UCLouvain

FACULTE DES SCIENCES

Faculté des sciences

Non-life insurance pricing under ethical constraints (interpretability, non-discrimination and fairness)

Author: **Maxime DOUMONT**
Supervisors: **Florian PECHON, Michaël LECUIVRE**
Reader: **Michel DENUIT**
Academic year 2023–2024
Master [120] en sciences actuarielles

Abstract

When implementing a predictive model for non-life insurance pricing, it is legitimate for this model to respect several ethical constraints. Firstly, having an interpretable model will make it easier for the people working on it to understand its behaviour, and it will also be easier to explain to policyholders why they are paying a certain premium. Moreover, non-discrimination and fairness are also ethical constraints that receive significant attention. An illustration is the European Council directive stating that a woman and a man having the same risk profile must pay the same premium for their insurance product.

Regarding these three ethical constraints, this master thesis focuses on two research questions. Does an interpretable model such as EBM (explainable boosting machine) have prediction performances similar to RF (random forests) and GBM (gradient boosting machine) models ? Is it possible to have a model for non-life insurance pricing that satisfies simultaneously the three ethical constraints stated earlier: interpretability, non-discrimination and fairness ? To answer these questions, this master thesis first studies the EBM models. Then, different non-discrimination methods and fairness criteria are implemented. This will allow to see the implications or not between the non-discrimination methods and the fairness criteria.

Acknowledgments

First of all, I would like to sincerely thank my supervisor at Reacfin, Michaël Lecuivre, for his precious support and his availability throughout my master thesis. He gave me valuable advice that helped me to complete this work.

My special thanks also go to my supervisor, Florian Pechon, for the fruitful discussions we had throughout the academic year. These have always helped to answer all my questions and improve my work.

Finally, I would like to take this opportunity to say a special thank you to my family and my girlfriend. They have always supported and encouraged me throughout my university journey and this master thesis.

Contents

Chapter 1

Introduction

When implementing a predictive model for non-life insurance pricing, it is legitimate for this model to respect several ethical constraints. It is of course important to look at the accuracy of the model and ensure that it has good predictive performances but another important characteristic that a predictive model can possess is interpretability. In the context of insurance, having an interpretable model offers several advantages. Firstly, it helps actuaries and other professionals working with the model in understanding its behavior. Moreover, an interpretable model makes it easier to explain the premium to policyholders, thereby increasing their confidence in the insurer.

However, predictive performance and interpretability are two concepts that often conflict in predictive models. Indeed, generally speaking, there are two types of predictive models. The first type are models such as generalized linear model (GLM) or generalized additive model (GAM). These models have an explicit score and are therefore easily interpretable. However, they tend to have lower prediction performances compared to the second type of models regrouping models such as gradient boosting machine (GBM) and random forests (RF). This type of models is more flexible and therefore performs better, but loses in interpretability. To bridge the gap between these two types of models, the explainable boosting machine (EBM) model is investigated in this master thesis. As it will be detailed later, this model has the advantage to have an explicit score, but is also based on gradient boosting and bagging techniques and can therefore pretend to have better performances than the models like GLM and GAM. In this master thesis, the performances of the EBM model are compared to the ones of a GAM, a RF and a GBM model. The application used to compare these models is the prediction of the annual claims frequency based on a MTPL database.

Non-discrimination and fairness are also ethical constraints that receive significant attention in predictive models. In the field of actuarial science, this topic has recently received increasing attention in the literature with authors such as Lindholm [1], Wüthrich [2] and Charpentier [3]. Regarding the insurance context, several variables, called protected variables, are prohibited by laws and regulation from being used to distinguish between the policyholders. As an example, the European Council has issued a directive in 2004 stating that the use of sex as a factor in the calculation of premiums and benefits for the purposes of insurance and related financial services shall not result in differences in individuals [4]. This directive is commonly called 'the unisex rule'. There may have been exceptions to this directive in the past, but since 21 December 2012 and the 'Test-Achats'

ruling, the unisex rule must be applied without exception. According to the unisex rule, the gender of the policyholder cannot be used as rating factor to give a different premium to a woman and a man having exactly the same risk profile, except their gender. The direct discrimination, which consists of directly using a protected variable as rating factor, is therefore avoided.

However, regarding the guidelines of the unisex rule, the non-protected variables correlated to the gender can be used as long as they are true risk factors in their own right [5]. As a consequence, indirect discrimination can still occur. The notion of indirect discrimination refers to the case where policyholders seem to be treated only on the basis of the non-protected variables. However, due to the correlation between the protected and the non-protected variables, the model can capture information about the protected variables from the non-protected variables and therefore uses this information to distinguish the policyholders. Indirect discrimination can therefore occur even if the protected variables are not used.

In this master thesis, different methods used to remove discrimination are presented. Some of them eliminate direct discrimination but not indirect discrimination, while others eliminate both. Several methods already exist in the literature such as the unawareness model and the discrimination-free model [1]. They are detailed in this work. Moreover, in the context of this master thesis, another method aiming to remove both direct and indirect discrimination is proposed. This method uses an EBM model and takes advantage of the fact that EBM models have an explicit score.

Regarding the notion of fairness, the unisex rule requires that the individual fairness is respected. The notion of individual fairness refers to the idea of treating similar people similarly [3]. Therefore, a woman and a man having the same risk profile must pay the same premium according to this individual fairness criterion and it is what the unisex rule stipulates. However, the notion of individual fairness is not the only notion of fairness. Another major type of fairness criterion is group fairness. The notion of group fairness refers to the idea of separating the population into different groups according to a particular protected variable (gender, religious belief, ethnicity, ...) and wanting a certain statistical measure to be equal across all groups [6].

One of the aims of this master thesis is to look at the implications between the non-discrimination methods and the different notions of fairness. This will be done in the context of our application, which consists of predicting the annual claims frequency from a MTPL database. A result that we will see is that the non-discrimination methods will imply individual fairness but not group fairness. However, techniques of pre-processing and post-processing exist in order to improve the group fairness. Several methods of pre-processing and post-processing are therefore investigated in this master thesis. These methods consist in pre-processing the database before training the predictive model or post-processing the predictions made by the model. In particular, one of the pre-processing technique is called disparate impact remover [7]. In the context of this master thesis, an alternative to this method is proposed in order to improve the accuracy of the model while maintaining the group fairness of the model.

Regarding the three ethical constraints described in this introduction (interpretabil-

ity, non-discrimination and fairness), it is legitimate to ask whether it is possible to have a model for non-life insurance pricing that is simultaneously interpretable, non-discriminatory and fair. This question will also be addressed in this master thesis.

Chapter 2

Explainable boosting machine

In this chapter, the explainable boosting machine (EBM) model is presented based on articles [8] and [9]. The aim of this model is to bridge the gap between two types of predictive models. The first type are models such as generalized linear model (GLM) or generalized additive model (GAM). These models are known to be easily interpretable since they have an explicit score. Thanks to this explicit score, the effect of each explanatory variable on the model's prediction is directly observable. However, GLMs and GAMs do not automatically capture the interactions and the latter need to be integrated manually in the score to appear.

The second type of models are models such as gradient boosting machine (GBM) and random forests (RF). This type of models is more flexible because interactions can be captured automatically. Therefore, GBM and RF models tend to have better predictive performances than GLMs and GAMs. However, on the other hand, GBM and RF models do not have an explicit score and are therefore less easily interpretable than GLMs and GAMs. Regarding these two types of models, explainable boosting machine (EBM) models seem to have the advantages of the two worlds. Indeed, as we will see in this chapter, EBM models have an explicit score constructed using gradient boosting techniques and detect automatically pairwise interactions.

In this chapter, we will therefore describe in details the EBM algorithm. Then, we will use the EBM model in a MTPL context. Indeed, the aim of the EBM model will be to predict the annual claims frequency for a MTPL database. Finally, based on this application, we will compare the EBM model with GAM, RF and GBM models.

2.1 Description of the EBM model

The explainable boosting machine (EBM) is a particular case of a generalized additive model (GAM). The goal of the EBM model is to estimate for a policyholder i the mean μ_i of the response Y_i with the features $\mathbf{x}_i$. Formally, μ_i is defined as

$$\mu_i = \mu(\mathbf{x}_i) = \mathbb{E}[Y_i|\mathbf{X}_i = \mathbf{x}_i].$$

The EBM model has the following form

$$g(\mu_i) = \beta_0 + \sum f_j(x_{ij}) + \sum f_{jk}(x_{ij}, x_{ik})$$

with g the link function, β_0 the intercept, x_{ij} the j^{th} feature of the policyholder i and f_j the univariate feature function for the feature x_j. The particularity of the EBM model is that it learns each feature function f_j using gradient boosting and bagging techniques [8]. The boosting procedure is trained on one feature at a time in round-robin fashion. It is this particularity that allows us to have an explicit score. Moreover, a small learning rate is used such that the order of the features does not matter.

A second particularity is that the EBM model includes pairwise interaction terms that are detected automatically thanks to the FAST algorithm [9] as it is explained in this section. This allows to increase the accuracy while maintaining the explainability of the model since functions that use two features are still explainable. To learn the functions f_j and f_{jk}, a two-stage approach is adopted. First, the EBM model is built using only one-dimensional components. More precisely, the algorithm used to construct the one-dimensional functions is the following:

Algorithm 1 EBM algorithm for the one-dimensional components

$m = 0,\ r_i^m = y_i,\ f_j = 0\ \forall j$
for round $l = 1, ..., L$ **do**
    for feature $j = 1, ..., J$ **do**
        Construct function $F_{j,l}$ on residuals r_i^m with a tree using only feature j
        $f_j = f_j + \lambda F_{j,l}$
        $m = m + 1$
        Compute the residuals r_i^m wrt all f_j
    end for
end for

In this algorithm, only gradient boosting techniques are used. It is however possible to combine boosting and bagging. In this case, the functions $F_{j,l}$ are constructed using a forest instead of a tree. The trees composing the forest are constructed on sub-samples of data drawn with replacement.

Once the univariate functions f_j are built, they are fixed. Then, the pairwise interaction functions f_{jk} are learned. For this second step, the FAST algorithm [9] is used to detect automatically the M most important pairwise interactions. The FAST algorithm sorts the pairwise interactions. The latter are sorted based on their ability to explain the residuals obtained with the model containing the univariate functions. For each interaction function f_{jk}, a simple model is built using cuts on x_j and x_k. The simplest model we can build is to place one cut on each variable, i.e., we place one cut c_j and one cut c_k on x_j and x_k, respectively. Those cuts are parallel to the axes. The interaction predictor T_{jk} is constructed by taking the mean of all target points in each quadrant, where the target points are in this case the residuals of the model obtained with the fixed univariate functions. We search for all possible $(c_j; c_k)$ and pick the best T_{jk}, i.e. the one with the lowest residual sum of squares. This indicator is used to measure the strength of the interaction (x_j, x_k) and to sort all the pairwise interactions. The best M pairwise interactions are then selected based on this indicator. Finally, the same round-robin boosting algorithm as in the univariate case (Algorithm 1) is launched to fit the M interaction terms f_{jk}. The only difference is that the functions f_{jk} are constructed based on two variables and not

on one as in the univariate case.

2.2 Claims frequency prediction with an EBM model

In this section, we will use an EBM model to predict the annual claims frequency. This will allow to show the benefits of using such a model with a concrete application. The database used for this application is a motor third party liability (MTPL) insurance portfolio from a Belgian insurer in 1997 [10]. The dataset contains 163 231 policyholders. Each of the policyholder is observed during a period of time ranging from one day to one year. The database contains the following variables:

- `NClaims` : The number of claims filed by the policyholder.
- `Exp` : The fraction of the year 1997 during which the policyholder was exposed to the risk.
- `Coverage` : Type of coverage provided by the insurance policy: TPL = only third party liability, PO = partial omnium = TPL + limited material damage, FO = full omnium = TPL + comprehensive material damage.
- `Fuel` : Type of fuel of the vehicle: gasoline or diesel.
- `Use` : Main use of the vehicle: private or work.
- `Fleet` : The vehicle is part of a fleet: yes or no.
- `Sex` : Gender of the policyholder: male or female.
- `AgePH` : Age of the policyholder in years.
- `AgeC` : Age of the vehicle in years.
- `Power` : Horsepower of the vehicle in kilowatt.

The MTPL database is analyzed in terms of observed annual claims frequency and risk exposure in Appendix A. In our claims frequency prediction application, the response variable is the annual claims frequency. The annual claims frequency y_i for the policyholder i is obtained by dividing the observed claim number N_i of the policyholder i by its risk exposure e_i

$$y_i = \frac{N_i}{e_i}.$$

The explanatory variables are

$$\mathbf{X} = [\texttt{Coverage, Fuel, Use, Fleet, Sex, AgePH, AgeC, Power}].$$

The first five explanatory variables are categorical and the last three are continuous. Since the response variable is the annual claims frequency, the EBM model takes as weight the risk exposure. Since the number of claims is assumed to follow a Poisson distribution, the loss function used to optimize the EBM model is the Poisson deviance.

The MTPL database is separated into a training set and a testing set. The training set is used to train the EBM model while the testing set is used to evaluate the prediction

performances of this latter. The training set contains 80% of the database and the testing set 20%. The training and testing sets are separated by stratifying the data on the basis of the number of claims so that there is a similar proportion of policies that have $N = 0, 1, 2, 3, ...$ in the training and testing sets.

In the EBM model, there are several hyperparameters that can be adjusted in order to improve the model. In this master thesis, we will focus on three of them:

- `max_leaves`: This parameter determines the maximum number of leaves in each tree. Since EBM uses gradient boosting techniques, the maximum number of leaves is limited to have small trees and this parameter will allow to do so.
- `outer_bags`: This is the number of times the whole EBM algorithm is run. Each EBM is fit on a different subsample of the training set. The final feature functions are an average of the ones learned in each outer bag. This procedure is used to smooth the functions and determine the size of error bars for these functions.
- `inner_bags`: This is the number of subsamples drawn with replacement on which trees are constructed during one step of the boosting procedure. Indeed, during this latter, when the EBM algorithm examines a particular feature, it generates `inner_bags` subsets of the training set and builds trees on each subset. Then, these trees are averaged together to obtain the final update used in the gradient boosting process. While including inner bagging will allow to improve the predictive performances, it will also increase the training time of the EBM model.

Moreover, the maximum number of rounds in the EBM algorithm is set to 5000. There is an early stop if there are 50 rounds with no improvement of the Poisson deviance. The number of interactions included in the EBM model is equal to 10.

To obtain the best values for the hyperparameters, the EBM model is cross-validated with a 5-folds cross-validation on the training set. The values tested for the hyperparameters are

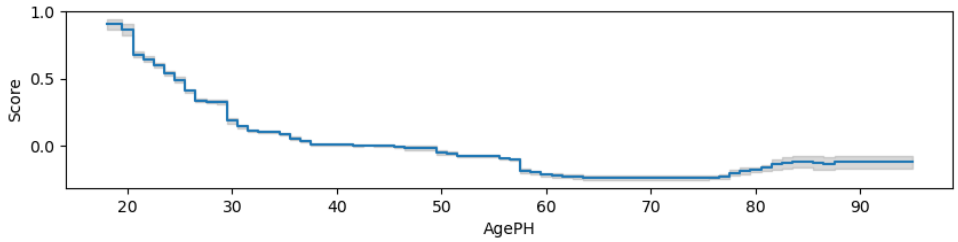
$$\texttt{max\_leaves} = [3, 4, 5], \qquad \texttt{outer\_bags} = [1, 8, 15, 25], \qquad \texttt{inner\_bags} = [0, 15, 25].$$

After cross-validation, the best hyperparameters are given by

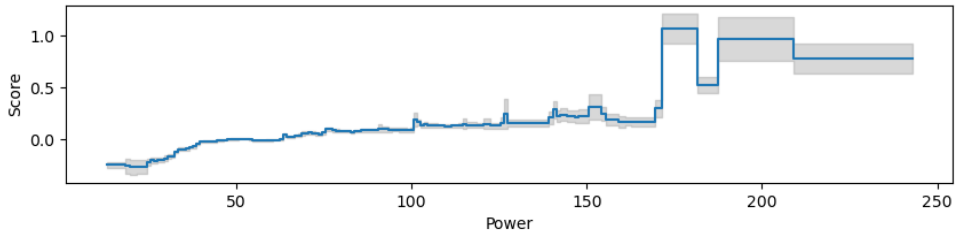
$$\texttt{max\_leaves} = 3, \qquad \texttt{outer\_bags} = 8, \qquad \texttt{inner\_bags} = 25.$$

Once the EBM has been cross-validated, since it has an explicit score, we can directly look at the part of the score related to each feature. We can for example begin by analysing the effect of the age of the policyholder (`AgePH`) as shown in Figure 2.1a. As expected, the predicted annual claims frequency will be much higher for younger policyholders than for older ones. Furthermore, the insured's age score decreases with age for ages between 18 and 40 years. Then, the score is almost constant between 40 and 50 years before decreasing again with age and finally increasing for elderly individuals.

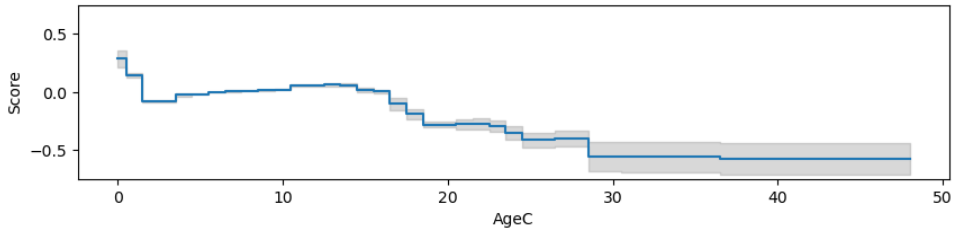
We can also look at the score of the power and of the age of the car (Figure 2.1b). For the power, the score generally increases with power, which was expected. For power ratings above 150 kW, the score is more volatile. This is because there are few cars with



(a) Univariate effect for the age of the policyholder.



(b) Univariate effect for the power of the car.



(c) Univariate effect for the age of the car.

Figure 2.1: Univariate effect of the variables `AgePH, Power` and `AgeC` in the EBM model.

a power rating above 150 kW. For the age of the car (Figure 2.1c), the trend is less clear, although the score seems to decrease as the age of the car increases. There is once again more volatility in the score for older cars, due to their low numbers.

Concerning the categorical variables, which are all nominal, the score for these variables is made up of a coefficient for each value of the categorical variable. For example, for the variable describing the fuel of the car, the diesel cars have a positive score and the gasoline cars a negative one. It was expected since the diesel cars tend to cover more kilometers and therefore tend to have more claims. The coefficient values for all the categorical variables can be found in Appendix B.1.

After the univariate effects, we can investigate the pairwise interactions. The most important interaction detected by the EBM model is the interaction between the age of the policyholder and its gender. As shown in Figure 2.2, for women, the part of the score related to this interaction is negative for young women and then increases with women's age. On the other hand, for men, the score is positive for young men and then decreases with men's age. There is therefore clearly an interaction between the gender and the age of the insured. Other graphs of pairwise interactions can be found in Appendix B.2.

All these graphs show that a big advantage of the EBM model is to have an explicit score allowing to show clearly and directly the relation between the explanatory variables and the predicted annual claims frequency. This is for example not the case for models

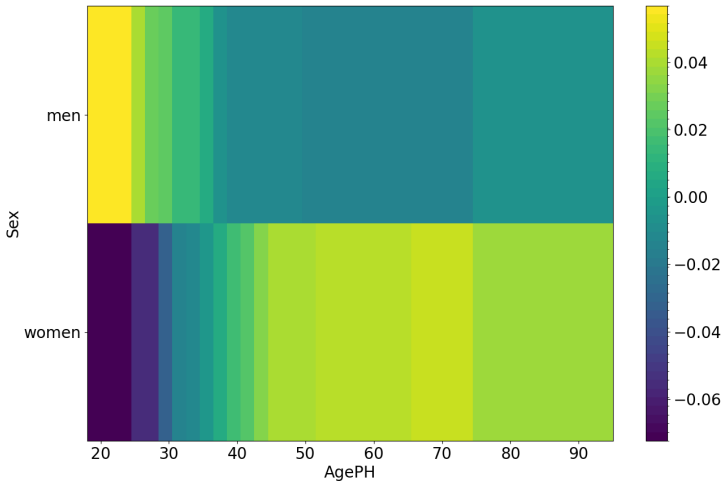


Figure 2.2: Interactions between the age of the policyholder (`AgePH`) and its gender (`Sex`) in the EBM model

like gradient boosting machine or random forests that do not have an explicit score. In the next section, we will therefore examine the prediction performance of all these models to see whether the EBM model can have good prediction performance like GBM and RF while having an explicit score.

To evaluate the prediction performances, we compute the Poisson deviance on the testing set. For our EBM model, the Poisson deviance is equal to 17687.33. It will be compared to the Poisson deviance obtained with the other models to see whether they have similar performances.

2.3 Comparison of the EBM model with other models

The goal of this section is to compare the EBM model with other models such as gradient boosting machine (GBM), random forests (RF) and generalized additive model (GAM).

2.3.1 Generalized additive model (GAM)

Generalized additive models (GAMs) allow to handle the continuous variables in flexible way. Indeed, GAMs have an additive score of the following form

$$g(\mu_i) = \beta_0 + \sum_{j=1}^{p_{cat}} \beta_j x_{ij} + \sum_{j=p_{cat}+1}^{p} f_j(x_{ij})$$

with p_{cat} the number of categorical variables and p the total number of explanatory variables [11]. The feature functions f_j are built using B-splines. This allows to model non-linear relationships between the features and the response.

As the aim of our MTPL application is to predict the annual claims frequency, we elaborate a GAM model with a Poisson distribution and the logarithm function for the link function g. The Python package used to elaborate our GAM model is `pyGAM`. As a remainder, the explanatory variables are

$$\mathbf{X} = [\texttt{Coverage, Fuel, Use, Fleet, Sex, AgePH, AgeC, Power}].$$

The first five variables are categorical and the last three are continuous. The categorical variables are dummified and there is one coefficient β_j for each level of the categorical variable, except for the reference level which goes in the intercept. Each continuous variable has a spline term, i.e. a continuous function f_j. The number of B-splines used to construct each feature function can be cross-validated. A 5-fold cross-validation is therefore made on the training set to obtain the best number of B-splines for each function f_j. Since the default number of B-splines in the package `pyGAM` is 20, the three values cross-validated are 15, 20 and 25 splines. This kind of cross-validation is selected for our GAM so that all the models of this master thesis is elaborated with the same type of cross-validation. After the cross-validation, the number of splines selected is 15 for the variable `AgePH`, 25 for `AgeC` and 15 for `Power`.

After cross-validating the GAM, we can compute the Poisson deviance on the testing set. The Poisson deviance is equal to 17709.89 which is higher than the one of our EBM model (17687.33). Our EBM model has therefore better prediction performances than our GAM. However, the performances of our GAM can perhaps be improved by including interaction terms in its score. The disadvantage is that the GAM does not automatically account for interactions while the EBM model does. The point here was therefore to show that EBM models can achieve better prediction performances than GAMs since EBM models detect pairwise interactions automatically.

2.3.2 Random Forests

The model called random forests is a method that combines a large number of trees in order to predict the mean μ of a response variable Y based on features $\mathbf{X}$. These trees are constructed on random samples of the training set taken with replacement and that have the same size as the training set. It is the bagging part of the random forests model. Moreover, the predictors (explanatory variables) are also randomized. Indeed, at each node of a tree constructed on a random sample of the training set, $m \leq p$ features are selected at random before each split and used as candidates for the split [12].

Since the aim is to design a RF model to predict the annual claims frequency, the criteria used to find the best split at each node of a tree is the Poisson deviance. The Python function used to develop our RF model is `RandomForestRegressor` from `sklearn.ensemble`. With this tool, several hyperparameters can be adjusted to improve the RF model:

- `min_samples_leaf`: This is the minimum number of samples required at any leaf node. Therefore, at each node of the tree, a split will only be considered if there are at least `min_samples_leaf` samples in the right and left branches.
- `max_features`: This is the number of features investigated when looking for the best split. This parameter is expressed as a fraction of the number of explanatory variables. A value of 1.0 for `max_features` corresponds therefore to bagged trees

and a value lower than 1.0 to the random forests case strictly speaking. This latter case allows more randomness and trees that are less correlated.

- `n_estimators`: This is the number of trees in the forest.

These hyperparameters are again cross-validated with a 5-folds cross-validation on the training set. The values tested for the hyperparameters are

$$\texttt{min\_samples\_leaf} = [250, 500, 1000], \qquad \texttt{max\_features} = [0.25, 0.5, 0.75, 1.0],$$

$$\texttt{n\_estimators} = [1000, 2500, 5000].$$

After cross-validation, the best hyperparameters are given by

$$\texttt{min\_samples\_leaf} = 500, \qquad \texttt{max\_features} = 0.5, \qquad \texttt{n\_estimators} = 5000.$$

It is then possible to compute the Poisson deviance of our cross-validated RF model on the testing set. The Poisson deviance is equal to 17686.53 while the one of our EBM model is equal to 17687.33. The prediction performances of our RF model are therefore very similar to the ones of our EBM model while this latter has an explicit score and thus gains in interpretability.

2.3.3 Gradient boosting

The boosting procedure works iteratively in order to predict the mean μ of a response variable Y based on features $\mathbf{X}$. At each iteration of the boosting, a new base learner T is added to the current score in order to decrease the loss function $L(Y, m)$, where $L(Y, m)$ represents the loss when estimating the mean of Y with m. The base learners are often small trees, but a little deeper than a stump with two leaves, so that they can capture interactions [11]. Gradient boosting is a boosting procedure executed by applying a least-squares principle to the gradients of the loss function.

In the context of our application consisting in predicting the claims frequency, the loss function to minimize is again the Poisson deviance. To construct our GBM model, we use the Python function `LGBMRegressor` from the package `LightGBM`. Among the hyperparameters that this function proposes, we focus here on 6 of them:

- `n_estimators`: The number of boosted trees to fit. Since the successive trees are correlated in the gradient boosting process, it will be an important hyperparameter to avoid overfitting.

- `max_depth`: The maximum depth of the tree. In GBM models, the maximum depth is often limited.

- `learning_rate`: The boosting learning rate.

- `min_child_samples`: The minimum number of samples required at any leaf.

- `subsample`: The proportion of the training set drawn without replacement which is used to construct a tree. Data subsampling in the gradient boosting procedure can be used to avoid overfitting.

- `subsample_freq`: The frequency of subsample. k means perform subsample at every k iteration of the gradient boosting process.

These hyperparameters are cross-validated with a 5-folds cross-validation on the training set. The values tested for the hyperparameters are

$$\texttt{n\_estimators} = [100, 500, 1000], \qquad \texttt{max\_depth} = [2, 3, 4],$$

$$\texttt{learning\_rate} = [0.01, 0.1], \qquad \texttt{min\_child\_samples} = [10, 20, 100, 500, 1000],$$

$$\texttt{subsample} = [0.25, 0.5, 0.75, 1.0], \qquad \texttt{subsample\_freq} = [0, 1],$$

After cross-validation, the best hyperparameters are given by

$$\texttt{n\_estimators} = 500, \qquad \texttt{max\_depth} = 2, \qquad \texttt{learning\_rate} = 0.1,$$

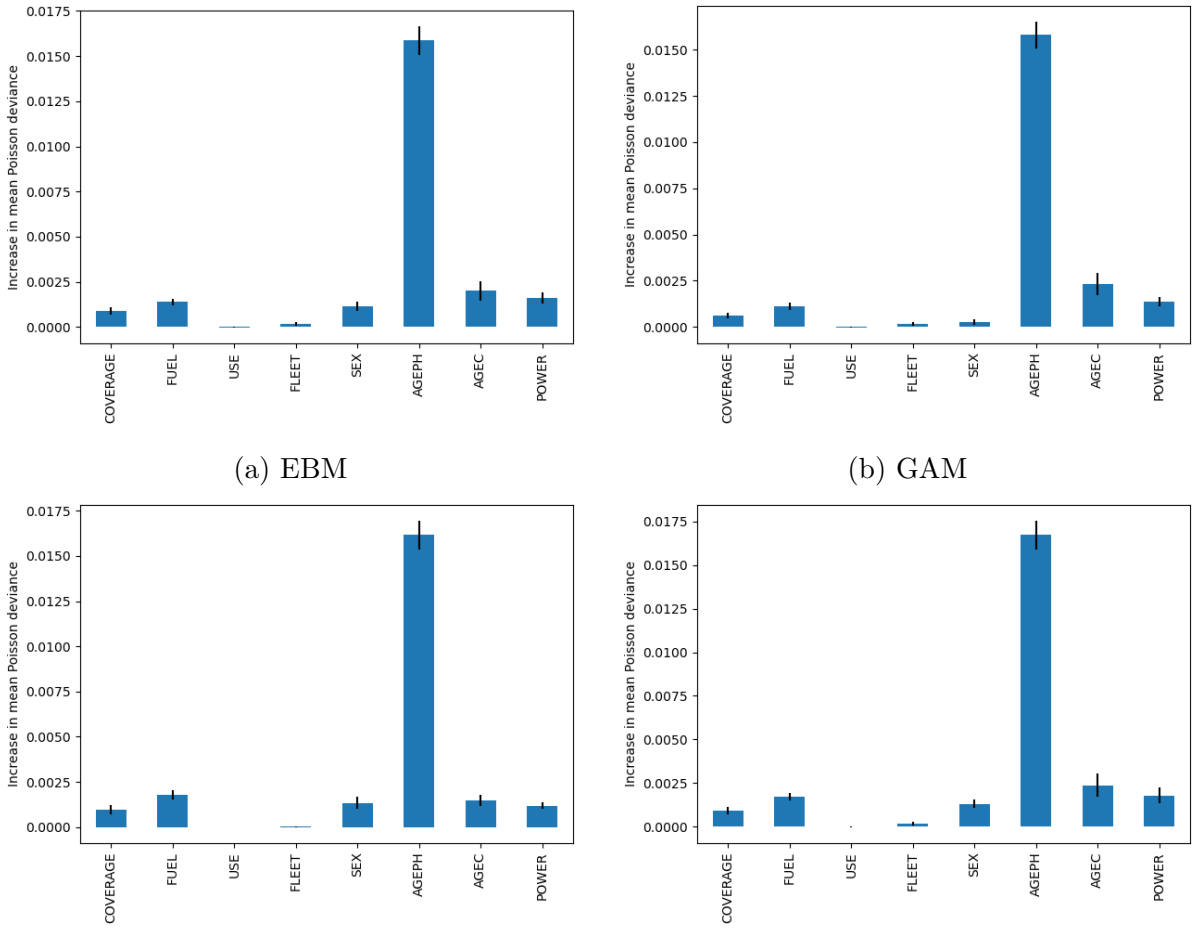
$$\texttt{min\_child\_samples} = 20, \qquad \texttt{subsample} = 0.5, \qquad \texttt{subsample\_freq} = 1.$$

Then, with the GBM model cross-validated, we can compute the Poisson deviance on the testing set. The Poisson deviance is equal to 17684.55, which is a bit lower but very similar to the one of our EBM model (17687.33). Our GBM model outperforms a little bit the prediction performances of our EBM model but this latter is more interpretable thanks to its explicit score.

2.3.4 Feature importance

The aim of the feature importance is to measure the importance of each variable in the model. One possible method is the permutation feature importance. This method evaluates how much the model's prediction error grows when the values of a particular feature are randomly shuffled, disrupting its connection with the actual outcome. In our case, the model's prediction error is the mean Poisson deviance. The permutation feature importance is computed on the testing set. This will allow to evaluate which variables are the most important to predict the claims frequency on unseen data and therefore to reduce the prediction error. The permutation feature importance is computed for our four models (EBM, GAM, GBM, RF) and is shown on Figure 2.3.

The permutation feature importance is similar for our four models. This suggests that features identified as important thanks to the feature importance are truly important for the prediction of the claims frequency as all models point in the same direction. The age of the policyholder is the most important variable in all our models. It was expected since it is known that there is a clear link between the age of the policyholder and the claims frequency. Indeed, the young drivers tend to have a higher claims frequency than the older ones. Then, there are 5 others variables that have relatively the same feature importance : the age of the car, its power, the type of fuel, the gender of the policyholder and the type of coverage he or she has underwritten. Finally, the type of use of the car and whether this latter belongs to a fleet or not are two variables that have almost no importance in predicting the claims frequency.



(a) EBM

(b) GAM

(c) RF

(d) GBM

Figure 2.3: Permutation feature importance computed on the testing set for the EBM (a), the GAM (b), the RF (c) and the GBM (d) model respectively.

2.3.5 Execution time

In this section, the execution time of each model is compared. To compute the execution time of a particular model, we take the execution time of the cross-validation for this model and we divide by the number of fits done during this cross-validation. We therefore obtain the mean execution time for the fitting procedure of each model.

Table 2.1 shows that the model with the lowest mean execution time in the GBM model, followed by the GAM, then the EBM model and finally the RF model. One of the reason why the RF model has the highest execution time is that the number of trees used to cross-validate our EBM model is quite high. Indeed, `n_estimators` = [1000,2500,5000]. The much longer execution time for the EBM model than for the GBM model can be explained by different reasons. Firstly, the EBM model begins by learning the univariate effects with a round-robin method, i.e using trees constructed with only one feature at a time. Then, in a second step, the pairwise interactions terms are constructed. For the GBM model, it is not the case. During the GBM procedure, the trees are constructed using several features at the same time. The interactions are therefore directly taken into account. This is one of the reason why the GBM model is faster than the EBM model. Secondly, the inner bagging present in the EBM model also increases the execution time.

	EBM	GAM	RF	GBM
Mean execution time [s]	107.08	7.89	124.72	1.96

Table 2.1: Mean execution time for the fitting procedure of the EBM, GAM, RF and GBM model respectively.

In conclusion, to have an explicit score, the EBM model needs to follow a more time-consuming procedure. Its execution time is therefore much longer than the one of the GBM model. However, the EBM model still has a lower execution time than the RF model since this latter is trained on a high number of trees.

2.4 Conclusion

The aim of this chapter was to study the EBM model. In particular, the performances of the EBM model was evaluated and compared to other models for our MTPL application, which consists of predicting the annual claims frequency from a MTPL database. This has allowed to show that the EBM model has prediction performances very similar to the RF and the GBM models. However, compared to these two models, the EBM model has the advantage of being easier to interpret thanks to its explicit score. In the remainder of this master thesis, EBM models will therefore be used.

Chapter 3

Non-discrimination in non-life insurance

In insurance pricing, there are several policyholder features that cannot be used to distinguish the insureds. These features are called protected features. On the contrary, non-protected features are those that can be used to distinguish between policyholders in the sense that two insureds can have different insurance tariffs if they have different non-protected features. In this context, in 2004, the European Council has issued a directive stating that "Member States shall ensure that in all new contracts concluded after 21 December 2007 at the latest, the use of sex as a factor in the calculation of premiums and benefits for the purposes of insurance and related financial services shall not result in differences in individuals' premiums and benefits." [4]. This directive is commonly referred to as 'the unisex rule'. There may have been exceptions to this directive in the past, but since 21 December 2012 and the 'Test-Achats' ruling, the unisex rule must be applied without exception. According to this rule, the gender is therefore clearly a protected feature and it is this feature that we are going to focus throughout this master thesis.

The notion of protected and non-protected variable leads to two different concepts: the direct discrimination and the indirect discrimination. The direct discrimination corresponds to the use of protected variables as a rating factor [13]. In this case, the protected variables will have a direct impact on the pricing of a given insurance product since they are used directly in the pricing process to differentiate between policyholders.

In the case of indirect discrimination, policyholders appear to be treated solely on the basis of non-protected variables, since the actuary has taken care not to use protected variables as rating factors. However, due to the correlation between the protected and the non-protected features, it is possible to infer information about protected variables from the non-protected variables [13]. The model will therefore indirectly discriminate the policyholders based on this information. Discrimination can therefore still occur, even if the protected variables are not used in the model as rating factors. Discrimination then appears in the form of indirect discrimination.

3.1 Non-discrimination methods

The aim of this section is to present different methods used to remove discrimination in the insurance pricing. Some methods will avoid both direct and indirect discrimination,

while others will only be able to avoid direct discrimination. Several non-discrimination methods already exist in the literature like the unawareness model and the discrimination-free model introduced by Lindholm et al. [1]. These methods are detailed below. Moreover, another non-discrimination method called No Protected Effect (NPE) model is proposed in this master thesis.

In order to introduce the different non-discrimination methods, we first begin by defining the best estimate model [1]. Let Y be the response variable, $\mathbf{X}$ the non-protected variables and P the protected variable. The goal of the best estimate model is to estimate the mean of the response variable Y knowing that $\mathbf{X} = \mathbf{x}$ and $P = p$, i.e.

$$\mu_{BE}(\mathbf{x}, p) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}, P = p].$$

Therefore, the best estimate model can be expressed as

$$\hat{\mu}_{BE}(\mathbf{x}, p) = f_{BE}(\mathbf{x}, p). \tag{3.1}$$

This model allows for direct discrimination since the protected variable is directly used as rating factor.

A first method to try to obtain a model without discrimination is to remove the protected variable from the database on which the model is trained. This corresponds to the notion of unawareness model [1]. The unawareness model aims to estimate the mean of Y knowing that $\mathbf{X} = \mathbf{x}$, i.e.

$$\mu_U(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}].$$

The unawareness model can therefore be expressed as

$$\hat{\mu}_U(\mathbf{x}) = f_U(\mathbf{x}). \tag{3.2}$$

In this case, the direct discrimination is avoided. However, as already mentioned above, if there is dependence between the protected and the non-protected variables, the model will capture information about the protected variable from the non-protected variables and there will be indirect discrimination.

The unawareness model can also be expressed in a more analytical way as introduced by Lindholm et al. [1]. Indeed, $\mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ can be rewritten as

$$\mathbb{E}[Y|\mathbf{X} = \mathbf{x}] = \mathbb{E}[\mathbb{E}[Y|\mathbf{X}, P]|\mathbf{X} = \mathbf{x}] = \sum_p \underbrace{\mathbb{E}[Y|\mathbf{X} = \mathbf{x}, P = p]}_{\mu_{BE}(\mathbf{x},p)} \mathbb{P}(P = p|\mathbf{X} = \mathbf{x})$$

if the protected variable is discrete, which will be the case in this master thesis. Therefore, the unawareness model can be expressed in a more analytical way and will be called "analytical unawareness model":

$$\hat{\mu}_{AU}(\mathbf{x}) = \sum_p \hat{\mu}_{BE}(\mathbf{x}, p)\ \mathbb{P}(P = p|\mathbf{X} = \mathbf{x}) \tag{3.3}$$

where $\hat{\mu}_{BE}(\mathbf{X}, P)$ is the prediction of the best estimate model. The analytical unawareness model corresponds therefore to a weighted average of the predictions of the best estimate

model across the conditional probability $\mathbb{P}(P = p|\mathbf{X} = \mathbf{x})$.

Thanks to this formulation, we clearly see that indirect discrimination can appear with the conditional probability $\mathbb{P}(P = p|\mathbf{X} = \mathbf{x})$. Indeed, this latter enables inference of the protected variable P from the non-protected variables $\mathbf{X}$. There is however one case in which there is no indirect discrimination. If the protected and the non-protected variables are independent, the model cannot deduce information about the protected variable from non-protected ones. Mathematically, this corresponds to $\mathbb{P}(P = p|\mathbf{X} = \mathbf{x}) = \mathbb{P}(P = p)$.

The last observation leads to the notion of discrimination-free model [1] which can be expressed as

$$\hat{\mu}_{DF}(\mathbf{x}) = \sum_{p} \hat{\mu}_{BE}(\mathbf{x}, p)\ \mathbb{P}(P = p) \tag{3.4}$$

The discrimination-free model is still a weighted average of the predictions of the best estimate model but now across the unconditional probability $\mathbb{P}(P = p)$. Therefore, there is no inference of the protected variable from the non-protected variables and there is no indirect discrimination.

In this master thesis, an alternative method of non-discrimination is proposed. This method, that I call the No Protected Effect (NPE) model, works in the following way. The first step is to calibrate a model based on all the non-protected and protected variables of the database. The first step corresponds therefore to the best estimate model. The model must however have an explicit score. Since the protected variable is directly used in the calibration, a part of the model score is dedicated to this variable. This part of the score is then set to zero. With this method, there is no direct discrimination since the effect of the protected variable is eliminated in the score and therefore the protected variable is not a rating factor.

Moreover, there is no indirect discrimination if we assume that the model has captured all the effect of the protected variable in the part of the score dedicated to it. Indeed, under this assumption, the non-protected variables do not capture any information on the protected variable and there is no direct and no indirect discrimination.

This method can therefore have a limitation. In case of strong correlation between a protected and a non-protected variable, it can be difficult to assess the effect of each variable. In this case, removing the part of the score related to the protected variable does not ensure that all the effect of the latter has been removed. However, for this master thesis and more specifically in the context of our application, it was checked that the non-protected variables are not too highly correlated to the protected variable, which is the gender. In the case of our MTPL database, the highest correlation between a non-protected variable and the gender is equal to -0.16. Therefore, there is no non-protected variable which is too highly correlated to the protected variable.

Once we have the model score where we have removed the part related to the protected variable, we can predict the expected response $\hat{\mu}_{NPE}(\mathbf{x})$ where NPE stands for "No Protected Effect" to say that the prediction is based on the new score where we have removed the part related to the protected variable. Since a part of the score has been removed, a correction must be made so that the sum of the response variables is equal to

the sum of the expected responses. The corrected expected response for the policyholder i is therefore given by

$$\hat{\mu}^*_{NPE}(\mathbf{x}_i) = \hat{\mu}_{NPE}(\mathbf{x}_i)\frac{\sum_i Y_i}{\sum_i \hat{\mu}_{NPE}(\mathbf{x}_i)}. \tag{3.5}$$

3.2 Results

The first aim of this section is to compare the different methods of non-discrimination that are presented in Section 3.1. These methods are implemented with EBM models and will be compared in the context of models predicting the annual claims frequency based on the MTPL database described in Section 2.2. The non-protected variables are therefore

$$\mathbf{X} = [\texttt{Coverage, Fuel, Use, Fleet, AgePH, AgeC, Power}]$$

and the protected variable is P = `Sex`.

More precisely, we will first compare the two unawareness models ((3.2) and (3.3)) and then the two methods which allow to remove both direct and indirect discrimination ((3.4) and (3.5)). Each time, we will check whether the two methods give similar results.

3.2.1 Comparison of the two unawareness models

To compute the two unawareness models, we will use EBM models. For the first unawareness model (3.2), the EBM model is fitted on the training set of the MTPL database using only the non-protected variables in order to obtain the predicted annual claims frequency $\hat{\mu}_U(\mathbf{x})$. The annual claims frequency is then predicted on a simulated database. This simulated database contains, for any policyholder's age between 18 and 80 and for each gender, the following risk profile:

- `Coverage`: TPL
- `Use`: private
- `AgeC`: 5
- `Fuel`: gasoline
- `Fleet`: no
- `Power`: 50

This database is constructed for graphical purposes, so that we can observe the evolution of the predicted annual claims frequency according to the age of the policyholder. To select our risk profile, we take, for each variable, a value for which there are a lot of policyholders to have a representative risk profile. However, what follows in this section has been tested for several risk profiles and the conclusions are the same for all of them. The annual claims frequency predicted with the first unawareness model is shown in Figure 3.1 and will be compared to the one obtain with the second unawareness model.

The second unawareness model is the analytical unawareness model given by

$$\hat{\mu}_{AU}(\mathbf{x}) = \sum_p \hat{\mu}_{BE}(\mathbf{x}, p)\ \mathbb{P}(P = p|\mathbf{X} = \mathbf{x}) \tag{3.6}$$

In our case, the protected variable P is the gender of the policyholder and can therefore take the value $P = w$ for woman or $P = m$ for man. In the literature, the conditional probability $\mathbb{P}(P = p|\mathbf{X} = \mathbf{x})$ is computed empirically based on the database of interest. However, if there are many variables, the empirical calculation of this probability can

become unstable, as there are not always many policyholders for each type of profile. This is why, in this master thesis, we compute the conditional probability with a logistic regression.

The logistic regression is made with a GBM with the Python package LightGBM. The response variable of the model is the variable `Sex` of the MTPL database (man = 1 and woman = 0). The explanatory variables are all the non-protected variables:

$$\mathbf{X} = [\text{Coverage, Fuel, Use, Fleet, AgePH, AgeC, Power}]$$

The loss function is the binomial deviance and the link function is the logit function. The GBM model is cross-validated with a 5-fold cross-validation on the training set of the MTPL database in order to obtain the best hyperparameters.

Regarding the analytical unawareness model (3.6), the term $\hat{\mu}_{BE}(\mathbf{x}, p)$ corresponds to the annual claims frequency predicted with the best estimate model. This latter corresponds to the model of Section 2.2 which is also cross-validated on the training set with a 5-fold cross-validation. Once we have the two models necessary to compute the two terms composing the analytical unawareness model, the annual claims frequency can be predicted on the simulated database based on this model.

Figure 3.1 shows the comparison between the claims frequency predicted with the unawareness model (3.2) and the one predicted with the analytical unawareness model (3.6) on the simulated database. We observe that the two predicted claims frequencies are very close. In the remainder of this master thesis, we will use the unawareness model (3.2) so that the results of the unawareness model is not linked to the ones of the best estimate model as it is the case for the analytical unawareness model (3.6).

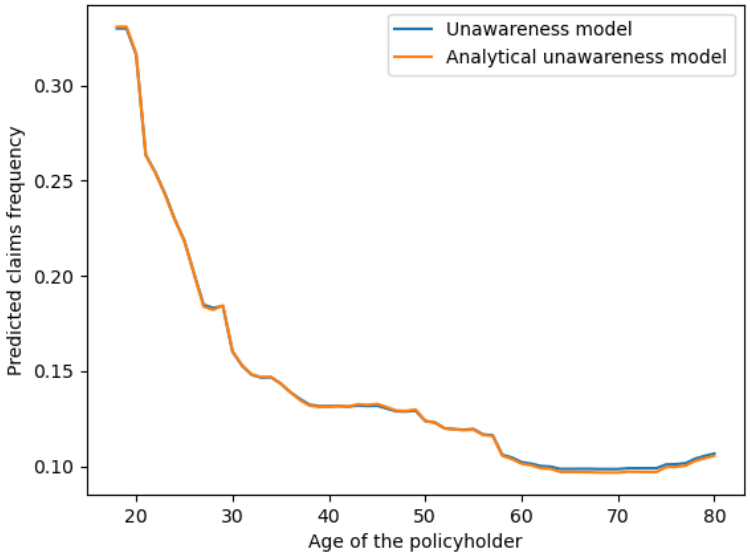


Figure 3.1: Comparison of the annual claims frequency predicted with the unawareness model and the analytical unawareness model on the simulated database.

3.2.2 Comparison of the two models without any discrimination

The first model without direct and indirect discrimination is the discrimination-free model (3.4). Once again, this model depends on the term $\hat{\mu}_{BE}(\mathbf{x}, p)$, which is the annual claims frequency predicted with the best estimate model. The second term of the discrimination-free model is $\mathbb{P}(P = p)$. In our case, this corresponds to the probability of being a woman or a man. This probability is computed empirically on the training set of the MTPL database since the discrimination-free model is trained on the training set. The probability to be a man, taking into account the exposure of the policyholders, is equal to 0.74. Once we have the two terms composing the discrimination-free model, we can predict the annual claims frequency on the simulated database as shown in Figure 3.2.

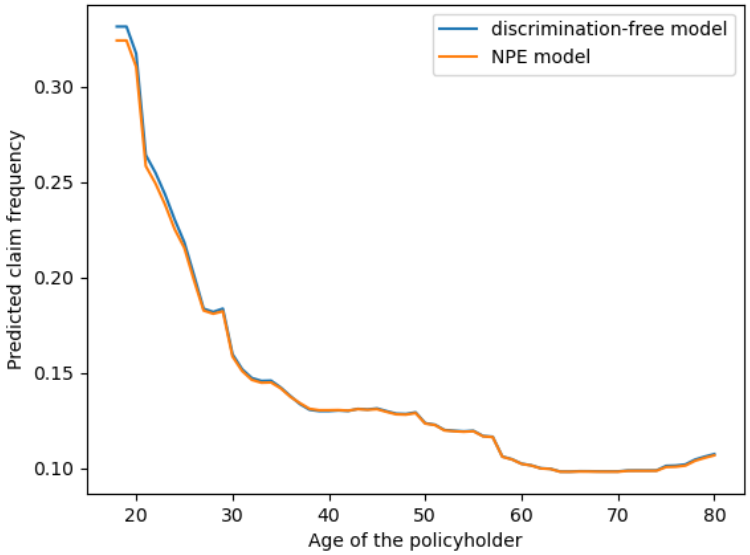


Figure 3.2: Comparison of the annual claims frequency predicted with the discrimination-free model and the NPE model on the simulated database.

The second model without direct and indirect discrimination is the model where the part of the score related to the protected variable is set to zero. The first step is therefore to start with the best estimate model fitted on the training set. Since the best estimate model uses directly the protected variable, which is the gender, to predict the annual claims frequency, several terms of the model score will be dedicated to the gender variable. In our case, the model score contains one univariate term for the gender variable and 3 pairwise interaction terms: one for the interaction between the gender and the age of the car, one between the gender and the policyholder's age and one between the gender and the power of the car. These terms are therefore set to zero.

Once these terms are removed, we can predict the annual claims frequency. We then compare on the training set the total number of claims predicted by this model to the total number of observed claims and we apply a correction to make them equal. It is expected that a correction must be done since some terms of the fitted model score are removed. Once this correction is computed based on the training set, we can predict the annual claims frequency with the new score on the simulated database and apply the

correction.

Figure 3.2 compares the annual claims frequency predicted with the discrimination-free model (3.4) and the one with the "No Protected Effect" (NPE) model (3.5). The two predicted claims frequencies are again very close. This can be explained analytically in the following way. As a remainder, the annual claims frequency predicted with the discrimination-free model is given by

$$\hat{\mu}_{DF}(\mathbf{x}_i) = \sum_p \hat{\mu}_{BE}(\mathbf{x}_i, p)\ \mathbb{P}(P = p).$$

Based on the fact that we are using EBM models, it can be rewritten as

$$\begin{aligned}
\hat{\mu}_{DF}(\mathbf{x}_i) &= \exp\left(\beta_0 + \sum_j f_j(x_{ij}) + \sum_{jk} f_{jk}(x_{ij}, x_{ik}) + g(P = m) + \sum_k g_{pk}(P = m, x_{ik})\right)\mathbb{P}(P = m) \\
&\quad + \exp\left(\beta_0 + \sum_j f_j(x_{ij}) + \sum_{jk} f_{jk}(x_{ij}, x_{ik}) + g(P = w) + \sum_k g_{pk}(P = w, x_{ik})\right)\mathbb{P}(P = w) \\
&= \underbrace{\exp\left(\beta_0 + \sum_j f_j(x_{ij}) + \sum_{jk} f_{jk}(x_{ij}, x_{ik})\right)}_{A}\left[\exp\left(g(P = m) + \sum_k g_{pk}(P = m, x_{ik})\right)\mathbb{P}(P = m)\right. \\
&\quad \left. + \exp\left(g(P = w) + \sum_k g_{pk}(P = w, x_{ik})\right)\mathbb{P}(P = w)\right]
\end{aligned}$$

where $\mathbf{x}_i$ is the vector of non-protected variables for the policyholder i, f_j and f_{jk} are the univariate and the interaction functions that do not depend on the protected variable P (the gender) and g and g_{pk} are the univariate and the interaction functions that depend on P.

For the NPE model, the annual predicted claims frequency is given by

$$\hat{\mu}^*_{NPE}(\mathbf{x}_i) = \hat{\mu}_{NPE}(\mathbf{x}_i)\frac{\sum_i N_i}{\sum_i \hat{\mu}_{NPE}(\mathbf{x}_i)\ e_i}$$

where N_i corresponds to the number of observed claims for the policyholder i and e_i is his risk exposure, measuring the fraction of the year in which he is insured. This frequency can be rewritten as

$$\hat{\mu}^*_{NPE}(\mathbf{x}_i) = \underbrace{\exp\left(\beta_0 + \sum_j f_j(x_{ij}) + \sum_{jk} f_{jk}(x_{ij}, x_{ik})\right)}_{A}\frac{\sum_i N_i}{\sum_i \hat{\mu}_{NPE}(\mathbf{x}_i)\ e_i}.$$

We therefore observe that the annual claims frequency predicted with the discrimination-free model or with the NPE model both correspond to the same term A multiplied by a correction.

As the annual claims frequency predicted with the discrimination-free model and the NPE model is very close (Figure 3.2), for the remainder of this master thesis, the NPE model is selected since this latter does not depend on another model. Indeed, on the other hand, the discrimination-free model is a weighted average of the predictions of the best estimate model and depends therefore on this latter.

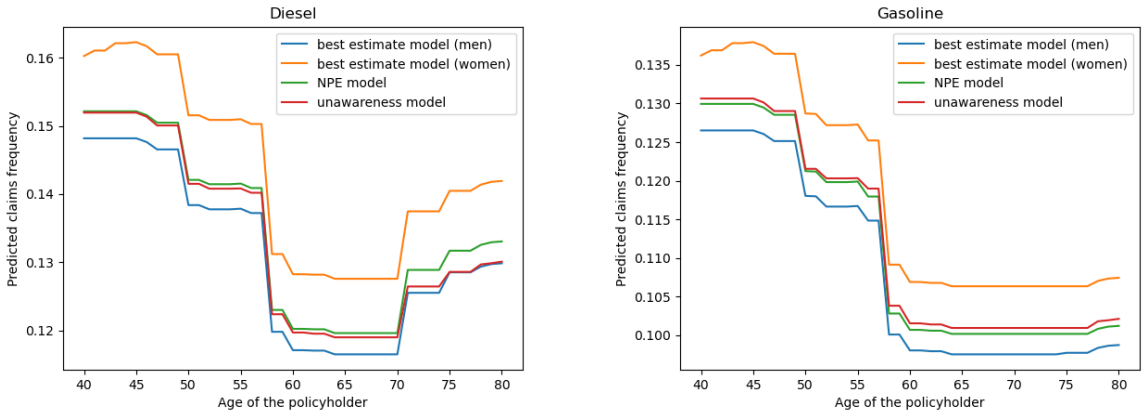
3.2.3 Illustration of the indirect discrimination in a MTPL database

In this section, three different models are compared in the context of the prediction of the claims frequency with the MTPL database described in Section 2.2. This will allow to highlight the impact of the indirect discrimination in a concrete MTPL example. The first model is the best estimate model (3.1) which involves direct discrimination. The second is the unawareness model (3.2) which eliminates direct discrimination but not indirect discrimination. Finally, the third model is the No Protected Effect model (3.5) which eliminate both direct and indirect discrimination.

In order to explain clearly the impact of the indirect discrimination, the MTPL database will be restricted to keep 3 explanatory variables. There are two non-protected variables: the age of the policyholder and the fuel of the car. The third explanatory variable is the protected variable which is the gender of the policyholder. Each of the three models is first calibrated on the training set of this database. Then, the annual claims frequency is predicted for each model on a simulated database containing all the possible combinations of the following variables:

- Age of the policyholder: 40, 41, ..., 80
- Fuel: diesel or gasoline
- Gender: woman or man

The predicted annual claims frequencies are shown separately for people with diesel cars (Figure 3.3a) and those with gasoline cars (Figure 3.3b). First of all, there is the evolution of the best estimate claims frequency for women and for men. Indeed, since the best estimate model uses directly the variable `Sex`, the predicted claims frequency is different for women and men. In both graphs, the best estimate claims frequency for women is higher than for men.



(a) Diesel cars

(b) Gasoline cars

Figure 3.3: Claims frequency predicted with the best estimate, the unawareness and the NPE model for diesel and gasoline cars.

For diesel cars, the unawareness claims frequency is below the one predicted by the NPE model. Moreover, the unawareness claims frequency and the best estimate claims

frequency for men are closer than the latter and the one of the NPE model. This can be explained by Figure 3.4. Indeed, in our MTPL database, we observe that in proportions, the people using diesel fuel are more likely to be men. This means that the variable `Fuel` has explanatory power to predict the gender of the policyholder. The unawareness model therefore gives a lower claims frequency to people using diesel compared to the NPE model because they are in proportion more likely to be men and men have a lower claims frequency as we observe on Figure 3.3.

For gasoline cars, the unawareness claims frequency is above the one predicted with the NPE model. Moreover, the unawareness claims frequency and the best estimate claims frequency for women are closer than the latter and the one of the NPE model. This can also be explained by Figure 3.4 since we observe that, in proportions, the people using gasoline fuel are more likely to be women. As already said, this shows that the variable `Fuel` has explanatory power to predict the gender. The unawareness model therefore gives a higher claims frequency for the gasoline cars compared to the NPE model because they are in proportion more likely to be women and women have a higher claims frequency as shown in Figure 3.3.

To sum up, for diesel cars, the unawareness model gives a lower claims frequency than the NPE model since diesel cars are more likely to be driven by men and men have a lower claims frequency. On the other hand, for gasoline cars, the unawareness model gives a higher claims frequency than the NPE model since gasoline cars are more likely to be driven by women and women have a higher claims frequency. Although the unawareness model do not use directly the variable `Sex` to predict the claims frequency, it takes into account the gender effect indirectly, whereas a model where no discrimination is desired cannot take into account the gender effect, either directly or indirectly. This example therefore shows that removing the gender from the database (unawareness model) is not sufficient to obtain a model without any discrimination and that this model induces indirect discrimination.

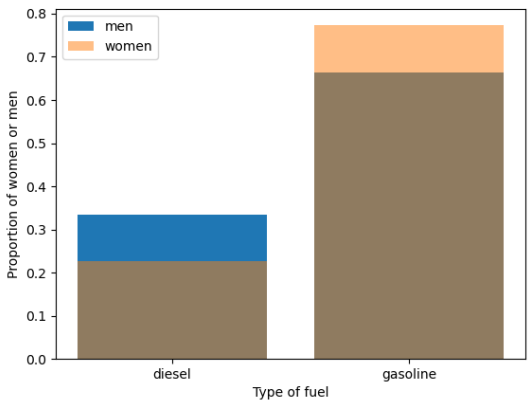


Figure 3.4: Comparison of the proportion of women and men owning a diesel or a gasoline car in terms of risk exposure.

3.3 Conclusion

In this chapter, we have studied various non-discrimination techniques. After comparing these different methods, we have selected two of them for the remainder of this master thesis: the unawareness model (3.2) and the No Protected Effect (NPE) model (3.5). The NPE model allows to eliminate both direct and indirect discrimination while the unawareness model allows to eliminate the direct but not the indirect discrimination as shown with our MTPL application in Section 3.2.3.

The non-discrimination methods can also be compared in terms of prediction performances. For that, Table 3.1 shows the Poisson deviance on the testing set of the MTPL database for the best estimate model (direct discrimination), the unawareness model (no direct but indirect discrimination) and the NPE model (no direct and indirect discrimination). The model with the lowest Poisson deviance is the best estimate model. It was expected since this model uses all the variables available to it, even the protected variable (the gender) to differentiate the policyholders. The second model with the lowest Poisson deviance is the unawareness model. This model is trained on a database that does not contain the gender of the policyholder to predict the claims frequency. However, the unawareness model still captures information on the gender due to the correlation between the gender and the non-protected variables. This is why the unawareness model has a lower deviance compared to the NPE model, where both direct and indirect discrimination are eliminated.

	Poisson deviance
Best estimate model	17687.33
Unawareness model	17704.94
No Protected Effect (NPE) model	17707.87

Table 3.1: Comparison of the Poisson deviance for the best estimate model, the unawareness model and the NPE model on the testing set of the MTPL database.

Due to the unisex rule stated by the European Council and described above, the best estimate model cannot be used in non-life insurance pricing. There is therefore a trade off between the unawareness model and the NPE model. Indeed, the unawareness model has a lower Poisson deviance but does not eliminate the indirect discrimination while the NPE model has a higher Poisson deviance but eliminate the indirect discrimination.

Chapter 4

Fairness in non-life insurance

4.1 Fairness criteria

Due to the growing interest in recent years in tackling the problem of fairness in predictive modeling and statistical learning, numerous definitions of fairness have appeared in the literature. These definitions can be separated in two different categories: group fairness and individual fairness. The notion of group fairness refers to the idea of separating the population into different groups according to a particular protected variable (gender, religious belief, ethnicity, ...) and wanting a certain statistical measure to be equal across all groups [6]. On the other hand, the notion of individual fairness refers to the idea of treating similar people similarly [3].

4.1.1 Group fairness

There are various criteria in the literature for measuring the group fairness of a model: the statistical/demographic parity [14], the equalized odds [15] and the bounded group loss [14]. In this master thesis, we focus on the most used group fairness criterion which is the statistical parity, introduced by Agarwal et al. [14]. This criterion is presented in the next section.

4.1.1.1 Statistical Parity

The statistical parity criterion is defined differently for classification and regression problems. In case of binary classification, a predictor $\hat{\mu} = f(\mathbf{X}, P)$ satisfies statistical parity if

$$\mathbb{P}(\hat{\mu} = 1|P = a) = \mathbb{P}(\hat{\mu} = 1|P = b) \qquad \forall a, b \in \mathcal{P}$$

where $\mathcal{P}$ is the set of values that the protected variable P can take [16].

In case of regression [2], a predictor $\hat{\mu} = f(\mathbf{X}, P)$ satisfies statistical parity if

$$\mathbb{P}(\hat{\mu} \leq m|P = a) = \mathbb{P}(\hat{\mu} \leq m|P = b) \qquad \forall a, b \in \mathcal{P} \text{ and } \forall m \in \mathbb{R}. \tag{4.1}$$

This criterion requires that the distribution of the predictions must be statistically the same in all groups where each group contains the people having a particular value for the protected variable P. In other words, this criterion wants that the prediction is independent of the protected variable:

$$\hat{\mu} \perp\!\!\!\perp P.$$

In this master thesis, we focus on the regression case as our application is the prediction of the claims frequency in the context of an MTPL insurance. This is why, in the remainder, we develop the fairness criteria for the regression case.

In practice, to evaluate if a certain predictor is statistical parity fair, we proceed in the following way. First, for a given database, we can obtain the predictions of this predictor and compute the cumulative distribution function (CDF) of these predictions. It is then possible to obtain the quantiles q_α of this CDF for all $\alpha \in [0, 1]$. Assuming that the protected variable P is a categorical variable with two levels a and b, we compute for each quantile q_α:

$$\left|\mathbb{P}(\hat{\mu} \leq q_\alpha | P = a) - \mathbb{P}(\hat{\mu} \leq q_\alpha | P = b)\right|.$$

Finally, we take the maximum value among all quantiles:

$$\text{MASP} = \max_{q_\alpha} \left|\mathbb{P}(\hat{\mu} \leq q_\alpha | P = a) - \mathbb{P}(\hat{\mu} \leq q_\alpha | P = b)\right|. \tag{4.2}$$

This latter expression is called Maximum Absolute Statistical Parity (MASP). This MASP criterion will allow to evaluate if a predictor $\hat{\mu}$ is statistical parity fair or not. Indeed, if the MASP is equal to zero, it means that

$$\mathbb{P}(\hat{\mu} \leq q_\alpha | P = a) = \mathbb{P}(\hat{\mu} \leq q_\alpha | P = b)$$

for all quantiles q_α. Therefore, the predictions are distributed exactly in the same way for people having $P = a$ or people having $P = b$ and the predictor $\hat{\mu}$ is perfectly statistical fair.

One characteristic of the statistical parity criterion is that, to be satisfied, this criterion imposes that the distribution of the prediction must be the same across all the groups defined by the levels of the protected variable P, no matter the other explanatory variables (i.e. the non-protected variables). However, it could also be legitimate to look at the statistical parity criterion for people having the same value for certain non-protected variables. This is why the conditional statistical parity criterion is introduced next.

4.1.1.2 Conditional statistical parity

The idea of the conditional statistical parity is to look at the statistical parity criterion for people having the same value for a subset $\mathbf{X}_S$ of the non-protected variables $\mathbf{X}$. Compared to the statistical parity criterion (4.1), a condition is therefore added so that the statistical parity is analyzed for people having the same values for $\mathbf{X}_S$. In case of regression, a predictor $\hat{\mu} = f(\mathbf{X}, P)$ satisfies conditional statistical parity if

$$\mathbb{P}(\hat{\mu} \leq m | P = a, \mathbf{X}_S = \mathbf{x}_S) = \mathbb{P}(\hat{\mu} \leq m | P = b, \mathbf{X}_S = \mathbf{x}_S)$$

$\forall a, b \in \mathcal{P}$ and $\forall m \in \mathbb{R}$.

To evaluate the conditional statistical parity of a predictor $\hat{\mu}$ in practice, we adopt the same reasoning that for the statistical parity. A predictor $\hat{\mu}$ is perfectly conditional statistical fair if the conditional MASP defined as

$$\text{conditional MASP} = \max_{q_\alpha} \left|\mathbb{P}(\hat{\mu} \leq q_\alpha | P = a, \mathbf{X}_S = \mathbf{x}_S) - \mathbb{P}(\hat{\mu} \leq q_\alpha | P = b, \mathbf{X}_S = \mathbf{x}_S)\right|$$

is equal to zero.

4.1.2 Individual fairness

The individual fairness refers to the concept of treating similar people similarly [3]. In the context of insurance pricing, this means that a woman and a man having the same risk profile must pay the same premium for their insurance. To derive mathematically the criterion allowing to evaluate if a predictor is fair according to the individual fairness, it is possible to start from the concept of conditional statistical parity. Indeed, in case of the individual fairness, we also look at the fairness of a predictor conditionally on the values of the non-protected variables $\mathbf{X}$. However, in case of individual fairness, we compare individuals that have the same risk profile. Therefore, instead of conditioning with respect to a subset $\mathbf{X}_S$ of the non-protected variables, we condition with respect to all non-protected variables $\mathbf{X}$.

In case of regression, a predictor $\hat{\mu} = f(\mathbf{X}, P)$ satisfies individual fairness if

$$\mathbb{P}(\hat{\mu} \leq m | P = a, \mathbf{X} = \mathbf{x}) = \mathbb{P}(\hat{\mu} \leq m | P = b, \mathbf{X} = \mathbf{x})$$

$\forall a, b \in \mathcal{P}$ and $\forall m \in \mathbb{R}$. The individual fairness is therefore a particular case of the conditional statistical parity where we condition with respect to all the non-protected variables.

Since the prediction is conditional on all the variables involved in the predictor, for each condition, we have only one value for the prediction and not a distribution. Therefore, the individual fairness criterion can be rewritten as

$$f(P = a, \mathbf{X} = \mathbf{x}) = f(P = b, \mathbf{X} = \mathbf{x}) \qquad \forall a, b \in \mathcal{P}. \tag{4.3}$$

This individual fairness criterion corresponds to the unisex rule already presented in this chapter. Indeed, this directive states that the use of sex in the calculation of premiums and benefits shall not result in differences in individuals' premiums and benefits. In other words, two individuals with the same risk profile but different genders should have the same premium and this corresponds exactly to the individual fairness criterion (4.3). The individual fairness criterion is therefore the criterion that insurers must meet under the current regulation.

One drawback of the individual fairness is that it does not take into account the potential interdependence between the non-protected variables $\mathbf{X}$ and the protected variable P [16]. Therefore, an unawareness model, where we only remove the protected variable before fitting the model, will satisfy the individual fairness criterion. However, as we have already seen, an unawareness model can imply indirect discrimination if there is correlation between the protected and the non-protected variables. Since the individual fairness does not check this point, we can have a model with indirect discrimination which satisfy the individual fairness.

4.2 Pre-processing techniques

As we have already discussed previously in this master thesis, there can be a dependence between the protected variable P and the non-protected variables $\mathbf{X}$. As a consequence, if we use the unawareness model as a non-discrimination method, this can cause

indirect discrimination. Moreover, since $\mathbf{X}$ and P are dependent, $\mathbf{X}$ is not distributed in the same way for all the values of P. Therefore, the probability that the model's prediction is not distributed in the same way for all values of P is high, no matter if we use the unawareness model or the No Protected Effect (NPE) model as non-discrimination method. This implies that the statistical parity (group fairness criterion) will not be satisfied.

The aim of the pre-processing methods is to remove the dependence between the protected variable P and the non-protected variables $\mathbf{X}$. Thanks to that, it limits the two disadvantages mentioned just above. Indeed, assuming that the pre-processing methods allow to remove all the dependence between $\mathbf{X}$ and P, there will be no indirect discrimination with the unawareness model. Moreover, a sufficient condition to satisfy statistical parity is that $\mathbf{X}$ and P are independent and that P is not used in the predictive model. Therefore, in this master thesis, we investigate two pre-processing techniques that are used in order to improve the statistical parity criterion: the correlation remover and the disparate impact remover (DIR) techniques.

4.2.1 Correlation remover

The aim of the correlation remover technique [17] is to remove from the non-protected variables their correlation, i.e. their linear dependence, with the protected variable. To do so, for each quantitative non-protected variable, a linear regression model is fitted to explain the non-protected variable X with the centered protected variable P:

$$X = \beta(P - \bar{P}) + \epsilon$$

Then, the new non-protected variable is defined as the residual $X - \beta(P - \bar{P})$ of the regression :

$$X_{new} = X - \beta(P - \bar{P})$$

The idea is that the model has captured all the variations in X that can be attributed to P. Consequently, the residual, i.e. the new non protected variable, does not contain linear dependence with the protected variable P. This correlation remover technique is only applied to the quantitative variables.

One drawback of this pre-processing technique is that it only tries to remove the correlation between the protected and the non-protected variables. Since correlation is only a measure of linear dependence, it does not guarantee independence between these variables. This pre-processing technique is implemented with the function `CorrelationRemover` of the Python package `Fairlearn` [17].

4.2.2 Disparate impact remover

The Disparate impact remover (DIR) technique, introduced by Feldman et al. [7], is a pre-processing technique designed to remove the dependence between the non-protected variables and the protected one. Let X a quantitative non-protected variable from which we want to remove its dependence with a protected variable P. We define the conditional CDF of X as $F_{X|P}(x) = \mathbb{P}(X \leq x|P = p)$ and the corresponding quantile function is denoted as $F^{-1}_{X|P}(u)$ where u is the CDF level. For each value p of the protected variable

P, we will therefore compute the quantile function $F^{-1}_{X|P}(u)$. Then, we compute the median of these quantile functions:

$$F^{-1}_{DIR} : F^{-1}_{DIR}(u) = \text{median}_{p \in P}\ F^{-1}_{X|P}(u)$$

Finally, the non-protected variable X is adjusted in the following way to obtain the new non-protected variable after pre-processing:

$$x_{new} = F^{-1}_{DIR}(F_{X|P}(x))$$

The DIR technique is only applied to the quantitative variables. As shown, the idea of this pre-processing technique is to adjust the values of the non-protected variable X such that X has the same distribution in each group defined by the values of P (for example the men's group and the women's group in case P is the gender). Moreover, the rank within each group is preserved. Indeed, in a certain group p, if $x_a \leq x_b$, then $x_{new,a} \leq x_{new,b}$.

In our case, the protected variable is the gender. The variable gender is therefore $P = w$ for women or $P = m$ for men. Therefore, the median of the two quantile functions is defined as:

$$x_{new} = F^{-1}_{DIR}(u) = \frac{1}{2}\ F^{-1}_{X|P=w}(u) + \frac{1}{2}\ F^{-1}_{X|P=m}(u)$$

This formula works in the following way. Let a woman/man having the value x before pre-processing for a certain non-protected variable X and let this value x corresponding to the level u for the women/men CDF $F_{X|P=w}(x)/F_{X|P=m}(x)$. Then, the value of X for this woman/man after pre-processing, x_{new}, corresponds to the median between the value of the women quantile function for the CDF level u ($F^{-1}_{X|P=w}(u)$) and the value of the men quantile function for the same CDF level u ($F^{-1}_{X|P=m}(u)$).

4.2.2.1 Weighted DIR

In this master thesis, a first alternative to the classical DIR technique is proposed and is called weighted DIR. With the weighted DIR technique, the adjusted non-protected variable is not obtain by taking the median of the two quantile functions but a weighted average of the latter. Indeed, the weighted DIR technique consists in computing the weighted average of the two quantile functions with the probability to be a woman and a man as weights.

With this alternative, the adjusted non-protected variable after pre-processing is given by:

$$x_{new} = F^{-1}_{DIR}(u) = \mathbb{P}(P = w)\ F^{-1}_{X|P=w}(u) + \mathbb{P}(P = m)\ F^{-1}_{X|P=m}(u)$$

Indeed, it seems a good idea that if there is more men than women in the database, the weight of the men's quantile function must be higher. This means that we will retain more information about the value of the variable before pre-processing for men than for women, as there are more men than women in the database. This weighted DIR method will therefore allow to have a new database after pre-processing which is more aligned with the original one than it was with the new pre-processed database obtained with the classical DIR method. As the database pre-processed with this alternative is more aligned to the initial database, a model fitted with this pre-processed database should have a lower Poisson deviance than the same model fitted with the database pre-processed thanks to the classical DIR. We will see in Section 4.4 if it is the case or not.

4.2.2.2 Conditional weighted DIR

A second alternative corresponding to an extension of the first one is also implemented. In the first alternative, i.e. the weighted DIR technique, the probability to be a woman or a man is taken into account. However, depending on the non-protected variable on which the pre-processing technique is applied and depending on the value of this variable, the probability to be a woman or a man can be different. To take this into account, we elaborate a second alternative, called conditional weighted DIR, in the following way. Let X the non-protected variable on which we apply the conditional weighted DIR technique. To take into account the fact that the probability to be a woman/man can vary with the value of X while keeping probabilities that are stables, two different probabilities are computed:

$$\mathbb{P}(P = m | x \leq \text{median}_X) \quad \text{and} \quad \mathbb{P}(P = m | x > \text{median}_X)$$

The probability to be a woman is then deduced from the one of the man:

$$\mathbb{P}(P = w | x \leq \text{median}_X) = 1 - \mathbb{P}(P = m | x \leq \text{median}_X)$$
$$\mathbb{P}(P = w | x > \text{median}_X) = 1 - \mathbb{P}(P = m | x > \text{median}_X)$$

Finally, with the conditional weighted DIR method, the adjusted non-protected variable after pre-processing is given by:

If $x \leq \text{median}_X$,

$$x_{new} = F^{-1}_{DIR,1}(u) = \mathbb{P}(P = m | x \leq \text{median}_X)\, F^{-1}_{X|P=m}(u) + \mathbb{P}(P = w | x \leq \text{median}_X)\, F^{-1}_{X|P=w}(u)$$

If $x > \text{median}_X$,

$$x_{new} = F^{-1}_{DIR,2}(u) = \mathbb{P}(P = m | x > \text{median}_X)\, F^{-1}_{X|P=m}(u) + \mathbb{P}(P = w | x > \text{median}_X)\, F^{-1}_{X|P=w}(u)$$

This second alternative will allow to have a new database after pre-processing which is even more aligned with the one before pre-processing. As the database pre-processed with the conditional weighted DIR is even more aligned to the initial database, a model fitted with this pre-processed database should have a lower Poisson deviance than the same model fitted with the database pre-processed thanks to the classical DIR or the weighted DIR technique. We will see in Section 4.4 if it is the case or not.

A drawback with this second alternative is that a women which has for the variable X a certain value x_1 corresponding to a level u for the CDF $F_{X|P="w"}(x_1)$ can have its new variable after pre-processing $x_{1,new}$ which is different from the new variable $x_{2,new}$ of a men having the same level u for the CDF $F_{X|P="m"}(x_2)$. Indeed, if $x_1 \leq median_X$, $x_{1,new}$ is obtained with $F^{-1}_{DIR,1}(u)$. While if $x_2 > median_X$, $x_{2,new}$ is obtained with $F^{-1}_{DIR,2}(u)$. This second alternative can thus introduce some differences between the distribution of the variable after pre-processing for women and men but it concerns a limited proportion of the portfolio.

4.3 Post-processing technique

In this section, a post-processing technique is presented. The aim of this technique is to post-process the model so that the prediction of this latter is distributed in the same

way in each group defined by the values of the protected variable P. For example, there are the men's group and the women's group in case P is the gender. This post-processing technique can therefore be used to improve the statistical parity criterion (group fairness).

The post-processing technique, introduced by Chzhen et al. [18], works as follows. Let's consider the case where the protected variable is binary: without loss of generality, $P = \{w, m\}$, where w is for 'women' and m for 'men'. For an individual in group w with the non-protected features $\mathbf{x}$, the fair prediction g is given by

$$g(\mathbf{x}, w) = \mathbb{P}(P = w) f(\mathbf{x}, w) + \mathbb{P}(P = m) t(\mathbf{x}, w) \tag{4.4}$$

with

$$t(\mathbf{x}, w) = \inf\{t \in \mathbb{R} : F_{f|m}(t) \geq F_{f|w}(f(\mathbf{x}, w))\}$$

and $f(\mathbf{x}, w)$ the prediction of the best estimate model before the post-processing method for an individual in group w with the non-protected features $\mathbf{x}$. Moreover, $F_{f|w}(f(\mathbf{x}, w))$ is defined as

$$\mathbb{P}(f(\mathbf{X}, P) \leq f(\mathbf{x}, w) | P = w)$$

and $F_{f|m}(t)$ is defined as

$$\mathbb{P}(f(\mathbf{X}, P) \leq t | P = m).$$

The formula (4.4) can be understood as follows. The fair prediction $g(\mathbf{x}, p)$ is a combination of the initial prediction $f(\mathbf{x}, p)$ and the adjusted prediction $t(\mathbf{x}, p)$ which is computed in the following way. If $p = w$, we first compute the fraction of individuals from group w whose prediction is at most $f(\mathbf{x}, w)$, that is, we compute $\mathbb{P}(f(\mathbf{X}, P) \leq f(\mathbf{x}, w) | P = w)$. Then, we find a candidate $\bar{\mathbf{x}}$ in group m, such that the fraction of individuals from group m whose prediction is at most $f(\bar{\mathbf{x}}, m)$ is the same, that is, $\bar{\mathbf{x}}$ is chosen to satisfy $\mathbb{P}(f(\mathbf{X}, P) \leq f(\bar{\mathbf{x}}, m) | P = m) = \mathbb{P}(f(\mathbf{X}, P) \leq f(\mathbf{x}, w) | P = w)$. Finally, the prediction of $\bar{\mathbf{x}}$ is exactly the adjustment for $\mathbf{x}$, that is, $t(\mathbf{x}, w) = f(\bar{\mathbf{x}}, m)$.

The idea is the following: if individuals $(\mathbf{x}, w)$ and $(\bar{\mathbf{x}}, m)$ share the same group-wise prediction ranking (for example, the two individuals have a median prediction within their respective group), then they should have the same prediction determined by the fair prediction

$$g(\mathbf{x}, w) = g(\bar{\mathbf{x}}, m) = \mathbb{P}(P = w) f(\mathbf{x}, w) + \mathbb{P}(P = m) f(\bar{\mathbf{x}}, m).$$

Thanks to this post-processing method, the prediction of the model will be distributed in the same way for $P = w$ and $P = m$. This will therefore allow to satisfy the statistical parity criterion (group fairness).

4.4 Results

4.4.1 Group fairness

In this section, we evaluate the statistical parity criterion in the context of our MTPL application. The aim is to see, for different non-discrimination methods presented in Chapter 3, whether the annual claims frequency predicted by these methods is fair according to the statistical parity criterion (4.1). This amounts to assess whether the predicted claims frequency is distributed in the same way for women and for men thanks to the

MASP criterion (4.2).

We look at the statistical fairness for the three levels of discrimination (direct discrimination, indirect discrimination and no discrimination) and we take for each level the model/method retained in Chapter 3. To represent the models with direct discrimination, we take the best estimate model (3.1). For the indirect discrimination, it is the unawareness model (3.2) and we take the No Protected Effect (NPE) model (3.5) to represent models with no discrimination.

For all models, the statistical fairness criterion is evaluated on the testing set of the MTPL dataset. To do so, the model, which has been trained on the training set, predicts the annual claims frequency on the testing set. The statistical parity criterion is then evaluated based on this prediction.

As explained in Section 4.1.1, to evaluate the statistical parity (group fairness) of a predictive model, we compute the Maximum Absolute Statistical Parity (MASP) criterion (4.2) on the predictions of this model. This criterion is always positive and the closer it is to zero, the fairer the model. The MASP criterion for the best estimate model, the unawareness model and the No Protected Effect (NPE) model are given in Table 4.1.

	MASP
Best estimate model	0.1681
Unawareness model	0.0416
No Protected Effect (NPE) model	0.0317

Table 4.1: Maximum Absolute Statistical Parity (MASP) criterion computed on the testing set for the best estimate, unawareness and NPE model.

For the best estimate model, the MASP criterion is equal 0.1681. It is the highest value among the three models. The best estimate model is therefore the less fair model. It was expected as there is direct discrimination in this model, whereas there is none in the other two.

For the unawareness model, the MASP criterion is equal to 0.0416 and has therefore well decreased compared to the best estimate model. Eliminating the direct discrimination by removing the protected variable, i.e. the gender of the policyholder, from the dataset before the fitting procedure allows to reduce the MASP criterion quite significantly. However, as we have seen, the unawareness model can involve indirect discrimination and this is one of the reason why the MASP criterion is not equal to zero.

The No Protected Effect (NPE) model is the one with the lowest MASP criterion which is equal to 0.0317. This can be explained by the fact that in this model, there is no direct and no indirect discrimination. However, the MASP criterion is still not equal to zero and we will explain why. The reason is that, by analyzing in more details the results, we observe that the CDF of the annual predicted claims frequency for men is almost always above the one for women (Figure 4.1). Women therefore tend to have a higher claims frequency than men. This can be explained by the age distribution of the policyholders in our MTPL database. Indeed, in this latter, there is in proportion more

women than men at younger ages and less at older ages (Figure 4.2). Since the predicted claims frequency is higher at younger ages, the predicted claims frequency is higher for women than men if we look at our portfolio in an aggregate way.

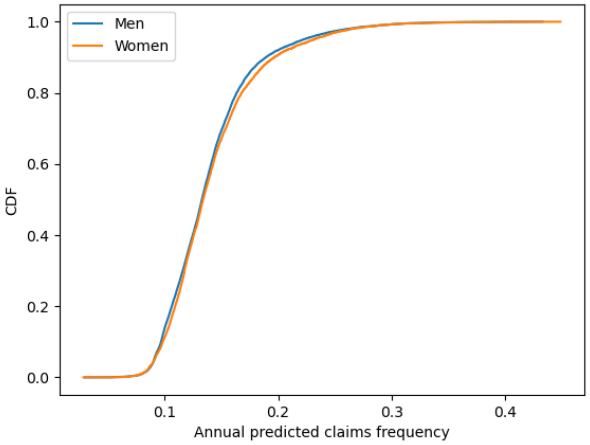


Figure 4.1: CDF of the annual claims frequency for women and men predicted with the NPE model.

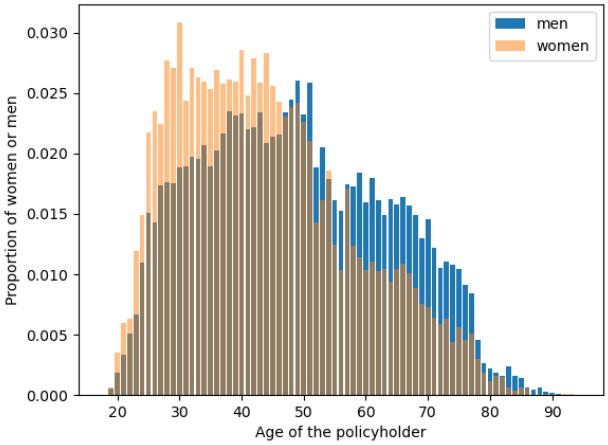


Figure 4.2: Proportion of women and men for each policyholder age.

In conclusion, the MASP criterion is not equal to zero even with the model without discrimination (NPE model) since the age distribution of the policyholder is not same for women and men. As a consequence, the distribution of the predicted claims frequency is not the same for women and men whereas it is what we need to have a MASP criterion exactly equal to zero. In view of this last observation, it can be a good idea to look at the conditional statistical parity conditionally on the age of the policyholders.

To evaluate the conditional statistical fairness of the NPE model, we compute the conditional MASP on the testing set for policyholders aged 45. It is equal to 0.2388. The reason why the conditional MASP is not equal to zero is also that some non-protected variables are not distributed in the same way for women aged 45 and men aged 45. For example, for people aged 45, there is in proportion more men than women with a diesel car, while it is the opposite for the gasoline cars. Since diesel cars tend to have a higher claims frequency, men aged 45 tend to have a higher predicted claims frequency than women aged 45 and the conditional MASP is therefore not equal to zero. Instead of continuing the reasoning and look at the conditional MASP for people aged 45 and for a certain type of fuel, in the next section, we will directly look at the individual fairness where the fairness of the model is assessed conditionally on the values of all non-protected variables.

4.4.2 Individual fairness

The individual fairness criterion (4.3) is assessed for the three following models: the best estimate, the unawareness and the No Protected Effect (NPE) model. To evaluate this criterion, we need to specify entirely the risk profile of the policyholder as we have to verify if two individuals that only differ by their gender have the same predicted claims

frequency. However, the results obtained are the same no matter the risk profile chosen.

For the best estimate model, the predicted claims frequency for a woman and a man having the same risk profile is not the same since the protected variable, which is the gender of the policyholder, is directly used to differentiate the insureds. The best estimate model is therefore not fair according to the individual fairness criteria.

For the unawareness model where we remove the gender before the fitting procedure, a woman and a man having the same risk profile will have the same predicted claims frequency. In the unawareness model, the gender does not intervene directly but there may be an indirect effect due to the correlation between the protected and the non-protected variables. However, as for the individual fairness criterion we condition on all non-protected variables, this indirect effect cannot be visible since the woman and the man of interest have exactly the same values for all non-protected variables. This is why the individual fairness criterion is verified with the unawareness model, even though there may still be indirect discrimination in this model. It is a drawback of the individual fairness criterion.

For the No Protected Effect (NPE) model, the predicted claims frequency for a woman and a man having the same risk profile will be the same. Indeed, in this model, we set the part of the score related to the gender to zero so that the effect of the gender does not intervene. Since all the other variables of the model are the same, the predicted claims frequency is automatically the same for a woman and a man having the same risk profile. The NPE model verifies therefore the individual fairness criterion.

4.4.3 Implication between non-discrimination methods and fairness criteria

To sum up our results about non-discrimination and fairness until now, we make a summary table showing the implications or not between the non-discrimination methods and the fairness criteria. The first non-discrimination method is the unawareness model, i.e. the technique removing the protected variable of the database before the fitting procedure. This allows to eliminate the direct discrimination but not the indirect discrimination. The second non-discrimination method is the No Protected Effect (NPE) model which consists to remove in the model score all the parts related to the protected variable. This allows to eliminate both direct and indirect discrimination.

Table 4.2 shows the implications or not between these two non-discrimination methods and the group or the individual fairness criterion. The table must be read as row i implies or not the column j.

- (1) : The unawareness model does not imply statistical parity (group fairness). One of the reason is that the unawareness model can induce indirect discrimination since it captures information on the protected variable thanks to the non-protected variables due to the correlation between them. Moreover, the non-protected variables are not especially distributed in the same way for each value of the protected variable (women and men). This implies that the prediction, i.e. the annual claims frequency, is not distributed in the same way for women and men and the statistical parity is therefore not satisfied.

	Unawareness model (only ind. discr.)	NPE model (no discr.)	Group fairness (Statistical parity)	Individual fairness
Unawareness model (only ind. discr.)	/	/	X (1)	V (3)
NPE model (no discr.)	/	/	X (2)	V (4)
Group fairness (Statistical parity)	X (5)	X (6)	/	X (7)
Individual fairness	X (8)	X (9)	X (10)	/

Table 4.2: Implications or not between the non-discrimination methods and the fairness criteria.

- (2) : The NPE model does not imply statistical parity. This is due to the fact that the non-protected variables are not distributed in the same way for each value of the protected variable as we have just explained.
- (3)-(4) : The unawareness and the NPE model imply individual fairness as we have seen in Section 4.4.2.
- (5)-(6)-(7) : First of all, we can imagine the hypothetical case of a model which uses the protected variable (the gender) directly to differentiate between insureds and which therefore gives a different prediction for a woman and a man with the same risk profile. However, this difference is compensated by a different distribution of the non-protected variables for women and men, so that when we look at the predictions globally, they are distributed in the same way for women and men. Therefore, the statistical parity (group fairness) is satisfied but it is not a model containing only indirect discrimination (unawareness model) or without discrimination (NPE model) and the individual fairness is not satisfied.

  Another argument is that, to satisfy the statistical parity criterion (group fairness), we have to act on the dependence between the non-protected variables $\mathbf{X}$ and the gender P (thanks to pre-processing) or on the dependence between the prediction $\hat{\mu}$ and P (thanks to post-processing). Indeed, the idea of the pre-processing or post-processing techniques is to adjust the explanatory variables or the model's prediction so that they are distributed in the same way for women and men. This can imply that women and men are treated differently based on their gender. This is a form of direct discrimination since pre-processing or post-processing methods use directly the gender variable P and treat women and men differently based on this variable so that $\mathbf{X} \perp\!\!\!\perp P$ or $\hat{\mu} \perp\!\!\!\perp P$. Therefore, a model satisfying statistical parity does not imply that this model is without discrimination or only with indirect discrimination. Moreover, as we have just said, to satisfy statistical parity, it can be necessary to treat women and men belonging to the same subgroup, i.e. having the same risk profile, differently. Therefore, satisfying statistical parity (group fairness) does not imply individual fairness.

- (8)-(9) : The unawareness model and the NPE model imply both individual fairness. Therefore, a model satisfying individual fairness does not imply that it is necessarily a unawareness model or necessarily a NPE model. It can be one or the other. However, what is certain is that a model satisfying individual fairness has no direct discrimination.

- (10) : A model satisfying individual fairness does not imply that it satisfies statistical parity (group fairness). Indeed, our unawareness and our NPE model satisfy individual fairness but not statistical parity.

4.4.4 Pre-processing

In this section, we analyse the results of our two pre-processing techniques: the correlation remover and the disparate impacte remover (DIR) techniques.

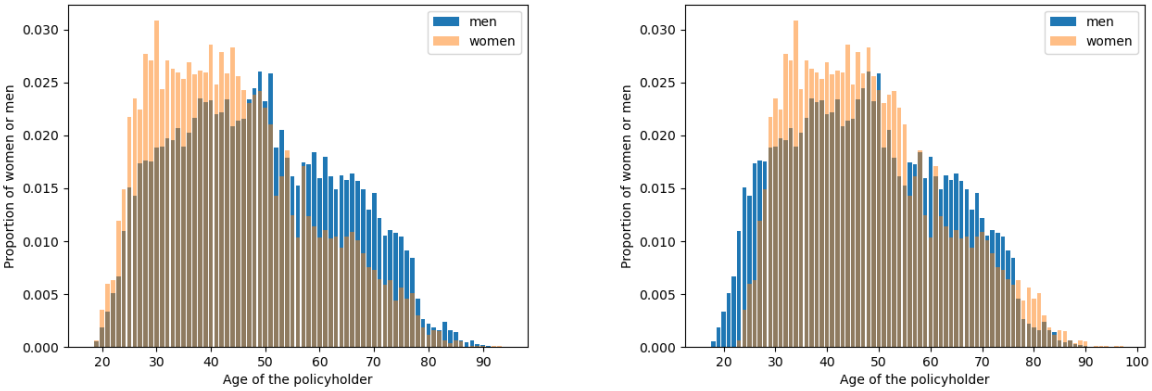
4.4.4.1 Correlation remover

As described in Section 4.2.1, the aim of the correlation remover technique is to remove from the non-protected variables their linear dependence with the protected variable. This technique is only applied to the quantitative variables. In the context of our MTPL application, this means that we aim to remove the linear dependence between the gender and the three quantitative non-protected variables: the age of the policyholder, the age of the car and its power. The correlation remover technique is first trained on the training set of the MTPL database and then applied on the testing set.

Figures 4.3a and 4.3b show the distribution of the age of the policyholder on the testing set before and after the correlation remover technique respectively. Before pre-processing, there is in proportion more women than men at younger ages and less at older ages. Then, the pre-processing technique allows to shift upwards the age of women and downwards the age of men so that the mean of the age is as close as possible for women and men. This reduces the tendency to have at younger ages more women than men in proportion and at older ages more men than women. The results are shown here for the age of the policyholder because it is the variable for which the results of the correlation remover technique are the clearest but the technique applied in the same way for the two other variables (age of the car and its power).

Once we have obtained our MTPL pre-processed database for the training and the testing set, the aim is to predict the annual claims frequency. For that, we use an unawareness EBM model, i.e. an EBM model where we remove the gender variable before the fitting procedure. The unawareness EBM model is first trained on the pre-processed training set and then the annual claims frequency can be predicted based on the pre-processed testing set.

As explained in Section 4.2, the aim of the pre-processing methods and in this case of the correlation remover technique is to improve the statistical parity criterion (4.1). We therefore compare our unawareness model with and without pre-processing based on the maximum absolute statistical parity (MASP) criterion (4.2) on the testing set as shown in Table 4.3. We also compare these two models to the best estimate model, i.e. the model where nothing is done to improve the fairness or the non-discrimination of the models.



(a) Before the correlation remover

(b) After the correlation remover

Figure 4.3: Distribution of the policyholder age before and after the correlation remover technique on the testing set.

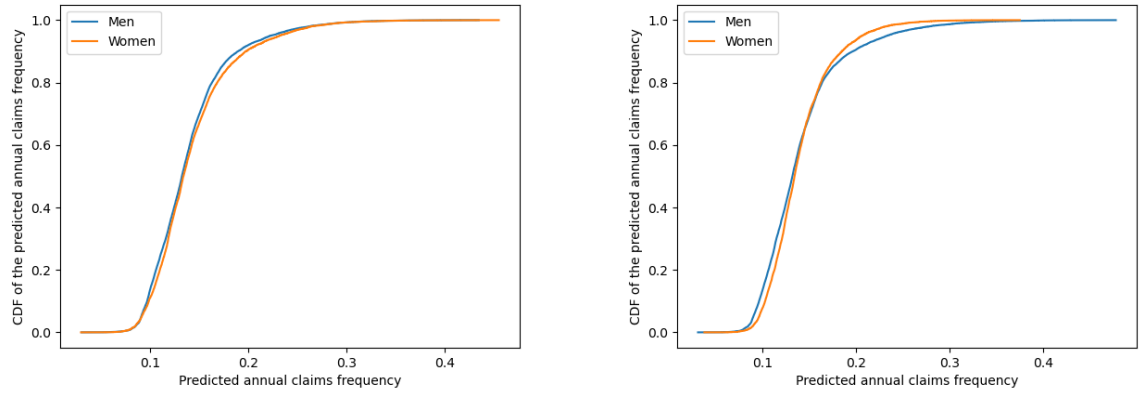
	MASP
best estimate EBM	0.1681
unawareness EBM	0.0416
unawareness EBM with correlation remover	0.0749

Table 4.3: Comparison of the MASP criterion for the best estimate model and the unawareness model with and without correlation remover.

First of all, as already observed, the unawareness models allows to improve the MASP criterion and therefore the group fairness compared to the best estimate model. However, we observe that the MASP increases with the correlation remover technique. The unawareness model is therefore less fair in the sense of the MASP criterion when we use this pre-processing technique. We can, however, look at the distribution of the annual predicted claims frequency in the women and men groups before and after the pre-processing technique (Figures 4.4a and 4.4b). Before the pre-processing technique, the CDF of the annual predicted claims frequency for the women is always below the CDF for the men, so women tend to have higher claims frequencies. After the pre-processing technique, there's a moment when the women's CDF rises above the men's CDF. But then we see that at certain times the two CDFs are further apart than without pre-processing, which explains why we have a higher MASP whit pre-processing.

The fact that the group fairness is not improved with the correlation remover technique may be due to several reasons. Firstly, this pre-processing is only applied to quantitative variables and only removes the correlation, i.e. the linear dependence. Furthermore, the highest correlation between an explanatory variable (`Power`) and the gender is only -0.1597. As a consequence, the MASP for the unawareness model without the pre-processing is therefore equal to 0.0416. This implies that the capacity of the MASP to decrease is limited in our case since the MASP is already low without pre-processing. Indeed, in the literature, a MASP criterion below 0.1 is considered as acceptable [19]. In the next section, we will therefore simulate a database so that the correlation between a certain explanatory variable and the gender is higher, and so that the MASP before the

pre-processing is higher.



(a) Without the correlation remover technique (b) With the correlation remover technique

Figure 4.4: CDF of the predicted annual claims frequency for the unawareness model with and without the correlation remover technique.

Another fairness criterion at which we can look is the expected prediction ratio (EPR)

$$\frac{\mathbb{E}[\hat{\mu}|P=a]}{\mathbb{E}[\hat{\mu}|P=b]}$$

The model is fair in the sense of this criterion if the ratio is equal to 1.

Table 4.4 gives the expected prediction ratio (EPR) for the same three models as before: best estimate model and unawareness model with and without correlation remover. The EPR is computed on the testing set of the MTPL database. We observe that the EPR is closer to 1 for the unawareness model with pre-processing compared to the case of the unawareness model without pre-processing. This can be explained by the fact that the correlation remover technique aims to translate the quantitative non-protected variables so that for each of these variables, their mean is as close as possible for women and men. As a consequence, with pre-processing, the mean of the annual predicted claims frequency in the men's group is more closer to the one of the women's group than without pre-processing. Therefore, with the EPR criteria, the unawareness model is fairer with pre-processing than without.

	EPR
best estimate EBM	0.9213
unawareness EBM	0.9760
unawareness EBM with correlation remover	1.0077

Table 4.4: Comparison of the ERP criterion for the best estimate model and the unawareness model with and without correlation remover.

Simulated dataset

The idea of this section is to see whether the correlation remover technique will reduce the MASP criterion in the case where there is a higher correlation between a certain

explanatory variable and the gender variable. Indeed, in the case of our MTPL database, one of the reason why the MASP does not decrease with this pre-processing technique could be that the explanatory variables are too poorly correlated with the gender variable. Therefore, we decide to increase the correlation between two variables of our MTPL database: the age of the policyholder and its gender. We therefore simulate the gender of the policyholder and correlate it with age in the following way. If the policyholder's age is below the median age in the MTPL database, then its gender is simulated as follows:

$$Gender = Ber(\epsilon_1)$$

where Ber corresponds to a Bernoulli distribution, Gender = 1 corresponds to 'man' and Gender = 0 corresponds to 'woman'. If the policyholder's age is above the median age in the database, then

$$Gender = Ber(\epsilon_2)$$

ϵ_1 and ϵ_2 are chosen so that $\frac{\epsilon_1+\epsilon_2}{2} = 0.74$. This keeps the database at around 74% men, as in the initial MTPL database. We look at different scenarios by changing $\epsilon_{1,2}$. The idea is to take ϵ_1 lower than ϵ_2 so that there is in proportion more women than men at younger age and less at older ages, as in the initial database. Moreover, the further apart the values of ϵ_1 and ϵ_2 are, the higher the correlation between the age of the policyholder and its gender will be. Table 4.5 shows the correlation between the age and the gender variable, the MASP criterion and the EPR criterion in the case of the unawareness model without the correlation remover technique. Table 4.6 shows the same elements in the case of the unawareness model with the correlation remover technique.

	$\epsilon_{1,2} = 0.53/0.95$	$\epsilon_{1,2} = 0.63/0.85$	$\epsilon_{1,2} = 0.73/0.75$
Correlation Age-Gender	-0.3977	-0.2114	-0.0202
MASP	0.3172	0.1641	0.0172
EPR	0.8376	0.9111	0.9940

Table 4.5: Correlation Age-Gender, MASP and ERP criterion for the unawareness model without correlation remover.

	$\epsilon_{1,2} = 0.53/0.95$	$\epsilon_{1,2} = 0.63/0.85$	$\epsilon_{1,2} = 0.73/0.75$
Correlation Age-Gender	1.51 e-14	-2.17 e-14	-1.33 e-15
MASP	0.1926	0.1114	0.0239
EPR	0.9566	0.9913	0.9944

Table 4.6: Correlation Age-Gender, MASP and ERP criterion for the unawareness model with correlation remover.

We observe that it is when the correlation between Age and Gender is too small in the database before pre-processing (last column of Tables 4.5 and 4.6) that the correlation remover fails to reduce the MASP since the MASP is already small even without pre-processing. Moreover, the higher the correlation, the easier it is to reduce the MASP criterion with the correlation remover technique. Therefore, the correlation remover technique allows to obtain a fairer model when the correlation between the protected and the non-protected variables is high enough, but not when it is too low.

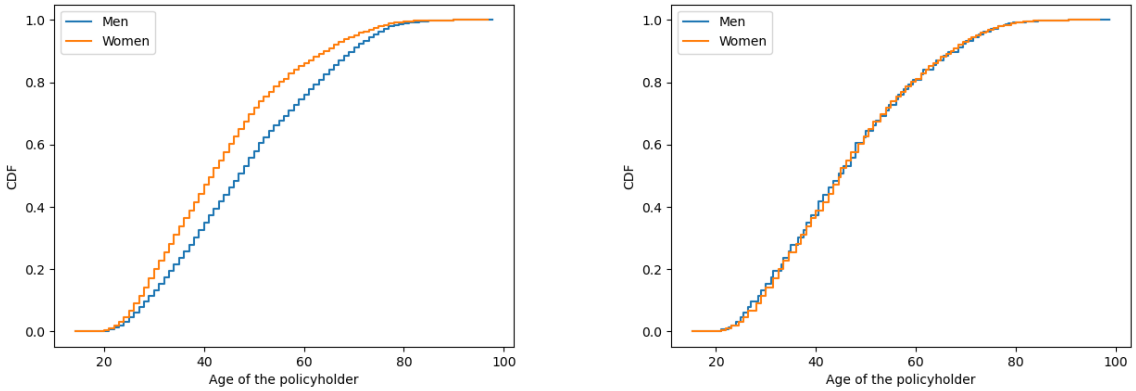
Finally, if we look at the expected prediction ratio (EPR), the EPR is always closer to 1 after the correlation remover even when the correlation between Age and Gender is small. This can be explained by the fact that the correlation remover technique aims to translate the quantitative non-protected variables so that for each of these variables, their mean is as close as possible for women and men. As a consequence, the average of the predicted claims frequency for men and women can be closer after pre-processing than before pre-processing. This can explain why, according to the ERP criterion, the model is always fairer after the correlation remover technique even in case of small correlations.

4.4.4.2 Disparate impact remover

The aim of the disparate impact remover (DIR) technique is to adjust the values of the quantitative non-protected variables such that they have the same distribution in each group defined by the values of the protected variable P. In the case of our MTPL application, the three quantitative non-protected variables (`AgePH`, `AgeC` and `Power`) are adjusted so that they are distributed in the same way for women and men. As a consequence, this should allow to have an unawareness model giving predictions distributed in a more similar way for women and men and therefore a fairer model in the sense of the group fairness criterion. Moreover, in this master thesis, two alternatives of the DIR technique are proposed to reduce the deviance of the model. In this section, we therefore verify whether the three DIR techniques allow to obtain a fairer model and whether the two alternatives allow to decrease the deviance in the case of our MTPL application.

The DIR technique is implemented thanks to the function `disparate_impact_remover` from the R package `fairmodels` [20]. This function is modified to take into account the two DIR alternative techniques. Regardless of the DIR technique used, the method is first trained on the training set and then applied on the testing set. Figures 4.5a and 4.5b show the CDF of the Age of the policyholder (`AgePH`) on the testing set before and after the classical DIR technique respectively. The CDFs of `AgePH` for women and men are very close after the classical DIR technique. The aim of the DIR method has therefore well been achieved. The conclusion is the same for the two alternative DIR techniques (weighted DIR and conditional weighted DIR) no matter the variable (`AgePH`, `AgeC` and `Power`). These results can be found in Appendix C.

We can also compare the policyholder's age before and after pre-processing for each DIR method: the classical DIR, the weighted DIR and the conditional weighted DIR. This is done thanks to Figure 4.6. First of all, we note that, for each method, the age of women increases with pre-processing and the age of men decreases. It was expected since before pre-processing, women tend to be younger than men and we want the age to be distributed in the same way after pre-processing. Moreover, with the two alternative DIR techniques (weighted DIR and conditional weighted DIR), we observe that the women age changes more than in the case of the classical DIR and the men age changes less as we see in the zoom of Figure 4.6. This can be explained in the following way. In Section 4.2.2, we have seen that the aim of the two alternative DIR methods is to obtain an age after pre-processing that retains more information about the age of men before pre-processing than the age of women before pre-processing because there are more men than women in the MTPL database.



(a) Before the classical DIR technique (b) After the classical DIR technique

Figure 4.5: CDF of `AgePH` before and after the classical DIR technique.

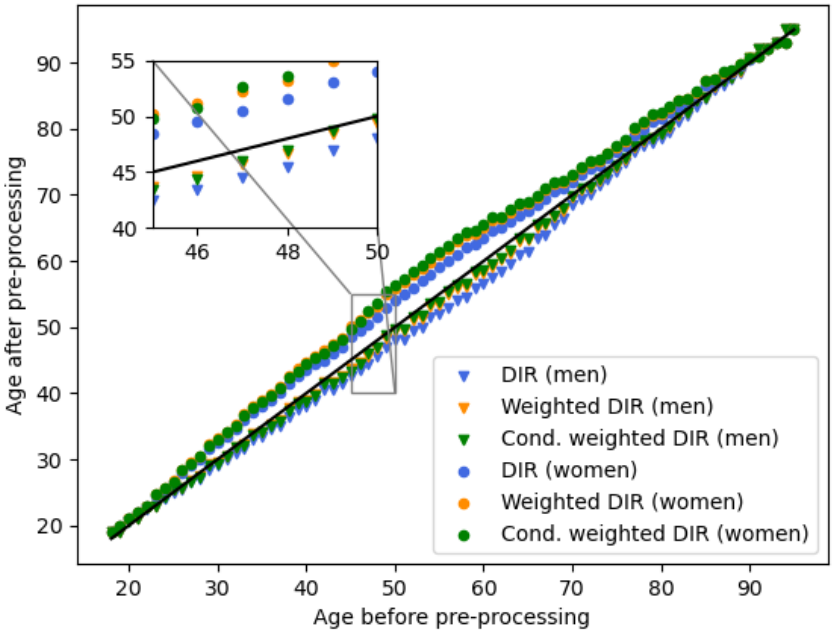


Figure 4.6: Comparison of the policyholder's age before and after pre-processing for each DIR method: classical DIR, weighted DIR and conditional weighted DIR.

We now compare the Poisson deviance on the training and the testing set obtained with the unawareness model depending on the DIR technique used (classical DIR, weighted DIR and conditional weighted DIR). They are also compared to the deviance obtained for the best estimate model and the unawareness model without pre-processing. The results are first analyzed for the training set. However, all the observations made for the training set are applicable to the testing set, except one. As shown in Table 4.7, the model with the lowest Poisson deviance is the best estimate model, which is coherent since this model directly uses the policyholder's gender to predict the claims frequency, whereas the other models do not. Moreover, on the training set, the unawareness model without pre-processing has a lower deviance than the ones where a DIR technique is applied. This

can be explained by the fact that without pre-processing, the database is not modified.

Then, the Poisson deviance for the three DIR techniques are compared. First, the Poisson deviance is exactly the same for the classical DIR and the weighted DIR technique. This can be explained by the fact that the only difference between the classical DIR and the weighted DIR is that the linear combination of the quantile functions $F^{-1}_{X|P=w}$ and $F^{-1}_{X|P=m}$ is not the same as shown in Section 4.2.2. Indeed, in the case of the classical DIR, we have

$$x_{new} = \frac{1}{2}\ F^{-1}_{X|P=w}(u) + \frac{1}{2}\ F^{-1}_{X|P=m}(u)$$

and in the case of the weighted DIR, we have

$$x_{new} = 0.26\ F^{-1}_{X|P=w}(u) + 0.74\ F^{-1}_{X|P=m}(u)$$

since the probability to be a man in our MTPL database is equal to 0.74. A lot of different combinations of the quantile functions have been tested and the prediction of the model is invariant to the linear combination used provided that

$$x_{new} = a\ F^{-1}_{X|P=\text{"}m\text{"}}(u) + (1-a)\ F^{-1}_{X|P=\text{"}w\text{"}}(u)$$

with $0 < a < 1$. Since the prediction of the model is invariant to the combination used, the Poisson deviance is exactly the same.

Finally, we observe that the Poisson deviance is lower with the conditional weighted DIR than with the two other DIR techniques. This is because the conditional weighted DIR better takes into account that the probability to be a woman or a man in the DIR procedure. Indeed, as explained in Section 4.2.2.2, the value of this factor a will in this case depend whether the value of the variable to be pre-processed is, before pre-processing, below or above the median of this variable. The database after pre-processing is therefore more aligned with the initial database and as a consequence the Poisson deviance is lower.

All the observations made for the training set are valid for the testing set. The only difference is that the Poisson deviance is lower for the unawareness model with the DIR techniques than without. However, it seems that the result is more a matter of chance.

	Poisson deviance train	Poisson deviance test
Best estimate EBM	70335.59	17687.33
Unawareness EBM	70415.77	17704.94
Unawareness EBM and classical DIR	70426.95	17700.36
Unawareness EBM and weighted DIR	70426.95	17700.36
Unawareness EBM and conditional weighted DIR	70419.57	17699.14

Table 4.7: Comparison of the Poisson deviance for the best estimate model, the unawareness model and the unawareness model with the different DIR techniques.

Since the aim of the DIR techniques is to improve the group fairness of the predictive models, we now compare in Table 4.8 the MASP criterion for the same models as before. The MASP criterion is lower on the training and the testing set for the unawareness model

with the DIR techniques compared to the unawareness model without pre-processing and the best estimate model. This is true no matter the DIR techniques used. These latter allow therefore to improve the group fairness of the models. Moreover, the MASP criterion is extremely closed for the three DIR techniques, especially on the training set. In conclusion, the DIR techniques allow to improve the group fairness of the models and the conditional weighted DIR allow to decrease the Poisson deviance compared to the two other techniques while maintaining the group fairness.

	MASP train	MASP test
Best estimate EBM	0.1724	0.1681
Unawareness EBM	0.0416	0.0416
Unawareness EBM and classical DIR	0.0172	0.0244
Unawareness EBM and weighted DIR	0.0172	0.0244
Unawareness EBM and conditional weighted DIR	0.0171	0.0181

Table 4.8: Comparison of the MASP criterion for the best estimate model, the unawareness model and the unawareness model with the different DIR techniques.

4.4.5 Post-processing

As described in Section 4.3, the aim of the post-processing method is to post-process a predictive model so that the prediction of this latter is distributed in the same way in each group defined by the values of the protected variable. In the context of our MTPL application, the goal is to have the predicted annual claims frequency distributed in the same way for women and men. This will allow to improve the statistical parity criterion, i.e. the group fairness of the model.

The post-processing technique is implemented thanks to the Python library `holisticai` [21]. To apply the post-processing technique, the annual claims frequency is first predicted with the best estimate EBM model on the whole MTPL database (after the best estimate model has been trained on the training set). The best estimate model is chosen because it is the simplest model, i.e. the model on which no non-discrimination method or fairness technique has already been applied, and for which the post-processing will allow to improve the statistical parity of the model.

The post-processing technique is then trained on the predictions of the training set and finally applied to the testing set. We can now compare the group fairness of the best estimate model with and without post-processing thanks to the MASP criterion computed on the testing set. As shown in Table 4.9, the MASP really decreases thanks to the post-processing. The best estimate model is therefore more fairer in the sense of the statistical parity after the post-processing as expected. However, there is a trade-off between the fairness and the accuracy of the model. Indeed, the best estimate model without post-processing has a lower Poisson deviance than the one with post-processing.

	MASP	Poisson deviance
Best estimate EBM	0.1681	17687.33
Best estimate EBM and post-processing	0.0222	17704.16

Table 4.9: Comparison on the testing set of the MASP criterion and the Poisson deviance for the best estimate model with and without post-processing.

4.5 Conclusion

This chapter has first looked at the implications between the non-discrimination methods and the different fairness criteria. This enabled us to show that the two non-discrimination methods, the unawareness model and the NPE model, imply the individual fairness but not the group fairness.

However, pre-processing and post-processing can be implemented to improve the group fairness of the predictive models. In this master thesis, two pre-processing techniques (the correlation remover technique and the disparate impact remover technique) and one post-processing technique have been studied.

In particular, for the pre-processing, we have shown that the disparate impact remover (DIR) technique reduces the MASP criterion, which means that it improves the group fairness of the models. The idea of the pre-processing technique is to adjust the explanatory variables of the model so that they are distributed in the same way for all the groups defined by the protected variable (women's and men's groups if the protected variable is the gender) in order to obtain model predictions distributed in the same way for women and men. Moreover, in this master thesis, an alternative to this DIR method has been proposed. This alternative has allowed to improve the deviance of the predictive model while maintaining its group fairness.

Finally, the post-processing technique studied is based on a similar principle to the DIR technique but is applied on the prediction of the model and not on the explanatory variables. This post-processing technique also allowed to reduce the MASP criterion and therefore to improve the group fairness of the model. Comparing the pre-processing (DIR technique) and the post-processing technique, both have allowed to decrease the MASP criterion around 0.02. The two techniques are therefore comparable in terms of improvement of the group fairness.

Chapter 5

Conclusion

This master thesis has studied different ethical constraints that a predictive model designed for non-life insurance pricing may have to satisfy: interpretability, non-discrimination and fairness. Regarding the model interpretability, the explainable boosting machine (EBM) model was presented. This model is a special case of a GAM and has therefore an explicit score. The EBM model has however the particularity to include pairwise interactions that are automatically detected. Moreover, the univariate and the interaction terms are estimated using bagging and boosting techniques applied on one feature at a time in round-robin fashion. These particularities enable the EBM model to have similar prediction performances compared to the RF and GBM models. It is indeed what this master thesis has shown thanks to the MTPL application which consists of predicting the annual claims frequency of the policyholders based on a MTPL database. As well as having similar performance to models such as RF and GBM, the EBM model is easier to interpret since it has an explicit score. Thanks to the latter, the effect of each explanatory variable on the model's prediction is directly observable, which is a major advantage of the EBM model.

The aim of this master thesis was also to answer to the following question: Is it possible to obtain a model that is simultaneously interpretable, non-discriminatory and fair ? The interpretability of the models has been addressed above. For the non-discrimination, we have seen that several non-discrimination methods exist. However, not all methods remove the direct and the indirect discrimination. Indeed, the unawareness model, which consist of removing the protected variables before the training procedure, eliminates the direct but not the indirect discrimination. This is due to the fact that the unawareness model can still infer information about the protected variables from the non-protected variables since protected and non-protected variables are correlated.

In this master thesis, a non-discrimination method is proposed to eliminate both the direct and the indirect discrimination. This method takes advantage of the fact that the EBM model has an explicit score. Indeed, this method, called No Protected Effect (NPE) model, sets the part of the EBM score model related to the protected variable to zero. Since the protected variable is initially in the model, the effect related to this variable is captured in the part of the score related to the protected variable. Therefore, the non-protected variables does not capture information about the protected variable and there is no indirect discrimination. Moreover, by removing the part of the score related to the protected variable, the direct discrimination is eliminated. In conclusion, concerning the

non-discrimination, it is better to use the NPE model than the unawareness model to eliminate both the direct and the indirect discrimination.

Finally, for the fairness of the model, we have seen that the non-discrimination method which consists of using the NPE model imply that the individual fairness of the model is satisfied. Indeed, since we remove the part of the score related to the protected variable, the model's prediction for a woman and a man having the same risk profile is the same. In the case of our MTPL application, it means that the predicted annual claims frequency is the same.

We therefore have all the elements to answer the question of whether it is possible to obtain a model that is interpretable, non-discriminatory and fair. It is possible if we consider the individual fairness criterion. Indeed, in this case, we can take an EBM model with the technique of the NPE model as non-discrimination method to obtain a model that is interpretable, non-discriminatory and fair according to the individual fairness.

However, we cannot answer in the same way with the group fairness criterion. Indeed, we have seen in this master thesis that the non-discrimination methods do not imply group fairness. Pre-processing and post-processing techniques can then be used to improve the group fairness criterion. However, it will have an impact on the non-discrimination of the models. Indeed, the idea of the pre-processing or post-processing techniques is to adjust the explanatory variables or the model's prediction so that they are distributed in the same way in all groups. For example, a pre-processing method can be used to adjust the age of the policyholders so that it is distributed in the same way for women and men. If initially, women are in general younger than men, the pre-processing method will increase the age of women and decrease the age of men. Therefore, the pre-processing method uses directly the gender of the policyholder, i.e. the protected variable, to treat in a different way women and men. This is therefore a form of direct discrimination. As a consequence, it shows that it is much more difficult to obtain a model that is interpretable, non-discriminatory and fair according to the group fairness.

This master thesis has addressed several questions regarding ethical constraints in non-life insurance but further work can be made. First, in this work, we have examined a MTPL application where one protected variable, the gender, was considered. It might however be interesting to look at scenarios involving multiple protected variables simultaneously and see whether the non-discrimination methods and the fairness criteria can be adapted to take this into account. Moreover, the gender variable is a binary protected variable but the protected variable could also be a categorical variable with more than two levels or a continuous variable. Finally, our MTPL application focuses on the prediction of the annual claims frequency but it could be interesting to look at the three ethical constraints of this master thesis for the claims severity and ultimately for the pure premium using the frequency/severity approach.

Bibliography

[1] Mathias Lindholm et al. "Discrimination-free insurance pricing". In: *ASTIN Bulletin: The Journal of the IAA* 52.1 (2022), pp. 55–89.

[2] Mathias Lindholm et al. "What is fair? Proxy discrimination vs. demographic disparities in insurance pricing". In: *Proxy Discrimination vs. Demographic Disparities in Insurance Pricing (May 2, 2023)* (2023).

[3] Arthur Charpentier. *Insurance, Biases, Discrimination and Fairness*. Springer, 2023. Chap. Individual Fairness.

[4] European Union. "Council Directive 2004/113/EC of 13 December 2004 implementing the principle of equal treatment between men and women in the access to and supply of goods and services". In: *Official Journal of the European Union* 47 (2004), pp. 37–43.

[5] European Union. "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)". In: *Official Journal of the European Union* 55 (2012), pp. 1–11.

[6] Xi Xin and Fei Huang. "Antidiscrimination insurance pricing: Regulations, fairness criteria, and models". In: *North American Actuarial Journal* (2023), pp. 1–35.

[7] Michael Feldman et al. "Certifying and removing disparate impact". In: *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 2015, pp. 259–268.

[8] Harsha Nori et al. "Interpretml: A unified framework for machine learning interpretability". In: *arXiv preprint arXiv:1909.09223* (2019).

[9] Yin Lou et al. "Accurate intelligible models with pairwise interactions". In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2013, pp. 623–631.

[10] Katrien Antonio. *Hands-on Machine Learning R Module 1*. https://github.com/katrienantonio/hands-on-machine-learning-R-module-1.

[11] Denuit Michel. *LACTU2110 : Modélisation prédictive et apprentissage statistique en assurance*. Slides. UCLouvain, École de Statistique, Biostatistique et Sciences Actuarielles (LSBA). 2022-2023.

[12] M Denuit, D Hainaut, and J Trufin. *Effective statistical learning methods for actuaries - tree-based methods*. Springer Actuarial Series, 2019.

[13] Katrien Antonio. *Bias, fairness and discrimination-free insurance pricing*. Slides. LRisk - KU Leuven and ASE - University of Amsterdam. 2022.

[14] Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. "Fair regression: Quantitative definitions and reduction-based algorithms". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 120–129.

[15] Moritz Hardt, Eric Price, and Nati Srebro. "Equality of opportunity in supervised learning". In: *Advances in neural information processing systems* 29 (2016).

[16] Alessandro Castelnovo et al. "A clarification of the nuances in the fairness metrics landscape". In: *Scientific Reports* 12.1 (2022), p. 4209.

[17] Fairlearn Development Team. *Fairlearn API Reference: CorrelationRemover*. https://fairlearn.org/main/api_reference/generated/fairlearn.preprocessing.CorrelationRemover.html.

[18] Evgenii Chzhen et al. "Fair regression with wasserstein barycenters". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 7321–7331.

[19] Alexander Stevens et al. "Explainability and fairness in machine learning: Improve fair end-to-end lending for kiva". In: *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE. 2020, pp. 1241–1248.

[20] Jakub Wiśniewski and Przemysław Biecek. "fairmodels: A flexible tool for bias detection, visualization, and mitigation". In: *arXiv preprint arXiv:2104.00507* (2021).

[21] Holistic AI Development Team. *Holistic AI Library*. https://holisticai.readthedocs.io/en/latest/.

Appendix A

Analysis of the MTPL database

In this appendix, the MTPL database used in this master thesis is analyzed. We first describe the categorical variables and then the continuous one. Figure A.1 shows that more people drive a gasoline car than a diesel car. Moreover, diesel cars tend to cause more claims. This can be explained by the fact that diesel cars tend to cover more kilometres. We observe on Figure A.2 that there are more men than women in the database, and women tend to have a higher claims frequency. Concerning the type of cover policyholders have, Figure A.3 shows that the majority have a TPL cover. The claims frequency seems to be quite independent of the cover taken but policyholders with TPL cover have a slightly higher claims frequency. For the use of the car, the vast majority of the policyholders have a private car and not a work car. Moreover, Figure A.4 shows that the claims frequency is independent of the car use. Finally, we observe on Figure A.5 that the vast majority of policyholders' cars do not belong to a fleet and that the cars belonging to a fleet tend to have a lower claims frequency.

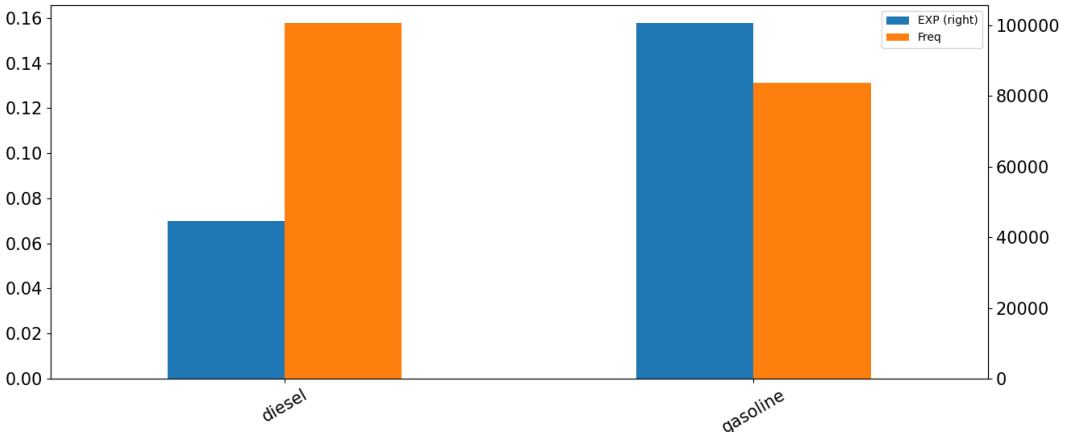


Figure A.1: Analysis of the risk exposure and the annual observed claims frequency for the fuel of the car (`Fuel`).

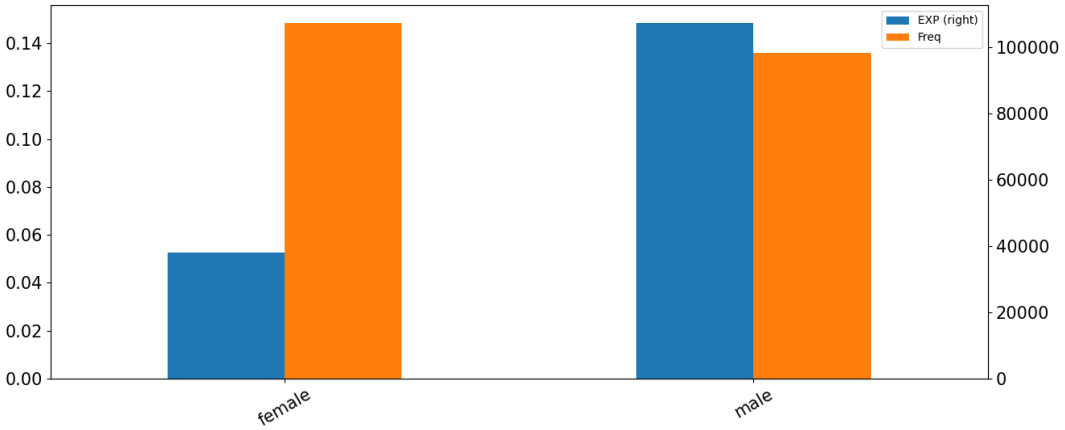


Figure A.2: Analysis of the risk exposure and the annual observed claims frequency for the gender of the policyholder (`Sex`).

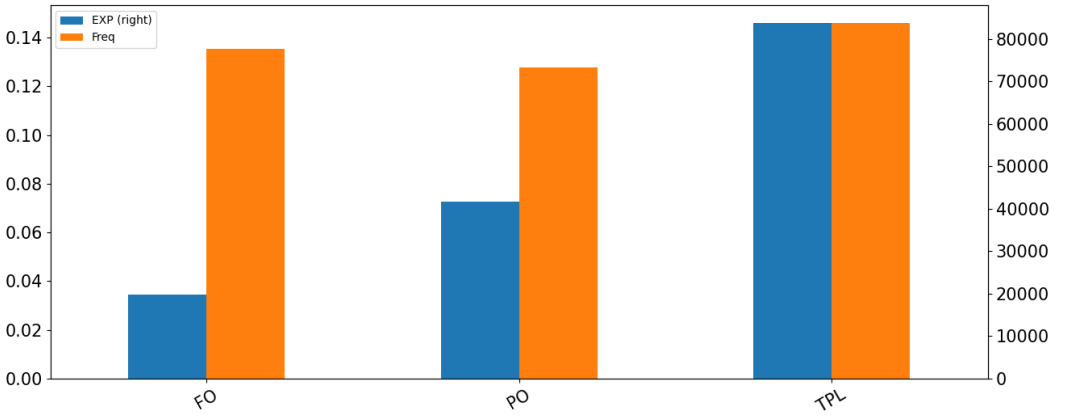


Figure A.3: Analysis of the risk exposure and the annual observed claims frequency for the coverage of the policyholder (`Coverage`).

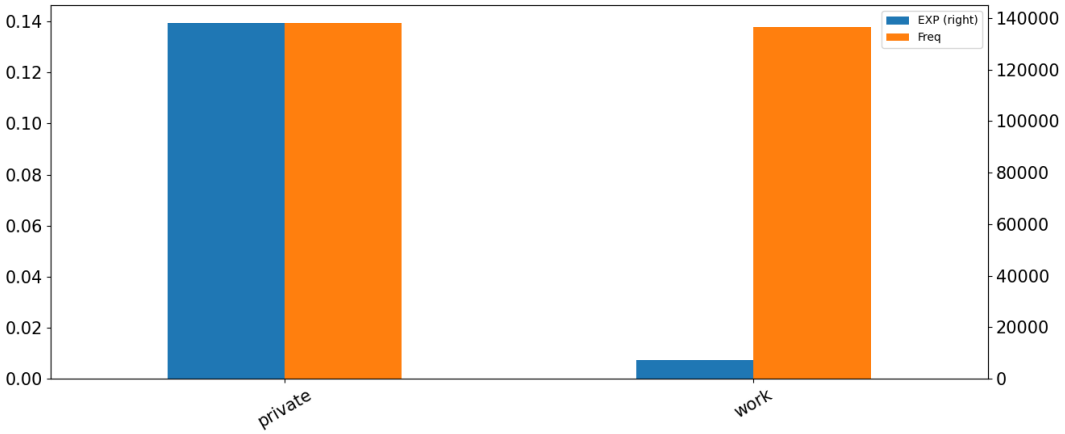


Figure A.4: Analysis of the risk exposure and the annual observed claims frequency for the use of the car (`Use`).

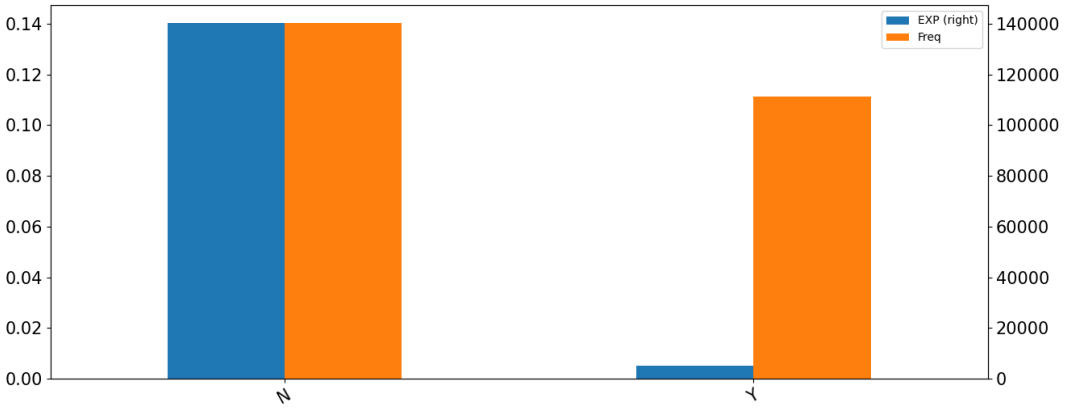


Figure A.5: Analysis of the risk exposure and the annual observed claims frequency depending on whether or not the car belongs to a fleet (`Fleet`).

For the policyholder's age, Figure A.6 shows that the number of insured increases with age until around the age of 50, before decreasing. Moreover, the claims frequency decreases with age before re-increasing at older ages. Concerning the car's age, the vast majority of cars are between 0 and 15 years old as shown on Figure A.7. There is no clear relationship between the age of the car and the claims frequency. Finally, we observe on Figure A.8 that the vast majority of cars have a power between 25 and 85 kilowatts. Moreover, the claims frequency increases with the power of the car.

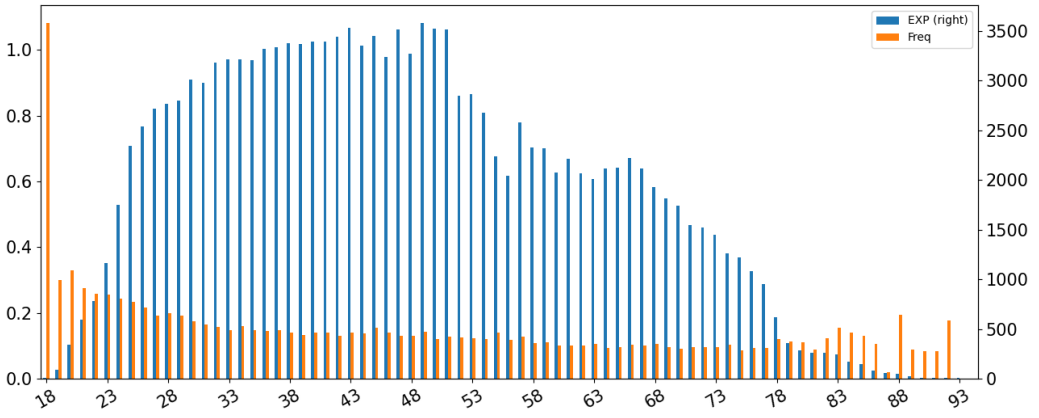


Figure A.6: Analysis of the risk exposure and the annual observed claims frequency for the age of the policyholder (`AgePH`).

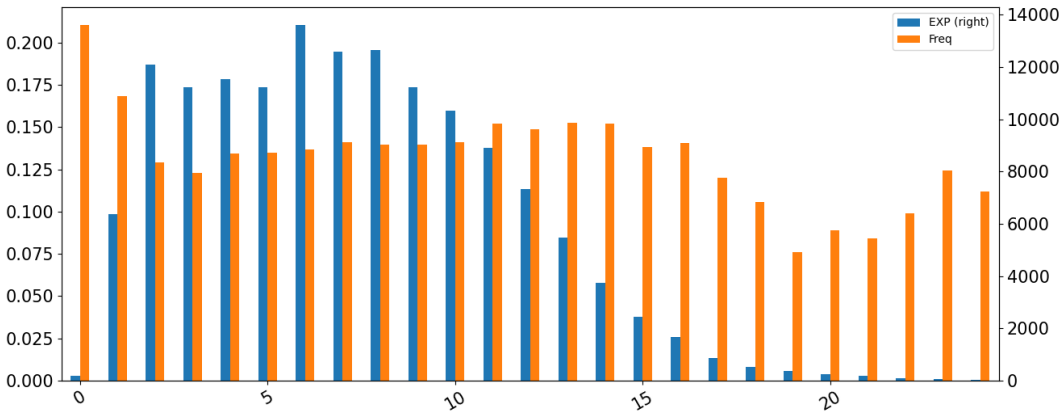


Figure A.7: Analysis of the risk exposure and the annual observed claims frequency for the age of the car (`AgeC`).

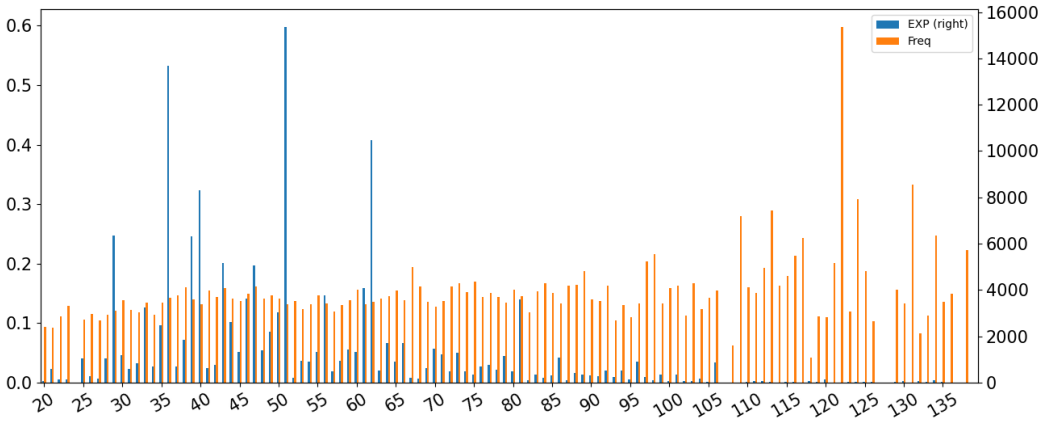


Figure A.8: Analysis of the risk exposure and the annual observed claims frequency for the power of the car (`Power`).

Appendix B

The EBM model score for our MTPL application

B.1 Score of the categorical variables

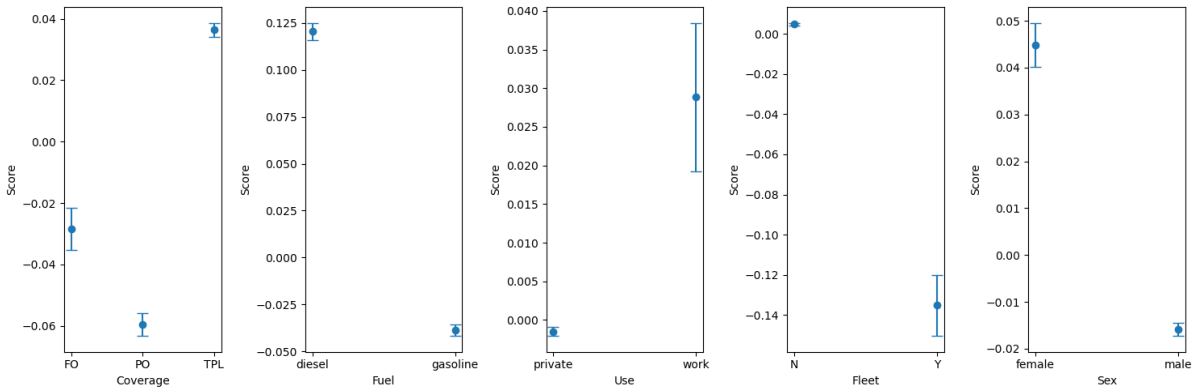


Figure B.1: Score of the categorical variables in the EBM model.

B.2 Interactions in the EBM model

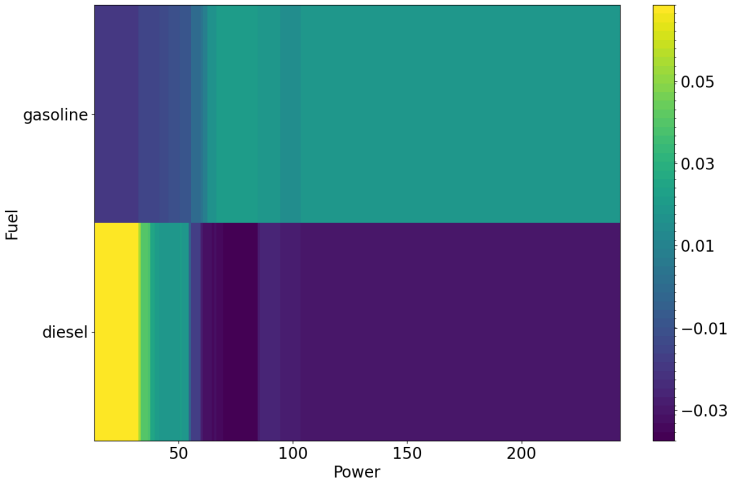


Figure B.2: Interactions between the fuel of the car (`Fuel`) and its power (`Power`) in the EBM model.

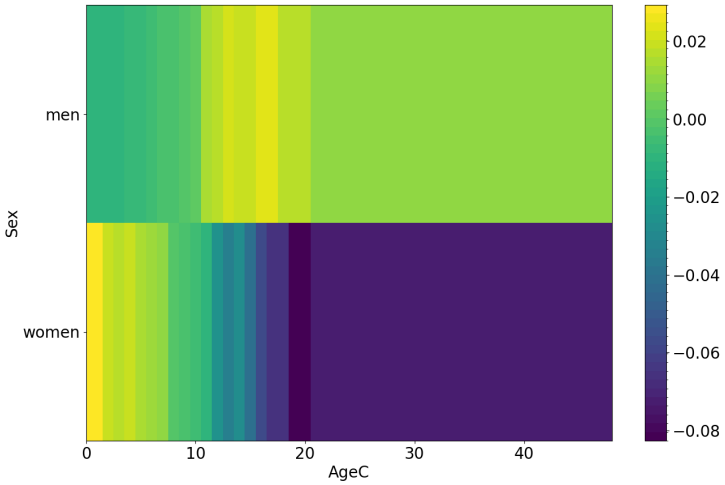
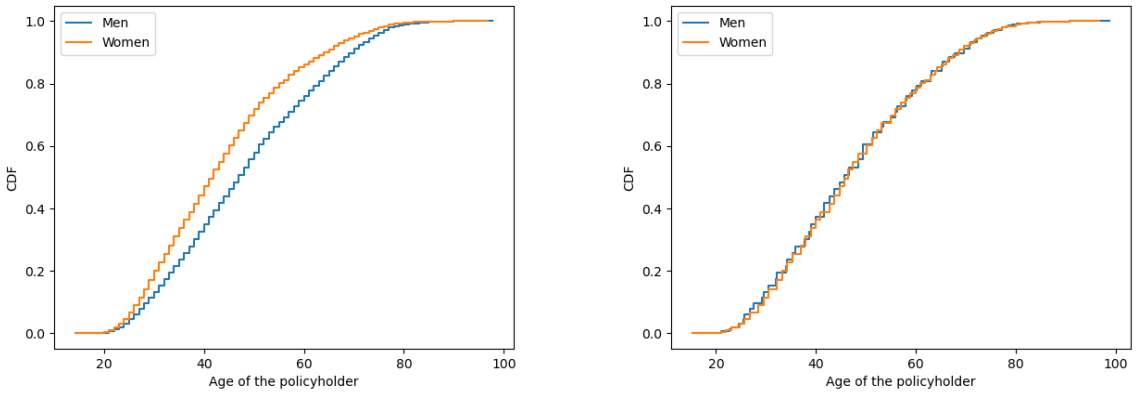


Figure B.3: Interactions between the gender of the policyholder (`Sex`) and the age of the car (`AgeC`) in the EBM model.

Appendix C

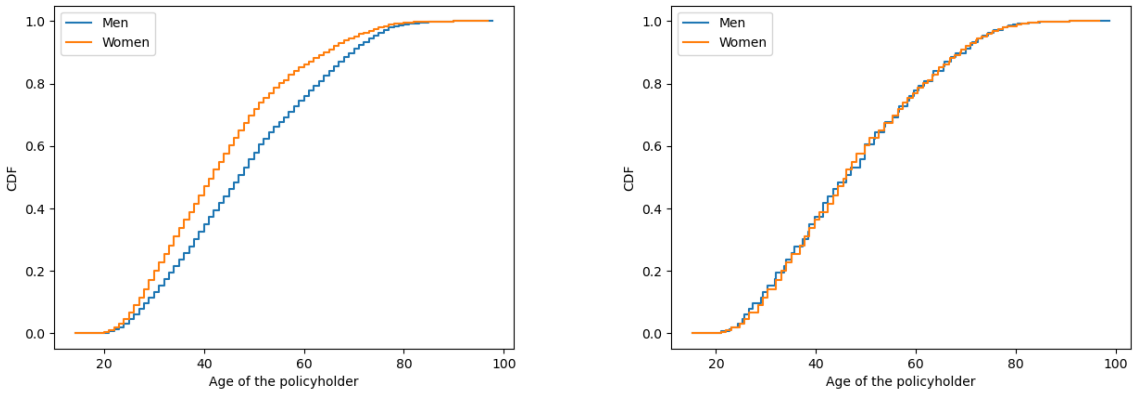
Results for the weighted DIR and the conditional weighted DIR

C.1 Distribution of the policyholder's age before and after pre-processing



(a) Before the weighted DIR technique (b) After the weighted DIR technique

Figure C.1: CDF of `AgePH` before and after the weighted DIR technique.



(a) Before the conditional weighted DIR technique

(b) After the conditional weighted DIR technique

Figure C.2: CDF of `AgePH` before and after the conditional weighted DIR technique.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Faculté des sciences
Place des Sciences, 2 bte L6.06.01, 1348 Louvain-la-Neuve, Belgique | www.uclouvain.be/sc