of fairness concepts, as demographic parity is a classical notion. Its contribution is a practical way to enforce demographic parity with tree-based machine learning methods, in a situation where the actuary has decided that demographic parity was the right fairness notion for the situation under consideration.

The remainder of this paper is organized as follows. Section 2 formalizes discrimination issues in insurance premiums. Section 3 describes the newly proposed DPTree approach, as well as the extension to random forests DPForest. Section 4 illustrates our new approach using motor insurance data. The final Section 5 briefly summarizes our main findings and concludes the paper with a discussion.

2. Demographic parity

Consider a claim response Y and a set of non-discriminatory features $X_1, \ldots, X_p$ gathered in the random vector X as well as a binary protected variable D. In this paper, we are interested in policyholder's gender D, supposed to take values in $\{0, 1\}$. The random vector (X, D) gathers all features at the insurer's disposal.

The dependence structure inside the random vector (Y, X, D) is exploited for extracting information about the expected response Y contained in (X, D). Without worrying about discriminatory features, the pure premium would be the conditional expectation $\mu(X, D) = \mathrm{E}[Y|X, D]$. The latter is called the best-estimate price (BEP) by Lindholm *et al.* (2022). Removing the protected variable D from the analysis means that the resulting premium is the conditional expectation $\mu(X) = \mathrm{E}[Y|X]$. The latter is called the unawareness price and is further considered in Section 4.4.1.

Classical group-fairness criteria include independence, separation, and sufficiency. See e.g. Corbett *et al.* (2017), Barocas *et al.* (2023), Steinberg *et al.* (2020), Caton and Haas (2020), and Charpentier (2022) for a discussion. In the literature, group-fairness criteria are often stated either with the model's score or with the model's predictions. Working with regression trees, model's scores are equivalent to model's predictions, so that we only mention model's predictions throughout this text to avoid any confusion.

Let $\widehat{\mu}(X)$ denote the model's prediction for the expected claim response. The independence criterion focuses on the independence between the model's predictions $\widehat{\mu}(X)$ and the protected variable D. This is often referred to as demographic parity, as formally stated next.

Definition 2.1 (Demographic parity). *A predictor $\widehat{\mu}(X)$ satisfies demographic parity if $\widehat{\mu}(X)$ and D are mutually independent.*

Notice that Definition 2.1 does not require X and D to be independent, only that the candidate pure premiums $\widehat{\mu}(X)$ are independent from D. This means that the distribution of $\widehat{\mu}(X)$ must remain the same for male and female policyholders. Therefore, the knowledge of the premium $\widehat{\mu}(X)$ charged to a policyholder does not bring any information about his or her gender. Denoting as F_0 the distribution function of $\widehat{\mu}(X)$ given $D = 0$ and as F_1 the distribution function of $\widehat{\mu}(X)$ given $D = 1$, we thus see that

$$\text{Demographic parity} \Leftrightarrow F_0 = F_1.$$

Equality of the distribution functions can be formally tested using Kolmogorov–Smirnov procedure, for instance.

Let us mention that the relevance of demographic parity as group fairness notion in insurance has been questioned by Baumann and Loi (2023). As pointed out by these authors in the

https://doi.org/10.1017/S1748499525100092 Published online by Cambridge University Press

introduction to their paper, there is no consensus about the fairness criterion and different contexts may call for different criteria. Baumann and Loi (2023) then specify that their analysis assumes that insurers implement chance solidarity and that there is no need for risk solidarity. In other words, premium transfers from over-priced low risks to under-priced high risks are prohibited. This point of view thus conflicts with the situation considered in the present paper, where it is assumed that the use of a risk factor is prohibited, which necessarily results in risk solidarity, or premium transfers. The equal treatment despite potentially different risk enforced in the EU in relation to gender is referred to by Baumann and Loi (2023) as another approach.

While we agree with the analysis conducted by these authors and share their conclusion about the suitability of sufficiency, or well-calibration, as fairness condition in a context where "chance solidarity, and no other form of solidarity is meant to be achieved" (Baumann and Loi, 2023, page 45, bottom), this is not the context of the present study where risk solidarity automatically results from equal treatment despite different risk. The reason why Baumann and Loi (2023) rule out demographic parity, as stated in the second paragraph of Section 3.3.1 in their paper, does not apply here since risk solidarity is present.

Notice that the definition of demographic parity adopted in this paper differs from the one stated in equation (3) of Baumann and Loi (2023). We consider here the stronger independence condition corresponding to Definition 8.3 in Charpentier (2024). This does not modify the reasoning held in the preceding paragraphs, showing that demographic parity may be a desirable property under equal treatment despite different risk and that the latter requirement rules out well-calibration. As another example of proper use of demographic parity in an insurance context, we refer to Fröhlich and Williamson (2024), who consider the independence condition retained in the present paper in their Definition 4.1 with "having migrant background" as sensitive feature D. There is thus no contradiction between the interesting study conducted by Baumann and Loi (2023) and the one in the present paper, which just applies to different contexts.

3. Demographic parity corrected regression tree

3.1 Regression tree

Let χ be the feature space spanned by the non-discriminatory features. Regression trees recursively partition the entire feature space χ into a set of disjoint subsets $\{\chi_t\}_{t\in\mathcal{L}}$ through successive binary splits, where $\mathcal{L}$ denotes the set of terminal nodes.

The database $\mathcal{D}$ records observations of the form (y_i, x_i, d_i) for a large number of cases i. It is divided into a train set $\mathcal{D}^{\texttt{train}}$ and a validation set $\mathcal{D}^{\texttt{valid}}$, that is, $\mathcal{D} = \mathcal{D}^{\texttt{train}} \cup \mathcal{D}^{\texttt{valid}}$. The train set $\mathcal{D}^{\texttt{train}}$ is itself divided into a training set $\mathcal{D}^{\texttt{training}}$ and a testing set $\mathcal{D}^{\texttt{testing}}$, that is, $\mathcal{D}^{\texttt{train}} = \mathcal{D}^{\texttt{training}} \cup \mathcal{D}^{\texttt{testing}}$. Henceforth, $i = 1, \ldots, n$ refers to the cases included in $\mathcal{D}^{\texttt{training}}$.

Let χ_t denote the subset of the feature space χ corresponding to node t, meaning that an observation i belongs to node t if $x_i \in \chi_t$. Let s be a candidate split for node t. The candidate split s generates two child nodes, denoted as $t_L^{(s)}$ and $t_R^{(s)}$. More precisely, $t_L^{(s)}$ and $t_R^{(s)}$ are the left and right child nodes of t resulting from split s. Let $\chi_{t_L^{(s)}}$ and $\chi_{t_R^{(s)}}$ denote the subsets of the feature space χ corresponding to nodes $t_L^{(s)}$ and $t_R^{(s)}$, respectively. By construction, we have $\chi_t = \chi_{t_L^{(s)}} \cup \chi_{t_R^{(s)}}$. The optimal split s_t at node t is obtained by minimizing a given loss function L, estimated on $\mathcal{D}^{\texttt{training}}$, over the set of candidate splits $\mathcal{S}_t$. Formally,

$$s_t = \underset{s\in\mathcal{S}_t}{arg\,min} \left\{ \sum_{i:x_i\in\chi_{t_L^{(s)}}} L\left(y_i, \bar{y}_{t_L^{(s)}}\right) + \sum_{i:x_i\in\chi_{t_R^{(s)}}} L\left(y_i, \bar{y}_{t_R^{(s)}}\right) \right\}, \tag{3.1}$$

https://doi.org/10.1017/S1748499525100092 Published online by Cambridge University Press

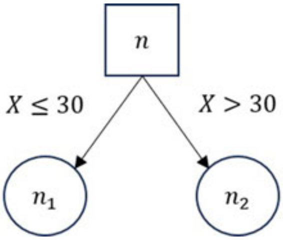

Figure 1. Regression tree "stump" with only one node and two leaves.

where $\overline{y}_{t_L^{(s)}}$ and $\overline{y}_{t_R^{(s)}}$ denote the average response, respectively, in the left and right candidate nodes. Often in insurance applications, the loss function L corresponds to the deviance associated with a distribution in the exponential dispersion family with the same support as Y.

Once the optimal split has been found at a node t, the feature space χ_t is divided into two disjoint subsets χ_{t_L} (left node) and χ_{t_R} (right node). We denote by $\overline{y}_{t_L}$ and $\overline{y}_{t_R}$ the associated average response, respectively, in the left and right nodes.

At the end of the partitioning process, the regression tree produces predictions of the form

$$\widehat{\mu}(x) = \sum_{t \in \mathcal{L}} \overline{y}_t \mathbb{1}[x \in \chi_t], \tag{3.2}$$

where $\mathbb{1}[\,\cdot\,]$ is the indicator function (equal to 1 if the event appearing within the brackets is realized and to 0 otherwise). In words, $\widehat{\mu}(x)$ is the average response $\overline{y}_t$ in the terminal leaf t corresponding to the risk profile x.

3.2 Demographic parity corrected regression tree

As the tree partitions χ into subsets χ_t, demographic parity holds if the distribution of D remains the same given that $X \in \chi_t$ for all t. This is the essence of the DPTree, where any split must result in subgroups where the proportion of cases with $D = 0$ remains equal to the one observed in the root node. The next example justifies this approach for the simple "stump" tree, which is known to be the basic building block of all trees.

Example 3.1. Consider the simple one-split regression tree depicted in Figure 1, often referred to as a stump. The unique split corresponds to the one which leads to the maximum loss function reduction on $\mathcal{D}^{\texttt{training}}$. To fix the ideas, we suppose in this example that this split is based on the numerical feature X and a threshold value of 30, say. The observations in the root node are then divided according to the decision rule $X \leq 30$ versus $X > 30$.

Let n_1 be the number of observations in the left node and n_2 in the right node, as shown in Figure 1. Also, denote as $n_{t,0}$ and $n_{t,1}$, $t \in \{0, 1, 2\}$, the number of observations in the root node ($t = 0$), in left node ($t = 1$) and in right node ($t = 2$) with protected variable $D = 0$ and $D = 1$, respectively. In the left leaf, the estimate of the unawareness price is given by

$$\widehat{\mu}(X \leq 30) = \overline{y}_{1,0}\frac{n_{1,0}}{n_1} + \overline{y}_{1,1}\frac{n_{1,1}}{n_1},$$

where $\overline{y}_{1,0} = \frac{\sum_{i=1}^{n} y_i \mathbb{1}[x_i \leq 30, d_i = 0]}{n}$ and $\overline{y}_{1,1} = \frac{\sum_{i=1}^{n} y_i \mathbb{1}[x_i \leq 30, d_i = 1]}{n}$. The corresponding estimate of the discrimination-free price maintaining the same proportion as in the root node writes

$$\mu^*(X \leq 30) = \overline{y}_{1,0}\frac{n_{0,0}}{n} + \overline{y}_{1,1}\frac{n_{0,1}}{n}.$$

The equality of the estimates of the unawareness price and discrimination-free price holds if

$$\frac{n_{1,0}}{n_1} = \frac{n_{0,0}}{n}. \tag{3.3}$$

https://doi.org/10.1017/S1748499525100092 Published online by Cambridge University Press

If we follow the same reasoning for the right leaf, we conclude that the estimates of the unawareness price and discrimination-free price coincide if

$$\frac{n_{2,0}}{n_2} = \frac{n_{0,0}}{n}. \tag{3.4}$$

As $n_{2,0} = n_{0,0} - n_{1,0}$ and $n_2 = n - n_1$, conditions (3.3) and (3.4) are in fact equivalent. Notice that we have silently assumed that $n_1 > 0$ and $n_2 > 0$ (one leaf cannot be empty).

Example 3.1 shows that constraining the proportion of observations with $D = 0$ in the left node to be equal to the proportion of observations with $D = 0$ in the root node is enough to characterize a DPTree. In general, a DPTree can thus be obtained by preserving the proportion of cases with $D = 0$ in the whole portfolio along each left node of the regression tree. This leads to the following definition.

Definition 3.2 (DPTree). *Let $p_{t_0} = \frac{\sum_{i=1}^{n} \mathbb{1}[d_i=0]}{n}$ be the proportion of cases with $D = 0$ in the root node t_0. The DPTree is built by reducing the set of candidate splits to*

$$\mathcal{S}_t^{(0)} = \{s \in \mathcal{S}_t | p_{t_L^{(s)}} = p_{t_0}\}$$

with

$$p_{t_L^{(s)}} = \frac{n_{t_L^{(s)},0}}{n_{t_L^{(s)}}} = \frac{\sum_{i=1}^{n} \mathbb{1}[x_i \in \chi_{t_L^{(s)}}, d_i = 0]}{\sum_{i=1}^{n} \mathbb{1}[x_i \in \chi_{t_L^{(s)}}]},$$

where $n_{t_L^{(s)},0}$ (resp. $n_{t_R^{(s)},0}$) denotes the number of cases in the left (resp. right) candidate node $t_L^{(s)}$ (resp. $t_R^{(s)}$) with $D = 0$. The optimal split is then determined according to (3.1) by replacing the set $\mathcal{S}_t$ of candidate splits with $\mathcal{S}_t^{(0)}$.

By ensuring that in every node, the proportion of cases with $D = 0$ is the same as in the root node, the resulting price produced by the regression tree is free of direct and indirect discrimination and is unbiased. By construction, the resulting predictions fulfill demographic parity from Definition 2.1. Indeed, for $j = 0, 1$, we have

$$\begin{aligned} F_j(u) &= \mathbb{P}[\widehat{\mu}(X) \le u | D = j] \\ &= \sum_{t \in \mathcal{L}} \mathbb{P}[\widehat{\mu}(X) \le u, X \in \chi_t | D = j] \\ &= \sum_{t \in \mathcal{L}} \mathbb{P}[\widehat{\mu}(X) \le u | X \in \chi_t, D = j] \mathbb{P}[X \in \chi_t | D = j]. \end{aligned}$$

Moreover, $\{\widehat{\mu}(X) | X \in \chi_t, D = j\} = \overline{y}_t = \{\widehat{\mu}(X) | X \in \chi_t\}$ and

$$\mathbb{P}[X \in \chi_t | D = j] = \frac{\mathbb{P}[D = j | X \in \chi_t] \mathbb{P}[X \in \chi_t]}{\mathbb{P}[D = j]}.$$

Hence, since by construction, $\widehat{\mathbb{P}}[D = j | X \in \chi_t] = \widehat{\mathbb{P}}[D = j]$ on the training set, it comes

$$\begin{aligned} \widehat{F}_j(u) &= \sum_{t \in \mathcal{L}} \widehat{\mathbb{P}}[\widehat{\mu}(X) \le u | X \in \chi_t] \widehat{\mathbb{P}}[X \in \chi_t] \\ &= \widehat{\mathbb{P}}[\widehat{\mu}(X) \le u], \end{aligned}$$

so that $\widehat{F}_0(u) = \widehat{F}_1(u)$.

https://doi.org/10.1017/S1748499525100092 Published online by Cambridge University Press

Strictly imposing the DPTree constraint may weaken the predictive power of the regression tree. This is why we propose to consider a margin m on this constraint, allowing for moderate departures to improve the DPTree performances.

Definition 3.3 (DPTree(m)). *The DPTree(m) is obtained by allowing for a margin m such that the set of candidate splits becomes*

$$\mathcal{S}_t^{(m)} = \left\{ s \in \mathcal{S}_t \;\middle|\; |p_{t_0} - p_{t_L^{(s)}}| \leq m, \;\; |p_{t_0} - p_{t_R^{(s)}}| \leq m \right\}.$$

The optimal split $s_t^{(m)}$ is then determined according to (3.1) by replacing the set $\mathcal{S}_t$ of candidate splits with $\mathcal{S}_t^{(m)}$.

The margin m in Definition 3.3 is generally expressed in relative terms, as

$$m = \varepsilon p_{t_0} \quad \text{with } \varepsilon = 0, 1\%, 2\%, \ldots$$

Henceforth, m and ε will be used interchangeably. Clearly,

$$\mathcal{S}_t^{(0)} \subset \mathcal{S}_t^{(m)} \subset \mathcal{S}_t = \mathcal{S}_t^{(1)}$$

and DPTree=DPTree(0). Thus, $\mathcal{S}_t^{(0)}$ corresponds to the constrained set of candidate splits with no margin ($m = 0$), while $\mathcal{S}_t^{(1)}$ corresponds to unconstrained regression trees. Let us now come back to Example 3.1 to see how the constraint is relaxed when growing regression trees.

Example 3.4 (Example 3.1 Ctd). When the margin m is allowed, equality is weakened into the following inequality:

$$\left| \frac{n_{1,0}}{n_1} - \frac{n_{0,0}}{n} \right| \leq m \implies |n_{1,0}n - n_{0,0}n_1| \leq mn_1 n. \tag{3.5}$$

Similarly,

$$|n_{2,0}n - n_{0,0}n_2| \leq mn_2 n \Leftrightarrow |(n_{0,0} - n_{1,0})n - n_{0,0}(n - n_1)| \leq m(n - n_1)n,$$

resulting in

$$|n_{1,0}n - n_{0,0}n_1| \leq mn(n - n_1). \tag{3.6}$$

Example 3.4 shows that inequality (3.6) is different from inequality (3.5), so the margin constraint should be respected in each node of the tree and not only in left nodes. The DPTree(m) approach has been implemented in `Python` following the approach of Breiman *et al.* (1984). The following algorithm describes the optimal split research at a node t of the DPTree(m):

3.3 Demographic parity corrected ensemble methods

As mentioned in Denuit *et al.* (2020), regression trees often suffer from instability. For instance, the tree may sometimes exhibit a completely different structure in case of only slight modification of the training set $\mathcal{D}^{\texttt{training}}$. Ensemble methods aim to improve stability by averaging predictions obtained from trees resulting from perturbed training set. The perturbations of $\mathcal{D}^{\texttt{training}}$ are obtained from resampling and random selection of features used in the splitting process of each individual tree. Precisely, consider an ensemble of B different training sets denoted as $\mathcal{D}^{\texttt{training,1}}$, $\mathcal{D}^{\texttt{training,2}}, \ldots, \mathcal{D}^{\texttt{training,B}}$. The predictions obtained from the random forest are expressed as

$$\widehat{\mu}^{rf}_{\mathcal{D}^{\texttt{training}}}(x) = \frac{1}{B} \sum_{b=1}^{B} \widehat{\mu}_{\mathcal{D}^{\texttt{training,b}}}(x)$$

where each $\widehat{\mu}_{\mathcal{D}^{\texttt{training,b}}}(x)$ is of the form (3.2).

https://doi.org/10.1017/S1748499525100092 Published online by Cambridge University Press

Algorithm 1. DPTree(m) algorithm at node t

```
Input: $(x_i, y_i, d_i) \in \mathscr{D}^{\texttt{training}}$, margin $m$.
Output: Optimal split $s_t$ at node $t$.
$\mathscr{S}_t$ ;                                  // Set of candidates splits $\mathscr{S}_t$
$\chi_t$ ;                                          // Feature space at node $t$
$p_{t_0} = \frac{\sum_{i=1}^{n} \mathbb{1}[d_i=0]}{n}$ ;          // Proportion $D=0$ in root node
$min\_loss \leftarrow +\infty$;
for each $s$ in $\mathscr{S}_t$ do
    $\chi_t = \chi_{t_L^{(s)}} \cup \chi_{t_R^{(s)}}$ ; // Split observations between candidate left
      and right nodes
    $p_{t_L^{(s)}} = \frac{n_{t_L^{(s)},0}}{n_{t_L^{(s)}}} = \frac{\sum_{i=1}^{n} \mathbb{1}[\boldsymbol{x}_i \in \chi_{t_L^{(s)}}, d_i=0]}{\sum_{i=1}^{n} \mathbb{1}[\boldsymbol{x}_i \in \chi_{t_L^{(s)}}]}$ ; // Proportion D=0 in left node
    $p_{t_R^{(s)}} = \frac{n_{t_R^{(s)},0}}{n_{t_R^{(s)}}} = \frac{\sum_{i=1}^{n} \mathbb{1}[\boldsymbol{x}_i \in \chi_{t_R^{(s)}}, d_i=0]}{\sum_{i=1}^{n} \mathbb{1}[\boldsymbol{x}_i \in \chi_{t_R^{(s)}}]}$ ; // Proportion D=0 in right node
    $loss \leftarrow \sum_{i: \boldsymbol{x}_i \in \chi_{t_L^{(s)}}} L\left(y_i, \overline{y}_{t_L^{(s)}}\right) + \sum_{i: \boldsymbol{x}_i \in \chi_{t_R^{(s)}}} L\left(y_i, \overline{y}_{t_R^{(s)}}\right)$
    if $loss < min\_loss$ and $|p_{t_0} - p_{t_L^{(s)}}| \leq m$ and $|p_{t_0} - p_{t_R^{(s)}}| \leq m$ then
        $min\_loss \leftarrow loss$;
        $s_t \leftarrow s$;
return $s_t$;
```

The approach leading to DPTree(m) described in Section 3.2 can be applied to any tree in the forest. An ensemble of discrimination-free regression trees can thus be obtained, depending on the number of trees B included in the ensemble, from the sampling rate s used to generate the B random training sets, and the number of features $k < p$ randomly selected at each node.

The predictions of the discrimination-free random forest are given by

$$\widehat{\mu}^{rf(m)}_{\mathcal{D}^{\texttt{training}}}(x) = \frac{1}{B} \sum_{b=1}^{B} \widehat{\mu}^{(m)}_{\mathcal{D}^{\texttt{training,b}}}(x),$$

where $\widehat{\mu}^{(m)}_{\mathcal{D}^{\texttt{training,b}}}(x)$ denotes the DPTree(m) prediction for risk profile x on the training set $\mathcal{D}^{\texttt{training,b}}$. When the number of features k is equal to the number of explanatory variables p, the random forest is equivalent to the bagging model. This approach is referred to as DPForest(m).

4. Numerical application

4.1 Database

The approach proposed in this paper is illustrated on a Belgian motor third-party liability portfolio, which comprises 234, 178 individual policies observed over one year. For each policy i, $p = 14$ non-discriminatory features are available. The response variable Y_i represents the number of reported claims for policy i. Gender is the protected variable D, with $D = 0$ for men and $D = 1$ for women. The data set is partitioned into a training set $\mathcal{D}^{\texttt{training}}$ composed of 64% of the observations, a testing set $\mathcal{D}^{\texttt{testing}}$ with 16% of the observations, and a validation set $\mathcal{D}^{\texttt{valid}}$ with the 20% remaining observations. Table 1 provides the summary statistics of the numerical

Table 1. Summary statistics of the numerical features in $\mathcal{D}^{\texttt{training}}$: power of the vehicle, weight of the vehicle, price of the vehicle, age of the vehicle, number of drivers mentioned in the contract, age of the main driver, and seniority of the contract

	Men		Women	
Features	`mean`	`std`	`mean`	`std`
veh_power	80.07	30.55	69.16	24.55
veh_weight	1907.46	901.33	1752.03	854.60
veh_value	20775.94	9611.04	17166.63	7783.80
veh_age	10.05	4.99	9.94	4.82
driv_number	1.21	0.50	1.25	0.55
driv_m_age	49.94	16.05	48.47	15.92
cont_seniority	6.77	9.08	6.43	8.12

features on the training set. Vehicle characteristics (power, weight, and value) tend to be higher for men than for women, whereas other variables remain similar for each gender. Tables 2–3 display the conditional distributions of the categorical features given gender. Table 2 shows little differences in terms of vehicle use (with more professional users among men). Premium cars (like Mercedes, BMW, Audi, or Volvo) are more frequent for men. Considering Table 3, we see some differences between men and women with respect to gasoil and petrol as fuel oils, as well as for monthly payments. These summary statistics suggest some degree of association between available features and gender, producing the kind of indirect discrimination discussed in the preceding sections.

4.2 Learning model

We assume that the observations (Y_i, X_i, D_i) are independent and identically distributed, sampled from the population of drivers in the portfolio. The expected number of claims is learned by assuming that, given $X_i = x_i$ and $D_i = d_i$, Y_i is Poisson distributed with mean $\mu(x_i)$, so that we adopt the Poisson deviance as loss function.

The procedure can be described as follows. In accordance with Section 4.7 in Denuit *et al.* (2020), where a similar data set has been studied, we set B to 1000 (this choice is supported by Figure 4.3 in that book) and check afterwards on the testing set that this value is large enough by considering predictive accuracy of the forest in function of the number of its constituent trees. Trees are grown on $\mathcal{D}^{\texttt{training}}$, and $\mathcal{D}^{\texttt{testing}}$ is used to fine-tune hyperparameters of constrained random forests with different margins m. Precisely, the hyperparameters are the maximum interaction depth (henceforth denoted as `id`) and the number of features randomly selected at each node (denoted as k). We consider models $\mathcal{M}(m, \texttt{id}, k)$ or $\mathcal{M}(\varepsilon, \texttt{id}, k)$ with

$$\texttt{id} \in \{1, 2, 3, 4, 5\}$$
$$k \in \{1, 2, \ldots 10\}$$

for increasing margins. Precisely, we consider regular random forests corresponding to $m = 1$. Then, we start from $m = 0$ and gradually increase the margin as long as the null assumption $F_0 = F_1$ is not rejected when tested on $\mathcal{D}^{\texttt{valid}}$. To this end, we use the Kolmogorov–Smirnov distribution-free test for differences in two continuous distribution functions, implemented in the R function `qKolSmirnLSA`. All margins m for which the null is not rejected at confidence level of 95% are said to be tolerable in the remainder of the text.

For each tolerable relative margin ε, best models are listed in Table 4 together with the null model (same prediction for every policy). Increasing the margin leads to lower out-of-sample

Table 2. Categorical features with associated levels and proportions according to gender on $\mathcal{D}^{\texttt{training}}$: use of the vehicle, ADAS (Advanced Driver Assistance Systems) equipped vehicle, and make of the vehicle

Features	Modalities	Prop. man (%)	Prop. woman (%)
veh_use	personal	84.73 %	85.25 %
	personal & commute	12.9 %	12.91 %
	professional	2.37 %	1.86 %
veh_adas	no	99.21 %	99.48 %
	yes	0.79 %	0.52 %
veh_make	volkswagen	12.0 %	12.18 %
	opel	10.21 %	10.86 %
	citroen	8.83 %	9.46 %
	peugeot	8.14 %	9.43 %
	renault	8.14 %	9.29 %
	ford	7.08 %	7.89 %
	mercedes	7.75 %	4.87 %
	bmw	6.73 %	3.76 %
	toyota	4.54 %	6.21 %
	audi	4.82 %	3.16 %
	fiat	3.18 %	4.3 %
	volvo	2.37 %	1.64 %
	skoda	1.74 %	1.95 %
	seat	1.79 %	1.66 %
	dacia	1.64 %	1.47 %
	nissan	1.2 %	1.62 %
	kia	1.18 %	1.22 %
	hyundai	0.92 %	1.07 %
	mini	0.53 %	1.44 %
	mitsubishi	0.82 %	0.8 %
	alfa romeo	0.89 %	0.59 %
	mazda	0.72 %	0.87 %
	land rover	0.79 %	0.44 %
	chevrolet	0.57 %	0.78 %
	suzuki	0.5 %	0.79 %
	other	0.61 %	0.34 %
	honda	0.54 %	0.45 %
	porsche	0.47 %	0.18 %
	smart	0.29 %	0.42 %
	lancia	0.18 %	0.34 %
	jaguar	0.26 %	0.13 %
	chrysler	0.2 %	0.15 %
	saab	0.18 %	0.14 %
	lexus	0.12 %	0.06 %
	other_luxe	0.09 %	0.02 %
	iveco	0.01 %	0.0 %

Table 3. Categorical features with associated levels and proportions according to gender on $\mathcal{D}^{\texttt{training}}$: parking in a garage, fuel of the vehicle, mileage limit specified in the policy, and premium payment

Features	Modalities	Prop. man (%)	Prop. woman (%)
veh_garage	no	92.28 %	93.09 %
	yes	7.72 %	6.91 %
veh_fuel	gasoil	65.85 %	49.52 %
	petrol	33.58 %	50.06 %
	gas	0.33 %	0.26 %
	hybrid	0.21 %	0.15 %
	electricity	0.02 %	0.02 %
	other	0.01 %	0.0 %
veh_mileage_limit	no	84.14 %	81.57 %
	yes	15.86 %	18.43 %
cont_paysplit	annual	45.82 %	44.24 %
	monthly	30.07 %	36.59 %
	biannual	14.17 %	10.49 %
	quartely	9.92 %	8.66 %
	other	0.02 %	0.02 %

Table 4. Best random forests and associated OOS deviances (OOS dev) computed on $\mathcal{D}^{\texttt{valid}}$ for different relative margins ϵ together with the OOS deviance of the null model and the normalized deviances (norm. dev). The null hypothesis H_0:$F_0 = F_1$ supporting demographic parity is rejected when $J^* > 1.358$

Parameters					
ϵ/m	`id`	k	OOS dev	norm. dev	J^*
$\epsilon = 0.01$	4	8	0.355880	39.52%	1.739
$\epsilon = 0.02$	5	7	0.355717	32.49%	2.330
$\epsilon = 0.03$	5	6	0.355557	25.60%	2.561
$\epsilon = 0.04$	5	5	0.355393	18.58%	**1.100**
$\epsilon = 0.05$	5	5	0.355336	16.09%	**1.063**
$\epsilon = 0.06$	5	5	0.355263	12.96%	1.675
$m = 1$	5	5	0.354962	0.00%	7.092
BE	5	6	0.354830	−5.64%	6.259
Null model			**0.357286**	100%	-

(OOS) deviance on $\mathcal{D}^{\texttt{valid}}$, as expected since the demographic parity constraint becomes less binding. Table 4 also displays the normalized deviance, defined as the difference between the OOS deviance and the OOS deviance of DPForest($m = 1$) divided by the difference between the OOS deviance of the null model and the OOS deviance of DPForest($m = 1$), so that the normalized deviance of the null model is 1 and the one of DPForest($m = 1$) is 0. We can see from Table 4 that the values of the test statistic J^* first lead to reject demographic parity up to $\varepsilon = 0.03$ before falling below the critical level for $\varepsilon = 0.04$ and 0.05. This may seem surprising since decreasing ε strengthens the demographic parity constraint, which should result in a smaller J^*. However, this is not the case here because the optimal choices of `id` and `k` also vary with ε. The null assumption supporting demographic parity is not rejected for $\varepsilon = 0.04$ and 0.05. The latter appears to be the largest tolerable relative margin, so we continue the analysis with a relative margin of 5%.

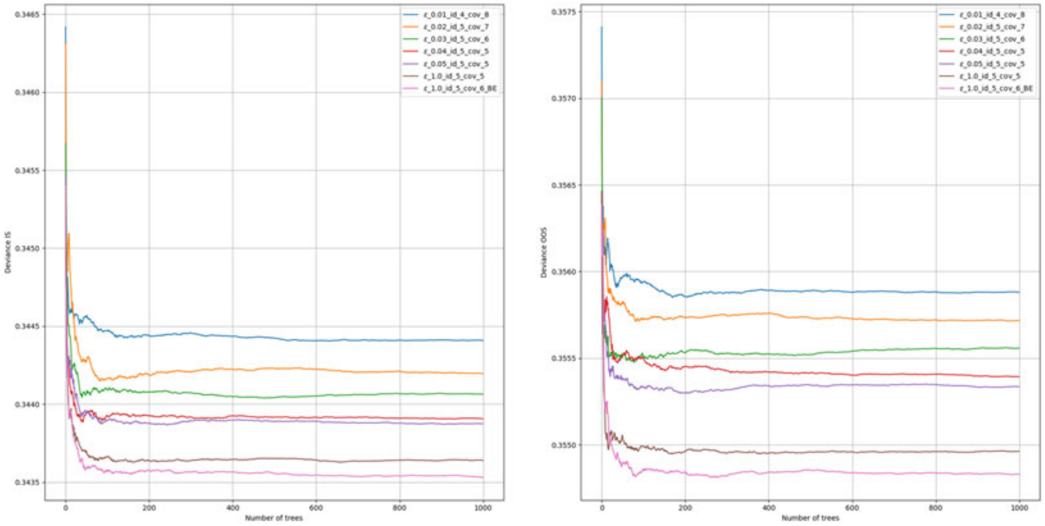

Figure 2. In-sample deviance (left panel) and out-of-sample deviance on $\mathcal{D}^{\texttt{testing}}$ (right panel) for each best model listed in Table 4.

The evolution of in-sample and OOS deviances according to the number of trees in the random forest is visible in Figure 2. All deviances stabilize when forests include about 400 trees and almost flatten when more trees are added, so that the choice $B = 1000$ appears to be large enough for the data set under consideration.

4.3 Impact of the demographic parity constraint

Figure 3 displays pairs of predicted claim frequencies for each individual for the largest tolerable relative margin $\varepsilon = 5\%$, along with those for the unconstrained model corresponding to $m = 1$. We see there that the majority of pairs are located near the 45-degree line of equality. The constrained random forest creates two groups: the first one, with a majority of women, lies above the main diagonal and is therefore overpriced compared to the unconstrained random forest, while the second group, with a majority of men, lies below the main diagonal and is therefore underpriced.

Feature importance is displayed in Figure 4 for both constrained and unconstrained random forests. The results are relatively similar between both models.

4.4 Comparison of DPForests with alternative methods

4.4.1 Discrimination-free insurance premiums

Often, insurance companies just remove the protected variable D considered to be discriminatory from the pricing process. This means that the resulting premium is the conditional expectation $\mu(X) = \mathrm{E}[Y|X]$. The latter is called the unawareness price. In our setting, the unawareness price writes

$$\mu(X) = \mathrm{E}[Y|X] = \mathrm{E}[Y|X, D=0]\mathbb{P}[D=0|X] + \mathrm{E}[Y|X, D=1]\mathbb{P}[D=1|X].$$

Thus, $\mu(X)$ is a weighted average of the BEPs $\mathrm{E}[Y|X, D=0]$ and $\mathrm{E}[Y|X, D=1]$, but the weights entering the calculation correlate with X, and thus possibly with the protected variable D.

The simple strategy leading to unawareness prices does not prevent indirect discrimination if there exists a dependence between D and X, that is, if the distribution of X given $D = 0$ differs from the distribution of X given $D = 1$. This generally induces dependence between $\mu(X)$ and D

https://doi.org/10.1017/S1748499525100092 Published online by Cambridge University Press

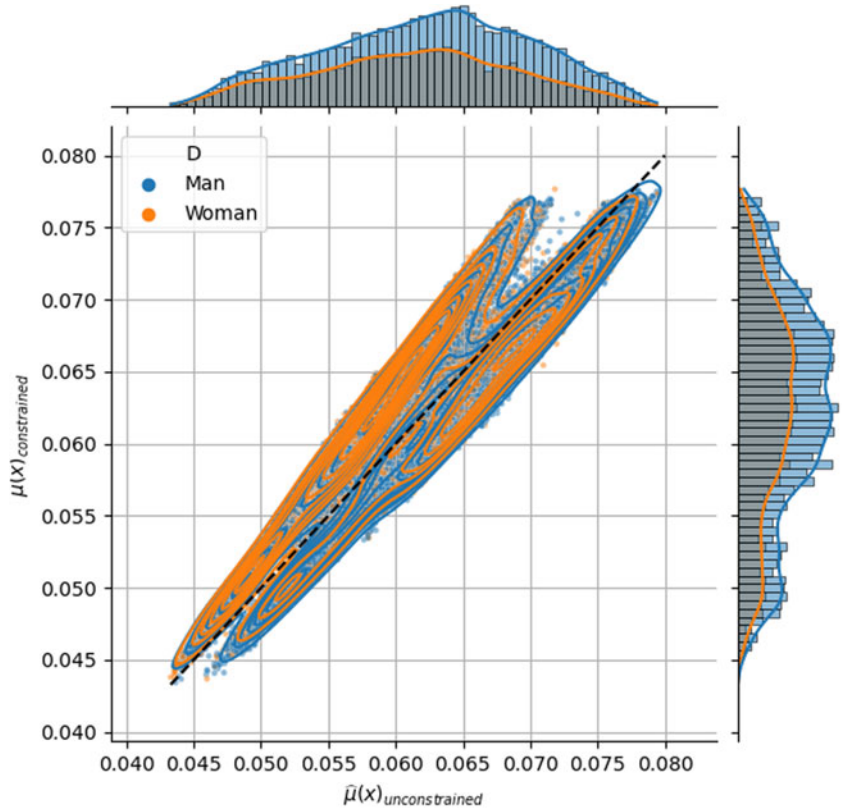

Figure 3. Pairs of predictions for unconstrained random forest ($m = 1$) and constrained random forest with the largest tolerable relative margin $\varepsilon = 5\%$ on $\mathcal{D}^{\texttt{valid}}$.

so that the protected variable implicitly enters the calculation of the unawareness price, resulting in indirect discrimination.

In order to remove indirect discrimination, Lindholm *et al.* (2022) suggested to derive the discrimination-free price from a weighted average using arbitrary weights w_0 and $w_1 = 1 - w_0$ both in the interval $[0, 1]$. The discrimination-free price is thus given by

$$\mu^*(X) = \mathrm{E}[Y|X, D = 0]w_0 + \mathrm{E}[Y|X, D = 1]w_1.$$

Lindholm *et al.* (2022) proved that the discrimination-free price is not unbiased in the sense that $\mathrm{E}[\mu^*(X)] \neq \mathrm{E}[Y]$ in general. Equality only holds in the special case where D and X are independent and the weights reflect the gender balance inside the portfolio, that is, $w_0 = \mathbb{P}[D = 0]$.

4.4.2 Wasserstein barycenters

Charpentier *et al.* (2023) suggested to restore the equality in distribution of $\widehat{\mu}(X)$ given $D = 0$ and $D = 1$ by using Wasserstein barycenters. Precisely, these authors define a discrimination-free predictor as

$$\widehat{\mu}_w(x, D = 0) = \widehat{\mathbb{P}}[D = 0]\widehat{\mu}(x, D = 0) + \widehat{\mathbb{P}}[D = 1]\widehat{F}_1^{-1} \circ \widehat{F}_0(\widehat{\mu}(x, D = 0))$$
$$\widehat{\mu}_w(x, D = 1) = \widehat{\mathbb{P}}[D = 0]\widehat{F}_0^{-1} \circ \widehat{F}_1(\widehat{\mu}(x, D = 1)) + \widehat{\mathbb{P}}[D = 1]\widehat{\mu}(x, D = 1).$$

It corresponds to averaging the model's predictions using the same quantile for both conditional distributions of the model given the value of the protected variable.

https://doi.org/10.1017/S1748499525100092 Published online by Cambridge University Press

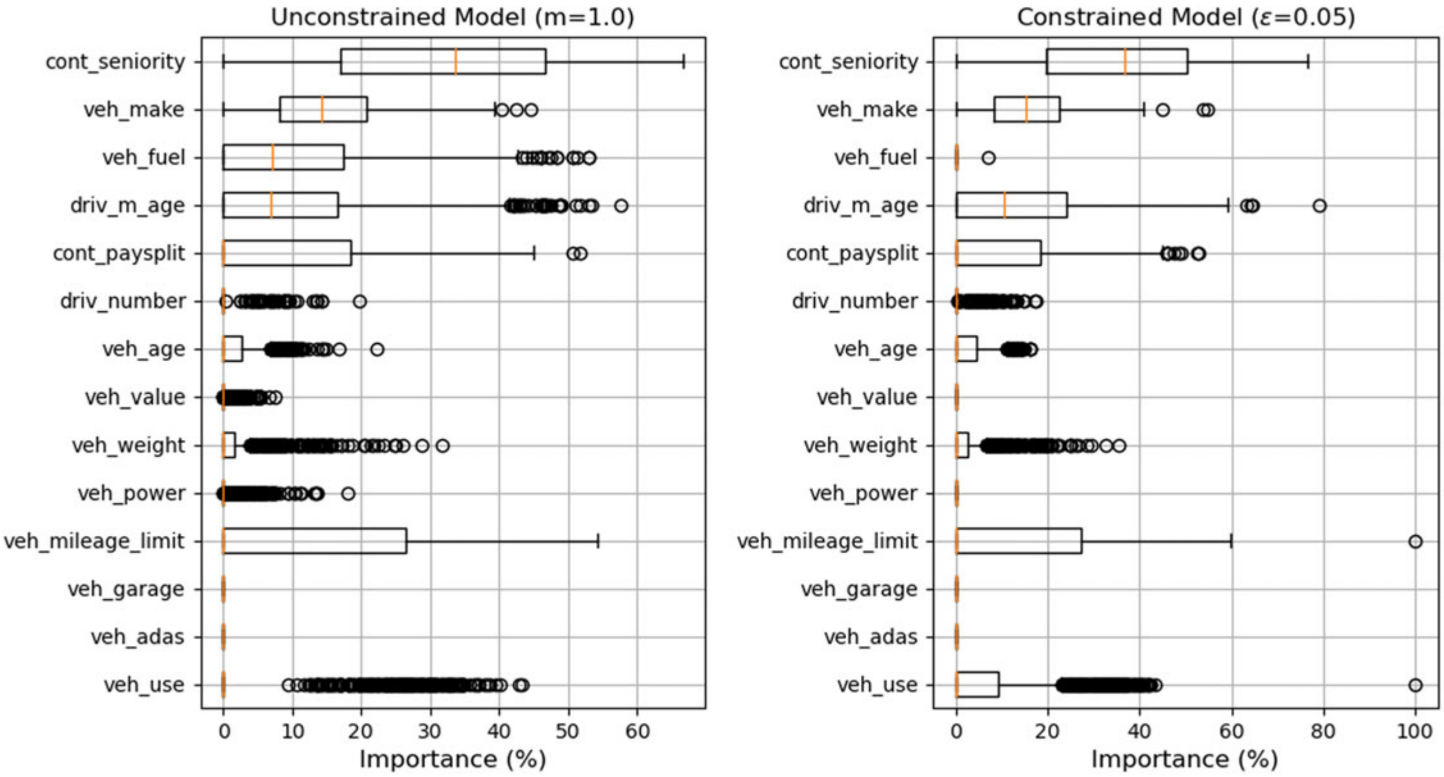

Figure 4. Feature importance for the unconstrained random forest (corresponding to $m = 1$) in the bottom panel and for the constrained random forest with the largest tolerable relative margin $\varepsilon = 5\%$ in the top panel.

https://doi.org/10.1017/S1748499525100092 Published online by Cambridge University Press

Table 5. Best and constrained random forests and associated OOS deviances (OOS dev) on $\mathcal{D}^{\texttt{valid}}$ for the largest tolerable relative margins $\epsilon = 0.05$ together with the OOS deviance obtained with discrimination-free prices (DFP) and Wasserstein barycenters correction (WBC). The null hypothesis H_0:$F_0 = F_1$ supporting demographic parity is rejected when $J^* > 1.358$

Parameters					
ϵ	id	k	OOS dev	norm. dev	J^*
0.05	5	5	0.355336	16.09%	**1.063**
BE	5	6	0.354830	−5.64%	6.259
DFP			0.354830	−5.65%	6.434
WBC			0.355012	2.17%	2.210
Null model			**0.357286**	100%	-

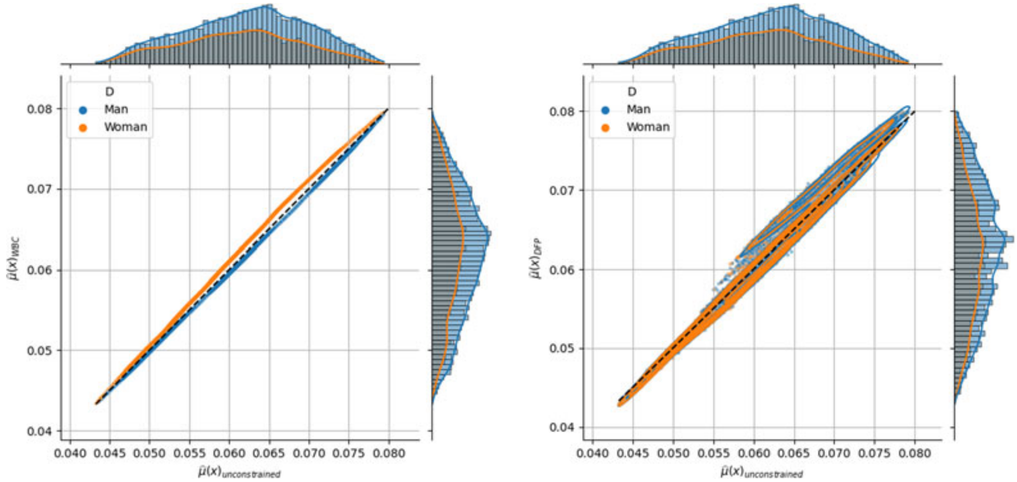

Figure 5. Pairs of predictions on $\mathcal{D}^{\texttt{valid}}$ for unconstrained random forest ($m = 1$) and those obtained with Wasserstein barycenters corrections in the left panel. Corresponding pairs with discrimination-free prices in the right panel.

4.4.3 Comparison

Table 5 compares the performances of the different models on the validation set, while Figure 5 displays the corresponding predictions. Remember that the normalized deviance percentages are expressed with respect to the unconstrained random forest model ($m = 1$). In terms of OOS deviance on $\mathcal{D}^{\texttt{valid}}$, we see in Table 5 that discrimination-free prices achieve the same OOS deviance as best estimates, despite their respective predictions differing (as will be seen below). Wasserstein barycenters correction and DPForest (0.05) exhibit higher OOS deviances as well as markedly different predictions. The 16.09% increase in OOS deviance appears to be the price to pay to achieve fairness in terms of demographic parity using constrained random forests. In terms of OOS deviance, DPForest is thus inferior to its competitors considered in this section. We nevertheless show next that the way predictions are modified may be preferable and that DPForest is the only approach fulfilling demographic parity with the data set under consideration.

The pairs of predictions between the unconstrained random forest and the Wasserstein barycenters correction are displayed in Figure 5. We observe that Wasserstein barycenters corrections overcharge (resp. undercharge) women (resp. men). Wasserstein barycenters corrections push predictions for men below the main diagonal and those for women above it. This is similar to DPForest(0.05) shown in Figure 3, except that this effect is more nuanced: the majority of women move in this way, but not all, and similarly for men.

https://doi.org/10.1017/S1748499525100092 Published online by Cambridge University Press

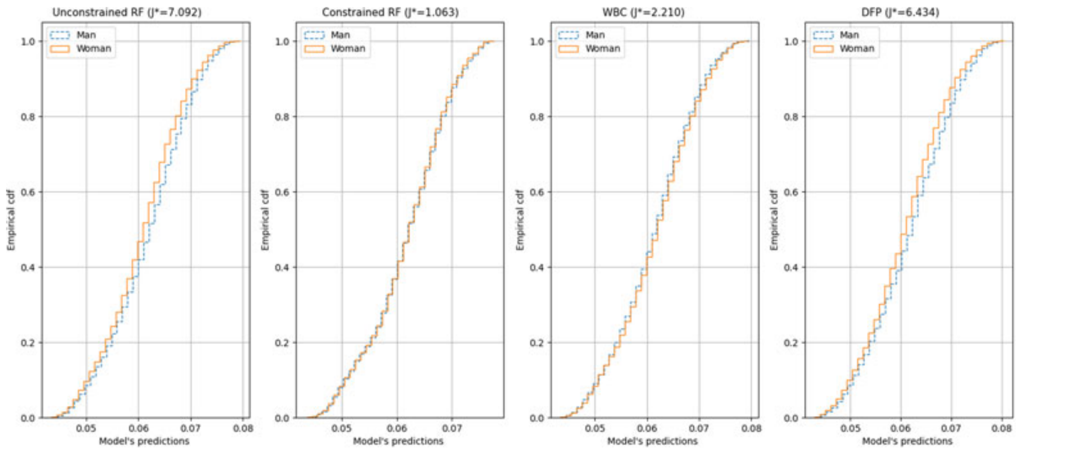

Figure 6. Empirical distribution functions $\widehat{F}_0$ and $\widehat{F}_1$ of $\widehat{\mu}(X)$ given $D=0$ and $D=1$, respectively, computed on $\mathcal{D}^{\texttt{valid}}$, for the unconstrained random forest ($m=1$, left), the constrained random forest ($\varepsilon=0.05$, middle left), the Wasserstein barycenters corrections (middle right), and the discrimination-free price (right).

Considering discrimination-free prices, we see that the predictions for women now tend to be located below the main diagonal.

The constrained random forest appears to be the only one respecting the demographic parity constraint on the data set under consideration. This is reflected in the empirical cumulative distribution functions of Figure 6. The reported Kolmogorov–Smirnov statistic J^* decreases from 7.092 to 6.434 when moving from the unconstrained DPForest($m=1$) to discrimination-free prices and to 2.210 when moving to the Wasserstein barycenters corrections. The minimum 1.063 is obtained with the constrained random forest DPForest(0.05). The latter is the only method for which demographic parity is not rejected.

5. Discussion

The DPTree approach proposed in this paper has been implemented in `Python`. Note that the libraries `Rpart` and `Scikit-learn,` often used by actuaries to build unconstrained regression trees, make use of native `C` for speeding up the best split research. The extension to DPForest, including DPTrees, has been implemented in `Python` as well. We aim to make use of native `C` for DPTree and DPForest in order to speed up the current computation time. This is currently under development. To the best of our knowledge, this is the first attempt to correct discrimination within the regression trees framework. In the proposed examples, Poisson deviance is the objective function to be minimized, as is typically the case when modeling expected claim frequency in actuarial sciences. Extensions to other deviance functions are straightforward. The main advantage of the methodology developed in this paper is that the margin parameter allows the actuary to define the desired degree of fairness.

The approach proposed in this paper has been compared with the alternative methods developed in Lindholm *et al.* (2022) and Charpentier *et al.* (2023). In all cases, reestablishing fairness has a cost for some risk profiles, especially when considering a small margin constraint in DPForest, but also has an impact on the model performance. Insurance companies must thus determine the optimal trade-off between accuracy and fairness.

Adverse selection certainly remains an issue. See e.g Thomas (2012) and Huang and Shimao (2025). Any change in the composition of the portfolio may induce indirect discrimination because gender balance does no more hold within risk classes, producing correlation between $\widehat{\mu}(X)$ and D. Since premiums resulting from DPTree and DPForest will be attractive for some risk profiles and not for others, we may expect that the introduction of the demographic parity corrected premiums by a single insurer leads to a change in the composition of its portfolio. See also Côté *et al.* (2024b) for a discussion of portfolio compared to market. The same phenomenon is likely to occur if insurers resort to the DPForest approach but learned on the basis of different sets of features. Hence, it seems to be extremely difficult to reach truly discrimination-free insurance pricing.

Among the strategies listed in Frees and Huang (2021), allowing insurers to use only a list of approved variables may be a good compromise, ensuring transparency and a reasonable degree of fairness. As documented in that paper, this is the strategy taken in the US individual health insurance market under the Affordable Care Act. Specifically, insurers may vary premium rates based on only four factors: (1) whether a plan covers an individual or family, (2) geographic area, (3) age, and (4) smoking status. If the regulator determines a limited number of rating factors and insurers use techniques as those derived in this paper or in Charpentier *et al.* (2023), then we may expect to protect consumers by ensuring high transparency and charge discrimination-free premiums.

Acknowledgements. The authors would like to express their gratitude to the anonymous referees and handling editor, whose comments have been extremely useful in revising a previous version of the present work.

Funding statement. This work was supported by P&V Assurances through the research chair "Telematics and Prevention." The authors thank their colleagues from P&V Assurances, especially Nikias De Feyter and Carolina Aldana, for interesting discussions that greatly contributed to the success of this research project.

Data availability statement. The data are not publicly available due to privacy and commercial restrictions. The code will be available in the near future once it makes use of native C.

Competing interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

Abraham, K. S. (1985). Efficiency and fairness in insurance risk classification. *Virginia Law Review*, **71**(3), 403–451.

Agarwal, A., Dudik, M., & Wu., Z. S. (2019). Fair regression: quantitative definitions and reduction-based algorithms. arXiv: 1905.12843.

Araiza Iturria, C. A., Hardy, M., & Mariott, P. (2024). A discrimination-free premium under causal framework. *North American Actuarial Journal*, **28**(4), 801–821.

Avraham, R. (2018). *The Routledge Handbook of the Ethics of Discrimination*. Kasper Lippert-Rasmussen, 335–347.

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

Barrio, D. E., Gamboa, F., Grodaliza, P., & Loubes, J.-P. (2019). Obtaining fairness using optimal transport theory. *Machine Learning Research*, **97**, 2357–2365.

Baumann, J., & Loi, M. (2023). Fairness and risk: An ethical argument for a group fairness definition insurers can use. *Philosophy & Technology*, **36**, 45.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stones, C. J. (1984). *Classification and Regression Trees*. Chapman & Hall/CRC, 55–89.

Calmon, F. P., Wei, D., Ramamurthy, K. N., & Varshney, K. R. (2017). Optimized data pre-processing for discrimination prevention. arXiv: 1704.03354.

Caton, S., & Haas, C. (2020). Fairness in machine learning: a survey. arXiv: 2010.04053.

Charpentier, A. (2022). Quantifying fairness and discrimination in predictive models. arXiv: 2212.09868.

Charpentier, A. (2024). *Insurance, Biases, Discrimination and Fairness*. Springer.

Charpentier, A., Hu, F., & Ratz., P. (2023). Mitigating discrimination in insurance with Wasserstein barycenters. arXiv: 2306.12912.

Chzhen, E., Denis, C., Hebiri, M., Oneto, L., & Pontil, M. (2020a). Fair regression with Wasserstein barycenters. *Advances in Neural information Processing Systems*, **33**, 7321–7331.

Chzhen, E., Denis, C., Hebiri, M., Oneto, L., & Pontil, M. (2020b). Fair regression via plug-Ii estimator and recalibration. *Advances in Neural information Processing Systems*, **33**, 19137–19148.

Cook, R. D., & Weisberg, S. (1997). Graphics for assessing the adequacy of regression models. *Journal of the American Statistical Association*, **92**(438), 490–499.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. Z. (2017). Algorithmic decision making and the cost of fairness. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Côté, O., Côté, M. P., & Charpentier, A. (2024a). A fair price to pay: Exploiting causal graphs for fairness in insurance. Available at SSRN 4709243.

Côté, O., Côté, M. P., & Charpentier, A. (2024b). Selection bias in insurance: why portfolio-specific fairness fails to extend market-wide. Available at SSRN 5018749.

Denuit, M., Hainaut, D., & Trufin, J. (2020). *Effective Statistical Learning Methods for Actuaries II: Tree-based Methods and Extensions*. Springer Actuarial Lecture Notes.

Dwork, C., Hardt, M., Pitassi, T., Reinglod, O., & Zemel, R. (2012). Fairness through awareness. Theoretical Computer Science Conference, pp. 214–226.

Frees, E. W., & Huang, F. (2021). The discriminating (pricing) actuary. *North American Actuarial Journal*, **27**(1), 2–24.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, **29**, 1189–1232.

Fröhlich, C., & Williamson, R. C. (2024). Insights from insurance for fair machine learning. Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, pp. 407–421.

Grari, V., Charpentier, A., & Detyniecki, M. (2022). A fair pricing model via adversarial learning. arXiv: 2202.12008.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, **29**, 3315–3323.

Huang, F., & Shimao, H. (2025). *Welfare Implications of Fair and Accountable Insurance Pricing*. UNSW Business School Research Paper Forthcoming.

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, **33**(1), 1–33.

Kamiran, F., Calders, T., & Pechenizkiy, M. (2010). Discrimination aware decision tree learning. *IEEE International Conference on Data Mining*, **33**(1), 1–33.

Kamishima, T., Akaho, S., & Sakuma, J. (2011). Fairness-aware learning through regularization approach. IEEE International Conference on Data Mining, pp. 643–650.

Kilbertus, N., Rojas Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., & Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. *Advances in Neural Information Processing Systems*, **30**, 656–666.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017) Counterfactual fairness. *Advances in Neural Information Processing Systems*, **30**, 4066–4076.

Lindholm, M., Richman, R., Tsanakas, A., & Wüthrich, M. V. (2022). Discrimination-free insurance pricing. *ASTIN Bulletin*, **52**(1), 55–89.

Lindholm, M., Richman, R., Tsanakas, A., & Wüthrich, M. V. (2024). What is fair ? Proxy discrimination vs. demographic disparities in insurance pricing. *Scandinavian Actuarial Journal*, **2024**(9), 935–970.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. 2nd edition, Cambridge University Press.

Petersen, F., Mukherjee, D., Sun, Y., & Yurochkin, M. (2021). Post-processing for individual fairness. arXiv: 2110.13796.

Prince, A. E. R., & Schwarcz, D. (2020). Proxy discrimination in the age of artificial intelligence and big data. *Iowa Law Review*, **105**, 1257–1318.

Raff, E., Sylvester, J., & Mills, S. (2017). Fair forests: regularized tree induction to minimize model bias. arXiv: 1712.08197.

Silvia, C., Ray, J., Tom, S., Aldo, P., Heinrich, J., & John, A. (2020). A general approach to fairness with optimal transport. *Artificial Intelligence*, **34**(04), 3633–3640.

Steinberg, D., Reid, A., & O'Callaghan, S. (2020). Fairness measures for regression via probabilistic classification. arXiv: 2001.06089.

Thomas, R. G. (2012). Non-risk price discrimination in insurance: market outcomes and public policy. *The Geneva Papers on Risk and Insurance - Issues and Practice*, **37**, 27–46.

Tschantz, M. C. (2022). What is proxy discrimination ? *ACM Conference on Fairness, Accountability, and Transparency*, pp. 1993–2003.

Zhao, Q., & Hastie, T. (2019). Causal interpretations of black-box models. *Journal of Business & Economic Statistics*, **39**(1), 272–281.

Cite this article: Simon P-A, Denuit M and Trufin J (2025). DPTree and DPForest: tree-based methods fulfilling demographic parity, *Annals of Actuarial Science*, 1–19. https://doi.org/10.1017/S1748499525100092

https://doi.org/10.1017/S1748499525100092 Published online by Cambridge University Press