Mémoire présenté devant l'ENSAE Paris
pour l'obtention du diplôme de la filière Actuariat
et l'admission à l'Institut des Actuaires
le 08/03/2023

Par : **Marguerite Saucé**

Titre : **AI and ethics in insurance: a new solution to mitigate proxy discrimination in risk modeling**

Confidentialité : ☒ NON ☐ OUI (Durée : ☐ 1 an ☐ 2 ans)

Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Membres présents du jury de la filière

Nom : Caroline Hillairet

Membres présents du jury de l'Institut des Actuaires

Entreprise : SCOR

Signature :

Directeur du mémoire en entreprise :

Nom : Antoine Chancel
Signature :

Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels *(après expiration de l'éventuel délai de confidentialité)*

Signature du responsable entreprise

Secrétariat:

Signature du candidat

Bibliothèque:

Ecole Nationale de la Statistique et de l'Administration Economique (ENSAE)
5, avenue Henry Le Chatelier - 91120 PALAISEAU, FRANCE

Résumé

Mots-clés : équité, discrimination indirecte, parité statistique, décorrélation, *Machine Learning*

L'essor de l'intelligence artificielle connaît un intérêt croissant de la part du grand public et, au cours des dernières années, de nombreux articles de presse ont remis en cause son objectivité en accusant des algorithmes d'être racistes, sexistes, etc. Incitée par l'attention croissante des régulateurs sur l'utilisation éthique des données en assurance, la communauté actuarielle se doit de repenser les pratiques de tarification et de sélection du risque pour une assurance plus équitable. L'équité est un thème philosophique qui possède différentes définitions dans chaque juridiction, liées les unes aux autres, sans atteindre de consensus à ce jour. En Europe, la Charte des Droits Fondamentaux définit les lignes directrices sur la discrimination, et l'utilisation des données personnelles à caractère sensible dans les algorithmes est réglementée. Si le simple retrait des variable protégées empêche toute discrimination dite « directe », les modèles sont toujours capables de discriminer « indirectement » les individus grâce aux interactions latentes entre variables qui apportent de meilleures performances (et donc une meilleure quantification du risque, segmentation des prix, etc.). Après avoir posé l'état des lieux des concepts clés gravitant autour de la discrimination, nous illustrons la complexité de les quantifier. Nous proposons ensuite une méthode innovante, non rencontrée dans la littérature, permettant de réduire les risques de discrimination indirecte à partir de concepts mathématiques d'algèbre linéaire. Cette technique est illustrée sur un cas de sélection de risque concret en assurance vie, démontrant sa simplicité d'usage et ses performances prometteuses.

Abstract

Keywords: fairness, indirect discrimination, statistical parity, decorrelation, Machine Learning

The development of Machine Learning is experiencing growing interest from the general public, and in recent years there have been numerous press articles questioning its objectivity: racism, sexism, ... Driven by the growing attention of regulators on the ethical use of data in insurance, the actuarial community must rethink pricing and risk selection practices for fairer insurance. Equity is a philosophy concept that has many different definitions in every jurisdiction that influence each other without currently reaching consensus. In Europe, the Charter of Fundamental Rights defines guidelines on discrimination, and the use of sensitive personal data in algorithms is regulated. If the simple removal of the protected variables prevents any so-called 'direct' discrimination, models are still able to 'indirectly' discriminate between individuals thanks to latent interactions between variables, which bring better performance (and therefore a better quantification of risk, segmentation of prices, and so on). After introducing the key concepts related to discrimination, we illustrate the complexity of quantifying them. We then propose an innovative method, not yet met in the literature, to reduce the risks of indirect discrimination thanks to mathematical concepts of linear algebra. This technique is illustrated in a concrete case of risk selection in life insurance, demonstrating its simplicity of use and its promising performance.

Note de Synthèse

Les concepts clés et la réglementation autour de la discrimination

L'équité actuarielle est un concept clé pour les assureurs, qui signifie que les individus sont traités équitablement en matière de risque. Cela permet leur répartition en classes de risques homogènes, assurant le bon déroulement de la segmentation et de la mutualisation. Mais de nos jours, avec le développement d'algorithmes complexes, des sources de données plus riches et l'amélioration des méthodes d'interprétabilité, de multiples sources de biais ont été exposées et l'objectivité des données et des modèles est remise en question.

Le secteur des assurances fait l'objet d'une attention croissante, le public et les régulateurs exigeant plus de transparence et de justification sur les questions d'équité. Mais il existe une multitude de points de vue sur ce sujet. Premièrement, d'un point de vue juridique, la discrimination est définie par la loi comme la différence de traitement entre des individus se trouvant dans des situations similaires en raison de critères prohibés. Ces critères sont également définis par la loi, mais dépendent de la juridiction. Par exemple, aux États-Unis, selon les États, les informations sur l'origine peuvent être utilisées dans toutes les lignes d'assurance, mais elles sont strictement interdites par la Charte des Droits Fondamentaux de l'Union Européenne. Lorsque les critères sont utilisés explicitement dans la prise de décision, on parle de discrimination 'directe' et lorsque la pratique est apparemment neutre mais conduit tout de même à des traitements différents, on parle de discrimination 'indirecte'. Ensuite, d'un point de vue statistique, il existe de nombreuses définitions différentes de l'équité. Toutes ces définitions tentent de traduire mathématiquement différentes visions du monde mais ne sont pas compatibles les unes avec les autres. Les régulateurs demandent aux actuaires d'avoir des modèles équitables mais ne précisent pas quelle définition utiliser, laissant de nombreuses questions sans réponse.

Jusqu'à présent, les actuaires ont empêché la discrimination directe en ne recueillant pas d'informations sensibles sur les individus. Cette méthode n'est pas une solution, car il peut encore y avoir une discrimination indirecte. En effet, si les variables non sensibles ont une relation de dépendance avec les variables sensibles, ce qui est presque toujours le cas, les modèles peuvent déduire ces dernières et maintenir un traitement injuste. Ces variables non sensibles qui permettent d'inférer les variables sensibles sont appelées proxys. De plus, si les informations sensibles ne sont pas collectées, il est pratiquement impossible de vérifier s'il y a discrimination.

Comme mentionné précédemment, de nombreux articles de recherche ont tenté de fournir une définition mathématique de l'équité. Il existe deux vision de l'équité : l'équité de groupe, qui vise à traiter différents groupes de manière égale, et l'équité individuelle, qui vise à traiter de manière similaire des individus similaires. Parmi la première catégorie, on peut citer la parité statistique, qui recherche l'indépendance entre la prédiction et les variables sensibles, l'égalité des opportunités, qui recherche l'indépendance entre la prédiction et les variables sensibles, conditionnellement à la vraie classe, qui se traduit par l'égalité des taux de vrais et des faux positifs, et enfin l'égalité des chances qui est la même chose que l'égalité des opportunités, mais uniquement pour les taux de vrais positifs. Aux États-Unis, le 'Disparate Impact' est une mesure populaire qui est une conséquence de la parité statistique et est utilisée dans les tribunaux pour prouver une allégation de discrimination, mais ne s'applique qu'à la classification binaire avec une variable protégée binaire également. Dans la plupart des cas, ces définitions s'appliquent aux problèmes de classification binaire et pour les problèmes de régression, il y a moins de définitions d'équité. Pour l'équité individuelle, les critères mathématiques ne sont pas aussi simples, car cela implique de définir une distance entre les individus pour mesurer leur similarité, ce qui n'est pas une question triviale.

Une méthode de prétraitement pour atténuer la discrimination indirecte

La problématique de ce mémoire d'actuariat est : comment atténuer la discrimination indirecte ? Les solutions consistent généralement soit à travailler directement sur les données (pré-traitement), sur le modèle (pendant le traitement) ou sur les prédictions (post-traitement). Nous avons décidé de rechercher une méthode de pré-traitement, basée sur l'une des définitions d'équité de groupe, la parité statistique. Cela permet ensuite l'utilisation de tout type de modèle car le problème est traité le plus en amont possible dans le processus, directement sur les données.

Nous nous sommes inspirés du processus de Gram-Schmidt, une méthode d'orthogonalisation d'un ensemble de vecteurs dans un espace avec un produit interne. La covariance est un produit scalaire dans l'espace des variables aléatoires centrées de variance finie. Pour en revenir à la définition de la parité statistique, nous recherchons une prédiction indépendante des variables protégées. Le but de notre méthode est donc de transformer les variables non sensibles de manière à ce qu'elles deviennent décorrélées des variables sensibles. Bien entendu, il s'agit d'une approximation, car la corrélation n'est que la composante linéaire de la dépendance. La non-corrélation équivaut à l'orthogonalité dans l'espace des variables aléatoires centrées à variance finie, cela nous a donc permis de poser le problème: avec $X_1, \ldots, X_s$ les s variables sensibles et $X_{s+1}, \ldots, X_n$ les variables non sensibles, nous cherchons la matrice de passage A qui donne $X' = AX$, donnant l'expression des nouvelles coordonnées en fonction des anciennes. Les s premières variables sont inchangées. Cela nous donne une matrice de transition de la forme

$$A = \begin{bmatrix} & & 0 & \ldots & 0 \\ & I_s & \vdots & & \vdots \\ & & 0 & \ldots & 0 \\ a_{s+1,1} & & \ldots & & a_{s+1,n} \\ \vdots & & & & \vdots \\ a_{n,1} & & \ldots & & a_{n,n} \end{bmatrix}$$

Nous cherchons les variables transformées $X'_{s+1}, \ldots, X'_n$ telles que $corr(X_i, X'_j) = 0$ pour $i = 1, \ldots, s$ et $j = s+1, \ldots, n0$. Cela nous donne un système de s équations à n inconnues, avec un nombre infini de solutions car $n > s0$. Nous devons poser $n - s$ contraintes de plus de manière à obtenir un système complet. Nous avons fait le choix d'exprimer chaque nouveau vecteur en fonction des vecteurs sensibles ainsi que de son homologue dans l'ancienne base :

$$X'_k = \sum_{j=1}^{s} a_{k,j} X_j + a_{k,k} X_k$$

Cela réduit le problème à s équations et $s+1$ inconnues. Une idée pour la dernière contrainte est de minimiser la distance entre les anciens et les nouveaux vecteurs non sensibles : $\min d(X_k, X'_k)0$. Le problème a une solution car la distance (correspondant à la variance de la différence des deux variables aléatoires) est positive. Nous avons enfin :

$$\min_{a_{k,1},\ldots,a_{k,s},a_{k,k}} d(X_k, X'_k) \text{ tel que } \begin{cases} \langle X_1, X'_k \rangle & = 0 \\ \ldots & \\ \langle X_s, X'_k \rangle & = 0 \end{cases} \text{ avec } X'_k = \sum_{j=1}^{s} a_{k,j} X_j + a_{k,k} X_k$$

En résolvant le problème pour tous les $k = s+1, \ldots, n$ nous obtenons A et trouvons $X' = AX0$.

Illustration sur des données simulées simples

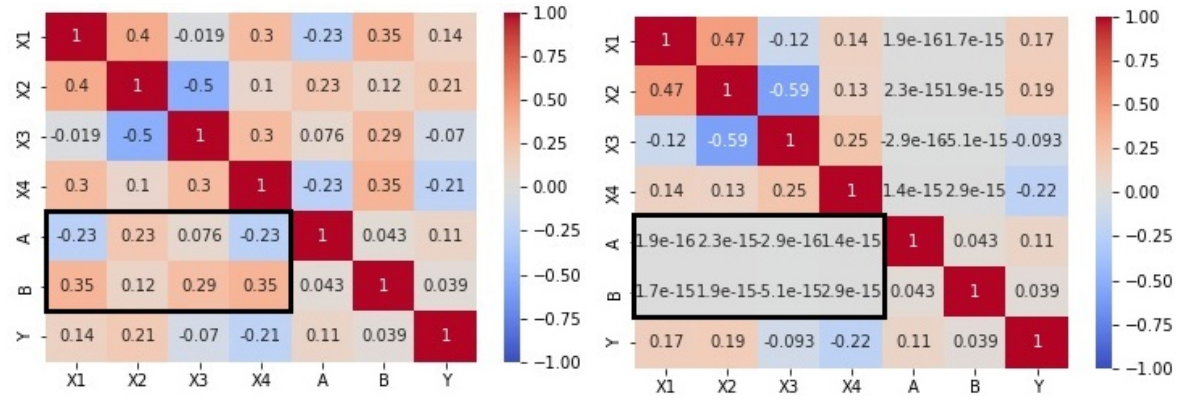
Pour illustrer la méthode, nous avons d'abord utilisé des données simulées. La raison est que nous voulons connaître les véritables relations entre les variables, ce qui n'est pas le cas avec des échantillons de données réelles. Le processus de simulation repose sur la théorie des copules. Nous avons créé un jeu de données avec deux variables sensibles binaires, A et B, quatre variables normales non-sensibles, $X^{(1)}, \ldots, X^{(4)}$ et une variable d'intérêt binaire. Toutes ces variables sont corrélées entre elles de manière controlée.

Nous appliquons ensuite un modèle de régression logistique, choisi pour sa simplicité et son interprétabilité, utilisant d'abord toutes les variables, puis uniquement les variables non sensibles et enfin les variables non sensibles transformées.

Le modèle utilisant toutes les variables est, sans surprise, injuste au regard des trois définitions de l'équité introduites précédemment, parité statistique, égalité des chances et égalité des opportunités. Selon les trois définitions, les groupes $A = 1$ et $B = 1$ sont désavantagés par le modèle par rapport aux groupes $A = 0$ et $B = 00$. Ce modèle fait preuve de discrimination directe car il y a une différence de traitement entre groupes suite à une utilisation explicite des variables sensibles.

Lorsque nous supprimons les variables protégées, il y a une légère baisse de la performance prédictive, mesurée par la précision et l'AUC. En ce qui concerne l'équité, la situation est pire lorsque l'on regarde la variable A, avec le groupe $A = 1$ encore plus désavantagé par le modèle qu'avant, et meilleure lorsque l'on regarde la variable B, avec le groupe $B = 1$ toujours le plus défavorisé mais moins qu'avec le modèle utilisant toutes les variables. En n'utilisant pas les variables sensibles, nous avons donc évité la discrimination directe, mais pas la discrimination indirecte car il existe toujours une différence de traitement entre les groupes.

Enfin, nous appliquons notre méthode de pré-traitement et transformons les variables non sensibles. Les matrices de corrélation avant et après transformation en figure 1 montrent le succès de notre méthode : il n'y a plus de corrélation entre les variables sensibles et les variables transformées.



(a) Before (b) After

Figure 1: Heatmaps des corrélations avant et après transformation des $X^{(i)}$

Nous appliquons ensuite le modèle aux variables transformées. Par rapport au modèle utilisant uniquement des variables non sensibles, il y a une baisse de la performance prédictive avec une précision et une AUC inférieures. Le modèle traite désormais presque parfaitement équitablement les groupes selon la définition de la parité statistique, ce qui était l'objectif de la méthode. Mais quand on regarde les deux autres définitions d'équité, pour la variable A, le modèle traite les deux groupes de manière plus juste qu'auparavant, mais maintenant c'est le groupe $A = 0$ qui est le plus défavorisé. Pour la variable B, le modèle est moins juste qu'avant

et encore une fois, c'est maintenant le groupe $B = 0$ qui est le plus défavorisé selon ces deux définitions.

Pour conclure, supprimer les variables protégées n'est pas une solution pour éviter la discrimination, et selon les relations entre les variables, cela peut soit améliorer, comme nous l'avons vu avec la variable A, ou détériorer l'équité, comme nous l'avons vu avec la variable B. Notre méthode a approché l'indépendance avec la non-corrélation, et nous avons réussi à approcher la parité statistique. Mais il y a quelques inconvénients : une baisse des performances, des problèmes d'interprétabilité concernant les variables transformées et une incompatibilité avec d'autres définitions d'équité.

Illustration sur des données réelles de mortalité

Nous avons appliqué la même méthode à un cas d'utilisation réel : la mortalité des personnes diagnostiquées avec un mélanome non métastatique, une forme de cancer de la peau. Pour réaliser cette étude, nous avons utilisé les données de la base de recherche publique SEER du *National Cancer Institute* aux États-Unis. C'est une source d'information très riche et complète, mais qui a nécessité un long traitement avant de pouvoir être utilisée.

L'analyse de survie est très spécifique car l'objectif est de modéliser la durée de survie, qui n'est souvent observée que partiellement en raison des phénomènes de censure et de troncature. Pour résoudre ce problème, nous devons prendre en compte l'exposition de chaque individu et l'utiliser comme poids dans le modèle de régression logistique standard.

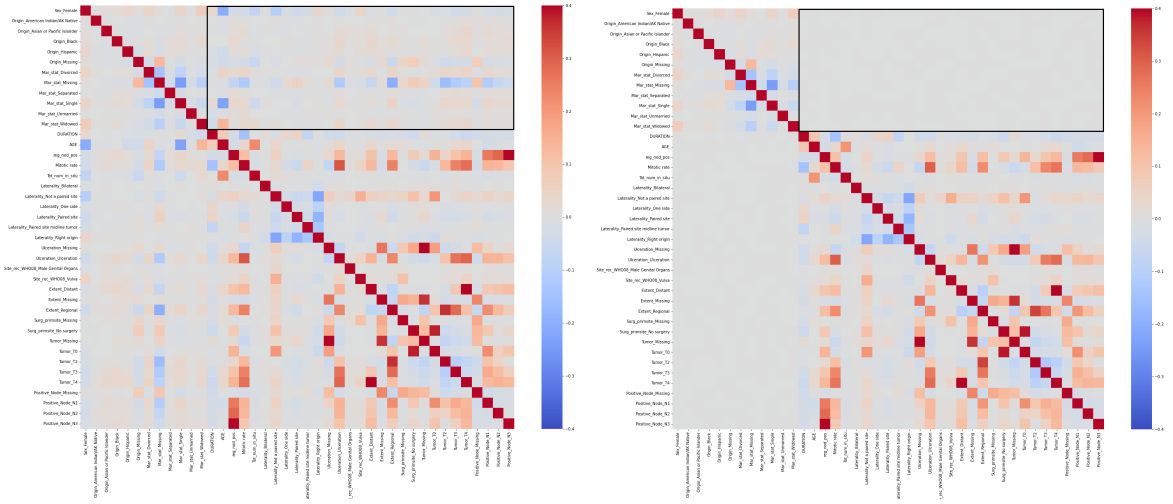
Avant cela, nous avons calculé les taux de mortalité sur cinq ans à partir des données et examiné les différences en fonction de certains critères. Tout d'abord, nous avons comparé les taux de mortalité des personnes atteintes de tumeurs de différentes tailles, stades et ampleurs. Nous avons constaté, en accord avec la littérature médicale, que des tumeurs plus grosses, la présence de métastases, une propagation plus étendue aux ganglions lymphatiques régionaux et des stades plus avancés entraînent des taux de mortalité plus élevés. Ensuite, en examinant les taux de mortalité en fonction des variables sensibles, nous avons constaté que les taux de mortalité varient selon le sexe, l'origine et le statut civil. Bien que cela puisse être causé par une plus grande représentation de certains autres facteurs de risque dans ces catégories, nous suspectons que des problèmes d'équité pourraient survenir lors de la modélisation des taux de mortalité à partir de ces données.

Une première étape de la modélisation a été la sélection de variables avec trois types de contraintes : médicale, statistique et de souscription. En effet, les variables qui ne sont pas pertinentes médicalement, statistiquement ou qui ne peuvent être obtenues au moment de la souscription ne doivent pas être utilisées dans le modèle. Après cette sélection, il nous reste trois variables sensibles, que sont le sexe, l'origine et l'état civil, ainsi que douze variables non sensibles.

Comme pour les données simulées, nous commençons par appliquer notre modèle de régression logistique à toutes les variables. Le modèle fonctionne très bien, avec une AUC de 0,8769. Nous examinons ensuite les mesures d'équité et, sans grande surprise, nous constatons que le modèle n'est équitable selon aucune des trois définitions d'équité. En regardant les taux d'acceptation par origine dans la figure 3a, nous constatons qu'il existe de grands écarts entre les taux d'acceptation, un groupe étant plus défavorisé par le modèle que les autres.

Lorsque nous supprimons les variables protégées, il y a, comme dans le cas simulé, une légère baisse des performances. Par sexe, le modèle est plus juste mais le même groupe reste défavorisé selon toutes les définitions. Par origine, selon la définition, le modèle fonctionne plus ou moins bien. Pour les taux d'acceptation, comme le montre la figure 3b, les niveaux ont changé mais il existe toujours des écarts entre les groupes et le même reste le plus défavorisé. Par état civil, le modèle est plus juste mais le groupe le plus défavorisé n'est plus le même qu'avant.

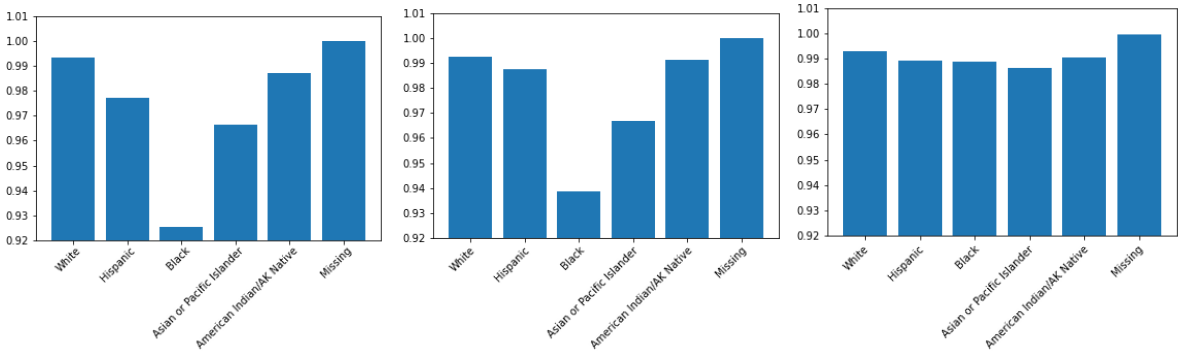
Nous appliquons ensuite notre méthode de décorrélation, et obtenons des vecteurs transformés non corrélés aux vecteurs sensibles. La figure 2 donne la matrice de corrélation avant et après transformation. Lors de l'application du modèle à ces variables transformées, nous avons encore une diminution des performances prédictives avec une AUC plus faible, à 0,8534. En regardant les taux d'acceptation par origine, les écarts sont désormais très faibles entre les groupes. Pour toutes les variables protégées, nous avons la même conclusion : nous avons presque atteint la parité statistique et l'égalité des chances mais nous sommes moins proches de l'égalité des opportunités, et le groupe le plus défavorisé a changé.



(a) Données originales (b) Données transformées

Figure 2: Corrélations, avant et après transformation,
corrélations entre les attributs sensibles et les autres encadrées en noir

Nous avons les mêmes conclusions que dans le cas simulé : il ne suffit pas de supprimer les variables protégées pour avoir un modèle juste, notre méthode nous a permis d'atteindre approximativement la parité statistique, il y a un compromis entre performance et équité, et toutes les définitions d'équité ne sont pas compatibles.



(a) Avec toutes les variables (b) Sans variables protégées (c) Variables transformées

Figure 3: Taux d'acceptation par origine

Executive Summary

Understanding key concepts and regulations around discrimination

Actuarial fairness is a key concept for insurers, meaning that individuals are treated fairly when it comes to risk. This allows their classification into homogeneous risk classes, ensuring the smooth running of segmentation and pooling. But nowadays, with the development of complex algorithms, the richer data sources and the improvement of interpretability methods, multiple sources of bias have been exposed and the objectivity of data and models is questioned.

The insurance sector is under close scrutiny as both the public and regulators are demanding more transparency and justification on fairness issues. But there are a multitude of points of view on this subject. First, on a legal perspective, discrimination is defined by the law as the difference in treatment between individuals in similar situations due to prohibited criteria. These criteria are also legally defined, but depend on the jurisdiction. For example, in the US, depending on the State, information on national origin can be used in all insurance lines, but it is strictly prohibited by the Charter of Fundamental Rights of the European Union. When the criteria are used explicitly in the decision-making, we talk about 'direct' discrimination and when the practice is apparently neutral but still leads to different treatments, we talk about 'indirect' discrimination. Secondly, on a statistical point of view, there are many different definitions of fairness. All of these definitions try to mathematically translate world views but are not compatible with each other. Regulators ask actuaries to have fair models but do not specify which definition to use, leaving many questions unanswered.

Up until now, actuaries have prevented direct discrimination by not collecting sensitive information about individuals. This method is not a solution, because there can still be indirect discrimination. Indeed, if the non-sensitive variables have some dependency relationship with the sensitive ones, which is almost always the case, models can infer the latter and maintain unfair treatment. These non-sensitive variables that allow to infer the sensitive ones are called proxies. Furthermore, if the sensitive information is not collected, it is virtually impossible to check for discrimination.

As mentioned previously, many papers have tried to provide mathematical definitions of fairness. They can be separated into two categories: group fairness, which aims to treat different groups equally, and individual fairness, which aims to treat similar individuals similarly. Among the first category, we can cite statistical parity, that seeks independence between the prediction and the sensitive variables, equalized odds, that seeks independence between the prediction and the sensitive variables, conditionally on the outcome, which translates into having equal true and false positive rates, and finally equal opportunity is the same as the latter, but only looks at true positive rates. In the US, the 'disparate impact' is a popular metric that results from the statistical parity definition and is used in courts of law to prove a discrimination allegation, but it is only for binary classification with a binary protected variable. In most cases, these definitions apply to binary classification problems but for regression problems, there are fewer definitions. For individual fairness, mathematical criteria are not as easy to find, because it requires the definition of a distance between individuals to measure their similarity, which is not an trivial question.

A pre-processing method to mitigate indirect discrimination

The problematic of this actuarial thesis is: how can we avoid indirect discrimination? Solutions usually consist either in working directly on the data (pre-processing), on the model (in-processing) or on the predictions (post-processing). We decided to look for a pre-processing method, based on one of the group fairness definitions, statistical parity. This then allows the

use of any type of model because the problem is dealt with as early as possible in the process, directly tackling the data.

We were inspired by the Gram-Schmidt process, a method for orthogonalizing a set of vectors in an inner product space. The covariance is an inner product in the space of centered random variables with finite variance. Going back on the statistical parity definition, we are looking for a prediction that is independent of the protected variables. The goal of our method is therefore to transform the non-sensitive variables such that they become uncorrelated with the sensitive ones. Of course, this is an approximation, as correlation is only the linear component of dependence. Uncorrelatedness is equivalent to orthogonality in the space of centered random variables with finite variance, so this allowed us to set the problem: with $X_1, \ldots, X_s$ the s sensitive variables and $X_{s+1}, \ldots, X_n$ the non-sensitive variables, we are looking for the transition matrix A giving the change-of-basis formula $X' = AX$, expressing the new coordinates in terms of the old ones. The first s variables stay the same. This gives a transition matrix of the shape

$$A = \begin{bmatrix} & 0 & \ldots & 0 \\ I_s & \vdots & & \vdots \\ & 0 & \ldots & 0 \\ a_{s+1,1} & \ldots & & a_{s+1,n} \\ \vdots & & & \vdots \\ a_{n,1} & \ldots & & a_{n,n} \end{bmatrix}$$

We are looking for $X'_{s+1}, \ldots, X'_n$ such that $corr(X_i, X'_j) = 0$ for $i = 1, \ldots, s$ and $j = s + 1, \ldots, n0$. This gives a system of s equations and n unknown variables, with an infinite number of solutions as $n > s0$. We need to set $n - s$ more constraints in order to have a complete system. We made the choice to express each new vector as a linear combination of the sensitive vectors and of its equivalent in the old basis:

$$X'_k = \sum_{j=1}^{s} a_{k,j} X_j + a_{k,k} X_k$$

This narrows the problem to a system of s equations and $s+1$ unknown variables. An idea for the last constraint is to minimize the distance between the old and the new vectors: $\min d(X_k, X'_k)0$. The problem has a solution as the distance (corresponding to the variance of the difference between the two variables) is positive. We finally have:

$$\min_{a_{k,1},\ldots,a_{k,s},a_{k,k}} d(X_k, X'_k) \text{ such that } \begin{cases} \langle X_1, X'_k \rangle & = 0 \\ \ldots & \\ \langle X_s, X'_k \rangle & = 0 \end{cases} \text{ with } X'_k = \sum_{j=1}^{s} a_{k,j} X_j + a_{k,k} X_k$$

Solving the problem for every $k = s + 1, \ldots, n$ gives us the transition matrix A, and we can compute $X' = AX0$.

Illustration on simple simulated data

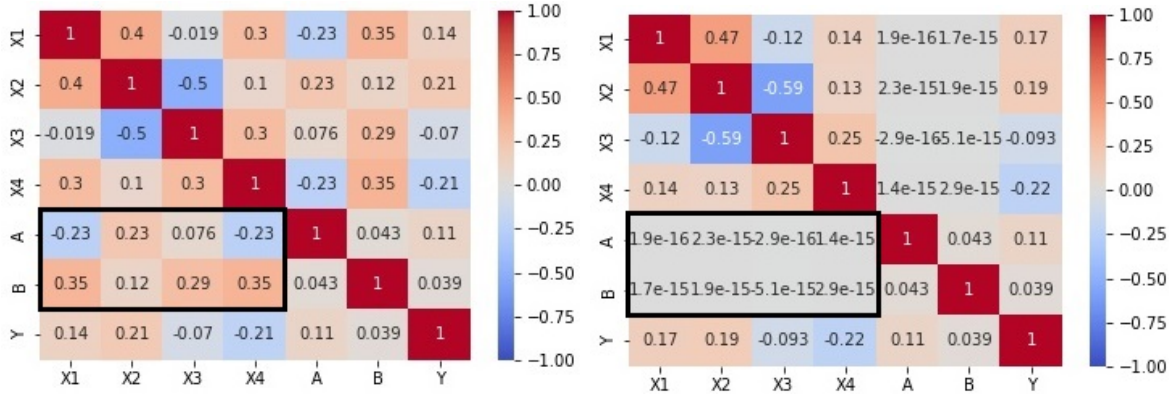
To illustrate the method, we first used simulated data. The reason for this is that we want to know the true relationships between variables, which is not the case with samples of real data. The simulation process relies on the theory of copula. We created a dataset of two binary sensitive variables, A and B, four non-sensitive normal variables, $X^{(1)}, \ldots, X^{(4)}$ and a binary variable of interest. All of these variables are correlated with each other.

We then apply a logistic regression model, chosen for its simplicity and interpretability, first using all variables, then only non-sensitive variables and finally transformed non-sensitive variables.

The model using all variables is, unsurprisingly, unfair under the three definitions of fairness introduced previously: statistical parity, equal opportunity and equalized odds. Under all three definitions, groups $A = 1$ and $B = 1$ are disadvantaged by the model. This model exhibits direct discrimination as the result is a difference in treatment after an explicit use of sensitive variables.

When we remove protected variables, there is slight decline in performance, measured by the accuracy and the AUC. Looking at fairness, the situation if worse when looking at variable A, with group $A = 1$ even more disadvantaged by the model, and better when looking at variable B, with group $B = 1$ still the most disadvantaged but less than with the model using all variables. By not using the sensitive variables, we have avoided direct discrimination, but not indirect discrimination as there is still a difference of treatment between groups.

Finally, we apply our pre-processing method and transform the non-sensitive variables. The correlation matrices before and after transformation in figure 4 show the success of our method: there are no more correlations between the sensitive variables and the non-sensitive transformed variables.



(a) Before (b) After

Figure 4: Heatmaps of correlations before and after transformation of the $X^{(i)}$

We then apply the model to the non-sensitive transformed variables. Compared to the model using only non-sensitive variables, there is a decline in performance with a lower accuracy and AUC. The model now treats the protected groups almost perfectly fairly under the statistical parity definition, which was the goal of the method. But when we look at the other two definitions, for variable A, the model treats both groups in a fairer way than before, but now it is group $A = 0$ that is the most disadvantaged. For variable B, the model is less fair than before and again, it is now group $B = 0$ that is the most disadvantaged under these two definitions.

To conclude, simply removing protected variables is not a solution to avoid discrimination, and depending on the relationships between variables, it can both improve, like we saw with variable A, or deteriorate fairness, like we saw with variable B. Our method approximated independence to uncorrelatedness, and we have managed to approximate statistical parity. But there are multiple drawbacks: a decrease in performance, interpretability issues concerning the transformed variables and incompatibility with other definitions of fairness.

Illustration on real mortality data

We finally applied the same method to a real-life use case: the mortality of individuals diagnosed with non-metastatic melanoma of the skin cancer. In order to perform this study, we used data from the public research SEER database of the National Cancer Institute in the United States. It is a huge source of information, but required lengthy processing.

Survival analysis is very specific as the goal is to model survival duration, which is often observed only partially because of censuring and truncation phenomena. To tackle this issue, we need to take into account the exposure of each individuals and use it as a weight in the logistic regression model.

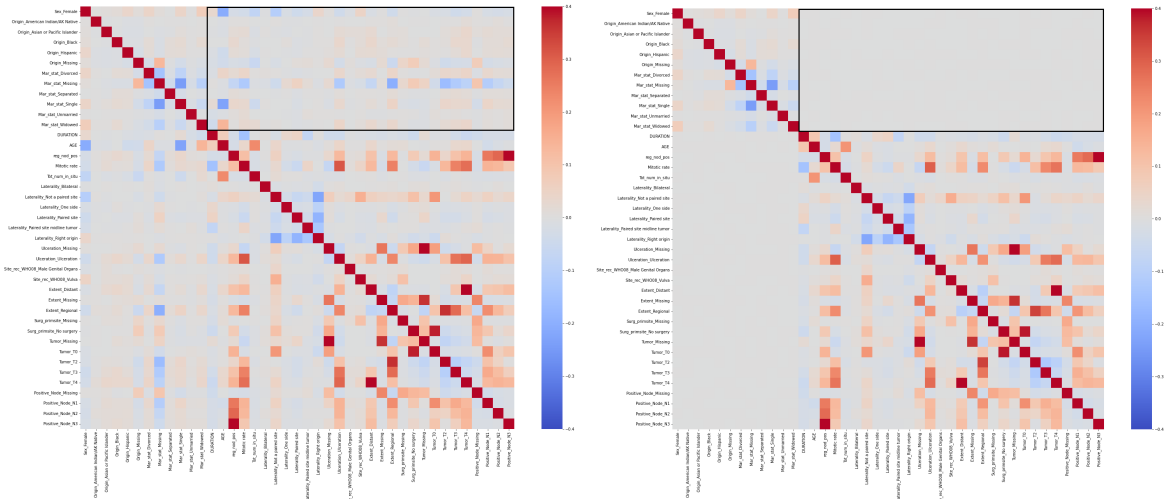
Before doing so, we computed five-year mortality rates from the data, and looked at the differences depending on some criteria. First, we compared mortality rates for individuals with tumors of different sizes, stages and extents. We found, consistently with medical literature, that bigger tumors, the presence metastasis, more extensive spread to regional lymph nodes and higher stages result in higher mortality rates. Then, looking at mortality rates depending on sensitive variables, we found that mortality rates vary on sex, origin and marital status. Although this could be caused by greater representation of certain risk factors in some of these categories, it gives us a first hint at fairness issues that could arise when modeling mortality rates using this data.

A first step for this modeling was variable selection under three types of constraints: medical, statistical and underwriting. Indeed, variables that are not relevant medically, statistically or cannot be obtained at the underwriting stage should not be used in the model. We are left with three sensitive variables, Sex, Origin and Marital Status, and twelve non-sensitive variables.

Like for the simulated data, we begin by applying our logistic regression model to all the (selected) variables. The model performs very well, with an AUC of 0.8769. We then look at fairness metrics, and with little surprise we see that the model is not fair under any of the three fairness definitions. Looking at acceptance rates by Origin in figure 6a, we see that there are large gaps between acceptance rates, with Black the most disadvantaged group.

When we remove the protected variables, there is, like in the simulated case, a slight decrease in performance. By Sex, the model is fairer but the same group is still disadvantaged under all definitions. By Origin, depending on the definition, the model performs better or worse. For acceptance rates, as shown in figure 6b, the levels have changed but there are still gaps between groups and Black remains the most disadvantaged group. By Marital Status, the model is fairer but the most disadvantaged group is not the same as before.

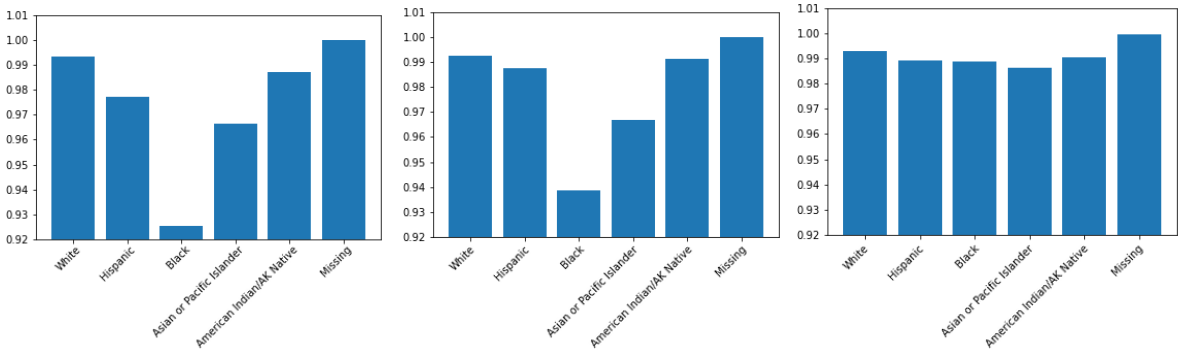
We then apply our change of basis method, and obtain transformed non-sensitive vectors uncorrelated to the sensitive ones. Figure 5 gives the correlation matrix before and after transformation. When applying the model to these transformed variables, we have another decrease in performance with a lower AUC, at 0.8534. Looking at acceptance rates by Origin in figure 6c, the gaps are now very small between groups, and Asian or Pacific Islander is now the most disadvantaged group, even if not by much. For all protected variables, we have the same conclusion: we have almost reached statistical parity and equal opportunity but are further away for equalized odds, and the most disadvantaged group has changed.



(a) Original data (b) Transformed data

Figure 5: Heatmap of correlations, before and after transformation, correlations between the sensitive attributes and the others framed in black

We have the same conclusions as in the simulated case: simply removing the protected variables is not enough to have a fair model, our method allowed us to approximately reach statistical parity, there is a performance-fairness trade-off, and not all fairness definitions are compatible.



(a) With all variables (b) Without protected (c) Transformed variables

Figure 6: Acceptance rates by Origin

Acknowledgments

I would like to thank SCOR for giving me the opportunity of writing this thesis. I could not have undertaken this journey without my tutors, Antoine Chancel and Antoine Ly, who generously provided knowledge and expertise, and were always up for a brainstorming session in front of the whiteboard.

I am also grateful to all the fine people I have met and collaborated with on this journey. Many thanks to Jules Chancel and Eloise Sorin, who gave me insights on data law early on in my works. I would like to express my deepest appreciation to Arthur Charpentier, who shared his research and ideas, and gave wonderful feedback on the study. Many thanks to the Inclusive Medical Underwriting team, including but not limited to Professor Eric Raymond, Antoine Moll and Danis Jiogue, for their help in understanding the SEER cancer database. Lastly, thanks should also go to Fabien Faivre with whom I had the opportunity to introduce my research at a conference organized by the French Institute of Actuaries.

Thanks should also go to Caroline Hillairet and to the teachers of ENSAE who gave me the theoretical framework needed to conduct this extensive actuarial analysis.

Contents

Introduction

In many sectors, Machine Learning models have been exposed as unintentionally discriminatory, leading to unfair decisions that can have drastic consequences. The insurance industry has always been under close scrutiny when it comes to the use of personal data and discrimination issues, but in the light of recent denunciations in all industries, the attention on fairness issues has grown.

Fairness is a complex philosophical matter, and there is no single definition for it. Researchers have managed to define it statistically, but the definition depends on underlying assumptions and ideologies. Furthermore, as of today, many methods have been proposed to come closer to some definitions of fairness, but none can ensure perfect fairness.

Regarding previous points, we wonder to what extent fairness can be approximated when applying a Machine Learning model to insurance mortality data, and more specifically to what extent proxy discrimination can be avoided. This thesis aims at illustrating the complexity of quantifying discrimination and finding a way to mitigate it. For this purpose, we propose a simple and promising method that relies on linear algebra to mitigate indirect discrimination.

All along the thesis, we have provided the reader with summaries and key points in the form of light blue frames.

1 Understanding key concepts and regulations around discrimination

1.1 The life insurance industry

Historically, life insurers have used classic statistical models to assess risks for their products, covering Mortality, Critical Illness[1], Disability, Longevity and Medical Expenses. But with the development of Machine Learning and the now richer than ever data sources, new techniques are becoming more and more popular. These methods imply new challenges for actuaries, mostly due to the richness of information and the complexity of algorithms.

1.1.1 Actuarial fairness

One of the challenges concerns the notion of equity, which has always been a key issue for insurers. 'Actuarial fairness' means that risky insureds should contribute more and pay a higher premium. Actuaries have to determine how to classify policyholders between risky and non-risky, and more specifically what attributes are good indicators of risk. They then rely on historical data to estimate losses [26]. It is sometimes complicated to know if an attribute is directly related to risk or not. As we will see later on, there are country-specific regulations concerning the use of certain attributes for risk assessment.

1.1.2 Segmentation, pooling and adverse selection

Before insurance, the only way of hedging against risk was individual prudence. Pooling offered a new way to deal with uncertainty: losses were the collective responsibility of the pool. This created insurance solidarity with an understanding of fairness [20]. Today, insurers offer contracts at large levels, counting on the compensation between policyholders who file claims and those who do not. In order to ask for premiums accordingly to a policyholder's risk profile, insurers do a segmentation of the insureds [15]. Segmentation consists in creating homogeneous classes and estimating the risk on average [5], so that premiums are adapted to the risk profile.

The pure premium is the expected loss of the insured over the coverage period. Since pooling is based on the law of large numbers, risks have to be homogeneous, which is why insurers need to classify the risks properly. The classification is based on observable factors, which should indicate what the risk is [26].

If groups are heterogeneous, policyholders could cross-subsidy. This leads to adverse selection: lower risks are asked to pay more than their expected loss because they are not classified in the right pool, so they are attracted by a competitor who will offer lower prices.

The classification of insureds into groups with similar risk profiles is at the core of the insurance business. This classification is based on observable factors, supposedly good indicators of risk.

1.1.3 Risk modeling

Life insurers model biometric risks, which relate to human life conditions, and more specifically the duration until the occurrence of an event. In our case, we will be studying the mortality of cancer patients.

[1] Critical illness insurance compensates insureds with a lump sum payment upon diagnosis in order to cover treatment costs [9].

Individuals with cancer history are considered 'aggravated risks' and are often offered coverage with deterrent premiums, based on limited and imprecise criteria. But each cancer is specific and the evolution of treatment possibilities has tremendously increased survival odds in the past few years. At SCOR, the medical underwriting team is in charge of providing inclusive solutions for cancer patients. By using all available information about individuals, we can precisely estimate mortality rates thanks to Machine Learning models. These precise rates help insurers offer fair premiums to individuals with a history of cancer.

For cancer patients, insurers model the duration before death in the context of mortality or longevity products. Information about these individuals can also be used for critical illness products, to model the duration before the occurrence of a cancer. Once the R&D department has studied the rates linked to the covered condition, the pricing team takes over to put the add-on cover or product on the market. The underwriters can also benefit from the study to better assess risk profiles.

To do this modeling, we need to take into account a few constraints:

- Underwriting constraints: the variables used by the model need to be available to the underwriter, ie be included in the medical file of the individual. Depending on the local legislation, models must not be discriminatory against certain population, so as a 'solution', some variables are simply deleted. The insurer also has business constraints as he is in a competitive market.
- Medical constraints: the variables and the coefficients they are attributed need to be coherent with medical literature. For example, variables that are not medically relevant, such as the address, cannot be used by the model. Another example of variable coherence is if a larger tumor size implies a shorter life expectancy, it must be reflected by the model.
- Modeling constraints: the variables must be statistically relevant, not too numerous nor strongly correlated with each other.

To model mortality risk, the selected variables need to be available at the underwriting stage, non-discriminatory (according to the local regulation), coherent with medical literature and statistically relevant. These constraints are of three types: underwriting, medical and statistical.

1.2 Fairness and bias

As we saw with the constraints of variable selection, depending on the local regulation, selected variables must not lead to discriminatory results. Fairness and bias is not a new subject in Machine Learning, but it has received increasing attention these past years, as numerous examples of unfair and biased outcomes were revealed in various fields. The most famous one is the COMPAS Recidivism Algorithm, which was proved biased against black defendants [35]. As actuaries use high dimensional data and complex models, to price contracts for example, they need to check that the outcomes are not biased. We need to define fairness and bias, see how it impacts insurance and find a method to detect it.

1.2.1 The need for fairness in insurance

Reputation Fairness is a key issue in insurance, because actuaries need to explain to underwriters how their models work, so that in turn policy applicants can understand and trust the process to be fair. Insurance is not a well-seen sector in the public opinion and by the authorities in general, which is why fairness and transparency are crucial subjects.

Regulation In the EU, the GDPR (General Data Protection Regulation) gives individuals the right to control their personal data, and specifically 'the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her' [25]. This means that a human intervention is required in any automated decision process. Individuals also have a right to erasure, or 'right to be forgotten' [24], which means that datasets might not be complete, independently of the collection process.

But there are specific regulations for the insurance sector. The French Supervisory Authority, the ACPR, requires appropriate data management, with ethical considerations such as fairness of processing and absence of discriminatory bias [1].

In April 2021, the EU proposed the first-ever legal framework on Artificial Intelligence (AI), the AI Act. The proposal is to define four levels of risk, from unacceptable to none, concerning AI systems. Each category will be subject to requirements and specific obligations such as conformity assessments and registration in a database. It could become applicable as soon as the second half of 2024 [16]. Insurers do not know how they will be impacted by this new regulation, but if their AI systems are classified as risky, they will be under strict regulation.

All in all, there is a regulation stacking that needs to be understood. The real question is: which regulation will be predominant?

With the increased attention on Machine Learning by both the public and regulators, fairness is a key issue for insurers. There are new regulations being developed, like the AI Act in Europe. Insurers need to prepare for more and more scrutiny around their use of Machine Learning and the outcomes produced by models.

1.2.2 A multitude of points of view

Fairness is a relative notion that has many different definitions. A. Narayanan gave 21 definitions for fairness in classification tasks [40]. It is a legal requirement, but also an ethical concept [15]. Definitions can be categorized into individual or group fairness. Individual fairness aims to treat similar individuals similarly and group fairness treats different groups equally [37]. But group fairness might appear unfair at the individual level and a generalization based on group membership may be wrong [7]. There are therefore two opposite worldviews regarding fairness with respect to a specific task: We're All Equal (WAE), meaning all groups have

similar abilities, and What You See Is What You Get (WYSIWYG), meaning observations reflect similar abilities [6]. From this, many statistical definitions have been created to measure fairness for model outcomes, mostly for classification. We will study some of them in detail later on.

Machine Learning is a statistical discrimination by nature, but it becomes objectionable when there is a systematic advantage for some privileged group. All definitions of fairness cannot be reconciled and cannot satisfy all aspects. There needs to be a legal study to provide an official definition of fairness and an associated metric.

There is no unique definition for fairness, as it is a legal and philosophical issue. Some statistical definitions have been proposed, and we can classify them into two groups: individual and group fairness. They are not all compatible, and regulators have not given any guidance on which ones need to be respected.

1.2.3 The different types of bias

Statistical fairness is strongly related to bias, which can be defined differently depending on the sector. In statistics, it is a systematic error in prediction outcomes [6]. For Machine Learning models, it can be introduced by users, come from the data or be amplified by algorithms. And it is a vicious circle, as algorithms learning from biased data give biased outcomes which will be fed into and amplified by future algorithms [37].

There are three types of biases that appear in classification problems [5]:

- Type 1: classes do not reflect the reality of the risk,
- Type 2: classes reflect a correlation with risk that is non-causal,
- Type 3: classes reflect a causal statistical reality, but are unacceptable because of ethical reasons.

Bias, strongly related to statistical fairness definitions, can come from many sources: users, data, algorithms... Three types of it exist in classification problems: when classes do not reflect risk, when classes reflect a non-causal correlation with risk and when classes reflects a causal statistical reality but are unacceptable for ethical reasons. The second type is the most difficult to detect, because causality is a complex and relatively new research subject.

1.2.4 Unfair and fair discrimination

To better understand and scope the concept from a scientific point of view, we discussed with different lawyers specialized in data law. Those exchanges helped us focus on the different key concepts exposed below.

Legal definition Discrimination is the difference in treatment between individuals in similar situations due to prohibited criteria. In France, the Penal Code [22] defines these criteria in the Article 225-1:

> 'their origin, their sex, their family situation, their pregnancy, their physical appearance, the particular vulnerability resulting from their economic situation, apparent or known to its author, their surname, their place of residence, their state of health, their loss of autonomy, their disability, their genetic characteristics, their morals,

their sexual orientation, their gender identity, their age, their political opinions, their trade union activities, their ability to express themselves in a language other than French, their membership or non-membership, real or supposed, of an ethnic group, a nation, a so-called race or a determined religion'

These general prohibitions to fight against discrimination are supplemented by the French Insurance Code, and the Article L117-2 states that there can be absolutely no distinction between individuals based on:

- age for access to insurance guarantees or termination of insurance benefits, with the exception of pricing for life insurance contracts with mortality tables;
- pregnancy and motherhood for premium and benefit computation;
- sex for premium and benefit computation, except for mandatory supplementary pension schemes.

The Article 16-13 of the Civil Code [21] defines an absolutely prohibited criteria that is applicable to insurance:

'No one may be discriminated against because of their genetic characteristics.'

Other criteria can be used if they are a justified business necessity, or to modulate premiums and guarantees.

The nature of insurance As we saw in section 1.1.2, insurance is based on segmentation and pooling. It is about treating different risks differently. Classification and regression tasks are by definition a form of discrimination: the aim is to distinguish individuals based on a statistical similarity [4]. This is a form of discrimination that is justified and deemed acceptable if it does not systematically put a protected group at a disadvantage.

Unfair discrimination Discrimination is unfair when a certain group is treated unequally based solely on their affiliation to it [32]. Direct discrimination happens when protected attributes are explicitly used to make the decision. Indirect discrimination happens when the treatment appears neutral and depends on non-protected attributes, but protected groups get treated unjustly [37]. Harmful discrimination in insurance can happen at several stages: for the decision to insure, during the underwriting or marketing phases, to renew or cancel policies, for coverage offer or for pricing [20].

Discrimination based on specific criteria is prohibited by the law, and is supplemented by the Insurance Code for this specific sector. By nature, because of segmentation and pooling, insurers discriminate, but it is deemed acceptable if this discrimination is based on true risk factors. Unfortunately, bias can be introduced at several stages in the process.

1.2.5 Interpretability: a first step to tackling bias

Interpretability is the degree to which a human can understand the cause of a decision. Some models are directly interpretable and others, qualified as 'black box' models, need interpretability methods. As the need to explain Machine Learning models become more and more important, a new research field XAI, for eXplainable Artificial Intelligence, was created. Understanding why a model predicted specific outcomes is a first step in bias detection [39].

Interpretable models Some models are directly interpretable because of their structure. Linear Regressions predict the target as a weighted sum of the feature inputs, but assume linear relationships. Generalized Additive Models take into account non linear effects. Decision trees and decision rules are easy to interpret and capture feature interaction.

Model-agnostic explanation methods Global methods explain how features affect the prediction on average. Techniques include feature interaction detection, prediction function decomposition, feature importance measure and representative data points choice.

Local methods explain the individual predictions. Techniques include the description of how changing a feature changes the prediction, of which features anchor a prediction, of which features would need to be changed to change a prediction and of which individual features are attributed to a prediction.

Detecting bias Having more information on which variables play an important role in predictions, which variables interact with each other and more generally what causes a prediction can help detect bias. The outcome might be biased if a protected attribute has a great importance in the prediction or if an important variable for the prediction has a strong correlation with a protected attribute. This can help find out where the bias actually comes from.

Ways to detect bias are crucial and a first step towards fairness. In recent years, interpretability methods have been developed and eXplainable Artificial Intelligence is its own research field. Some models are more straight-forward than others, the latter sometimes being black boxes.

1.3 The expression of fairness and bias in data

Feature selection is the process of choosing what attributes are observed and taken into account for analysis. They are necessarily a reduction of reality and fail to capture real-world phenomena [4]. But obtaining sufficiently rich information is not always possible, and insurers have to rely on reductive data.

Because of the underwriting process, data available to insurers is necessarily a reduction of reality. Any conclusions can only be applied to a similar portfolio and not generalized to a larger population.

1.3.1 Sensitive variables

Insurers are not allowed to use all variables that are available to them, because they are viewed as protected or sensitive, and using them might lead to biased outcomes. The choice of which variables are allowed depends on regulators, but also on society as a whole, because it is both a legal and ethical concern. Often, attributes that are not under the control of individuals are not accepted. Attributes that change over time are accepted, because individuals are on both sides at different stages in their life. Acceptable variables should be good predictors of risk. If attributes are known to cause a risky event they are accepted [20].

Regulation

EU The European Union has anti-discrimination legislation, in the Charter of Fundamental Rights, that applies to the insurance sector, but it can have the effect of restricting flexibility of risk management and raising costs and legal insecurity, which means that insurers offer less effective coverage. Lawmakers decide which factors are determining for risk assessment but they are not specialists. With these directives, insurers cannot perfectly prevent adverse selection and moral hazard, and as a consequence, consumers can end up penalized, paying higher premiums and deductibles [42].

France There is an evolving conflict between insurance and anti-discrimination standards: on the one hand, insurers classify risks and on the other, the law prohibits the differentiation of individuals based on criteria that deny equal dignity. Insurers can select risks as long as they demonstrate the objectivity and statistical foundations of the data they rely on to do so. The Insurance Code completely forbids discrimination based on pregnancy and motherhood, risk selection for supplementary pension and the use of genetic information [44]. In 2011, the EU court ruled that insurers offering different prices to men and women violated gender equality laws, and this affected car, term life, health insurance and annuities [34]. This means that there is no possible discrimination on these criteria, even though there are some technical and pragmatic arguments for using them.

US Recently, unfair discrimination has become closely connected to disparate impact (see section 1.4.1). It is a measure of how a practice affects a group more than another, even when it appears neutral. But as of 2009, no court had applied the disparate impact standard to evaluate insurance rates [38]. Unfair discrimination in insurance is indeed not exactly equivalent to disparate impact. In State law, unfair discrimination happens when similar risks are treated differently for determining rates, coverage, benefits and terms and conditions of policies. Depending on States, certain factors are prohibited for risk classifications and in underwriting decisions. But laws are not as restrictive as one could believe: in fifteen States, it is only

prohibited to use race as the only factor regarding a decision to issue or continue a policy and in four States there are no restrictions on the use of race for underwriting personal automobile insurance [49].

In July 2021, the governor of Colorado signed a Senate Bill on insurers' use of external consumer data [47]. Insurers in the State are prohibited from unfairly discriminating on several variables. Unfair discrimination is defined as including

> 'the use of one or more external consumer data and information sources, as well as algorithms or predictive models using external consumer data and information sources, that have a correlation to race, color, national or ethnic origin, religion, sex, sexual orientation, disability, gender identity, or gender expression, and that use results in a **disproportionately negative outcome** for such classification or classifications, which negative outcome exceeds the reasonable correlation to the underlying insurance practice, including losses and costs for underwriting.'

This means that insurers using external consumer data in Colorado will need to provide a disparate impact analysis [33], even though it is not a synonym for unfair discrimination. The following question is: will other States follow in the same direction as Colorado?

What are they?

Forbidden variables in France are pregnancy, motherhood and genetic information in all processes. For car, term life, health insurance and annuities, gender is a forbidden variable. In the US, there are some rules that come from federal law: insurers cannot consider pre-existing health conditions or gender in the underwriting process, genetic information in coverage availability or premium charging, or housing practices that have a disparate impact on protected classes. There are no other federal laws regulating what criteria can be taken into account. Historically, States are responsible for regulating insurance discrimination. These regulations strongly depend on which State and insurance line are in question: nine States completely prohibit the use of race and national origin in all lines, 7 States religion, one State gender and five States sexual orientation. Louisiana explicitly allows the use of race for life insurance. No State completely bans the use of age, credit score, genetic testing or ZIP Code [3].

Sensitive variables in France are those defined in the anti-discrimination law that are not forbidden, as we saw in section 1.2.4. They can only be used if statistical data proves their relevance and objectivity for risk analysis. Information on ethnicity, religion, sex, gender, sexual orientation, disability, and age can be viewed as sensitive. Less obvious sensitive variables include parenthood, military service, political party, socioeconomic status, or involvement in the criminal justice system.

Depending on the jurisdiction, forbidden variables are not the same. Preventing the use of some of them can sometimes have adverse effects, for both the insurer and the insured: it can lead to adverse selection and as a consequence prices go up for all consumers, regardless of their characteristics.

1.3.2 Proxy variables

Proxies are unprotected variables that are strongly correlated to protected variables but also contain strong predictive information. Using proxy variables may result in indirect discrimination [4]. Insurers seek good segmentation of risk and profit maximization, but this might lead

to perpetuating inequalities in society if outcomes are biased because of the variables that are used for prediction.

Name and surname Accuracy of guessing the national origin of names varies significantly by the individual perception of national origin. Numerous characteristics matter, such as gender, popularity and the average level of educations of mothers who gave that name [23]. A name depends on the culture of group, trends, the social level and time. With a name, you could identify the sex and the national origin of a person. You could also estimate the average age of people who have that name based on trends and popularity, but it would be difficult to infer the exact age [15].

Address In the US, Black and Hispanic segregation and spacial isolation is still very active in some metropolitan areas [45]. This means that addresses are good proxies for ethnicity, and classifiers using this variable will exhibit discriminatory behavior [31]. The term redlining comes from the 1930s when residential security maps were created to indicate which parts f a city were safe to invest in: neighborhoods outlined in red were the riskiest [36]. From the 1960s, in the context of the struggle for Black Civil Rights, the use of redlining for risk classification was strongly criticized. The address is a non-causal variable, but it is strongly correlated with non-observable risk factors and with ethnicity [5].

There is a strong correlation between income and ethnicity and between income inequality and income segregation in the US. This is partly due to housing discrimination after World War II, forcing Black families with lower incomes to live in proximity in urban areas [43]. Addresses are therefore strongly correlated with income, which is strongly correlated with ethnicity.

Night lighting and wealth are correlated, and it is possible to approximately estimate how wealthy a neighborhood is from satellite imagery with a strong predictive power [30]. Address (and its associated satellite images) is a proxy for wealth.

With Google Street View, it is also possible to use the number and type of cars to infer wealth, ethnicity, education level and political preferences. It is also easy to detect the presence of handicap access ramps [27] or a flag indicating national origins, political preferences or sexual orientations [15].

Occupation Despite efforts for parity in the workplace, numerous occupations are still dominated by one of the two genders. In 2016 in a representative French region, sectors such as social working, healthcare and teaching are mostly feminine and industry, construction and transportation are mostly masculine. Another visible trend is that very feminine sectors tend to become even more so [18].

In 2018 in France, 18% of employees worked part-time, 78% of which were women [17]. Knowing an individual works part-time means it is three times more likely that it is a woman.

Credit score This variable cannot be measured directly, as it comes from the problem definition of creditworthiness. It is a non-arbitrary definition, not a given. The definition process can already itself be biased [4], as it relies on measurable attributes that are available. The choice of which variables to use can introduce discrimination. Credit scores also create a vicious circle in terms of poverty, and using this variable introduces a disparate impact on racial minorities and low-income households, who then have to pay a higher premium [15].

Face With the boom of facial recognition softwares, there are numerous opportunities of application. It is now possible to use facial recognition tools for health assessment by using measurements and proportions of facial attributes. They may be considered biometric data,

so there are ethical issues surrounding their use [10]. Facial recognition can accurately predict gender and ethnicity, which raises moral questions.

Speech The way an individual talks can indicate origin if he or she has difficulties with pronunciation, has a strong accent or speaks in a dialect. Linguistic profiling is the identification of an individual's ethnicity based on how their voice sounds and using that information for discrimination [48].

Current Natural Language Processing tools are trained on traditional written sources, which are different from spoken language, and even more from dialectal spoken language. The latter are more likely to be incorrectly classified, so bias can arise, with an incorrect representation of ideas and opinions from minority groups [8].

Chatbots are rule-based, information retrieval or learning-based systems that are widely used today. In March 2016, a Microsoft chatbot was supposed to improve its small-talk capabilities by learning from conversations with human users. In less than a day, it was displaying racist and sexist abusive content [46]. This shows that technical difficulties must be tackled in order to avoid such outcomes, especially when black box algorithms are used for treating voices.

In insurance, writing or speaking chatbots are used to report insurance claims. For example, Izzy Constat is a tool for amicable reports after a car accident. The chatbot asks for drivers' identities and context and generates a sketch representing the incident [13]. What if this chatbot were biased against a group that had a specific dialect? This could result in understating the severity of impact and lower benefits.

Network Who you know either gets you access to resources or makes you guilty by association. Recommendation systems are based on similarity between individuals, and if insurers had access to this kind of information, they could find customers and limit their financial risks. But this kind of practice may not be ethical, as people who are already marginalized can be even more affected [11].

There are many other variables with more or less predictive power that could be used as proxies for protected attributes. We will not be able to list them all, but the conclusion is that a study of their meaning and how they are linked to protected attributes is essential before using them as inputs in a prediction model.

Many variables can act as a proxy for a forbidden variable and their use can lead to a discriminatory decision.

1.4 How to measure fairness

Numerous metrics have been created to measure bias and fairness, the two notions not being distinct in most works. The general setting is the following:

- $X \in \mathcal{X} \subset \mathbb{R}^n$ is the set of n non-protected variables,
- $A \in \mathcal{A}$ is the protected or sensitive variable,
- $Y \in \mathcal{Y}$ is the true outcome,
- f is the predictor, classifier or regressor,
- $\hat{Y} = f(X, A) \in \mathcal{Y}$ is the predicted outcome.

1.4.1 Binary classification

In the binary classification setting, $\mathcal{Y} = \{y^-, y^+\}0$. $Y = y^-$ means the outcome is negative and $Y = y^+$ means the outcome is positive. In this problem, it is possible to compute the confusion matrix between the true and predicted outcomes for each protected group, see table 1.

		Predicted outcome	
		Positive	Negative
True outcome	Positive	TP	FN
	Negative	FP	TN

Table 1: Confusion matrix
TP: True Positive, FN: False Negative, FP: False Positive, TN: True Negative

The most common metrics in the literature are the following [2].

Statistical parity (or demographic parity) requires the likelihood of a positive outcome to be the same for all protected groups ie $\hat{Y} \perp\!\!\!\perp A$:

$$\forall a \in \mathcal{A}, \mathbb{P}(\hat{Y} = y^+ | A = a) = p$$

This is equivalent to having the same predicted acceptance rates AR for all protected groups:

$$\text{AR} = \frac{\text{TP} + \text{FP}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Equalized odds requires all protected groups to have the same probabilities of being correctly assigned a positive outcome and of being incorrectly assigned a positive outcome ie to have the same true and false positive rates ie $\hat{Y} \perp\!\!\!\perp A|Y$:

$$\forall (y, a) \in \mathcal{Y} \times \mathcal{A}, \mathbb{P}(\hat{Y} = y^+ | Y = y, A = a) = p$$

This is equivalent to having the same true positive rates TPR and false positive rates FPR for all protected groups:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$
$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

Remark: the true positive rate is also called recall or sensitivity.

Equal opportunity is the same as equalized odds but only requires all protected groups to have the same probability of being correctly assigned a positive outcome ie requires the same true positive rates TPR for all protected groups:

$$\forall a \in \mathcal{A}, \mathbb{P}(\hat{Y} = y^+ | Y = y^+, A = a) = p$$

Disparate Impact is a popular metric in the US to measure bias. It is defined as the ratio in probability of favorable outcomes between groups $A = 0$ and $A = 1$ [6]:

$$\text{DI} = \frac{\mathbb{P}(\hat{Y} = 1 | A = 0)}{\mathbb{P}(\hat{Y} = 1 | A = 1)}$$

It is a consequence of the statistical parity definition, in the case where the probabilities are non-null. A Disparate Impact of 1 would mean that the model is fair, lower than 1 that the model is unfair to group $A = 0$ and above 1 that the model is unfair to group $A = 10$. This Disparate Impact is only defined in the case of a binary classification with a binary protected variable. Its estimation is not as easy as we could think, because of its definition as a ratio: we can have robust estimators for both probabilities, but the estimator of a ratio is not the ratio of estimators. This is why we decided not to use this fairness metric and to keep the original definition of statistical parity which is not as restrictive.

Metric	Definition	Meaning
Statistical parity	$\hat{Y} \perp\!\!\!\perp A$	Same likelihood of positive outcome
Equalized odds	$\hat{Y} \perp\!\!\!\perp A \mid Y$	Same likelihood of being correctly and incorrectly assigned a positive outcome
Equal opportunity	Same true positive rates	Same likelihood of being correctly assigned a positive outcome
Disparate impact	$\frac{P(\hat{Y} = 1 \mid A = 0)}{P(\hat{Y} = 1 \mid A = 1)} = 1$	A consequence of the statistical parity definition

In the following sections of this thesis, we will focus on the statistical parity definition, as it is the most frequently used in the literature. We will compute the other two to illustrate compatibility issues and for comparison purposes.

1.4.2 Extension of the metrics to other settings

Most of these definitions can be extended to a multi-class classification ($\mathcal{Y} \subset \mathbb{N}$) or a regression ($\mathcal{Y} \subset \mathbb{R}$) setting: by defining a subset $\mathcal{Y}^+ \subset \mathcal{Y}$ (respectively $\mathcal{Y}^- \subset \mathcal{Y}$) of values reflecting a positive (respectively negative) outcome, we adapt the definitions from the previous section:

Statistical parity: $\forall a \in \mathcal{A}, \mathbb{P}(\hat{Y} \in \mathcal{Y}^+ | A = a) = p$

Equalized odds: $\forall a \in \mathcal{A}, \mathbb{P}(\hat{Y} \in \mathcal{Y}^+ | Y \in \mathcal{Y}^+, A = a) = \mathbb{P}(\hat{Y} \in \mathcal{Y}^+ | Y \in \mathcal{Y}^-, A = a) = p$

Equal opportunity: $\forall a \in \mathcal{A}, \mathbb{P}(\hat{Y} \in \mathcal{Y}^+ | Y \in \mathcal{Y}^+, A = a) = p$

This supposes that we can categorize every value of output as either positive or negative.

1.4.3 Model evaluation

We will compare models using the confusion matrix and metrics deriving from it:

- The accuracy is the proportion of correct classifications among all classifications.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

- The true positive rate is the proportion of correct positive classifications among actual positive values.

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- The false positive rate is the proportion of wrong positive classifications among actual negative values. It is sometimes called the probability of false alarm.

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

- The acceptance rate is the proportion of positive classifications among all classifications.

$$\text{Acceptance rate} = \frac{\text{TP} + \text{FP}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

In order to evaluate a model, we need to take into account the facts that:

- An acceptable accuracy threshold depends on the context of the prediction: do we need to be perfectly accurate in order for the decision to be accepted by society? Do mistakes cost a lot to the company? Furthermore, if there is a large class imbalance, accuracy is not the best metric as it can be very high while the model only fits the majority population.

Looking only at the accuracy is not enough to evaluate a model: we need to pay attention to class imbalance, to the context of the decision and to the cost of mistakes.

- Missclassification errors are more or less acceptable depending on the context. For example, in a risk selection decision, underwriters want to select 'good' risks. It is generally more acceptable and less costly to falsely reject good risks than falsely accept bad risks.

It is important to know if a high number of false positives or negatives has a large cost, and it depends on the context of the decision.

2 Simulated data

We will apply discrimination detection and correction methods to simulated data before tackling a real use case. The reason for this is that we wish to know the answer to the question 'Is my prediction discriminatory towards a group?' while knowing how the outcome was computed, so that we can look for a solution to a well-defined problem, instead of having to make assumptions.

We will begin by generating our data, which consists of two sensitive variables A and B, a set of non-protected variables $X = \{X_i\}_{i=3,...n}$, $n \in \mathbb{N}$ and a variable of interest $Y0$. All variables can be correlated with each other, but depending on the country and its regulation, insurers are not always allowed to use the sensitive variables as explanatory variables, and sometimes, with the GDPR for example, cannot even collect the information. In this section, however, we will suppose that the variable is available, because it is the only way to measure discrimination.

To link with a real-life example, if we were in a pricing context for automobile insurance, A could represent gender, B marital status, the X_i other variables such as age or car value and Y the claim occurrence. Gender is not the cause of an accident, but statistically we observe that gender and claim occurrence are correlated. Intuitively, we should not discriminate based on gender, because it would be unfair as it is a stereotype, but we do not have access to a 'fairer' variable, which could be driving behavior.

This analysis on simulated data in a controlled framework is necessary because in practice the collected information is only partial and the observed correlations do not imply causality. We have to study discrimination mechanisms with this limitation in mind. Understanding the observable effects is important so as not to draw hasty conclusions or even be biased in our analysis.

For simplification reasons, we will suppose that the protected variables and the outcome are binary. We will begin by giving the theoretical framework, then illustrate the simulation process with two variables, and finally create the dataset.

2.1 A reminder on statistical tools used to set the framework

In this section, we will pose important mathematical concepts that will help understand the observed effects on the fairness metric.

We want to generate the variables while controlling the relationship they have with each other. To do so, we will use the theory of copulas.

The goal is to generate n random variables of chosen laws while controlling their correlation with each other. This is the reason why we chose to use the theory of copulas.

Notations

First, we need to introduce some notations:

- f is a probability density function and F the associated cumulative distribution function: $F' = f$

- $\mathbb{E}$ is the expected value and Var the variance
- Σ is the covariance matrix: $\Sigma_{i,j} = cov(X_i, X_j)$
- R is the correlation matrix: $R_{i,j} = corr(X_i, X_j)$
- ϕ is the probability density function and Φ the cumulative distribution function of a standard normal distribution $\mathcal{N}(0, 1)$
- ϕ_R is the probability density function and Φ_R the cumulative distribution function of a standard multivariate normal distribution $\mathcal{N}_n(0, R)$

Probabilities [14]

Definition 1. *A $X = (X_1, \ldots, X_d)^T$ d-dimensional real random vector is a standard normal random vector if all of its components $X_i, i = 1, \ldots, d$ are independent and follow a standard normal distribution.*

Definition 2. *A $X = (X_1, \ldots, X_d)^T$ d-dimensional real random vector follows a multivariate normal distribution if there exists a random k-vector Z, which is a standard normal random vector, a d-vector μ and a $k \times d$ matrix A of full rank such that $X = AZ + \mu 0$.*
We denote $X \sim \mathcal{N}_d(\mu, \Sigma)$ where $\mu = \mathbb{E}[X] = (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_d])^T$ is the mean vector and $\Sigma = AA^T$ is the covariance matrix.

Definition 3. *The multivariate normal distribution is non-degenerate when the covariance matrix Σ is positive definite, in which case it has the following density*

$$f_X(x_1, \ldots, x_d) = \frac{exp(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))}{\sqrt{(2\pi)^k|\Sigma|}}$$

where $x = (x_1, \ldots, x_d)^T \in \mathbb{R}^d$ and $|\Sigma| = det(\Sigma)$ is the determinant of $\Sigma 0$.

Proposition 1. *For any univariate cumulative distribution function F,*

$$U \sim \mathcal{U}([0, 1]) \Rightarrow F^{-1}(U) := X \sim F$$

Definition 4. *Set X and Y two continuous random variables. X and Y are independent if and only if $F_{X,Y}(x, y) = F_X(x)F_Y(y)0$.*

Definition 5. *Set X and Y two continuous random variables with finite variance. Pearson's linear correlation coefficient is defined as*

$$corr(X, Y) = \frac{cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

Remark: Pearson's correlation coefficient is linear:

$$corr(X, Y) = \pm 1 \iff \exists\ a, b, Y = aX + b \text{ a.s.}$$

Proposition 2.

$$X \perp\!\!\!\perp Y \Rightarrow corr(X, Y) = 0$$
$$corr(X, Y) = 0 \not\Rightarrow X \perp\!\!\!\perp Y$$

Remark: This proposition means that uncorrelated data is not independent in general. Furthermore, correlation does not imply causation, meaning that we cannot deduce a cause-and-effect relationship between variables solely based on their correlation. However, in the case of a multivariate normally distributed random vector, variables that are uncorrelated are independent.

Proposition 3. *Set X and Y two random variables. For any measurable function Ψ such that $\Psi(Y)$ is square-integrable,*

$$\mathbb{E}[\Psi(Y)X] = \mathbb{E}[\Psi(Y)\mathbb{E}[X|Y]]$$

Matrices [28]

Definition 6. *A real symmetrical matrix A is positive definite (respectively semi-definite) if and only if for any non zero real column vector z, $z^T Az$ is positive (respectively non negative).*

Proposition 4. *A matrix is positive definite (resp. semi-definite) if and only if all of its eigenvalues are positive (resp. non-negative).*

Definition 7. *The Cholesky decomposition of a symmetrical real positive-definite matrix is a unique decomposition of the form $A = LL^T$ where L is a lower triangular matrix with real and positive diagonal entries.*

Proposition 5. *If A is positive definite (resp. semi-definite), it can be written as $A = LL^T$ with L a lower triangular matrix with a positive (resp. non negative) diagonal.*

Remark: This is the unique (resp. non unique) Cholesky decomposition.

Proposition 6. *If a matrix can be eigendecomposed and all its eigenvalues are non null then it is invertible.*

Copulas [14]

Copulas are used to study multi-hazard risks, ie random vectors $X = (X_1, \dots, X_n)0$. Most of the time, the marginal laws are known. Copulas are then used to get the joint law and model the dependence between variables [19].

Definition 8. *A d-dimensional copula is a cumulative distribution function $C : [0,1]^d \to [0,1]$ whose margins are uniform on $[0,1]0$.*

Theorem 1 (Sklar's theorem)**.** *For any multivariate cumulative distribution function H with marginals $F_1, \dots, F_d$, there exists a d-dimensional copula C such that*

$$H(x_1 \dots, x_d) = C(F_1(x_1), \dots, F_d(x_d))$$

If $F_1, \dots, F_d$ are all continuous, then C is unique.

Definition 9. *The independence copula is defined as*

$$C^{\perp}(u_1, \dots, u_n) = u_1 \times \cdots \times u_n$$

Remark: We will note $X^{\perp}$ a random vector which has copula $C^{\perp}0$.

Definition 10. *For some correlation matrix R, the n-dimensional Gaussian copula with parameter R is defined as*

$$C_R(u) = \Phi_R(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_n)), u \in [0,1]^n$$

Using the Gaussian copula in dimension n allows us to control the action of the parameters in the interactions between variables and thus the simulation of correlation effects. We will then be able to interpret the discrimination metric.

2.2 Simulation process

We will now see how to simulate copula models. It all relies on proposition 1. Our goal is to simulate $X = (X_1, \dots, X_n)$, which is characterized by its marginal distributions $F_1, \dots, F_n$ and copula C_{R_Z}, chosen to be the Gaussian copula. The procedure is the following:

1. Draw $Z \sim \mathcal{N}_n(0, R_Z)$
2. Compute $U = (\Phi(Z_1), \dots, \Phi(Z_n))$
3. Compute $X = (F_1^{-1}(U_1), \dots, F_n^{-1}(U_n) = (F_1^{-1}(\Phi(Z_1)), \dots, F_n^{-1}(\Phi(Z_n)))$

To simulate the wanted variables, we begin with normally distributed and correlated random variables. We then transform then into uniform variables. Finally, we apply the inverse distribution of the desired laws. We end up with variables of the desired distributions that are correlated with each other.

2.2.1 Illustration in dimension 2

We will look at the results in dimension n=2 before creating our final dataset. This will allow us to give a visual illustration. First, we generate $Z = (Z_1, Z_2) \sim \mathcal{N}_2(0, R)$ with $R = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ and $\rho = corr(Z_1, Z_2)0$. Then, we compute $X = (X_1, X_2) = \big(F_1^{-1}(\Phi(Z_1)), F_2^{-1}(\Phi(Z_2))\big)0$.

The correlation between the Z_i is

$$corr(Z_1, Z_2) = \rho = \frac{\mathbb{E}(Z_1 Z_2) - \mathbb{E}(Z_1)\mathbb{E}(Z_2)}{\sqrt{Var(Z_1)Var(Z_2)}} = \mathbb{E}(Z_1 Z_2)$$

because $Z_i \sim \mathcal{N}(0, 1)$, $i = 1, 20$. And the correlation between the X_i is

$$corr(X_1, X_2) = \frac{\mathbb{E}(X_1 X_2) - \mathbb{E}(X_1)\mathbb{E}(X_2)}{\sqrt{Var(X_1)Var(X_2)}}$$

First case: $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ with $\mu_i \in \mathbb{R}, \sigma_i^2 \in \mathbb{R}_*^+, i = 1, 2$
Then

$$F_i^{-1}(p) = \mu_i + \sigma_i \Phi^{-1}(p), p \in [0, 1], i = 1, 2$$

In order to find the correlation between the X_i, we need to compute:

$$\begin{aligned}
\mathbb{E}(X_1 X_2) &= \mathbb{E}[F_1^{-1}(\Phi(Z_1))F_2^{-1}(\Phi(Z_2))] \\
&= \mathbb{E}[(\mu_1 + \sigma_1 \Phi^{-1}(\Phi(Z_1)))(\mu_2 + \sigma_2 \Phi^{-1}(\Phi(Z_2)))] \\
&= \mathbb{E}[\mu_1\mu_2 + \mu_1\sigma_2 Z_2 + \mu_2\sigma_1 Z_1 + \sigma_1\sigma_2 Z_1 Z_2] \\
&= \mu_1\mu_2 + \mu_1\sigma_2\mathbb{E}(Z_2) + \mu_2\sigma_1\mathbb{E}(Z_1) + \sigma_1\sigma_2\mathbb{E}(Z_1 Z_2) \\
&= \mu_1\mu_2 + \sigma_1\sigma_2\rho
\end{aligned}$$

And so,

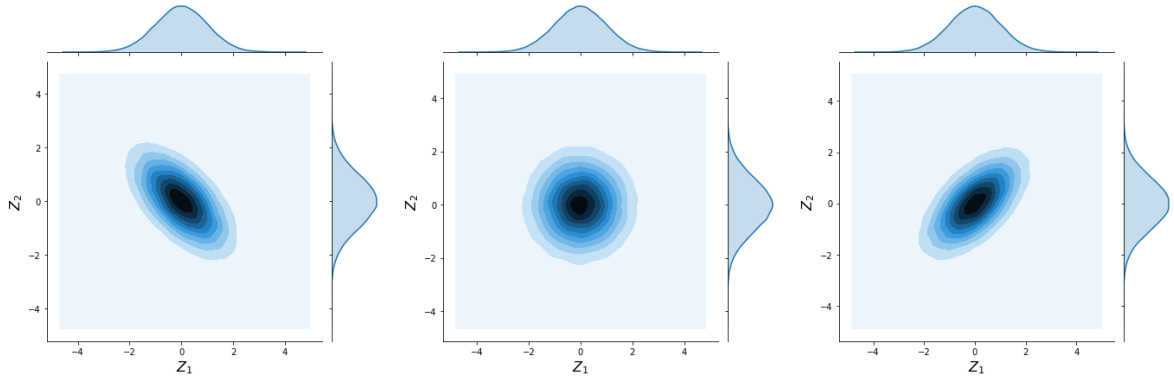
$$corr(X_1, X_2) = \frac{\mu_1\mu_2 + \sigma_1\sigma_2\rho - \mu_1\mu_2}{\sigma_1\sigma_2} = \rho \tag{1}$$

Illustration: We will take $X_1 \sim \mathcal{N}(0, 2)$ and $X_2 \sim \mathcal{N}(3, 1)0$.

As specified by step 1, we first draw

$$Z \sim \mathcal{N}_2\Big(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\Big)$$

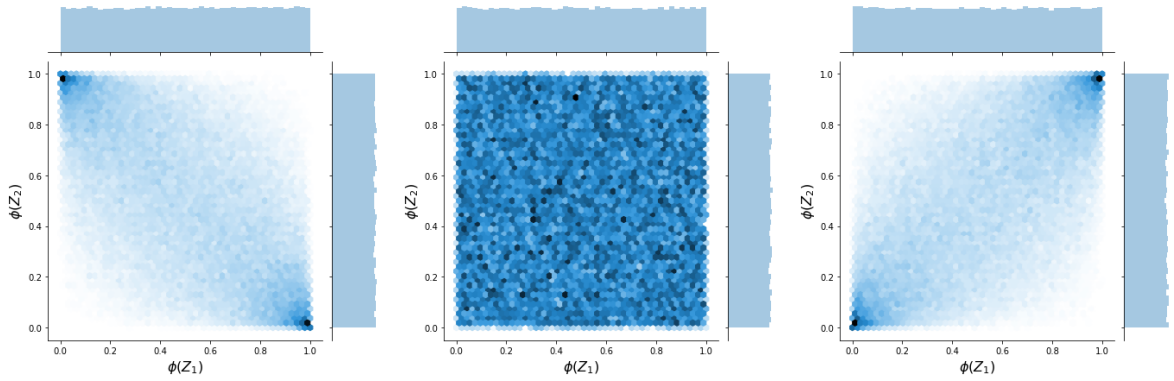
of size $2 \times 100,0000$. We will follow the rest of the simulation process for three values of ρ: -0.6, 0 and 0.6, so as to compare the results for different correlations. Figure 7 represents the joint distribution of $(Z_1, Z_2)0$. We see that increasing the degree of correlation positively (resp. negatively) between the marginal distributions concentrates the joint distribution around the line $y = x$ (resp. $y = -x$).



(a) $\rho = -0.6$ (b) $\rho = 0$ (c) $\rho = 0.6$

Figure 7: Step 1: joint distributions of (Z_1, Z_2)

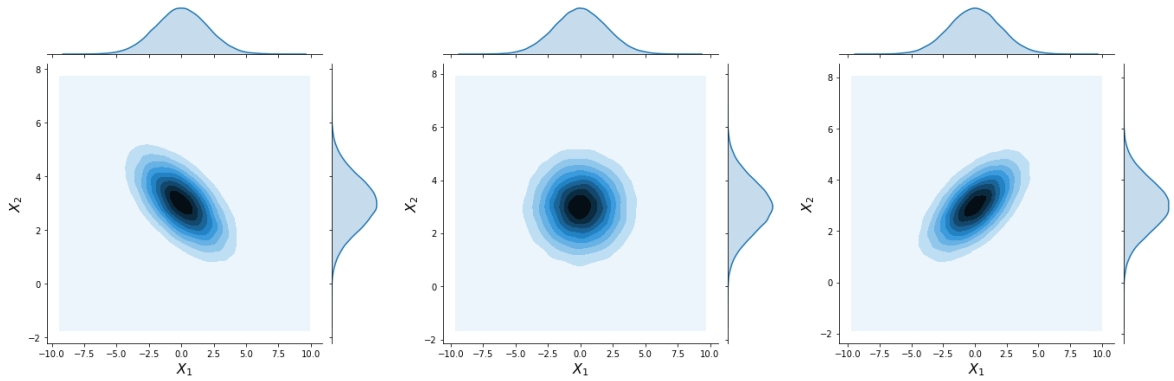
We then apply Φ to Z_1 and Z_2 (step 2) and get their joint cumulative distribution, as plotted in figure 8. As $Z_i \sim \mathcal{N}(0, 1)$, $\Phi(Z_i) \sim \mathcal{U}([0, 1])$, $i = 1, 2$, which is another formulation of proposition 1. So we get two uniform random variables that are correlated. Once again, the joint cumulative distribution concentrate around the line $y = \pm x$ when the correlation varies.



(a) $\rho = -0.6$ (b) $\rho = 0$ (c) $\rho = 0.6$

Figure 8: Step 2: joint cumulative distributions of (Z_1, Z_2)

We finally apply F_i^{-1} to $\Phi(Z_i)$ to obtain X_i, $i = 1, 2$ (step 3), as plotted in figure 9. We have the same conclusion as before on the correlation. We are well-aware that we could have directly generated a multivariate normal vector with the desired expected values and standard deviations, but this way allowed us to give an easy example of how to follow the simulation process.



(a) $\rho = -0.6$ (b) $\rho = 0$ (c) $\rho = 0.6$

Figure 9: Step 3: joint distributions of (X_1, X_2)

We wrote down the values of the estimated Pearson correlation coefficients in table 2. As computed in equation 1, we have $corr(X_1, X_2) = corr(Z_1, Z_2)$ for the three values of $\rho 0$.

ρ	-0.6	0	0.6
$corr(Z_1, Z_2)$	-0.6018	-0.0009	0.6005
$corr(X_1, X_2)$	-0.6018	-0.0009	0.6005

Table 2: $corr(Z_1, Z_2)$ and $corr(X_1, X_2)$

Intuitively enough, if we want normally distributed variables, the correlation is the same as for the first step standard multivariate normal vector:
$corr(X_1, X_2) = corr(Z_1, Z_2)$ when $(Z_1, Z_2) \sim N_2(0, R)$ and $X_i \sim N(\mu_i, {\sigma_i}^2)$

Second case: $X_1 \sim \mathcal{U}([a, b])$ and $X_2 \sim \mathcal{N}(\mu, \sigma^2)$ with $(a, b) \in \mathbb{R}^2, a < b, \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_*^+$
The computation of $corr(X_1, X_2)$ is not as direct as in the first case. As X_1 has a uniform distribution, we have

$$\mathbb{E}(X_1) = \frac{a+b}{2}$$
$$Var(X_1) = \frac{(b-a)^2}{12}$$
$$F_1^{-1}(p) = a + p(b-a), p \in (0, 1)$$

We can compute

$$\begin{aligned}
\mathbb{E}(X_1 X_2) &= \mathbb{E}[F_1^{-1}(\Phi(Z_1))F_2^{-1}(\Phi(Z_2))] \\
&= \mathbb{E}[(a + (b-a)\Phi(Z_1))(\mu + \sigma Z_2)] \\
&= a\mu + a\sigma\mathbb{E}(Z_2) + (b-a)\mu\mathbb{E}[\Phi(Z_1)] + (b-a)\sigma\mathbb{E}[\Phi(Z_1)Z_2] \\
&= a\mu + (b-a)\frac{\mu}{2} + (b-a)\sigma\mathbb{E}[\Phi(Z_1)Z_2] \\
&= \frac{a+b}{2}\mu + (b-a)\sigma\mathbb{E}[\Phi(Z_1)Z_2]
\end{aligned}$$

$\mathbb{E}[\Phi(Z_1)] = \frac{1}{2}$ because $Z_1 \sim \mathcal{N}(0, 1)$ and so $\Phi(Z_1) \sim \mathcal{U}([0, 1])$ (consequence of proposition 1).

So we have

$$\begin{aligned} corr(X_1, X_2) &= \frac{\mathbb{E}(X_1X_2) - \mathbb{E}(X_1)\mathbb{E}(X_2)}{\sqrt{Var(X_1)Var(X_2)}} \\ &= \frac{\frac{a+b}{2}\mu + (b-a)\sigma\mathbb{E}[\Phi(Z_1)Z_2] - \frac{a+b}{2}\mu}{\frac{b-a}{\sqrt{12}}\sigma} \\ &= 2\sqrt{3}\mathbb{E}[\Phi(Z_1)Z_2] \\ &= 2\sqrt{3}\mathbb{E}[\Phi(Z_1)\mathbb{E}[Z_2|Z_1]] \end{aligned}$$

because of proposition 3, with $\Phi(Z_1)$ square-integrable.
And $\mathbb{E}[Z_2|Z_1] = \rho Z_1$, because

$$Z_2|Z_1 = z_1 \sim \mathcal{N}(\rho z_1, 1-\rho^2) \tag{2}$$

Proof.

$$\begin{aligned} f_{Z_2|Z_1=z_1}(z_1, z2) &= \frac{f_{Z_1,Z_2}(z_1, z_2)}{f_{Z_1}(z_1)} \\ &= \frac{\frac{1}{2\pi\sqrt{1-\rho^2}}exp(-\frac{1}{2(1-\rho)^2}(z_1^2 - 2\rho z_1 z_2 + z_2^2))}{\frac{1}{\sqrt{2\pi}}exp(-\frac{z_1^2}{2})} \\ &= \frac{1}{\sqrt{2\pi(1-\rho^2)}}exp(-\frac{1}{2(1-\rho^2)}(z_1^2 - 2\rho z_1 z_2 + z_2^2 - (1-\rho^2)z_1^2)) \\ &= \frac{1}{\sqrt{2\pi(1-\rho^2)}}exp(-\frac{1}{2(1-\rho^2)}(z_2^2 - 2\rho z_1 z_2 + (\rho z_1)^2)) \\ &= \frac{1}{\sqrt{2\pi(1-\rho^2)}}exp(-\frac{(z_2-\rho z_1)^2}{2(1-\rho^2)}) \end{aligned}$$

By identification,

$$\begin{aligned} \mathbb{E}[Z_2|Z_1 = z_1] &= \rho z_1 \\ Var(Z_2|Z_1 = z_1) &= 1 - \rho^2 \end{aligned}$$

□

So we have

$$\begin{aligned} corr(X_1, X_2) &= 2\sqrt{3}\mathbb{E}[\Phi(Z_1)\rho Z_1] \\ &= 2\sqrt{3}\rho\mathbb{E}[\Phi(Z_1)Z_1] \\ &= 2\sqrt{3}\rho\int_{\mathbb{R}} z\Phi(z)\phi(z)\,dz \end{aligned}$$

We will do an integration by parts:

$$\int_a^b u'(x)v(x)\,dx = [u(x)v(x)]_a^b - \int_a^b u(x)v'(x)\,dx$$

We can set

$$\begin{aligned} u'(x) &= x\phi(x) = x\frac{e^{-x^2/2}}{\sqrt{2\pi}} \\ v(x) &= \Phi(x) \end{aligned}$$

So

$$u(x) = -\phi(x) = -\frac{e^{-x^2/2}}{\sqrt{2\pi}}$$
$$v'(x) = \Phi'(x) = \phi(x)$$

So we have

$$corr(X_1, X_2) = 2\sqrt{3}\rho\big[[-\phi(z)\Phi(z)]_{\mathbb{R}} - \int_{\mathbb{R}} -\phi^2(z)\,dz\big]$$

$\lim_{z\to-\infty}\Phi(z) = 0$ and $\lim_{z\to+\infty}\Phi(z) = 1$ because Φ is a cumulative distribution function
$\phi(z) = \frac{e^{-z^2/2}}{\sqrt{2\pi}}$ so $\lim_{z\to-\infty}\phi(z) = \lim_{z\to+\infty}\phi(z) = 0$
So we have $[-\phi(z)\Phi(z)]_{\mathbb{R}} = 0$ and

$$\begin{aligned}\int_{\mathbb{R}} \phi^2(z)\,dz &= \int_{\mathbb{R}} \frac{e^{-z^2}}{2\pi}\,dz \\ &= \int_{\mathbb{R}} \frac{e^{-y^2/2}}{2\pi\sqrt{2}}\,dy \\ &= \frac{1}{2\sqrt{\pi}}\end{aligned}$$

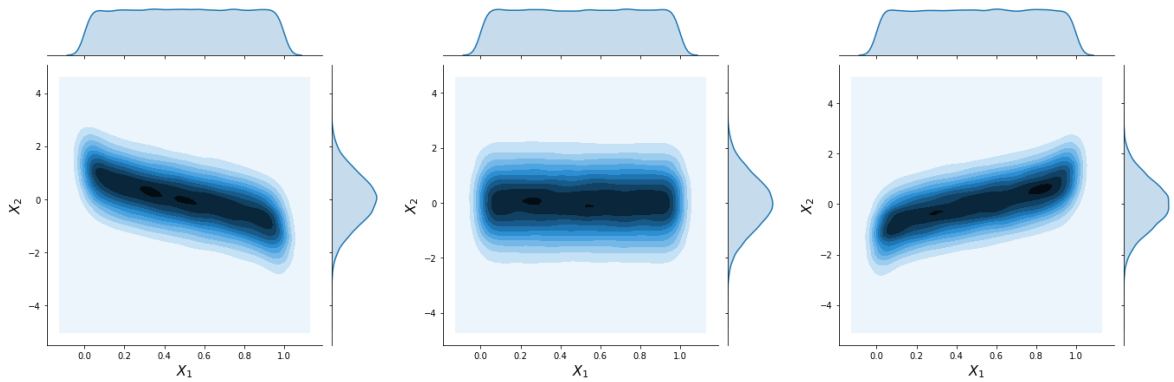
because $\int_{\mathbb{R}} \frac{e^{-y^2/2}}{\sqrt{2\pi}}\,dy = 1$, as it is the integral on the entire support ($\mathbb{R}$) of a probability density function (of the standard normal distribution). In the end,

$$corr(X_1, X_2) = 2\sqrt{3}\rho\frac{1}{2\sqrt{\pi}}$$

$$corr(X_1, X_2) = \sqrt{\frac{3}{\pi}}\rho \tag{3}$$

To conclude, as $\sqrt{\frac{3}{\pi}} \simeq 0.997$, we almost keep the same correlation between X_1 and X_2 as between Z_1 and Z_2, in the case where $X_1 \sim \mathcal{U}([a, b])$ and $X_2 \sim \mathcal{N}(\mu, \sigma^2)0$.

Illustration: We will take $X_1 \sim \mathcal{U}([0, 1])$ and $X_2 \sim \mathcal{N}(0, 1)0$. The joint distributions are plotted in figure 10. Like in the previous case, increasing the degree of correlation positively (respectively negatively) between the marginal distributions shifts the joint distribution on the line $y = x$ (respectively $y = -x$).



(a) $\rho = -0.6$ (b) $\rho = 0$ (c) $\rho = 0.6$

Figure 10: Joint distributions of $(X_1,\, X_2) = \big(F_1^{-1}(\Phi(Z_1)), F_2^{-1}(\Phi(Z_2))\big)$

We computed the estimated Pearson correlation coefficients between the simulated Z_1 and Z_2, and X_1 and X_2 in table 3. As found with equation 3, $corr(X_1, X_2) = \sqrt{\frac{3}{\pi}}\rho$ for the three values of $\rho 0$.

ρ	-0.6	0	0.6
$corr(Z_1, Z_2)$	-0.5983	-0.0031	0.6006
$\sqrt{3/\pi}\rho$	-0.586	0	0.586
$corr(X_1, X_2)$	-0.5838	-0.0029	0.5874

Table 3: $corr(X_1, X_2)$

If one of the variables is instead uniformly distributed, the correlation is not preserved but is proportional:

$$corr(X_1, X_2) = \sqrt{\frac{3}{\pi}} corr(Z_1, Z_2)$$

when $(Z_1, Z_2) \sim N_2(0, R)$, $X_1 \sim U([a, b])$ and $X_2 \sim N(\mu, \sigma^2)$

Third case: $X_1 \sim \mathcal{B}(p)$ and $X_2 \sim \mathcal{N}(\mu, \sigma^2)$, $p \in (0, 1), \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_*^+$
As F_1 is not continuous, we are not under the assumptions of Sklar's theorem (theorem 1), so the copula C might not be unique.

We will place ourselves in the second case and simulate $X_1 \sim \mathcal{U}([0, 1])$ and $X_2 \sim \mathcal{N}(\mu, \sigma^2)$ with $\mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_*^+$, then transform X_1 into a Bernoulli random variable using the function

$$\begin{aligned} h :&[0, 1] \to \{0, 1\} \\ u \mapsto h(u) &= \begin{cases} 0 & \text{if } u < \tau \\ 1 & \text{if } u \geq \tau \end{cases} = \mathbb{1}_{u \geq \tau} \end{aligned}$$

We then have $h(X_1) \sim \mathcal{B}(p) 0$.

Proof.

$$\mathbb{P}(h(X_1) = 1) = \mathbb{P}(\mathbb{1}_{X_1 \geq \tau} = 1) = \mathbb{P}(X_1 \geq \tau) = 1 - F_1(\tau) = 1 - \tau$$
$$\mathbb{P}(h(X_1) = 0) = \mathbb{P}(\mathbb{1}_{X_1 \geq \tau} = 0) = 1 - \mathbb{P}(\mathbb{1}_{X_1 \geq \tau} = 1) = \tau$$

By identification, $h(X_1) \sim \mathcal{B}(p)$ with $p = 1 - \tau$ □

As the cumulative distribution function of a Bernoulli random variable is not invertible, we cannot go back to the generative copula from the set $(h(X_1), X_2) 0$. In a sense, we have broken the correlation structure. But our transformation h allows the computation of the correlation as a function of ρ:

$$corr(h(X_1), X_2) = \frac{\mathbb{E}[h(X_1)X_2] - \mathbb{E}[h(X_1)]\mathbb{E}[X_2]}{\sqrt{Var(h(X_1))Var(X_2)}}$$

with

$$\begin{aligned} \mathbb{E}[h(X_1)X_2] &= \mathbb{E}[h(X_1)F_2^{-1}(\Phi(Z_2))] \\ &= \mathbb{E}[h(X_1)(\mu + \sigma Z_2)] \\ &= \mu\mathbb{E}[h(X_1)] + \sigma\mathbb{E}[h(X_1)Z_2] \end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}[h(X_1)Z_2] &= \mathbb{E}[\mathbb{1}_{X_1\geq\tau}Z_2] \\
&= \mathbb{E}[\mathbb{1}_{F_1^{-1}(\Phi(Z_1))\geq\tau}Z_2] \\
&= \mathbb{E}[\mathbb{1}_{Z_1\geq\Phi^{-1}(F_1(\tau))}Z_2] \\
&= \mathbb{E}[\mathbb{1}_{Z_1\geq\Phi^{-1}(\tau)}Z_2] \\
&= \mathbb{E}[\mathbb{1}_{Z_1\geq\Phi^{-1}(\tau)}\mathbb{E}[Z_2|Z_1]] \text{ (proposition 3)} \\
&= \mathbb{E}[\mathbb{1}_{Z_1\geq\Phi^{-1}(\tau)}\rho Z_1] \text{ (equation 2)} \\
&= \rho\int_{\mathbb{R}}\mathbb{1}_{z\geq\Phi^{-1}(\tau)}z\phi(z)\,dz \\
&= \rho\int_{\Phi^{-1}(\tau)}^{+\infty}z\phi(z)\,dz \\
&= \rho[-\phi(z)]_{\Phi^{-1}(\tau)}^{+\infty} \\
&= \rho\phi(\Phi^{-1}(\tau))
\end{aligned}$$

So

$$\begin{aligned}
corr(h(X_1),X_2) &= \frac{\mu\mathbb{E}[h(X_1)]+\sigma\rho\phi(\Phi^{-1}(\tau))-\mathbb{E}[h(X_1)]\mu}{\sqrt{\tau(1-\tau)\sigma^2}} \\
&= \frac{\rho\phi(\Phi^{-1}(\tau))}{\sqrt{\tau(1-\tau)}} \\
&= \frac{\rho\frac{1}{\sqrt{2\pi}}e^{-\frac{(\Phi^{-1}(\tau))^2}{2}}}{\sqrt{\tau(1-\tau)}}
\end{aligned}$$

So we have

$$corr(h(X_1),X_2) = \frac{\rho e^{-\frac{(\Phi^{-1}(\tau))^2}{2}}}{\sqrt{\tau(1-\tau)2\pi}} \tag{4}$$

As the correlation coefficient is in $[-,1]$,

$$\begin{aligned}
\min_{\rho\in(-1,1)} corr(h(X_1),X_2) &= \min_{\rho\in(-1,1)}\frac{\rho e^{-\frac{(\Phi^{-1}(\tau))^2}{2}}}{\sqrt{\tau(1-\tau)2\pi}} = -\frac{e^{-\frac{(\Phi^{-1}(\tau))^2}{2}}}{\sqrt{\tau(1-\tau)2\pi}} \\
\max_{\rho\in(-1,1)} corr(h(X_1),X_2) &= \max_{\rho\in(-1,1)}\frac{\rho e^{-\frac{(\Phi^{-1}(\tau))^2}{2}}}{\sqrt{\tau(1-\tau)2\pi}} = \frac{e^{-\frac{(\Phi^{-1}(\tau))^2}{2}}}{\sqrt{\tau(1-\tau)2\pi}}
\end{aligned}$$

So $corr(h(X_1),X_2)$ has a minimal and maximal value:

$$corr(h(X_1),X_2)\in\Big[-\frac{e^{-\frac{(\Phi^{-1}(\tau))^2}{2}}}{\sqrt{\tau(1-\tau)2\pi}},\frac{e^{-\frac{(\Phi^{-1}(\tau))^2}{2}}}{\sqrt{\tau(1-\tau)2\pi}}\Big]$$

Figure 11 represents $\frac{e^{-\frac{(\Phi^{-1}(\tau))^2}{2}}}{\sqrt{\tau(1-\tau)2\pi}}$ for $\tau\in(0,1)0$.

$$\max_{\tau\in(0,1)}\Big(\frac{e^{-\frac{(\Phi^{-1}(\tau))^2}{2}}}{\sqrt{\tau(1-\tau)2\pi}}\Big)=0.798 \text{ for } \tau=\frac{1}{2}$$

So for all $\tau\in(0,1), |corr(h(X_1),X_2)|\leq 0.7980$.

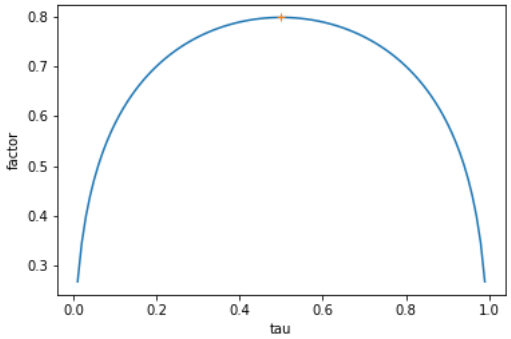


Figure 11: $corr(h(X_1), X_2)$ for $\tau \in (0, 1)$, with a maximum of 0.798 for $\tau = 0.5$

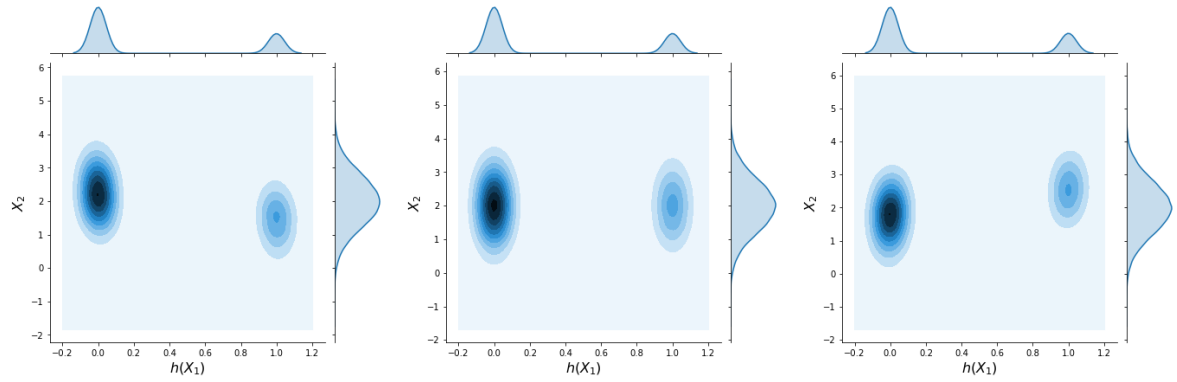
Illustration: We will take $X_1 \sim \mathcal{B}(0.8)$ and $X_2 \sim \mathcal{N}(2, 0.8)0$.
Then $\Phi^{-1}(\tau) = \Phi^{-1}(1 - 0.8) = \Phi^{-1}(0.2) \simeq -0.842$ (numeric computation) and

$$\frac{e^{-\frac{(\Phi^{-1}(\tau))^2}{2}}}{\sqrt{\tau(1-\tau)2\pi}} = \frac{e^{-\frac{(-0.842)^2}{2}}}{\sqrt{(1-0.8)\times 0.8\times 2\pi}} \simeq 0.759$$

So

$$corr(h(X_1), X_2) = 0.759\rho \in (-0.759, 0.759)$$

The joint distributions are plotted in figure 12. Like in the previous cases, increasing the degree of correlation positively (respectively negatively) between the marginal distributions shifts the joint distribution on the line $y = x$ (respectively $y = -x$).



(a) $\rho = -0.6$ (b) $\rho = 0$ (c) $\rho = 0.6$

Figure 12: Joint distributions of $(h(X_1), X_2)$

We computed the estimated Pearson correlation coefficients between the simulated $h(X_1)$ and X_2 in table 4. We can see that $corr(h(X_1), X_2) = \frac{\rho e^{-\frac{(\Phi^{-1}(\tau))^2}{2}}}{\sqrt{\tau(1-\tau)2\pi}}$ for the three values of ρ, as we computed in equation 4.

ρ	-0.6	0	0.6
$\frac{\rho e^{-\frac{(\Phi^{-1}(\tau))^2}{2}}}{\sqrt{\tau(1-\tau)2\pi}}$	-0.455	-0.001	0.455
$corr(h(X_1), X_2)$	-0.454	0.000	0.455

Table 4: $corr(h(X_1), X_2)$

For this last case, we look at a Bernoulli distributed variable, which will be often the case for the sensitive variables, for example for the sex of an insured. In this case, the correlation is proportional to the input correlation of the Gaussian vector, but also depends on the parameter of the Bernoulli law.

2.2.2 Illustration in dimension n: creating the dataset

We can now create the simulated dataset. For simplicity reasons, we will suppose that we are in the setting of section 1.4.1, with two binary protected variables and a binary outcome. We have

- a set of non-protected variables $\boldsymbol{X} = \{X^{(i)}\}_{i=1,...n-3}$, $n \in \mathbb{N}$ with $X^{(i)} \sim \mathcal{N}(\mu_i, \sigma_i^2)$, $\mu_i \in \mathbb{R}$, $\sigma_i \in \mathbb{R}_*^+$
- two protected variables $A \sim \mathcal{B}(p_a), p_a \in (0,1)$ and $B \sim \mathcal{B}(p_b), p_b \in (0,1)$
- the output variable $Y \sim \mathcal{B}(p_y), p_y \in (0,1)$
- $\boldsymbol{Z} \sim \mathcal{N}_n(0, R_Z)0$. For $(i,j) \in [\![1,n]\!]^2$,

$$R_{Z,ij} = \begin{cases} 1 & \text{if } i = j \\ corr(Z_i, Z_j) & \text{if } i \neq j \end{cases}$$

with $corr(X^{(i)}, X^{(j)}) = corr(Z_i, Z_j)$, $(i,j) \in [\![1, n-3]\!]$ (equation 1)

$$corr(X^{(i)}, A) = \frac{e^{-\frac{(\Phi^{-1}(1-p_a))^2}{2}}}{\sqrt{p_a(1-p_a)2\pi}} corr(Z_i, Z_{n-2}), i \in [\![1, n-3]\!] \text{ (equation 3)}$$

$$corr(X^{(i)}, B) = \frac{e^{-\frac{(\Phi^{-1}(1-p_b))^2}{2}}}{\sqrt{p_b(1-p_b)2\pi}} corr(Z_i, Z_{n-1}), i \in [\![1, n-3]\!] \text{ (equation 3)}$$

$$corr(X^{(i)}, Y) = \frac{e^{-\frac{(\Phi^{-1}(1-p_y))^2}{2}}}{\sqrt{p_y(1-p_y)2\pi}} corr(Z_i, Z_n), i \in [\![1, n-3]\!] \text{ (equation 3)}$$

For the choice of R_Z, there are constraints. It must be a square matrix and:

- symmetrical, because $corr(Z_i, Z_j) = corr(Z_j, Z_i)$
- with a unity diagonal, because $corr(Z_i, Z_i) = 1$
- all values must be in $[-1, 1]$ by definition of the correlation
- positive semi-definite
- invertible, otherwise the distribution is degenerate and does not have a density, as we saw in definition 3

In order to have an invertible and positive semi-definite matrix, it needs to be positive definite. Indeed, proposition 6 gives the equivalence between an invertible matrix and its eigenvalues being non null, and proposition 4 gives the equivalence between a positive definite (respectively semi-definite) matrix and its eigenvalues being positive (respectively positive or null).

Using proposition 5 (Cholesky decomposition), we will compute $R_Z = LL^T$, choosing L as a lower triangular matrix with a positive diagonal. That way we will have a square symmetrical, positive definite (so positive semi-definite and invertible) matrix.

$$\begin{aligned} R_Z &= \begin{bmatrix} L_{11} & 0 & 0 & \dots \\ L_{21} & L_{22} & 0 & \dots \\ \vdots & & \ddots & \end{bmatrix} \times \begin{bmatrix} L_{11} & L_{21} & \dots \\ 0 & L_{22} & \dots \\ \vdots & 0 & \ddots \end{bmatrix} \\ &= \begin{bmatrix} L_{11}^2 & L_{11}L_{21} & \dots \\ L_{11}L_{21} & L_{21}^2 + L_{22}^2 & \\ \vdots & & \ddots \end{bmatrix} \end{aligned}$$

We have

$$\begin{aligned} R_{Z,ii} &= \sum_{k=1}^{i} L_{ik}^2 \\ R_{Z,ij} = R_{Z,ji} &= \sum_{k=1}^{i} L_{ik}L_{jk}, i > j \end{aligned}$$

As we want the diagonal of R_Z to be unity, we need

$$\begin{aligned} R_{Z,ii} = 1 \text{ ie } &\sum_{k=1}^{i} L_{ik}^2 = 1 \\ \text{ie } L_{ii} &= \sqrt{1 - \sum_{k=1}^{i-1} L_{ik}^2} \text{ (the diagonal has to be positive)} \end{aligned}$$

The computation of the L_{ii} implies constraints on the values of the L_{ij}:

$$\forall i = 1, \dots, n, 1 - \sum_{k=1}^{i-1} L_{ik}^2 > 0 \text{ ie } \sum_{k=1}^{i-1} L_{ik}^2 < 1$$

Finally, to set R_Z, we set values for the $L_{ij}, i > j$, check that the constraint above is verified, compute the L_{ii} accordingly and then $R_Z 0$. We also need to check that we have $R_{Z,ij} \in [-1, 1]0$.

Finally, the procedure is:

- set the matrix L with the constraints mentioned above
- compute $R_Z = LL^T$
- generate $Z \sim \mathcal{N}_n(0, R_Z)$
- set the parameters of the laws of

  $X^{(i)} \sim \mathcal{N}(\mu_i, \sigma_i^2)$
  $A \sim \mathcal{B}(p_a)$
  $B \sim \mathcal{B}(p_b)$
  $Y \sim \mathcal{B}(p_y)$

- compute

$$\boldsymbol{X} = (F_1^{-1}(\Phi(Z_1)), \dots, F_{n-3}^{-1}(\Phi(Z_{n-3})))$$
$$A = h_a(F_{n-3}^{-1}(\Phi(Z_{n-3})))$$
$$B = h_b(F_{n-2}^{-1}(\Phi(Z_{n-2})))$$
$$Y = h_y(F_n^{-1}(\Phi(Z_n)))$$

Reminder: $h_a(u) = \mathbb{1}_{u \geq 1-p_a}$

Final dataset We set $n = 7$ and generated 100 datasets with the same parameters:

$$X^{(1)} \sim \mathcal{N}(2, 0.6)$$
$$X^{(2)} \sim \mathcal{N}(0.2, 0.3)$$
$$X^{(3)} \sim \mathcal{N}(-0.3, 2)$$
$$X^{(4)} \sim \mathcal{N}(0.7, 0.4)$$
$$A \sim \mathcal{B}(0.3)$$
$$B \sim \mathcal{B}(0.9)$$
$$Y \sim \mathcal{B}(0.2)$$

$$R_{X,A,B,Y} = \begin{bmatrix} 1 & 0.395 & -0.018 & 0.297 & -0.230 & 0.350 & 0.139 \\ 0.395 & 1 & -0.501 & 0.103 & 0.226 & 0.111 & 0.209 \\ -0.018 & -0.501 & 1 & 0.294 & 0.076 & 0.294 & -0.066 \\ 0.297 & 0.103 & 0.294 & 1 & -0.227 & 0.348 & -0.208 \\ -0.230 & 0.226 & 0.076 & -0.227 & 1 & 0.043 & 0.105 \\ 0.350 & 0.111 & 0.294 & 0.348 & 0.043 & 1 & 0.039 \\ 0.139 & 0.209 & -0.066 & -0.208 & 0.105 & 0.039 & 1 \end{bmatrix}$$

The reason for that is that we want to take into account the instability of results. Instability can come from the generative process: not all datasets will have the exact same variable distribution because of the limited sample size (100,00 lines here). It can also be caused later on by the train test split or sampling which depend strongly on the execution. Generating 100 datasets allows us to average and get confidence intervals on metrics and results. In reality, we often do not have access to the data generator so, to construct confidence intervals, we use bootstraping, which estimates the sampling distribution of statistics such as sample mean thanks to random sampling with replacement. Figure 13 shows the head of one of the datasets.

	X1	X2	X3	X4	A	B	Y
0	2.988190	0.075263	-1.141215	0.227668	0	1	0
1	1.355945	0.156263	-0.605319	0.285155	0	0	0
2	2.027057	0.082588	-0.280549	0.692738	0	1	0
3	2.203643	0.131036	-1.817976	0.454448	0	1	0
4	1.955881	0.384338	-2.316403	0.433489	1	1	1

Figure 13: Head of a simulated dataset

As mentioned previously, we will take advantage of having access to the data generation process. It means we can produce confidence intervals for every value: the mean of a variable, the number of observations of a certain class, the weights of the regression etc. This is the reason why we generated 100 datasets: we will look at the average and standard deviation over

these datasets to produce the confidence interval of the value we are looking at. We gave a reminder on confidence intervals in appendix A.

The simulated dataset consists of two Bernoulli sensitive variables, 4 normal non-sensitive variables and one Bernoulli variable of interest. We control their relationships through the input correlation matrix of the first step multivariate standard normal vector. We simulated 100 datasets of this format so as to compute confidence intervals when computing fairness metrics.

2.3 Descriptive statistics

Before heading into applying methods, we need to prepare and explore our data. As we have built our datasets ourselves, we already know a lot about them.

2.3.1 Variable identification and univariate analysis

The explanatory variables are the $X^{(i)}, i = 1, \dots, 4$, A and B, and the target variable is $Y0$. As expected, we have 100,000 non-null values for each variable. The $X^{(i)}$ are continuous, and A, B and Y are categorical, taking values in $\{0, 1\}0$.

X By construction, the $X^{(i)}$ have means and standard derivations as defined in section 2.2.2, and it is verified by the computation of the sample means and standard derivations in table 5. We can notice that the confidence intervals are of size 0. This means that over our 100 datasets, we have a probability of 95% that they all have the means and standard deviations as defined in data generation process.

i	1	2	3	4
mean	2.00 ± 0.00	0.20 ± 0.00	-0.30 ± 0.00	0.70 ± 0.00
std	0.60 ± 0.00	0.30 ± 0.00	2.00 ± 0.00	0.40 ± 0.00

Table 5: Means and standard derivations of the $X^{(i)}$

A We computed it to follow a Bernoulli distribution of parameter $p_a = 0.3$, so we find as planned that

$$\begin{aligned} \text{mean(A)} &= 0.30 \pm 0.00 = p_a \\ \text{std(A)} &= 0.46 \pm 0.00 = \sqrt{0.3(1-0.3)} = \sqrt{p_a(1-p_a)} \end{aligned}$$

Table 6 gives the number of observations for each value of $A0$. By construction, as $A \sim \mathcal{B}(0.3)$, there is an imbalance: about 30% individuals have $A = 1$ and 70% individuals have $A = 00$. So the group imbalance ratio is

$$IR_A = \frac{\text{Number of majority observations}}{\text{Number of minority observations}} = 2.33 \pm 0.00$$

meaning there are 2.33 times more observations of $A = 0$ than $A = 10$.

A	Observations
0	$70,002.22 \pm 290.73$
1	$29,997.78 \pm 290.73$

Table 6: Number of observations by value of A

B We computed it to follow a Bernoulli distribution of parameter $p_b = 0.9$, so we find as planned that

$$\text{mean(B)} = 0.90 \pm 0.00 = p_a$$
$$\text{std(B)} = 0.30 \pm 0.00 = \sqrt{0.9(1-0.9)} = \sqrt{p_b(1-p_b)}$$

Table 7 gives the number of observations for each value of $B0$. By construction, as $B \sim \mathcal{B}(0.9)$, there is an imbalance: about 90% individuals have $B = 1$ and 10% individuals have $B = 00$. So the group imbalance ratio is

$$IR_A = 8.99 \pm 0.02$$

meaning there are almost 9 times more observations of $B = 0$ than $B = 10$.

B	Observations
0	$89,980.70 \pm 207.11$
1	$10,019.30 \pm 207.11$

Table 7: Number of observations by value of B

Y We computed it to follow a Bernoulli distribution of parameter $p_y = 0.2$, so as expected

$$\text{mean(Y)} = 0.20 \pm 0.00 = p_y$$
$$\text{std(A)} = 0.40 \pm 0.00 = \sqrt{0.2(1-0.2)} = \sqrt{p_y(1-p_y)}$$

This results in an imbalance too: table 8 gives the number of observations for each value of Y, and the imbalance ratio is

$$IR_Y = 4.00 \pm 0.01$$

meaning there are 4 times more observations of $Y = 0$ than $Y = 10$.

Y	Observations
0	$80,010.49 \pm 262.28$
1	$19,989.51 \pm 262.28$

Table 8: Number of observations by value of Y

Going back on imbalance In real life, imbalance can be explained either by the way the data was collected or by the natural domination of one class. The collection of data can lead to an imbalance if the sampling is biased or if mistakes are made, for examples writing down the wrong labels on observations. In insurance, there are sampling biases: conclusions about risks only concern individuals who have been accepted at the underwriting stage. As underwriters aim at selecting 'good' risks, most of the time, the insurer's portfolio will be very specific and predictions on claims frequency, for example, cannot be generalized to a different population.

The underwriting stage introduces a sampling bias, so we cannot draw conclusions about another population than the one that was selected.

2.3.2 Multivariate analysis

An important part of data exploration consists in studying the relationships between variables. To do so, we will analyze how they are correlated to each other. As a reminder, correlation is how linearly related two variables are, and is only the first order approximation of dependence, as seen previously.

Correlations with Y A heatmap of the correlations can help understand which variables are correlated with each other. Figure 14 shows the heatmap of correlations. The strongest positive correlations are colored in bright red and the strongest negative correlations are colored in bright blue. The strongest correlations to Y in absolute value are with $X^{(4)}$, $X^{(2)}$, $X^{(1)}$, A, $X^{(2)}$, $X^{(3)}$ and then $B0$. This is coherent with the theoretical values of the correlation matrix $R_{X,A,B,Y}$ in section 2.2.2.

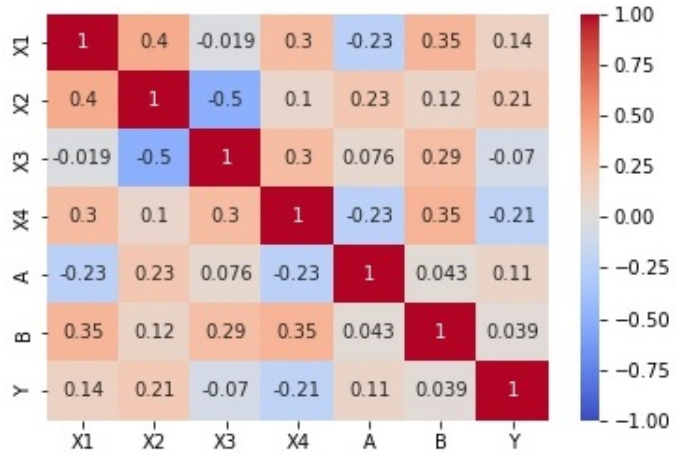


Figure 14: Heatmap of correlations

Correlations with A We saw in the previous paragraph that A has the third strongest absolute value correlation to $Y0$. It is also strongly correlated to the $X^{(i)}$, as we see in figure 14. We noted in the previous section that there are imbalances in the number of observations for both A and Y, and we see in table 9 that among group A there is an imbalance in output values too. If we compute the imbalance ratio for the output, we get

$$\text{A=0: } IR_Y = 4.86 \pm 0.01$$
$$\text{A=1: } IR_Y = 2.73 \pm 0.01$$

meaning that within group $A = 0$, there are 4.86 times more observations of $Y = 0$ than $Y = 1$ and within group $A = 1$, there are 2.73 times more observations of $Y = 0$ than $Y = 10$. In general there are a lot more $Y = 0$ outputs than $Y = 1$ ones, and when we zoom in on protected groups, the imbalance ratio is larger for group $A = 0$ than for $A = 1$, meaning that the former has a larger proportion of $Y = 0$ outputs than the latter.

	Y=0	Y=1
A=0	$58,048.36 \pm 33.08$	$11,953.86 \pm 21.76$
A=1	$21,962.13 \pm 26.37$	$8,035.65 \pm 15.49$

Table 9: Number of observations by values of A and Y

Correlations with B We saw that B has the weakest absolute value correlation to Y, but it is still strongly correlated to the $X^{(i)}0$. Within groups, as for A, there can be imbalances, as we see in table 10. The imbalance ratios by group are

$$\text{B=0: } IR_Y = 5.53 \pm 0.02$$
$$\text{B=1: } IR_Y = 3.88 \pm 0.01$$

meaning that within group $B = 0$, there are 5.53 times more observations of $Y = 0$ than $Y = 1$ and within group $B = 1$, there are 3.88 times more observations of $Y = 0$ than $Y = 10$. The imbalance ratio is larger for group $B = 0$ than for $B = 10$.

	Y=0	Y=1
B=0	$8,484.30 \pm 19.82$	$1,535.00 \pm 9.06$
B=1	$71,526.19 \pm 29.28$	$18,454.51 \pm 24.25$

Table 10: Number of observations by values of B and Y

Correlations between the $\mathbf{X^{(i)}}$ By construction the $X^{(i)}$ are related to each other, through the correlation matrix. Figure 15 gives the pairwise relationships between the $X^{(i)}$, and the diagonal is their marginal distribution. We observe that $X^{(1)}$ is positively correlated with $X^{(2)}$ and $X^{(3)}$, $X^{(2)}$ is negatively correlated with $X^{(3)}$, and $X^{(3)}$ is negatively correlated with $X^{(4)}0$. Correlations between other variables are less obvious.

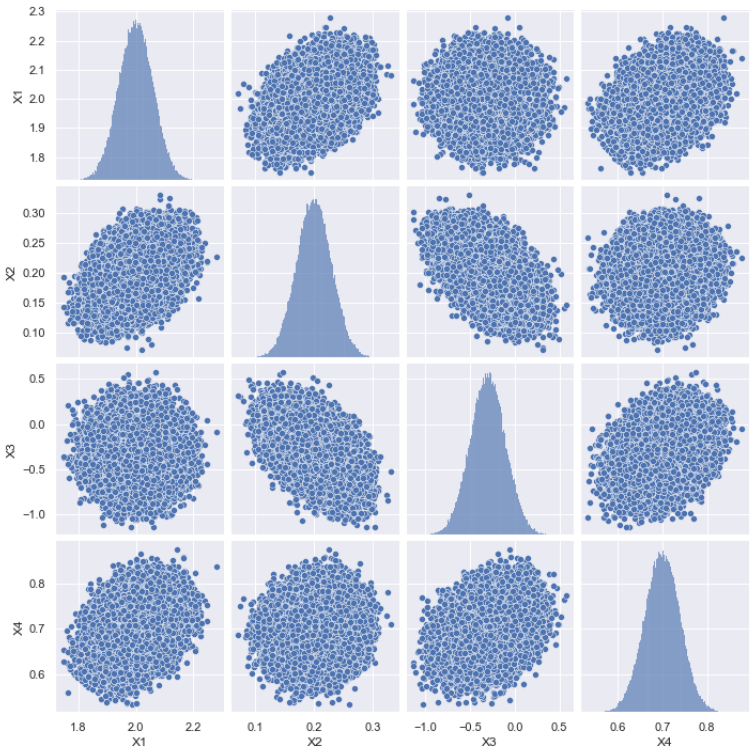


Figure 15: Pairplots of the $X^{(i)}$

2.3.3 Outliers

As the $X^{(i)}$ were computed to follow normal distributions, and A, B and Y to take values in $\{0, 1\}$, we do not expect to have any outliers. This is verified in our datasets: there are no unexpected values for the $X^{(i)}$ and we have values of A, B and Y in $\{0, 1\}0$.

3 Discrimination mitigation applied to the simulated data

The goal of the section is to compare different pre-processing steps and see how they influence our fairness metrics. After this, we will apply a logistic regression model to the simulated explanatory variables predict the variable of interest. Appendix B gives a reminder on the logistic regression.

The reason why we chose the logistic regression is that it is interpretable, which is a major issue with Machine Learning, and an important characteristic to simplify our study on fairness. It also presents other advantages, such as its simplicity with a low number of parameters. The main drawback is that a lot of preprocessing must be done: we need to select variables that are not strongly correlated with each other, and to transform most continuous variables into categorical variables, or only their general effect will be captured by the coefficients. For this simulated dataset, we will keep all variables as there are only 5 of then, and they are very simple.

The goal is to highlight the mechanisms of discrimination, which are shown through the fairness metrics rather than the model itself. This is why we chose to rely on the logistic regression model, many insurers choosing GLMs, even though the choice of the model can influence the outcome of a risk analysis.

3.1 Regression model with no pre-processing step

In this section, we will simply predict the outcome with all the variables. We standardized the variables to have mean zero and standard deviation 1, then randomly separated the dataset into a train (80%) and a test dataset (20%), and applied a logistic regression model.

Table 11 gives confidence intervals of the weights of this logistic regression, their standard errors and the associated p-values. All variables have p-values below 0.05, so they are all significant to the model.

Variable	Coefficient	Standard Error	P-value
Intercept	-1.82 ± 0.01	0.05 ± 0.00	0.00 ± 0.00
X_1	0.55 ± 0.00	0.02 ± 0.00	0.00 ± 0.00
X_2	2.81 ± 0.01	0.05 ± 0.00	0.00 ± 0.00
X_3	0.25 ± 0.00	0.01 ± 0.00	0.00 ± 0.00
X_4	-2.53 ± 0.01	0.03 ± 0.00	0.00 ± 0.00
A	-0.17 ± 0.01	0.03 ± 0.00	0.00 ± 0.00
B	0.41 ± 0.01	0.04 ± 0.00	0.00 ± 0.00

Table 11: Coefficients of the model using all variables

Performance evaluation

- We have 81.05% correct classifications on average. The accuracy acceptability depends on the business context. As we saw previously, if mistakes have a high cost then it might not be sufficient.
  If there is a large class imbalance, accuracy is not the best metric as it can be very high while the model only fits the majority population. As a reminder, the output imbalance ratio is $IR_Y = 4.000$. So we cannot rely solely on accuracy to evaluate our model.

- As a reminder, the ROC (Receiver Operating Characteristic) curve plots the true positive rate against the false positive rate for varying classification thresholds. A random classifier

will exhibit a linear ROC curve as the one plotted in a dashed orange line (figure 16). Above that line, the model performs better than the random classifier, and below, worse. The perfect classifier has a ROC curve that is confined to the (0,1) point. Here, our model performs a lot better than the random classifier: the AUC (Area Under the ROC Curve) is of 0.7568.

(%)	Global
Accuracy	81.05 ± 0.05

Table 12: Global metrics (all variables)

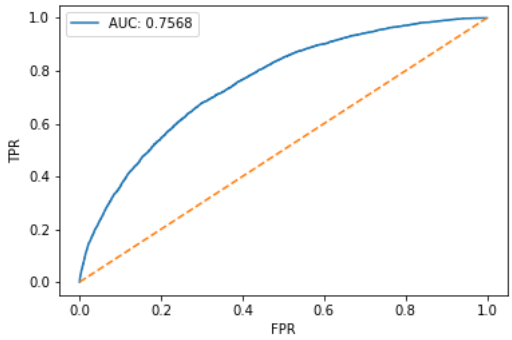


Figure 16: ROC curve for the model with all variables

Fairness evaluation We will compare fairness under all three definitions, first comparing the metrics between values of protected groups A and B (table 13), then between subgroups of combinations of A and B (table 14).

- Statistical parity requires the same acceptance rates for all protected groups.
  - It is a lot higher for group $A = 0$ than for group $A = 1$, by 8.26 points. So group $A = 1$ is disadvantaged by the model under this definition.
  - It is higher for group $B = 0$ than for group $B = 1$, by 3.47 points. So group $B = 1$ is disadvantaged by the model under this definition.
  - Looking at subgroups, the most advantaged subgroup is for $(A = 0, B = 0)$ and the most disadvantaged is for $(A = 1, B = 1)0$. This shows that looking at combinations of protected variables reveals that groups with a certain combination of characteristics are even more disadvantaged.
- Equal opportunity requires the same true positive rates for all protected groups.
  - It is higher for group $A = 0$, so group $A = 1$ is disadvantaged by the model under this definition.
  - It is higher for group $B = 0$, although not by much, so group $B = 1$ is disadvantaged by the model under this definition.
  - It is highest for subgroup $(A = 0, B = 0)$ and lowest for group $(A = 1, B = 1)0$.
- Equalized odds requires the same true and false positive rates for both protected groups.
  - The false positive rate is higher for group $A = 0$ than group $A = 10$. Group $A = 1$ has lower true and false positive rates, so it is disadvantaged by the model under this definition.

– The false positive rate is higher for group $B = 0$ than $B = 10$. Group $B = 1$ has lower true and false positive rates, so it is disadvantaged by the model under this definition.

– The false positive rate is highest for subgroup $(A = 0, B = 0)$ and lowest for $(A = 1, B = 1)0$.

To conclude, the model is unfair and disadvantages groups $A = 1$ and $B = 1$ under all three fairness definitions, and the most disadvantaged subgroup is $(A = 1, B = 1)0$.

(%)	Global	A=0	A=1	Difference	B=0	B=1	Difference
AR	94.13 ± 0.05	96.60 ± 0.04	88.34 ± 0.10	8.26 ± 0.10	97.25 ± 0.09	93.78 ± 0.05	3.47 ± 0.10
TPR	96.98 ± 0.04	98.26 ± 0.03	93.60 ± 0.09	4.67 ± 0.09	98.62 ± 0.06	96.79 ± 0.04	1.83 ± 0.07
FPR	82.69 ± 0.15	88.55 ± 0.15	73.95 ± 0.24	14.60 ± 0.26	89.72 ± 0.39	82.11 ± 0.16	7.62 ± 0.42

Table 13: Fairness metrics globally and by protected group (all variables)

(%)	A=0		A=1	
	B=0	B=1	B=0	B=1
AR	98.47 ± 0.07	96.39 ± 0.04	93.72 ± 0.22	87.92 ± 0.12
TPR	99.21 ± 0.05	98.15 ± 0.03	96.58 ± 0.19	93.41 ± 0.09
FPR	93.65 ± 0.36	88.08 ± 0.16	83.15 ± 0.66	73.34 ± 0.26

Table 14: Fairness metrics by protected subgroups (all variables)

A model using all variables, including the protected ones, easily gives unfair outputs and we are dealing with direct discrimination. Under every fairness definition, the same groups are disadvantaged.

3.2 Removing protected variables to avoid direct discrimination

When sensitive variables are omitted, models can still learn stereotypes, because sensitive information is embedded in datasets even if it is not intentional. Leaving out sensitive variables forces the correlated variables to take on a greater importance. This is the omitted variable bias [51].

Removing sensitive attributes is problematic, because it becomes impossible to check for bias and discrimination. We cannot see if the most important variables for prediction are strongly correlated with a protected attribute or compute metrics. This is a problem related to data regulations: as we saw in section 1.2.1, the GDPR requires minimal data collection, but more data is needed to prove discrimination.

We will preprocess our data by removing the sensitive attributes, A and B, then predict the outcome with a logistic regression model. Table 15 gives the predicted weights of this logistic regression. As A and B are no longer used as explanatory variables, the weights of regression have changed. For example, the biggest change is for $X^{(3)}$, which had a coefficient of 0.25 and now has a coefficient of 2.46. Remembering the correlation matrix, $X^{(3)}$ is correlated with A and B with relatively high Pearson correlation coefficients: -0.23 and 0.35 respectively. This can indicate that A and B will still indirectly play a part in the model predictions.

With all variables

Variable	Coefficient	Standard Error	P-value
Intercept	-1.82 ± 0.01	0.05 ± 0.00	0.00 ± 0.00
X_1	0.55 ± 0.00	0.02 ± 0.00	0.00 ± 0.00
X_2	2.81 ± 0.01	0.05 ± 0.00	0.00 ± 0.00
X_3	0.25 ± 0.00	0.01 ± 0.00	0.00 ± 0.00
X_4	-2.53 ± 0.01	0.03 ± 0.00	0.00 ± 0.00
A	-0.17 ± 0.01	0.03 ± 0.00	0.00 ± 0.00
B	0.41 ± 0.01	0.04 ± 0.00	0.00 ± 0.00

Without protected variables

Variable	Coefficient	Standard Error	P-value
Intercept	-1.76 ± 0.01	0.05 ± 0.00	0.00 ± 0.00
X_1	0.65 ± 0.00	0.02 ± 0.00	0.00 ± 0.00
X_2	2.68 ± 0.01	0.05 ± 0.00	0.00 ± 0.00
X_3	2.46 ± 0.00	0.01 ± 0.00	0.00 ± 0.00
X_4	-2.41 ± 0.01	0.03 ± 0.00	0.00 ± 0.00

Table 15: Weights of the logistic regression

Performance evaluation

- As we can see in table 17, the accuracy has only decreased by 0.01 on average, which is negligible, especially considering the width of the confidence interval.
- Looking at the ROC curve in figure 17, the model still performs quite well, and the AUC has decreased from 0.7568 to 0.7466 compared to the model using all variables.

(%)	With all variables	Without protected variables
Accuracy	81.05 ± 0.05	81.04 ± 0.05

Table 16: Global metrics

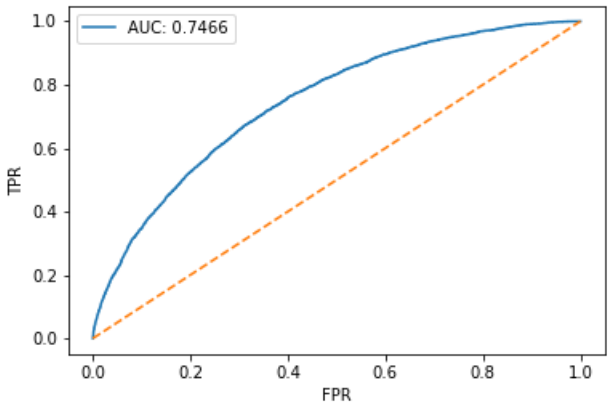


Figure 17: ROC curve for the model without protected variables

Fairness evaluation

- Acceptance rate
  - The global acceptance rate has slightly increased.
  - For groups A, the gap between acceptance rates has increased by more than one point. This was expected, as $X^{(3)}$, strongly correlated with A, has taken on a great importance in the model prediction.
  - For groups B, the difference in acceptance rates has decreased.
  - Looking at protected subgroups, the most advantaged subgroups are the same as in the model with all variables: the most advantaged subgroup is $(A = 0, B = 0)$ and the most disadvantaged is $(A = 1, B = 1)0$. We now have a lower gap between acceptance rates for subgroups $(A = 0, B = 0)$ and $(A = 0, B = 1)$ and for $(A = 1, B = 0)$ and $(A = 1, B = 1)$, which is coherent as acceptance rates between groups B are closer than previously.
- For true and false positive rates, we have the same conclusions as for the acceptance rates.

For groups A, the gaps between fairness metrics are wider when the protected variables are not used in the model, but for group B, they are smaller, although this phenomenon comes from the structure of the correlation matrix. The same groups remain disadvantaged.

To conclude, the performance metrics have not deteriorated too much compared to when using all variables. The fairness metrics are worse for groups A but better when looking at groups B. To conclude, simply ignoring the protected variables is not a solution.

With all variables

(%)	Global	A=0	A=1	Difference	B=0	B=1	Difference
AR	94.13 ± 0.05	96.60 ± 0.04	88.34 ± 0.10	8.26 ± 0.10	97.25 ± 0.09	93.78 ± 0.05	3.47 ± 0.10
TPR	96.98 ± 0.04	98.26 ± 0.03	93.60 ± 0.09	4.67 ± 0.09	98.62 ± 0.06	96.79 ± 0.04	1.83 ± 0.07
FPR	82.69 ± 0.15	88.55 ± 0.15	73.95 ± 0.24	14.60 ± 0.26	89.72 ± 0.39	82.11 ± 0.16	7.62 ± 0.42

Without protected variables

(%)	Global	A=0	A=1	Difference	B=0	B=1	Difference
AR	94.29 ± 0.05	97.05 ± 0.04	87.55 ± 0.10	9.50 ± 0.09	95.38 ± 0.10	94.07 ± 0.055	1.31 ± 0.09
TPR	97.02 ± 0.03	98.51 ± 0.03	93.07 ± 0.09	5.44 ± 0.08	97.41 ± 0.08	96.97 ± 0.04	0.44 ± 0.08
FPR	82.92 ± 0.12	89.94 ± 0.13	72.53 ± 0.23	17.41 ± 0.26	84.15 ± 0.46	82.82 ± 0.13	1.33 ± 0.48

Table 17: Fairness metrics

With all variables

(%)	A=0		A=1	
	B=0	B=1	B=0	B=1
AR	98.47 ± 0.07	96.39 ± 0.04	93.72 ± 0.22	87.92 ± 0.12
TPR	99.21 ± 0.05	98.15 ± 0.03	96.58 ± 0.19	93.41 ± 0.09
FPR	93.65 ± 0.36	88.08 ± 0.16	83.15 ± 0.66	73.34 ± 0.26

Without protected variables

(%)	A=0		A=1	
	B=0	B=1	B=0	B=1
AR	97.43 ± 0.08	97.00 ± 0.04	88.94 ± 0.27	87.43 ± 0.10
TPR	98.56 ± 0.06	98.50 ± 0.03	93.41 ± 0.25	93.04 ± 0.09
FPR	90.01 ± 0.44	89.92 ± 0.13	72.61 ± 0.87	72.53 ± 0.24

Table 18: Fairness metrics by protected subgroups

Removing the protected variables avoids direct discrimination, but not indirect one. Depending on the dependence structure between protected and non-protected variables, the outcome can be more (variable A) or less (variable B) discriminatory than in the previous case.

3.3 Transforming the non-protected variables to mitigate indirect discrimination

3.3.1 Theory

The idea Focusing on the definition of fairness as statistical parity, we can view it as an independence condition. Statistical parity requires $\hat{Y} \perp\!\!\!\perp A, B0$. Since we do not want A or B to impact the predicted output, the goal is firstly not to use them as explanatory variables and secondly to have explanatory variables that are independent of them. Dealing with independence is a complex problem, which is why we will tackle it on a linear level only - with correlation. We will try to obtain transformations of the variables $X^{(i)}$ uncorrelated with A and B, and we will then use them as inputs of a logistic regression model to predict $Y0$.

Drawing inspiration from the Gram-Schmidt process (a reminder is given in appendix C), the idea is to view the variables as vectors in a n-dimensional space (n being the number of variables in our dataset) and the covariance between them as a scalar product. Let us set the theoretical framework.

The following theory is necessary to explain the method which has not been found in the literature. It relies on linear algebra and was inspired by the Gram-Schmidt process, with the goal of transforming the data so that the new non-sensitive variables are uncorrelated with the sensitive ones.

Definition 11. *An inner product is a map*

$$\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$$

with V a real vector space, that is symmetric, bilinear and positive-definite.

Proposition 7. *Set X and Y real random variables with zero mean and finite variance. Their covariance*

$$\langle X, Y \rangle = cov(X, Y)$$

is an inner product (on the space of random variables with zero mean and finite variance).

Proof. Set X, Y, Z real random variables with zero mean and finite variance and $(a, b) \in \mathbb{R}^2 0$. Then $cov(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[XY]0$. We will check the three properties of the inner product:

- symmetry: $\mathbb{E}[XY] = \mathbb{E}[YX]$
- bilinearity: $\mathbb{E}[(aX + bY)Z] = a\mathbb{E}[XZ] + b\mathbb{E}[YZ]$
- positive-definiteness: $\mathbb{E}[XX] = \mathbb{E}[X^2] \geq 0$
  and $\mathbb{E}[X^2] = 0 \iff Var(X) + (\mathbb{E}[X])^2 = 0 \iff \begin{cases} Var(X) = 0 \\ \mathbb{E}[X] = 0 \end{cases} \iff X = 0$ a.s.

□

Definition 12. *Two vectors are orthogonal if their inner product is zero.*

As covariance is an inner product on the space of random variables with zero mean and finite variance, two random variables of this space are orthogonal if their covariance is zero ie if their correlation is zero, because as we saw in definition 5,

$$corr(X, Y) = \frac{cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

and their variance is finite. So we have an interpretation of random variables of this space as vectors with an inner product.

Going back to the initial goal, we wanted a transformation of our variables $X^{(i)}$ that is uncorrelated to both A and $B0$. With the framework we have set, it means that the transformed variables will be orthogonal to A and to $B0$. Suppose the (non-orthogonal) basis of our vector space is the linearly independent set $B = \{u_1, \ldots, u_n\}$ such that any vector Z of the space - which corresponds to the characteristic of one individual ie row of the dataset - can be uniquely written as a linear combination of the vectors of the basis:

$$Z = \sum_{i=1}^{n} x_i u_i \text{ with } X = \begin{bmatrix} x_1 \\ \ldots \\ x_n \end{bmatrix}$$

Our goal is to find a change of basis that gives us the new basis $B' = \{v_1, \ldots, v_n\}0$. We will then be able to write

$$Z = \sum_{j=1}^{n} x'_j v_j \text{ with } X' = \begin{bmatrix} x'_1 \\ \ldots \\ x'_n \end{bmatrix}$$

$A = (a_{i,j})_{i,j}$ is called the transition matrix and its jth row is formed by the coordinates of v_j:

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} & \ldots & a_{1,n} \\ a_{2,1} & a_{2,2} & \ldots & \\ \ldots & & & \\ a_{n,1} & \ldots & & a_{n,n} \end{bmatrix}$$

Then, the change-of-basis formula gives in matrix form

$$\begin{aligned} X' &= AX \\ &= \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n} \\ a_{2,1} & a_{2,2} & \dots & \\ \dots & & & \\ a_{n,1} & \dots & & a_{n,n} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix} \\ &= \begin{bmatrix} a_{1,1}x_1 + a_{1,2}x_2 + \dots + a_{1,n}x_n \\ a_{2,1}x_1 + a_{2,2}x_2 + \dots + a_{2,n}x_n \\ \dots \\ a_{n,1}x_1 + a_{n,2}x_2 + \dots + a_{n,n}x_n \end{bmatrix} \end{aligned}$$

So

$$x'_i = \sum_{j=1}^{n} a_{i,j}x_j$$

Our goal is to have a new basis in which the non-sensitive vectors are orthogonal to the sensitive ones. For this reason, we impose the following constraints for the construction of the new basis:

1. we do not want to transform the sensitive variables, so the first s vectors of the new basis will remain the same as in the old basis: $x_1 = x'_1, \dots, x_s = x'_s$

   $$\begin{aligned} x_i = x'_i \Leftrightarrow x_i &= \sum_{j=1}^{n} a_{i,j}x_j \\ &\Leftrightarrow \begin{cases} a_{i,i} = 1 \\ a_{i,j} = 0 \text{ for } j \neq i \end{cases} \end{aligned}$$

   As a result,

   $$A = \begin{bmatrix} & 0 & \dots & 0 \\ I_s & \vdots & & \vdots \\ & 0 & \dots & 0 \\ a_{s+1,1} & \dots & & a_{s+1,n} \\ \vdots & & & \vdots \\ a_{n,1} & \dots & & a_{n,n} \end{bmatrix}$$

   with I_s the identity matrix of size $s0$.

2. the new non-sensitive vectors will be orthogonal to the sensitive vectors: $\forall j \in [\![1, s]\!], \forall k \in [\![s+1, n]\!], \langle v_j, v_k \rangle = 0$
   This is equivalent to: for all $k \in [\![s+1, n]\!]$ a system of s equations with n unknown variables (the $a_{.,k}$):

   $$\begin{cases} \langle v_1, v_k \rangle &= 0 \\ \dots & \\ \langle v_s, v_k \rangle &= 0 \end{cases} \tag{5}$$

The rows of A are the v_j, so $v_j = \begin{cases} u_j & \text{if } j \in [\![1, s]\!] \\ \sum_{i=1}^{n} a_{j,i} u_i & \text{if } j \in [\![s+1, n]\!] \end{cases}$

$$(5) \Leftrightarrow \begin{cases} \langle u_1, \sum_{j=1}^{n} a_{k,j} u_j \rangle &= 0 \\ & \dots \\ \langle u_s, \sum_{j=1}^{n} a_{k,j} u_j \rangle &= 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} \sum_{j=1}^{n} a_{k,j} \langle u_1, u_j \rangle &= 0 \\ & \dots \\ \sum_{j=1}^{n} a_{k,j} \langle u_s, u_j \rangle &= 0 \end{cases}$$

For each row k of A, we are looking for the values of the $a_{k,1}, \dots, a_{k,n}0$. The previous constraints give for each row a system of s equations, but we have n unknown variables, so the system is underdetermined, with an infinite number of solutions. To simplify, we can set

$$\forall j \in \{s+1, \dots, k-1, k+1, \dots, n\}, a_{k,j} = 0 \text{ ie } v_k = \sum_{j=1}^{s} a_{k,j} u_j + a_{k,k} u_k$$

Meaning that each non-sensitive vector of the new basis is written as a linear combination of itself and of the sensitive vectors of the old basis. This gives a transition matrix of the shape

$$A = \begin{bmatrix} & & & 0 & \dots & & 0 \\ & I_s & & \vdots & & & \vdots \\ & & & 0 & \dots & & 0 \\ a_{s+1,1} & \dots & a_{s+1,s} & a_{s+1,s+1} & 0 & 0 \dots & 0 \\ a_{s+2,1} & \dots & a_{s+2,s} & 0 & a_{s+2,s+2} & 0 \dots & 0 \\ \vdots & & & & & & \\ a_{n,1} & \dots & a_{n,s} & 0 & 0 & \dots 0 & a_{n,s+1} \end{bmatrix}$$

We now have for each row, a system of s linear equations and $s+1$ unknown variables. We need one more constraints in order to have a unique solution. An idea is to minimize the distance between the old and the new basis (non-sensitive) vectors:

$$\min_{a_{k,1}, \dots, a_{k,n}} d(u_k, v_k)$$

As the distance is positive, it has a lower bound, so this problem has a solution. We have defined the inner product as the covariance between random variables with zero mean and finite variance, so the distance between two such random variables X and Y is

$$d(X, Y) = \langle X - Y, X - Y \rangle = cov(X - Y, X - Y) = Var(X - Y)$$

So

$$\begin{aligned} d(u_k, v_k) &= d(u_k, \sum_{j=1}^{s} a_{k,j} u_j + a_{k,k} u_k) \\ &= \langle (1 - a_{k,k}) u_k - \sum_{j=1}^{s} a_{k,j} u_j - a_{k,k} u_k, (1 - a_{k,k}) u_k - \sum_{j=1}^{s} a_{j,k} u_j \rangle \\ &= ((1 - a_{k,k}) u_k - \sum_{j=1}^{s} a_{k,j} u_j)^T ((1 - a_{k,k}) u_k - \sum_{j=1}^{s} a_{k,j} u_j) \\ &= \sum_{i=1}^{n} ((1 - a_{k,k}) u_{i,k} - \sum_{j=1}^{s} a_{k,j} u_{i,j})^2 \end{aligned}$$

As we had a system of s linear equations and $s+1$ unknown variables, we can express $a_{k,k}$ as a combination of the other $a_{.,k}0$. So the minimization of this distance gives a unique solution with the previous constraints. To summarize, for each $k = s+1, \dots, n$, we have the following minimization problem under constraints:

$$\min_{a_{k,1},\dots,a_{k,s},a_{k,k}} d(u_k, v_k) \text{ such that } \begin{cases} \langle v_1, v_k \rangle &= 0 \\ \dots & \\ \langle v_s, v_k \rangle &= 0 \end{cases} \text{ with } v_k = \sum_{j=1}^{s} a_{k,j} u_j + a_{k,k} u_k$$

Solving the problem for every $k = s+1, \dots, n$ gives us the transition matrix $A0$. Then, with X the coordinates of a vector in the base B and X' in the base B', we can write

$$X' = AX$$

Meaning that we will compute, for every observation, the transformation of each vector - corresponding to each individual - in the new basis.

Extreme-case scenarii We can wonder what would happen in the extreme-case scenario in which the non-sensitive variables are already uncorrelated with the sensitive ones. Then, we do not need to transform the non-sensitive variables: $A = I_n$ We have $\langle v_j, v_k \rangle = 0$ for $j = 1, \dots, s$ and $k = s+1, \dots, n$ and we have a minimal distance as $d(u_k, v_k) = d(u_k, u_k) = \langle u_k - u_k, u_k - u_k \rangle = 00$.

The other extreme-case scenario is the one in which all the variables are the non-sensitive variables are perfectly (positively or negatively) correlated with the sensitive ones. Then it means that the non-sensitive variables are a linear function of the sensitive ones, and consequently, of each other. It is therefore impossible to have non-sensitive vectors uncorrelated with the sensitive ones. Fortunately, in reality, when we have our datasets, we only have samples of 'true' distributions, meaning that variables are never perfectly correlated with each other (except if a variables appears twice, but we can delete the duplicate). In the worst case, if the non-sensitive variables are very correlated to the non-sensitive ones, we will end up with transformed non-sensitive variables that have a very low variance, meaning that they will not explain the output very well.

3.3.2 Results

Transition matrix The average transition matrix obtained on the 100 datasets is

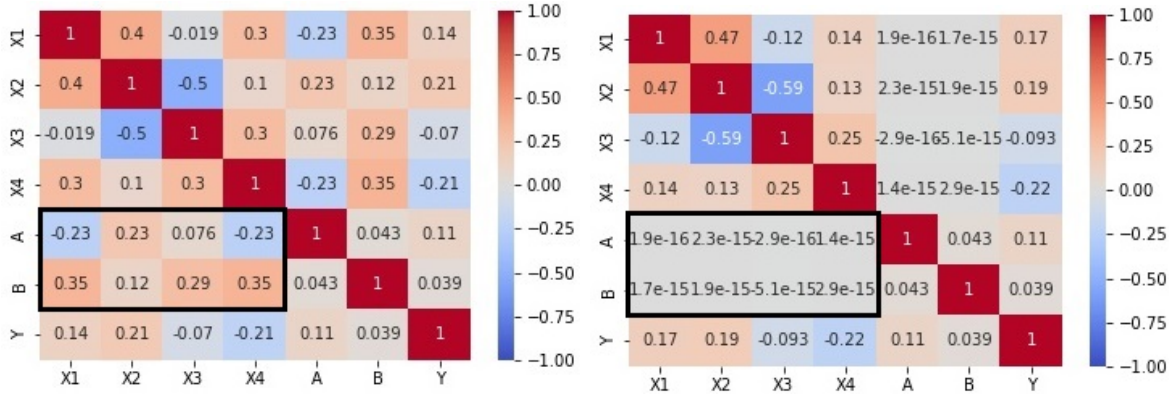
$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ -0.32 & 0.72 & 1 & 0 & 0 & 0 \\ 0.15 & 0.11 & 0 & 1 & 0 & 0 \\ 0.28 & 1.93 & 0 & 0 & 1 & 0 \\ -0.21 & 0.48 & 0 & 0 & 0 & 1 \end{bmatrix} \pm \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 3\text{e}{-3} & 6\text{e}{-3} & 0 & 0 & 0 & 0 \\ 2\text{e}{-3} & 3\text{e}{-3} & 0 & 0 & 0 & 0 \\ 1\text{e}{-2} & 2\text{e}{-2} & 0 & 0 & 0 & 0 \\ 2\text{e}{-3} & 3\text{e}{-3} & 0 & 0 & 0 & 0 \end{bmatrix}$$

This means that, on average,

$$A^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0.32 & -0.72 & 1 & 0 & 0 & 0 \\ -0.15 & -0.11 & 0 & 1 & 0 & 0 \\ -0.28 & -1.93 & 0 & 0 & 1 & 0 \\ 0.21 & -0.48 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$X'_1 = A^{-1}X_1 = .32A - 0.72B + X_1$$
$$X'_2 = A^{-1}X_2 = -0.15A - 0.11B + X_2$$
$$X'_3 = A^{-1}X_3 = -0.28A - 1.93B + X_3$$
$$X'_4 = A^{-1}X_4 = .21A - 0.48B + X_4$$

Transformed variables Figure 18 shows the correlations between variables before and after transforming the $X^{(i)}0$. We can see that the correlations between A (respectively B) and the $X(i)'$ have been reduced to zero, which was the goal of the procedure. Most correlations between other variables are close to before and after the change of basis, keeping the same signs and orders of magnitude. The most noticeable difference is that $corr(X^{(1)}, X^{(3)})$ went from -0.019 to -0.12.



(a) Before (b) After

Figure 18: Heatmaps of correlations before and after transformation of the $X^{(i)}$

The change in correlation matrix is in itself a summary of our works. The transformed variables no longer have any correlation with the sensitive ones.

Prediction We then apply the baseline model to predict the output, using only the transformed non-sensitive variables as explanatory variables.

Without protected variables

Variable	Coefficient	Standard Error	P-value
Intercept	-1.76 ± 0.01	0.05 ± 0.00	0.00 ± 0.00
X_1	0.65 ± 0.00	0.02 ± 0.00	0.00 ± 0.00
X_2	2.68 ± 0.01	0.05 ± 0.00	0.00 ± 0.00
X_3	2.46 ± 0.00	0.01 ± 0.00	0.00 ± 0.00
X_4	-2.41 ± 0.01	0.03 ± 0.00	0.00 ± 0.00

Transformed variables

Variable	Coefficient	Standard Error	P-value
Intercept	-1.65 ± 0.00	0.01 ± 0.00	0.00 ± 0.00
X_1	0.55 ± 0.00	0.02 ± 0.00	0.00 ± 0.00
X_2	2.72 ± 0.01	0.05 ± 0.00	0.00 ± 0.00
X_3	0.24 ± 0.00	0.01 ± 0.00	0.00 ± 0.00
X_4	-2.46 ± 0.01	0.03 ± 0.00	0.00 ± 0.00

Table 19: Weights of the logistic regression

Performance evaluation

- Table 20 gives the accuracy of the model. Compared to when simply deleting the protected variables, the accuracy decreases by only 0.23 points.
- Figure 19 gives the ROC curve of the model. As a reminder, the AUC of the model without protected variables was of 0.7466. The AUC for the model with transformed variables has only decreased by 0.0028.

(%)	Without protected variables	With transformed variables
Accuracy	81.04 ± 0.05	80.81 ± 0.05

Table 20: Global metrics

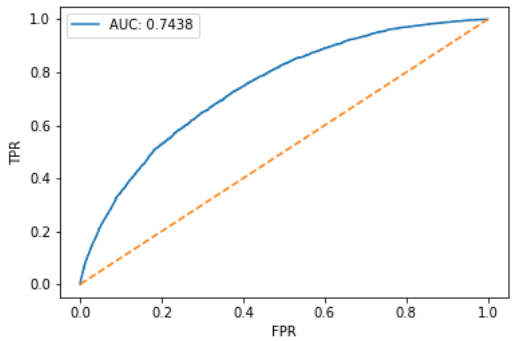


Figure 19: ROC curve for the model with transformed variables

Fairness evaluation

- Acceptance rate
  - The global acceptance rate has increased by 0.73 points on average, meaning that globally, more individuals get predicted the outcome $Y = 0$ with the model using the transformed variables compared to the model without protected variables.

  - For groups A: the difference in acceptance rates is now close to zero, which was the goal of the change of basis method. The reason why it is not exactly zero might come from the fact that we have approximated independence to correlation, so there might be some non linear dependence left between the transformed non-sensitive variables and A. It is interesting to note that the acceptance rate for group $A = 0$ has decreased and for $A = 1$ it has increased, and the advantageous position has shifted: the acceptance rate for group $A = 0$ is now slightly lower than the one for group $A = 10$.
  - For groups B: the difference in acceptance rates is now null, meaning that there we have $X^{(i)} \perp A$ ie independence.
  - Looking at protected subgroups in figure 22, not all subgroups are treated fairly by the model as we have slight gaps between subgroups with $A = 0$ and $A = 10$. The order of unfairness has also changed: now, the most disadvantaged subgroup is $(A = 0, B = 0)$ when it used to be the most advantaged one, and the most advantaged subgroup is $(A = 1, B = 1)$ when it used to be the most disadvantaged one.

- True positive rate
  - The global true positive rate has decreased compared to the model without protected variables.
  - Groups A: as for the acceptance rate, the difference in true positive rates is now closer to zero, and the sign has changed, meaning that group $A = 0$ now has a lower true positive rate than group $A = 10$.
  - Groups B: the difference in true positive rates is now very close to zero and has also changed signs.
  - As for the acceptance rate, the most advantaged subgroup now used to be the most disadvantaged and vice versa.

- False positive rate
  - Globally, the false positive rate has also decreased.
  - Groups A: the difference in false positive rates has also decreased and changed signs.
  - Groups B: surprisingly, the difference in false positive rates has increased and changed signs.
  - Looking at subgroups, there has also been a shift in which subgroup is the most and least advantaged.

To conclude, looking at the variable A, we have almost reached statistical parity, and for B, we have. We are closer to equal opportunity, although there was a shift in the advantage. For equalized odds, we are closer to fairness when looking at variable A but not B, and we have for both a shift in the advantage. All in all, the change of basis method has achieved the removal of linear dependence between the protected and non protected variables. We can also draw the conclusion that all three fairness definitions are not compatible.

Without protected variables

(%)	Global	A=0	A=1	Difference	B=0	B=1	Difference
AR	94.29 ± 0.05	97.05 ± 0.04	87.55 ± 0.10	9.50 ± 0.09	95.38 ± 0.10	94.07 ± 0.055	1.31 ± 0.09
TPR	97.02 ± 0.03	98.51 ± 0.03	93.07 ± 0.09	5.44 ± 0.08	97.41 ± 0.08	96.97 ± 0.04	0.44 ± 0.08
FPR	82.92 ± 0.12	89.94 ± 0.13	72.53 ± 0.23	17.41 ± 0.26	84.15 ± 0.46	82.82 ± 0.13	1.33 ± 0.48

Transformed variables

(%)	Global	A=0	A=1	Difference	B=0	B=1	Difference
AR	95.02 ± 0.04	94.91 ± 0.05	95.28 ± 0.06	-0.37 ± 0.06	95.02 ± 0.11	95.02 ± 0.04	0.00 ± 0.01
TPR	97.38 ± 0.03	97.15 ± 0.04	97.97 ± 0.05	-0.82 ± 0.05	97.07 ± 0.08	97.41 ± 0.03	-0.34 ± 0.08
FPR	85.58 ± 0.13	84.02 ± 0.17	87.92 ± 0.17	-3.90 ± 0.21	83.62 ± 0.51	85.74 ± 0.13	-2.13 ± 0.50

Table 21: Fairness metrics

Without protected variables

(%)	A=0		A=1	
	B=0	B=1	B=0	B=1
AR	97.43 ± 0.08	97.00 ± 0.04	88.94 ± 0.27	87.43 ± 0.10
TPR	98.56 ± 0.06	98.50 ± 0.03	93.41 ± 0.25	93.04 ± 0.09
FPR	90.01 ± 0.44	89.92 ± 0.13	72.61 ± 0.87	72.53 ± 0.24

With transformed variables

(%)	A=0		A=1	
	B=0	B=1	B=0	B=1
AR	94.87 ± 0.11	94.93 ± 0.05	95.18 ± 0.23	95.24 ± 0.06
TPR	96.75 ± 0.10	97.20 ± 0.04	97.60 ± 0.19	97.99 ± 0.04
FPR	82.62 ± 0.52	82.28 ± 0.18	86.34 ± 0.69	87.92 ± 0.17

Table 22: Fairness metrics by protected subgroups

Our method aimed at achieving statistical parity, and we have approximately. The approximation comes from the fact that we have focused on correlation, which is only a first order dependence.
Comparing with other metrics, we can see the incompatibility issue: the results in terms of fairness under the two other definitions are not optimal, as we have inverted which group was the most advantaged for one, and for the other the gap between metrics is worse than before.

3.4 Conclusion on the methods

We tested multiple preprocessing methods to mitigate unfairness. We compared the following: not doing anything, removing the protected variables, and transforming the non-sensitive variables in a way that they become uncorrelated to the sensitive ones. The goal of this last method was to approach fairness under the statistical parity, with equal acceptance rates for each protected group.

Comparing the model using all variables and the one using only non-protected variables, we can conclude that simply removing protected variables is not enough to reach fairness, under any definition. Depending on the correlation (and dependence) structure of the data, fairness can improve, with closer values of metrics for the different protected groups, as we saw is the

case when looking at the B variable. But it can also magnify unfairness as we saw is the case when looking at the A variable. All in all, it is not a solution.

With our change of basis method, we ensure that the variables that will be used by the model are uncorrelated with the protected variables. Of course, uncorrelatedness is only an approximation of independence and there could still be non-linear relationships between the protected and transformed variables, which is why acceptance rates are only approximately equal for protected groups.

With this method, we have a slight decrease in accuracy, because we have no access to the information about the protected attributes, which were correlated with the outcome.

Now that we have studied a simulated dataset in which the correlation structure was known, we will take on a real dataset.

4 Use case: mortality of individuals with melanoma of the skin

In this section, we will study a real-life use case: the mortality of individuals with melanoma skin cancer. It is the 17th most common cancer worldwide, with over 300,000 new cases in 2020 [41]. In the US, the 5-year survival rate is of 93.7% over the period of 2012 to 2018 [29].

4.1 Some information about skin cancer

The skin, the body's largest organ, consists of several lays. The two main ones are the epidermis and the dermis, as shown in figure 20. Skin cancer begins in the epidermis, which consists of three types of cells: squamous cells, basal cells and melanocytes. Melanocytes are cells that can make melanin, which is the pigment giving the skin its color.

Two different cancers can start in the skin: non-melanoma and melanoma, each representing about half of skin cancers. Non-melanoma skin cancer forms in the lower part of the epidermis or in squamous cells, but not in melanocytes. Melanoma forms in melanocytes and is more likely to spread out to other parts of the body. It can start in the skin, but also in mucous membranes such as parts of the eye. In this thesis, we focus on melanoma.

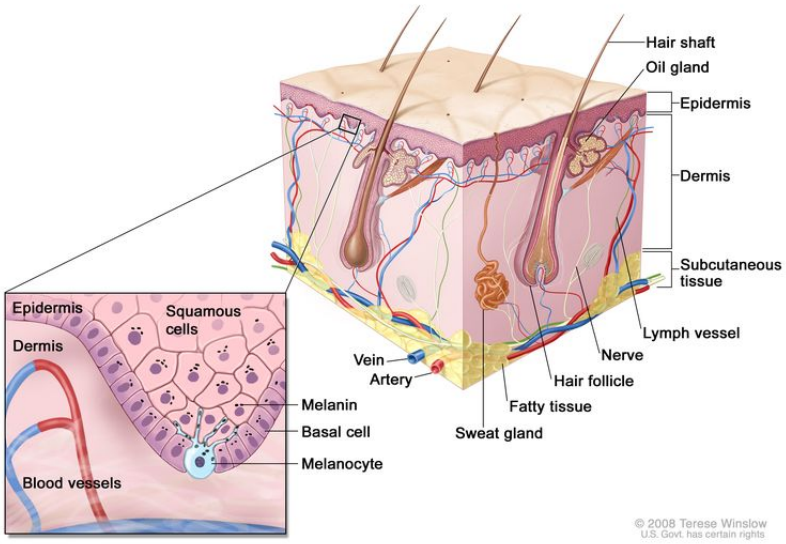


Figure 20: The anatomy of the skin [29]

Medical literature [29] identifies several risk factors for melanoma, including:

- Fair complexion
- Exposure to natural or artificial sunlight
- Exposure to certain environmental factors (radiation, solvents, ...)
- History of blistering sunburns
- Presence of several large or many small moles
- Family history of unusual moles or melanoma
- Weakened immune system

- Changes in genes linked to melanoma

Melanomas are mainly characterized by the location, thickness and ulceration (ie whether it has broken through the skin) of the tumor, the speed at which cancer cells divide, its spread to lymph nodes and other parts of the body. All of these factors can impact the severity of the condition. The TNM staging system is an internationally recognized standard for classifying cancers:

- T describes the size of the primary tumor, as shown in figure 21:
  - Tis: in situ
  - T1: less than 1 mm thick
  - T2: between 1 and 2 mm thick
  - T3: between 2 and 4 mm thick
  - T4: more than 4 mm thick

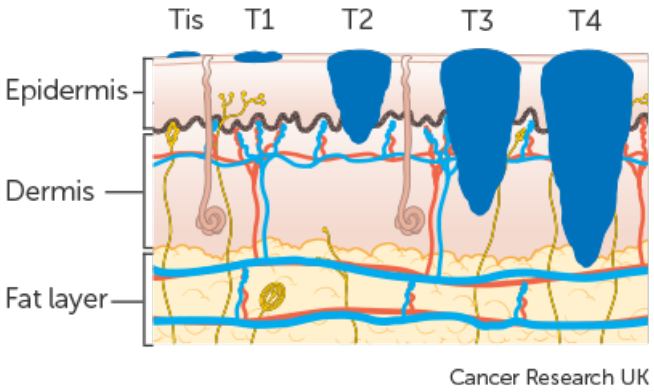


Figure 21: Tumor thickness [50]

- N describes the spread to regional lymph nodes:
  - N0: no melanoma cells in the nearby lymph nodes
  - N1: melanoma cells in one lymph node or in-transit, satellite or microsatellite metastases
  - N2: melanoma cells in 2 or 3 lymph nodes or in one lymph node and in-transit, satellite or microsatellite metastases
  - N3: melanoma cells in 4 or more lymph nodes or in 2 or 3 lymph nodes and in-transit, satellite or microsatellite metastases or in any number of lymph nodes stuck to each other (matted)
- M describes the presence of metastasis (M0 for non-metastatic and M1 for metastatic).

The more well-known overall stage grouping describes the progression of a cancer thanks to five categories:

- Stage 0: also called in situ, when abnormal melanocytes are found. They can become cancer and spread.
- Stage I: cancer has formed and is localized.

- Stage II: the cancer is locally advanced, in early stages.
- Stage III: the cancer is locally advanced, in late stages.
- Stage IV: the cancer has spread to other parts of the body which may be distant to the origin site, it is metastatic.

The link between the two staging systems is straightforward, as seen in table 23.

Stage	T	N	M
0	0	0	0
I	1-2	0	0
II	3-4	0	0
III	1-4	1-3	0
IV	1-4	1-3	1

Table 23: Staging and its relation to the TNM system

4.2 Database presentation and mapping

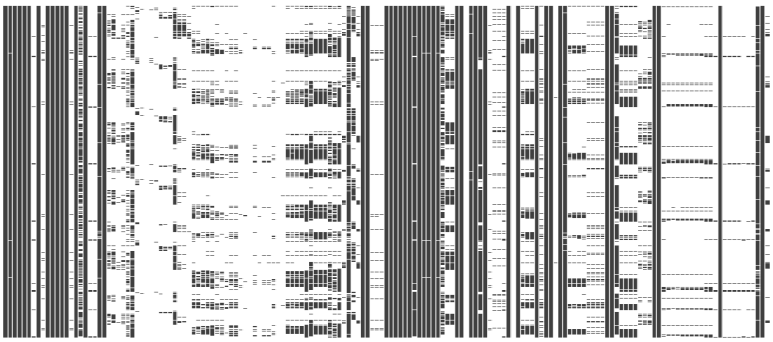
Presentation In order to model the mortality of melanoma of the skin cancer patients, we used the public research SEER database. SEER is the Surveillance, Epidemiology, and End Results program of the National Cancer Institute of the United States that collects and publishes cancer information about around 48% of the American population. This data collection process dates back to 1973, with around 400,000 new cases collected yearly in the most recent years. It is the largest and one of the most reliable cancer database, which makes it a dependable source of information for all types of studies. It is also representative of the US population in terms of measures of poverty and education.

The variables describe:

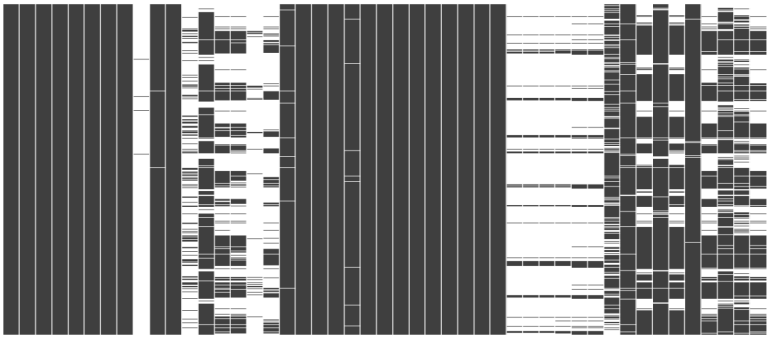
- the patient: ID number, sex, age at diagnosis, year of diagnosis, origin, marital status at diagnosis, ...
- the cancer (characteristics at diagnosis): type, site of origin, tumor size, spread to lymph nodes, metastatic state, stage, ...

Mapping This database is a huge source of information: there are 5,075,266 observations and 164 variables. Melanoma of the skin cancers represent about 4% of all observations. Along the years, the way cancers are described has drastically changed, mostly because of the changes in classification of diseases standards. We therefore had to preprocess the data by studying the significance of variables in order to have the same standards throughout the years. For example, we had to regroup 8 different variables to determine the size of the tumor.

Figure 22 shows how much preprocessing had to be done: the white spaces represent the missing values. In the raw data, we had more than 5 million rows and 164 columns. After keeping only patients with melanoma of the skin, mapping and deleting variables that had nothing to do with skin cancer (about breast cancer for example), we have 177,960 rows and 47 columns.



(a) Raw: 5,075,266 rows, 164 columns



(b) After mapping: 177,960 rows, 47 columns

Figure 22: Missing values (in white) in the raw and mapped databases

Missing values and initial variable selection After the mapping step, we still have many missing values. As we saw in section 1.1.3, there are many constraints to be taken into account when choosing which variables to keep for our model.

- Medical constraints:
  The probability of dying from cancer greatly depends on the metastatic state, so we cannot do without it. We remove individuals with missing metastatic state from the database.
  We also need to have the information on whether the individual is alive or not at the end of the observation period, so we remove individuals with missing information.
  Some variables are not medically relevant, for example the type of reporting source: if the information about an individual was collected by a hospital, a lab or another medical facility. So we delete this type of variable.

In this real use case, it is possible to capture indirect effects of discrimination because of proxy variables. This is why the medical constraints are necessary. Many variables were correlated with the mortality probability, but they were not medically relevant. Discussions with a doctor helped us sort them out.

As mortality rates evolve and change over the years, we need to set our time window so that the rates are still valid. Figure 23 shows the year of diagnosis distribution. We already have only recent years, mainly because we deleted individuals with unspecified metastatic state, and this variable was not collected in the past.

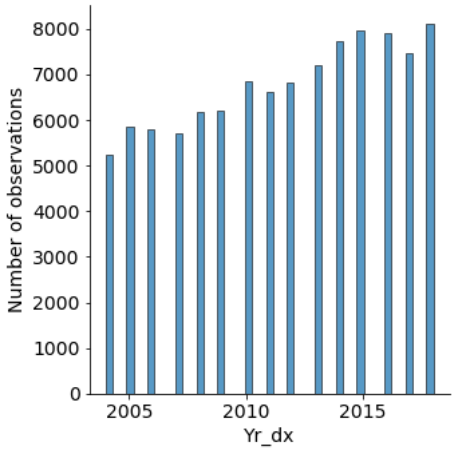


Figure 23: Number of new diagnoses per year

- Underwriting constraints:
  We only keep individuals who are between 18 and 80 years old, as they are the ones that are covered by most markets.
  Normally, underwriting constraints impose, depending on the local regulation, the deletion of sensitive variables, like the origin of the individual. But this thesis focuses on the subject of discrimination, so we have to keep these variables to be able to measure bias.

- Modeling constraints: we do not keep variables to are strongly correlated with each other. This is the case for variables representing redundant information: we have 3 different variables for age at diagnosis, 4 different variables for origin, ... We only keep one of each, the most complete one.

For the remaining missing values, we decided to delete the columns with almost only missing values, as they will not be useful for the model. Then, for numerical variables, the median of the series is assigned to any missing value. For categorical variables, we replace the missing values by a category 'Missing'. In the end, we are left with 101,797 rows and 28 columns.

4.3 Specificity of survival analysis

Most of the time, survival duration is observed partially. This can be due to the occurrence of the event of interest - in our case, death - outside the observation period, or to other events that result in the individual leaving the study (eg lapses, hospital transfers, recording systems failure). This censoring and truncation characterize survival data. If these effects were ignored, the probability of the event of interest would be underestimated. Another mistake would be to remove individuals for which the observation is incomplete from the study, because it would once again lead to biased estimations.

To deal with these issues, we either need to use specific models or modify the structure of the data to use standard models. We will use this second approach, which transforms the data by separating it into small time intervals. This allows to model the number of deaths by standard Machine Learning techniques using the exposure to risk as weights, taking into account the censoring.

4.3.1 Exposure

A first step is therefore to compute the exposure to risk of each individual. It can be computed differently depending on the hypothesis made on mortality. We compute the initial exposure

which represents the time each individual was exposed to risk in the interval. It is called initial because it is based on the information at the beginning of the interval.

The individual initial exposure $e_{i,j}$ for individual i in time interval $[\tau_j, \tau_{j+1}]$ takes value:

- 1 if the individual is alive during the entire time interval
- 1 if the individual dies during the time interval
- the fraction of the time interval he was observed if the individual is not observed during the entire time interval

Formally, denoting

$c_{i,j}$ the censoring time of individual i in time interval $[\tau_j, \tau_j + 1]$

$t_{i,j}$ the death time of individual i in time interval $[\tau_j, \tau_j + 1]$

w_j the number of individuals withdrawn from the study in time interval $[\tau_j, \tau_j + 1]$

l_j the number of individuals that are alive during the entire time interval $[\tau_j, \tau_j + 1]$

d_j the number of individuals that died during the time interval $[\tau_j, \tau_j + 1]$

we can write the individual initial exposure as:

$$\begin{aligned} e_{i,j} &= \begin{cases} 1 & \text{if } t_{i,j} > 1 \text{ and } c_{i,j} > 1 \\ 1 & \text{if } t_{i,j} < 1 \\ c_{i,j} & \text{if } c_{i,j} < 1 \end{cases} \\ &= \underbrace{\mathbb{1}_{t_{i,j}>1} \times \mathbb{1}_{c_{i,j}>1} + \mathbb{1}_{t_{i,j}<1}}_{1-\mathbb{1}_{c_{i,j}<1}} + c_{i,j}\mathbb{1}_{c_{i,j}<1} \end{aligned}$$

The global initial exposure for all individuals in time interval $[\tau_j, \tau_j + 1]$ is

$$\begin{aligned} E_j &= \sum_{i=1}^{l_j} e_{i,j} \\ &= \sum_{i=1}^{l_j} (1 - \mathbb{1}_{c_{i,j}<1} + c_{i,j}\mathbb{1}_{c_{i,j}<1}) \\ &= l_j - w_j + \sum_{i=1}^{w_j} c_{i,j} \end{aligned}$$

The number of deaths is the sum of the number of observed deaths and expected deaths (from censored individuals). Writing down

$q_j = \mathbb{P}(T < \tau_j + 1 | T > \tau_j)$ the mortality rate in time interval $[\tau_j, \tau_j + 1]$

${}_{c_{i,j}}q_j = \mathbb{P}(T < \tau_j + c_{i,j} | T > \tau_j)$ the mortality rate in time interval $[\tau_j, c_{i,j}]$

we can compute

$$\begin{aligned} d_j &= (l_j - w_j)q_j + \sum_{i=1}^{w_j} {}_{c_{i,j}}q_j \\ &= l_j q_j - \sum_{i=1}^{w_j} {}_{1-c_{i,j}}q_{j+c_{i,j}} \end{aligned}$$

The Balducci hypothesis supposes that mortality rates decrease over the interval and are defined as:

$$\begin{aligned} {}_{1-c_{i,j}}q_{j+c_{i,j}} &= \mathbb{P}(T_i \leq \tau_j + 1 | T_i > \tau_j + c_{i,j}) \\ &= (1 - c_{i,j})\mathbb{P}(T_i \leq \tau_j + 1 | T_i > \tau_j) \\ &= (1 - c_{i,j})q_j \end{aligned}$$

So we can write

$$d_j = l_j q_j - q_j \sum_{i=1}^{w_j} (1 - c_{i,j})$$

Solving the formula for q_j gives us:

$$\hat{q}_j = \frac{d_j}{l_j - \sum_{i=1}^{w_j}(1 - c_{i,j})} = \frac{d_j}{l_j - w_j + \sum_{i=1}^{w_j} c_{i,j}}$$

$$\boxed{\hat{q}_j = \frac{d_j}{E_j}}$$

We obtained the estimation for mortality rates on time interval $[\tau_j, \tau_j + 1]$, corrected for censoring, using the Balducci hypothesis. Although it is generally not verified because mortality rates increase with time, the errors can be ignored as withdrawals are small compared to the population.

Specificities of mortality data, such as censuring and truncating, mean that exposure needs to be taken into account when estimating mortality rates.

4.3.2 Five-year mortality rates: a first look into the influence of each variable

In order to better understand the influence of certain factors on the mortality of melanoma of the skin patients and to compare our data with medical literature, we will compute the five-year mortality rates depending on these factors. In relation to the previous section, we need to compute the number of deaths and the global initial exposure in time interval $[0, 5]$ (in years). For the number of deaths, we need to separate the cases of non-metastatic and metastatic patients, because metastasis implies the spread of the cancer to other sites, meaning that the cause of death can be a different cancer or disease that was caused by the initial cancer. For non-metastatic patients, the number of deaths will be computed with the number of deaths caused by skin melanoma and for metastatic patients, all causes of death will be taken into account. For individual i, we compute the death variable:

$$d_{i,5}^{M0} = \begin{cases} 1 & \text{if Dead (due to skin melanoma) and survival} < 5 \text{ years} \\ 0 & \text{else} \end{cases}$$

$$d_{i,5}^{M1} = \begin{cases} 1 & \text{if Dead (all causes) and survival} < 5 \text{ years} \\ 0 & \text{else} \end{cases}$$

Then, we compute the individual initial exposure as presented in the last section:

$$\begin{aligned} e_{i,5} &= \begin{cases} 1 & \text{if } t_{i,5} > 1 \text{ and } c_{i,5} > 1 \\ 1 & \text{if } t_{i,5} < 1 \\ c_{i,5} & \text{if } c_{i,5} < 1 \end{cases} \\ &= \begin{cases} 1 & \text{if Vital\_status=0 or Survival\_years} \geq 5 \text{ Years} \\ \frac{\text{Survival\_years}}{5} & \text{else} \end{cases} \end{aligned}$$

Then, the estimation of the 5-year mortality rate is:

$$\hat{q}_5 = \frac{\sum_{i=1}^{l_5} d_{i,5}}{\sum_{i=1}^{l_5} e_{i,5}}$$

We will compare the estimated 5-year mortality rates depending on the following factors: metastatic state (M), tumor size (T), spread to regional lymph nodes (N), stage, gender, origin and age. As we stated earlier, we need to separate the cases of non-metastatic and metastatic cancers, but without taking into account the metastatic state and looking at deaths caused by melanoma of the skin only, the 5-year mortality rate is of 7.59%, ie the 5-year survival rate is of 92.41% which is very close to the survival rate of 93.7% given by the National Cancer Institute.

For the following, it is important to note that looking at mortality rates in different categories is tricky as groups may have different age distributions. Mortality rates by sex might capture correlation with age, if for example there are more observations of one gender than the other in some age groups.

The five-year mortality rate by metastatic state is given in table 24. As expected, the mortality rate for metastatic patients is considerably higher than for non-metastatic patients.

	Melanoma of the skin	All causes
M0	5.28%	
M1		83.00%

Table 24: Five-year mortality rate by presence of metastasis

We expect mortality rates to go up with larger tumors. In the non-metastatic case, except for missing and T0 tumor sizes, the thicker the tumor, the higher the mortality rate. For a tumor size T1, the 5-year mortality rate is of 1.23%, going up to 28.13% for tumor size T4. The high mortality rate for T0 can be explained by the small number of observations: there are only 496 records of it, whereas for T1 for example there are 67,243 observations. A small number of observations results in a high variance in the estimation. Individuals with missing tumor sizes have a five-year mortality rate which is close to individuals with T2 tumors. In the metastatic case, we have the same problem for individuals with missing and T0 tumor sizes. Another unexpected result is that $\hat{q}_5^{M1,T1} > \hat{q}_5^{M1,T2}0$. This last result can come from the small sample sizes that result in a great estimation variance: there are respectively 202 and 154 individuals with tumors sizes T1 and T2. All mortality rates for metastatic cancers are in a smaller interval than for non-metastatic cancers. This might be interpreted as a lower predictive importance of the size of the primary tumor for metastatic cancers, and this might be due to the fact that these cancers have already spread to other sites.

		Melanoma of the skin	All causes
M0	Missing	8.41%	
	T0	32.56%	
	T1	1.23%	
	T2	7.00%	
	T3	16.53%	
	T4	28.13%	
M1	Missing		85.93%
	T0		82.09%
	T1		79.87%
	T2		75.54%
	T3		82.90%
	T4		83.52%

Table 25: Five-year mortality rate by tumor size

The spread to regional lymph nodes indicates how far a cancer has spread. The mortality rates are expected to go up with a more extensive spread. Only stages III and IV exhibit regional lymph node spread. For non-metastatic cases, we observe as expected an increase in mortality rates with greater spread. For the metastatic case, we have the same issues as in the previous paragraphs: mortality rates do not behave as expected. This can be due either to the little number of observations or the fact that with metastasis, the cancer has necessarily spread to regional lymph nodes, so the information collection process was faulty.

		Melanoma of the skin	All causes
M0	Missing	6.02%	
	N0	3.30%	
	N1	23.59%	
	N2	33.11%	
	N3	47.55%	
M1	Missing		86.49%
	N0		78.99%
	N1		85.06%
	N2		74.21%
	N3		84.38%

Table 26: Five-year mortality rate by spread to regional lymph nodes

Looking at the stage of the cancer, it is common knowledge that more advanced stages imply higher mortality rates. This is what we observe here. Non-metastatic cancers can have grades ranging from I to III, and stage IV is defined by the presence of metastases. There are no missing stage information for individuals with stage IV cancers, because we imputed the missing values: remembering table 23, metastatic cancers are stage IV, which is how the information was completed for the variable.

		Melanoma of the skin	All causes
M0	Missing	5.36%	
	I	1.44%	
	II	14.79%	
	III	29.87%	
M1	IV		83.00%

Table 27: Five-year mortality rate by stage

Medically speaking, skin cancer behaves the exact same way for both genders. Mortality rates for the general population in the US are higher for men than women, which can be explained by numerous factors such as behavior differences. In insurance, gender is a sensitive variable and risk selection or pricing cannot discriminate by gender, so the same mortality rates have to be used for both men and women. In the case of melanoma of the skin, mortality rates are also higher for men. In the case of non-metastatic cancers, the five-year mortality rate for men is 2.56 points higher than for women and for metastatic cancers, it is 2.83 points higher.

		Melanoma of the skin	All causes
M0	Women	3.82%	
	Men	6.38%	
M1	Women		81.09%
	Men		83.92%

Table 28: Five-year mortality rate by gender

When looking at the origin variable provided by the SEER database, which classifies individuals into five categories (Hispanic, American Indian/Alaska Native, Asian or Pacific Islander, Black, White), we observe great disparity. The five-year mortality rates vary greatly between the categories, both for non-metastatic and metastatic cancers. As we saw previously, skin cancer is more common in individuals with fair complexions because of the lower melanin production. As a result, individuals with darker complexions are less likely to get this type of cancer and it often goes undetected for a longer time, which explains higher mortality rates. It could also be due to other factors such as socioeconomic status, which is still very related to origin in the US, where this data comes from. Another reason for these gaps is the variance of the estimation, as not all categories contain enough observations to have a robust estimation. It is interesting to notice that individuals with missing information about race/origin are predicted very low five-year mortality rates compared to all classes. This is caused by the low number of such observations, leading to a less robust estimator.

		Melanoma of the skin	All causes
M0	Hispanic	7.34%	
	Missing	0.07%	
	American Indian/AK Native	9.86%	
	Asian or Pacific Islander	11.73%	
	Black	18.82%	
	White	5.21%	
M1	Hispanic		86.53%
	Missing		35.71%
	American Indian/AK Native		88.45%
	Asian or Pacific Islander		93.85%
	Black		81.85%
	White		82.71%

Table 29: Five-year mortality rates by origin

Marital status can impact health behavior, leading to differences in mortality rates. In our data, we indeed have different estimated five-year mortality rates for all marital status classes. The individuals with missing information about their marital status have the lowest mortality rates, for both non-metastatic and metastatic cancers. For both metastatic states, divorced and widowed individuals have the highest five-year mortality rates, but these categories also correspond to an older population.

		Melanoma of the skin	All causes
M0	Divorced	9.71%	
	Married	5.86%	
	Missing	1.75%	
	Separated	8.56%	
	Single	6.27%	
	Unmarried	7.71%	
	Widowed	11.17%	
M1	Divorced		85.15%
	Married		81.61%
	Missing		77.78%
	Separated		80.51%
	Single		82.94%
	Unmarried		74.38%
	Widowed		90.75%

Table 30: Five-year mortality rates by marital status

This study of the five-year mortality rates depending on some characteristic is coherent with medical literature when looking at non-sensitive attributes concerning the tumors. But for the three sensitive variables, all things equal otherwise, the rates should be the same for all groups. It is not the case here. This could be caused by an unequal representation of other characteristics among groups, like age for example. All in all, a model using this data has a good chance of giving biased outcomes.

4.3.3 A different data structure needed to use standard models: pseudo table

As mentioned previously, time discretization and data structure modification allow the use of standard Machine Learning models to predict the number of death in the desired time interval. This data structure modification is done by creating pseudo tables: for each individual, we create as many rows as the maximal duration. We consider times in years rather than in months, as it seems granular enough for our modeling purposes. In each row, we will have a death variable indicating if the individual died during that time interval and the initial exposure, which will then be used as a weight or an offset when applying the standard models to predict the number of deaths. Another advantage of the approach is that we can have time-varying information about an individual, such as the evolution of his marital status or of the size of his tumor. Unfortunately, in the SEER database, we do not have access to this kind of information, so we can question the relevance of using such variables.

The step-by-step process is as follows. We have for each individual i his age at diagnosis, year of diagnosis, survival time in months and a variable indicating if he died because of melanoma of the skin. The first step is to compute the survival time in years and the maximal duration, which is the ceiling of the survival time in years. Then, we create as many rows for each individual i as his/her maximal duration. We can then compute for each of the lines j the duration since diagnosis in years, and the age, year and death due to melanoma in that time interval. Finally, we can compute the individual exposure are we saw in the previous section. For each time interval j: $[\tau_j, \tau_j + 1]$ which corresponds to a year, we compute the variables of interest. The duration since diagnosis (in years) is

$$\text{Duration}_j = j - 1$$

the death variable (due to melanoma, in the time interval k) is

$$d_{i,j} = \begin{cases} 1 & \text{if Death\_melanoma} = 1 \text{ and } j = \text{Max\_duration} \\ 0 & \text{else} \end{cases}$$

the individual initial exposure is

$$e_{i,j} = \begin{cases} 1 & \text{if } Y_{i,j} = 1 \text{ or } j < \text{Max\_duration} \\ S_i - \text{Duration}_j & \text{else} \end{cases}$$

We will create the pseudo table step-by-step, starting from the raw data from table 31:

- Individual i=1:
  - For the survival time in years, we just convert the number of months into a number of years: $S_1 = 25/12 = 2.080$.
  - For the maximal duration in years, we take the ceiling of the survival time in years: Max_duration $= \lceil S_1 \rceil = \lceil 2.08 \rceil = 30$.
  - So we create 3 lines $j = 1, 2, 3$ for this individual:
    - j=1: $\text{Duration}_1 = j - 1 = 0$
      Age=Age_dx=25
      Year=Year_dx=2002
      Death_melanoma $d_{1,1} = 0$ because Death_melanoma $\neq 1$
      Exposure $e_{1,1} = 1$ because $j = 1 < 3 =$ Max_duration
    - j=2: $\text{Duration}_2 = j - 1 = 1$
      Age=Age_dx+1=26
      Year=Year_dx+1=2003
      Death_melanoma $d_{1,2} = 0$ because Death_melanoma $\neq 1$
      Exposure $e_{1,2} = 1$ because $j = 2 < 3 =$ Max_duration

* j=3: Duration$_3 = j - 1 = 2$
  Age=Age_dx+2=27
  Year=Year_dx+2=2004
  Death_melanoma $d_{1,3} = 0$ because Death_melanoma $\neq 1$
  Exposure $e_{1,3} = S_1 - D_{1,3} = 2.08 - 2 = 0.08$ because $Y_{1,3} \neq 1$ and $j = 3 =$ Max_duration

- Individual i=2:
  - $S_2 = 4/12 = 0.330$.
  - Max_duration $= \lceil S_2 \rceil = \lceil 0.33 \rceil = 10$.
  - So we create 1 line $j = 1$ for this individual:
    * j=1: Duration$_1 = j - 1 = 0$
      Age=Age_dx=37
      Year=Year_dx=2004
      Death_melanoma $d_{2,1} = 1$ because Death_melanoma $= 1$ and $j = 1 =$ Max_duration
      Exposure $e_{2,1} = 1$ because $Y_{2,1} = 1$

- Individual i=3:
  - $S_3 = 58/12 = 4.830$.
  - Max_duration $= \lceil S_3 \rceil = \lceil 4.83 \rceil = 50$.
  - So we create 5 lines $j = 1, 2, 3, 4, 5$ for this individual:
    * j=1: Duration$_1 = j - 1 = 0$
      Age=Age_dx=56
      Year=Year_dx=2010
      Death_melanoma $d_{3,1} = 0$ because $j = 1 \neq 5 =$ Max_duration
      Exposure $e_{3,1} = 1$ because $j = 1 < 5 =$ Max_duration
    * j=2: Duration$_2 = j - 1 = 1$
      Age=Age_dx+1=57
      Year=Year_dx+1=2011
      Death_melanoma $d_{3,2} = 0$ because $j = 2 \neq 5 =$ Max_duration
      Exposure $e_{3,2} = 1$ because $j = 2 < 5 =$ Max_duration
    * j=3: Duration$_3 = j - 1 = 2$
      Age=Age_dx+2=58
      Year=Year_dx+2=2012
      Death_melanoma $d_{3,3} = 0$ because $j = 3 \neq 5 =$ Max_duration
      Exposure $e_{3,3} = 1$ because $j = 3 < 5 =$ Max_duration
    * j=4: Duration$_4 = j - 1 = 3$
      Age=Age_dx+3=59
      Year=Year_dx+3=2013
      Death_melanoma $d_{3,4} = 0$ because $j = 4 \neq 5 =$ Max_duration
      Exposure $e_{3,4} = 1$ because $j = 4 < 5 =$ Max_duration
    * j=5: Duration$_5 = j - 1 = 4$
      Age=Age_dx+4=60
      Year=Year_dx+4=2014
      Death_melanoma $d_{3,5} = 1$ because Death_melanoma $= 1$ and $j = 5 =$ Max_duration
      Exposure $e_{3,5} = 1$ because $Y_{3,5} = 1$

In the end, we get table 32.

i	Age_dx	Year_dx	Survival (months)	Death_melanoma
1	25	2002	25	0
2	37	2004	4	1
3	56	2010	58	1

Table 31: Raw survival data

i	Survival (years) S_j	Max_duration (years)	j	Duration$_j$ (years)	Age	Year	Death_melanoma_j $d_{i,j}$	Exposure $e_{i,j}$
1	25/12=2.08	3	1	0	25	2002	0	1
			2	1	26	2003	0	1
			3	2	27	2004	0	0.08
2	4/12=0.33	1	1	0	37	2004	1	1
3	58/12=4.83	5	1	0	56	2010	0	1
			2	1	57	2011	0	1
			3	2	58	2012	0	1
			4	3	59	2013	0	1
			5	4	60	2014	1	1

Table 32: Pseudo table created from the raw survival data

In order to use standard Machine Learning models on mortality data, we need to take into account the exposure. One way of doing so is by transforming the data into a pseudo table in which each line represents one year of an individual's life with all of the associated characteristics, the exposure and an indicator of the occurrence of death within that year.

4.4 Descriptive statistics

We will perform the descriptive statistics analysis on the pseudo table. As mentioned before, our study is restricted to non-metastatic cancers.

4.4.1 Variable identification

The 28 variables available after mapping and applying underwriting and medical constraints are described in appendix D. They can be categorized into 3 groups: the explanatory variables, the variable of interest and the weight variable. Within the explanatory variables category, we have non-sensitive and sensitive variables.

Non-sensitive explanatory variables They are the ones that are pure descriptors of the medical situation, for example the age, the type of cancer (`AYA_site_recode20`), the time since diagnosis (`DURATION`) or how far the cancer has spread (`Extent`). Some of them are continuous and others categorical.

Sensitive explanatory variables They are the ones that can introduce discrimination. They directly concern the individual: sex, origin and marital status.

Table 33 gives the number of observations by sex. The group imbalance ratio is

$$\text{IR}_{Sex} = \frac{287,820}{249,118} = 1.16$$

which is close to 1.

Sex	Observations
Female	249,118
Male	287,820

Table 33: Number of observations by sex

Table 34 gives the number of observations by origin. The majority population is the 'White' category, with almost 45 times more observations than the second biggest category, 'Missing'.

Origin	Observations
Hispanic	11,034
Missing	11,510
American Indian/AK Native	1,157
Asian or Pacific Islander	3,597
Black	1,767
White	507,873

Table 34: Number of observations by origin

Table 35 gives the number of observations by marital status. The majority class is 'Married' and the second largest is 'Missing'.

Marital status	Observations
Divorced	26,160
Married	272,743
Missing	159,649
Separated	1,946
Single	63,729
Unmarried	674
Widowed	12,037

Table 35: Number of observations by marital status

Variable of interest It is the variable that we will predict with our model: the event of interest is the death due to skin melanoma of the individual in the year (`Death_skin`). Table 36 gives the number of observations for each value of `Death_skin`. There is a great imbalance between classes: the group imbalance ratio is

$$\text{IR}_{Death\_skin} = \frac{533,167}{3,771} = 141.39$$

meaning there are 141.39 more observations for `Death_skin`=0 than for `Death_skin`=1.

Death_skin	Observations
0	533,167
1	3,771

Table 36: Number of observations by death_skin

The weight variable As mentioned previously, because of the characteristics of survival data, we need to take into account the exposure of individuals to predict mortality. The exposure variable will therefore be used as a weight in the model.

4.4.2 Multivariate analysis

We will now study the relationships between variables. For that, we will focus on correlations. As we have already mentioned multiple times before, absence of correlation is not equivalent to independence.

As we have many variables and had to turn categorical features into binary data, we end up with a large correlation matrix. The keys points are:

- The bright blue colors are negative correlations between different categories of the same variables, for example between `Sex_Female` and `Sex_Male`.
- Generally speaking, there are correlations between `Tumor`, `Positive_Node`, `Stage`, `Extent`, `Ulceration`, `Mitotic rate`, and `Surg_LN/oth/primsite`, which are linked with each other as they are the medical variables that allow the diagnosis.
- Time variables are naturally correlated with each other: age with age at diagnosis, year with year of diagnosis and duration.
- There are a few light correlations with the variable of interest, all less than 0.13 (in absolute value).
- Concerning the missing values, the dummy variables for Missing values of `Tumor`, `Stage` and `Positive_Nodes` are strongly positively correlated with each other. It seems coherent, as the stage is defined thanks to the TNM system. Missing categories of other variables are correlated with these variables and with each other: `Surg_primsite`, `Surg_LN`, `Surg_oth`, `Extent`, `Ulceration`, and these variables also describe the gravity of the cancer. A Missing origin or marital status is lightly positively correlated with the previously mentioned variables. We can conclude that when there are Missing values for an individual, they are for all or most of the mentioned categories.

With this study of correlations, we can conclude that under the modeling constraints, not all variables can be kept.

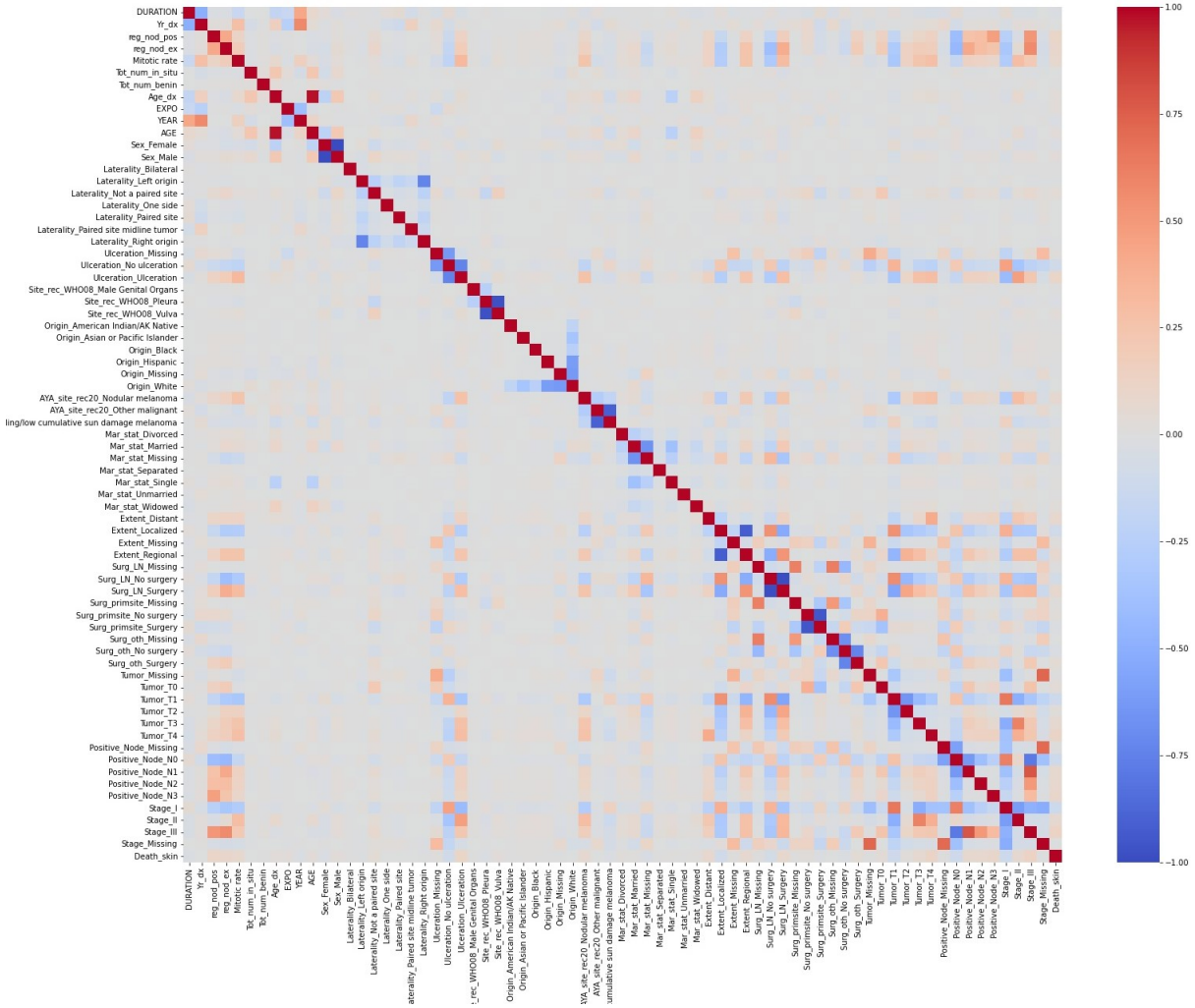


Figure 24: Heatmap of correlations

We observe that there are some correlations between the sensitive and non-sensitive variables, which indicates that a model will be discriminatory, even if it only uses the non-sensitive ones.

5 Discrimination mitigation applied to real mortality data

5.1 Adapting the logistic regression to survival data

If we considered a simple logistic regression to predict the event of death $\delta_{i,j} \sim \mathcal{B}(q_j)$, we would not take into account censoring. Instead, we assign weights to the sample with the individual initial exposure. This way, we estimate the mortality rate $q_j = \mathbb{P}(T \leq \tau_j + 1 | T > \tau_j)$ with $\hat{q}_j = \frac{\sum_i \delta_{i,j}}{\sum_i e_{i,j}} 0$.

Proof. The weighted likelihood of the model is

$$L = \prod_{i=1}^{l_j} q_j^{\delta_{i,j} e_{i,j}} (1 - q_j)^{(1-\delta_{i,j}) e_{i,j}}$$

The log likelihood to maximize is

$$log(L) = \sum_{i=1}^{l_j} \Big(\delta_{i,j} e_{i,j} log(q_j) + (1 - \delta_{i,j}) e_{i,j} log(1 - q_j) \Big)$$

Deriving with respect to q_j, we obtain

$$\frac{dlog(L)}{dq_j} = \sum_{i=1}^{l_j} \Big(\frac{\delta_{i,j} e_{i,j}}{q_j} - \frac{e_{i,j} - \delta_{i,j} e_{i,j}}{1 - q_j} \Big)$$

Equalizing with 0, we get

$$(1 - \hat{q}_j) \sum_{i=1}^{l_j} \delta_{i,j} e_{i,j} = \hat{q}_j \sum_{i=1}^{l_j} \big(e_{i,j} - \delta_{i,j} e_{i,j} \big)$$

$$\text{ie } \hat{q}_j = \frac{\sum_i \delta_{i,j} e_{i,j}}{\sum_i e_{i,j}} = \frac{\sum_i \delta_{i,j}}{\sum_i e_{i,j}} \text{ as } \delta = 1 \Rightarrow e = 1$$

□

Using the exposure of each line of the pseudo table as a weight in the logistic regression allows us to use the standard model computed in Python.

5.2 Variable selection

In order to select the variables that are relevant to the model performance, we begin with a model with all variables. Table 37 gives the coefficients of the regression, with their standard errors and p-values. We will only keep the variables with a p-value below or equal to 0.005, as it is good practice in modeling. The p-values that are above 0.005 are highlighted in orange. Variables that are not statistically significant are:

- `AGE`
- `reg_nod_ex`
- `Tot_num_benin`

- AYA_site_rec20
- Surg_LN
- Surg_oth

This does not necessarily mean that these variables are not medically relevant.

Variable	Coefficient	Standard error	P-value
const	174.9219	13.260	0.000
DURATION	-0.0191	0.003	0.000
Yr_dx	-0.0358	0.004	0.000
Age_dx	0.0207	0.002	0.000
YEAR	-0.0549	0.003	0.000
AGE	0.0016	0.002	0.396
Sex_Female	-0.3927	0.043	0.000
reg_nod_pos	0.0289	0.007	0.000
reg_nod_ex	0.0003	0.002	0.849
Mitotic rate	0.0528	0.007	0.000
Tot_num_in_situ	-0.0742	0.023	0.002
Tot_num_benin	0.0777	0.170	0.648
Laterality_Bilateral	0.6875	0.331	0.038
Laterality_Not a paired site	0.4738	0.061	0.000
Laterality_One side	0.0624	0.200	0.756
Laterality_Paired site	0.2974	0.093	0.001
Laterality_Paired site midline tumor	0.1334	0.128	0.296
Laterality_Right origin	-0.0520	0.044	0.236
Ulceration_Missing	0.1514	0.099	0.128
Ulceration_Ulceration	0.5980	0.048	0.000
Site_rec_WHO08_Male Genital Organs	0.0514	0.565	0.928
Site_rec_WHO08_Vulva	0.5930	0.183	0.001
Origin_American Indian/AK Native	0.3195	0.293	0.276
Origin_Asian or Pacific Islander	0.1691	0.156	0.280
Origin_Black	0.4672	0.167	0.005
Origin_Hispanic	0.2042	0.108	0.060
Origin_Missing	-3.0099	1.001	0.003
AYA_site_rec20_Nodular melanoma	0.1249	0.052	0.016
AYA_site_rec20_Superficial spreading /low cumulative sun damage melanoma	-0.0319	0.048	0.510
Mar_stat_Divorced	0.3849	0.064	0.000
Mar_stat_Missing	-0.4411	0.072	0.000
Mar_stat_Separated	0.3068	0.225	0.173
Mar_stat_Single	0.2578	0.055	0.000
Mar_stat_Unmarried	0.3127	0.426	0.463
Mar_stat_Widowed	0.3041	0.087	0.000
Extent_Distant	0.7494	0.078	0.000
Extent_Missing	0.6607	0.171	0.000
Extent_Regional	0.4792	0.057	0.000
Surg_LN_Missing	0.0226	0.347	0.948
Surg_LN_Surgery	-0.1323	0.057	0.020

Surg_primsite_Missing	0.7475	0.335	0.025
Surg_primsite_No surgery	0.5662	0.136	0.000
Surg_oth_Missing	-0.6858	0.474	0.148
Surg_oth_Surgery	0.2390	0.095	0.012
Tumor_Missing	1.0042	0.173	0.000
Tumor_T0	0.4037	0.203	0.047
Tumor_T2	0.9187	0.076	0.000
Tumor_T3	1.0184	0.094	0.000
Tumor_T4	1.2887	0.097	0.000
Positive_Node_Missing	0.9338	0.214	0.000
Positive_Node_N1	-0.1761	0.588	0.764
Positive_Node_N2	0.0468	0.590	0.937
Positive_Node_N3	0.4810	0.594	0.418
Stage_II	0.5843	0.083	0.000
Stage_III	1.5105	0.590	0.011
Stage_Missing	-0.3659	0.207	0.077

Table 37: Coefficients of the complete model,
p-values in orange when above the threshold of 0.005

For our final model, the selected variables need to be statistically relevant, consistent with medical literature and available at the underwriting stage. We have discussed with the oncologist of the team, drawing the following conclusions:

- For time variables, the most important ones are the current age and duration. As we had strong correlations between age and age at diagnosis, we will only keep age. The year does not influence mortality, as we only have data from recent years.
- Laterality does not influence mortality, but it statistically relevant so we will keep the variable in the model.
- Fairer complexions are at a higher risk of developing the cancer, but once the cancer has developed, it does not influence mortality, so all things equal otherwise, we should obtain the same mortality rates for all Origins. Unfortunately, and especially in the US, Origin is very often a proxy for social economic status, which influences access to healthcare for example, which in turn has consequences on mortality risks.
- Stage is a combination of the TNM staging variables, so there are strong correlations between them, and there is no use in keeping them all. So we will not keep the `Stage` variable.

To conclude, we will keep the `AGE` and `Laterality` variables. We note that we are left with the three sensitive variables, `Sex`, `Origin` and `Mar_stat`.

The analysis of the coefficients helped us select the statistically relevant variables and we confirmed their medical coherence with an oncologist. We have three statistically significant protected variables.

5.3 Regression model with no pre-processing step

As we have done in the simulated case, we begin by applying the regression model without any particular modifications. This will allow a comparison with other methods. As we are left with many variables, the coefficients of the regression are given in appendix E.

Figure 25 gives a visual understanding of the coefficients. We have an intercept and only kept the least represented classes for the categorical variables, so the coefficients for the remaining categorical variables represent the relative risk compared to the most represented class. For example, for the variable `Sex`, as we had more observations for Male, we only kept the variable `Sex_Female` in our model. The coefficient attributed to this variable by the logistic regression indicates how much more chances of dying a Female has, compared to a Male. For numerical variables, the coefficients indicate how much the chances of dying change for an increase of one in the variable value. We have also included the standard errors, represented by the black lines.

- `DURATION`: the higher the duration, the lower the probability of dying.
- `AGE`: the older the individual, the higher his probability of dying.
- `Sex`: compared to being a Male, being a Female decreases the probability of dying.
- `reg_nod_pos`: the higher the number of positive regional nodes, the higher the probability of dying.
- `Mitotic rate`: the higher the density of mitoses, the higher the probability of dying.
- `Laterality`: compared to the Left origin category, a Right origin lowers the probability of death and other categories increases it.
- `Ulceration`: compared to No ulceration, other categories increases the death probability, Ulceration more than Missing.
- `Site_rec_WHO08`: compared to tumors originating on the Pleura, tumors originated on the vulva and on male genital organs increase the probability of dying.
- `Origin`: compared to White, Missing largely decreases the probability of dying. Other categories increase it, in the increasing order we have Asian or Pacific Islander, Hispanic, American Indian/AK Native and Black.
- `Marital status`: compared to Married, Missing decreases the probability of death. Other categories increase it, in the increasing order we have Unmarried, Single, Separated, Widowed and Divorced. We must take into account the standard errors, which are larger than the coefficients for Unmarried and Separated.
- `Extent`: compared to Localized, all categories increase the probability of dying, in the increasing order we have Missing, Regional and Distant.
- `Surg_primsite`: compared to Surgery, other categories increase the probability of dying, No surgery more than Missing.
- `Tumor`: compared to T1, larger tumors increase the probability of dying in size order. Missing increases the same way as T2. But we have an anomaly for T0 as it seems it increases the probability of dying compared to T1. This might be due to the small sample size, as we saw when we were looking at the five-year mortality rates.
- `Positive_Node`: compared to N0, larger numbers of positives nodes increase the probability of dying. Missing increases it, but not at much as the larger sizes.

The coefficients are coherent with medical findings: a higher number of positive regional nodes, the presence of ulceration, a further extent, and a larger tumor all point towards a severe cancer, which implies lower survival odds.

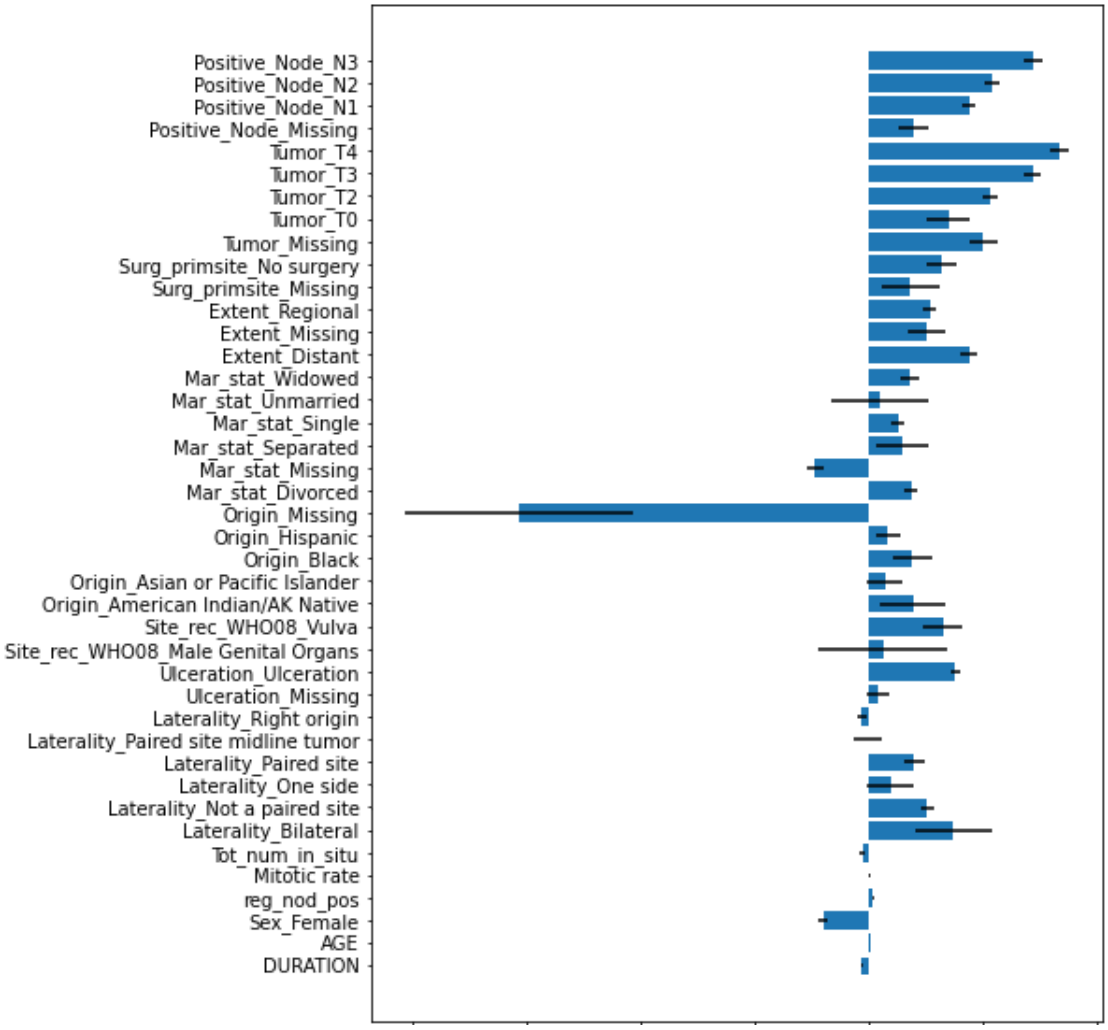


Figure 25: Coefficients and their standard errors for the model with selected variables

Performance evaluation Figure 26 gives the ROC curve. The performance of the model is almost the same as with all variables, with an AUC of 0.8769.

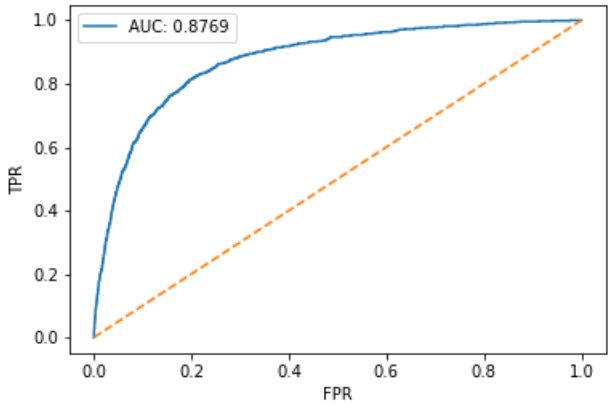


Figure 26: ROC curve for the model with selected variables

In order to compute fairness metrics, we need to classify individuals into two categories $\hat{Y} = 0$ or 1, $\hat{Y}$ being the event of death due to skin melanoma, 1 if the event occurred ie the individual died of this cause and 0 otherwise. As the probabilities of death are all very close to 0, we cannot use the standard threshold of 0.5 to classify individuals. We will select the threshold by looking at the true percentage of dead individuals. In the entire dataset, we have about 0.7% of dead individuals, as we saw in the section on descriptive statistics. We will therefore look for a threshold τ such that

$$\hat{Y} = 1 \text{ if } \mathbb{P}(\hat{Y} = 1) \geq \tau \text{ and 0 otherwise, giving 0.7\% of individuals with } \hat{Y} = 1$$

We find $\tau = 0.10880$. This value gives the confusion matrix as shown in figure 38. We have, as planned, about 0.7% individuals classified with $\hat{Y} = 10$. The class $\hat{Y} = 0$ is the positive class here.

	$\hat{Y}$=0	$\hat{Y}$=1
$Y = 0$	106,236	699
$Y = 1$	662	95

Table 38: Confusion matrix (model with selected variables)

Fairness evaluation We computed the acceptance rate and the true and false positive rates, and looked at them globally, by sex, by origin and by marital status. To conduct an extensive analysis, we should also look at subgroups, crossing information about the three sensitive attributes, like we did in the simulated case, but it becomes intricate as we have many different categories, and we risk having too few observations in one of the subcategories.

- Globally, we have values close to 100% for the acceptance and true positive rates, due to the numerous classifications into the positive class compared to the negative class. The false positive rate is quite high, as we have 6 times more false positives than true negatives.

(%)	Global
AR	99.26
TPR	99.35
FPR	87.45

Table 39: Global fairness metrics (model with selected variables)

- By sex
  - The acceptance rates are very close, only 0.76 points apart, but higher for Female.
  - The true positive rates are very close for both genders, only 0.69 points apart, but higher for Female.
  - The false positive rates are 3.87 points part, and higher for Female.

  All in all, looking at the three definitions of fairness, the Male sex is always disadvantaged by this model, although the gaps between metrics are not large.

(%)	Female	Male	Difference
AR	99.67	98.91	0.76
TPR	99.71	99.03	0.69
FPR	90.18	86.30	3.87

Table 40: Fairness metrics by sex (model with selected variables)

- By origin
  - The acceptance rates are between 92.53% and 100%. The lowest is for Black and the highest is for Missing.
  - The true positive rates are between 93.21% and 100%. The lowest is for Black and the highest for Missing.
  - The false positive rates are very distant. The lowest FPR is for Missing at 0%, then 89.20% for White. The highest is for American Indian/Alaska Native at 100.00%.

  It is interesting to note that the Missing category only has true positive classifications, which explains the values for the fairness metrics. Going back on the regression weights of figure 25, the `Origin_Missing` variable has a colossal negative coefficient compared to other dummy variables, and especially compared to other variables for `Origin`. Whenever the origin of an individual was not collected, his probability of dying within the year is null. We can wonder about the data collection process: perhaps individuals with zero to no chance of dying were not as 'interesting' for the database.
  Black is the most disadvantaged group under the statistical parity and equal opportunity definitions. For the equalized odds definition, Black has the lowest TPR but White has the lowest FPR, so we cannot conclude on the most disadvantaged group.

(%)	White	Hispanic	Black	Asian or Pacific Islander	American Indian /AK Native	Missing	Var
AR	99.33	97.70	92.53	96.63	98.72	100.00	6.10e−4
TPR	99.40	97.94	93.21	97.53	100.00	100.00	5.52e−4
FPR	89.20	97.94	93.21	97.53	100.00	0.00	8.35e−2

Table 41: Fairness metrics by origin (model with selected variables)

- By marital status
  - The acceptance rate is lowest for Separated and highest for Missing, then Married.
  - The true positive rates are between 97.54% and 99.99%, lowest for Widowed and highest for Missing, then Married.
  - The false positive rate is lowest for Widowed, at 78.79% and highest for Separated and Unmarried, at 100%.

  The Separated class is the most disadvantaged under the statistical parity definition, Widowed under the equal opportunity and equalized odds definitions.

(%)	Married	Missing	Single	Widowed	Divorced	Separated	Unmarried	Var
AR	99.23	99.99	98.82	97.29	97.47	95.75	98.44	1.72e−4
TPR	99.33	99.99	98.95	97.54	97.69	95.71	98.43	1.71e−4
FPR	88.16	97.10	84.30	78.79	82.19	100.00	100.00	6.76e−3

Table 42: Fairness metrics by marital status (model with selected variables)

To conclude, the most disadvantaged group is not the same depending on the definition of fairness that is used.

As expected, and like in the simulated case, a model using all variables is unfair to some groups, demonstrating effects of direct discrimination. Interestingly, the most disadvantaged group is not the same under all three definitions of fairness, showing once again their incompatibility.

5.4 Removing protected variables to avoid direct discrimination

We will now compare the model with the protected variables to the model without protected variables. As we saw in the simulated case, removing the protected variables is not the perfect solution as it does not prevent discrimination. Simply ignoring the sensitive attributes gives larger weights to variables that are correlated with them.

Here, we will apply a regression model to all variables except the sensitive ones. The coefficients are given in appendix F.

Figure 27 shows the importance of each variable. The main difference that we can observe, compared to the model with all variables, is that the variable `Site_rec_WHO08_Male Genital Organs` now has a negative coefficient, with a consequent standard error. Other than that, the weights stayed in the same order, with some amplitude variations.

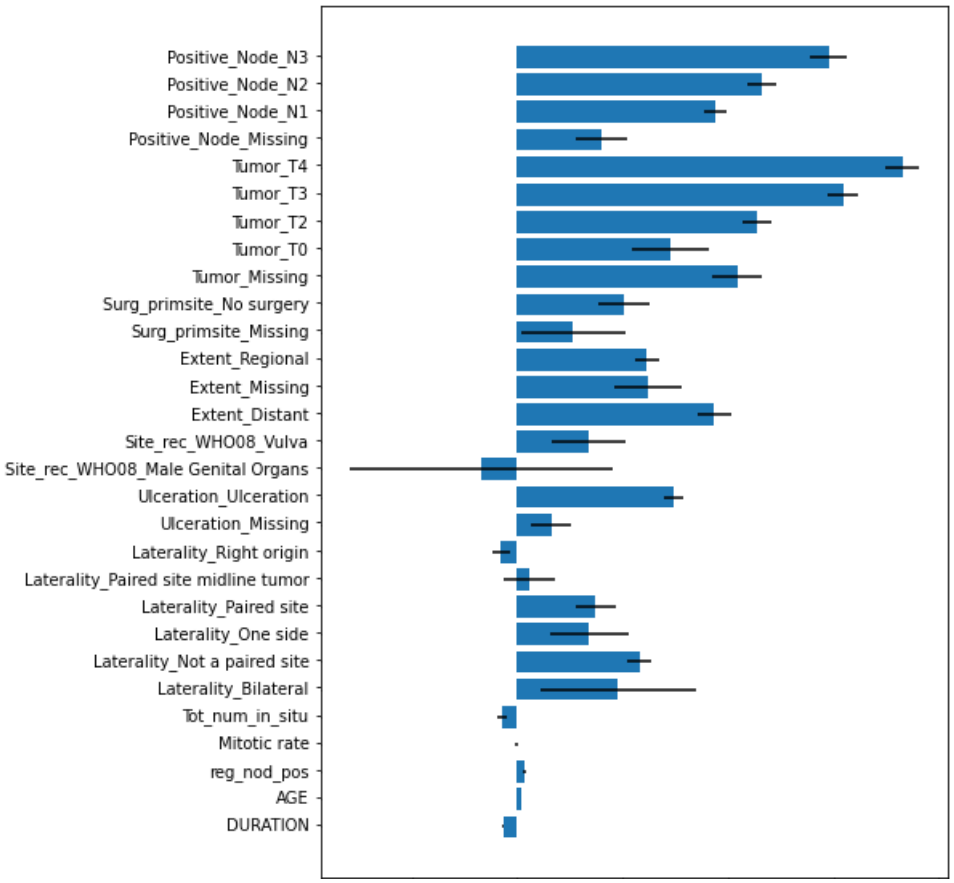


Figure 27: Coefficients and their standard errors (black lines), model without protected variables

Performance evaluation Figure 28 gives the ROC curve. The model with all selected variables had an AUC of 0.8769, and without protected variables the model has an AUC of 0.8778, which is slightly higher.

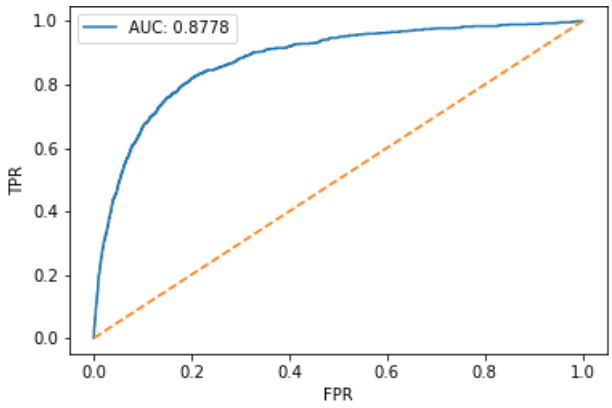


Figure 28: ROC curve for the model without the protected variables

Fairness evaluation

- Globally

  The acceptance rate, true positive rate and false positive rate have slightly decreased compared to the model using all variables.

With all variables		Without protected variables	
(%)	Global	(%)	Global
AR	99.26	AR	99.20
TPR	99.35	TPR	99.29
FPR	87.45	FPR	86.26

Table 43: Global fairness metrics

- By sex
  - The acceptance rates difference has decreased, and Female kept the higher value.
  - The true positive rates for Female and Male are closer than in the model with all variables. The Female category still has a higher TPR.
  - The false positive rates are slightly closer than previously. The Female group still has a higher FPR.

  Male is still disadvantaged under all three fairness definitions, although the gaps in metrics have narrowed.

With all variables				Without protected variables			
(%)	Female	Male	Difference	Sex	Female	Male	Difference
AR	99.67	98.91	0.76	AR	99.46	98.97	0.49
TPR	99.71	99.03	0.69	TPR	99.52	99.10	0.42
FPR	90.18	86.30	3.87	FPR	88.76	84.97	3.79

Table 44: Fairness metrics by sex

- By origin
  - The lowest acceptance rate has increased and the highest has stayed at 100%, for Missing. Aside from this category, the highest AR is still for White and the lowest still for Black.

- Black still has the lowest true positive rate, but it has increased. Missing still has the highest TPR, at 100%. The second highest is for American Indian/AK Native, which had a TPR of 100% in the previous model.
- Apart from Missing, White had the lowest false positive rate but now Black does. American Indian/AK Native had the highest FPR but now Hispanic does.

Black is still the most disadvantaged group under the statistical parity and equal opportunity definitions. With the previous model, we could not conclude on the most disadvantaged group under the equalized odds definition, but here without taking into account the value for Missing, Black is the most disadvantaged group.

With all variables

(%)	White	Hispanic	Black	Asian or Pacific Islander	American Indian /AK Native	Missing	Var
AR	99.33	97.70	92.53	96.63	98.72	100.00	6.10e−4
TPR	99.40	97.94	93.21	97.53	100.00	100.00	5.52e−4
FPR	89.20	97.94	93.21	97.53	100.00	0.00	8.35e−2

Without protected variables

(%)	White	Hispanic	Black	Asian or Pacific Islander	American Indian /AK Native	Missing	Var
AR	99.23	98.76	93.85	96.68	99.10	100.00	4.37e−4
TPR	99.31	98.88	95.32	97.05	99.54	100.00	2.71e−4
FPR	87.07	89.66	45.45	70.00	66.67	0.00	9.28e−2

Table 45: Fairness metrics by origin

- By marital status
  - Compared to the previous model, the acceptance rates are closer to each other, with the lowest value going from 95.75% for Separated to 97.48% for Unmarried and the highest value going from 99.99% to 99.89%, both for Missing.
  - The true positive rates are also closer to each other. Widowed had the lowest TPR, but now Unmarried does. Missing still has the highest.
  - The false positive rates are also closer to each other. Widowed still has the lowest FPR and Unmarried still has the highest. Separated had the highest with Unmarried, but now has a lower FPR.

For statistical parity, we are closer to fairness but the most disadvantaged group is not the same. For equal opportunity and equalized odds, we are a little closer to fairness and Widowed is still the most disadvantaged group.

With all variables

(%)	Married	Missing	Single	Widowed	Divorced	Separated	Unmarried	Var
AR	99.23	99.99	98.82	97.29	97.47	95.75	98.44	1.72e−4
TPR	99.33	99.99	98.95	97.54	97.69	95.71	98.43	1.71e−4
FPR	88.16	97.10	84.30	78.79	82.19	100.00	100.00	6.76e−2

Without protected variables

(%)	Married	Missing	Single	Widowed	Divorced	Separated	Unmarried	Var
AR	99.01	99.89	99.06	97.74	98.14	97.75	97.48	6.85e−5
TPR	99.12	99.91	99.15	98.13	98.29	97.96	97.44	6.23e−5
FPR	85.43	92.75	87.96	78.72	87.32	85.71	100.00	3.78e−3

Table 46: Fairness metrics by marital status

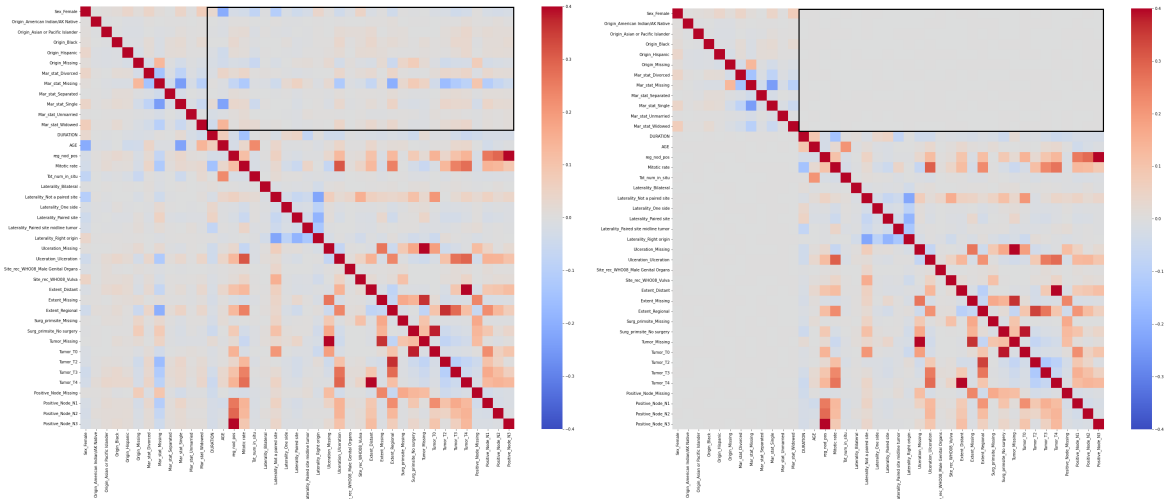
To conclude, the performance metrics have not deteriorated too much compared to the model using all variables. For Sex, the fairness is better under all three definitions and the same group remains in a disadvantage position. For Origin, we also have the same most disadvantaged group as in the model using all variables, for some metrics the fairness has improved and for others it has deteriorated. For marital status, fairness has improved but for one of the definitions the most diasadvantaged group is not the same as before.

Once again, we have the same mechanisms as in the simulated case: depending on the variables, and so on the dependence structure between sensitive and non-sensitive variables, for some definitions the fairness improves and for others it deteriorates. The predictive performance of the model does not vary by much.

5.5 Transforming the non-protected variables to mitigate indirect discrimination

As we have done for the simulated data, we will change the basis formed by the centered variables to obtain transformed non-sensitive explanatory variables uncorrelated with the sensitive ones. We are dealing with 12 sensitive variables, because the sensitive variables `Sex`, `Origin` and `Mar_stat` have been converted to dummy variables. The procedure is the exact same as in the simulated case, but with more variables.

Application to the data We apply the procedure to all explanatory variables (ie not to the exposure nor the variable of interest) and obtain the projected non-sensitive variables that are uncorrelated to the sensitive ones. Figure 29 shows the correlation matrix before and after orthogonal projection. The correlations between the sensitive and other variables are framed in blacked. On the left, we can see a few correlations (blue and red boxes) and on the right, they are all null (only grey boxes). As expected, these results are the same as in the simulated case.



(a) Original data (b) Transformed data

Figure 29: Heatmap of correlations, before and after change of basis, correlations between the sensitive attributes and the others framed in black

The correlation matrix shows the success of our method: the transformed non-sensitive variables are no longer correlated with the sensitive ones.

Applying the model We will now apply the logistic regression model to the projected non-sensitive variables, keeping the exposure as weights, to predict the probabilities of death due to melanoma of the skin. We will compare the performance and fairness metrics of this model to the ones of the model without protected variables. The coefficients are presented in appendix G.

Figure 30 helps us visualize the coefficients and their standard errors. The most noticeable difference with the coefficients of the model without protected variables concerns `Site_rec_WHO08_Male Genital Organs` which still has a large standard error but a positive coefficient instead of a negative one. We have the oppposite effect for `Laterality_Paired site midline tumor`.

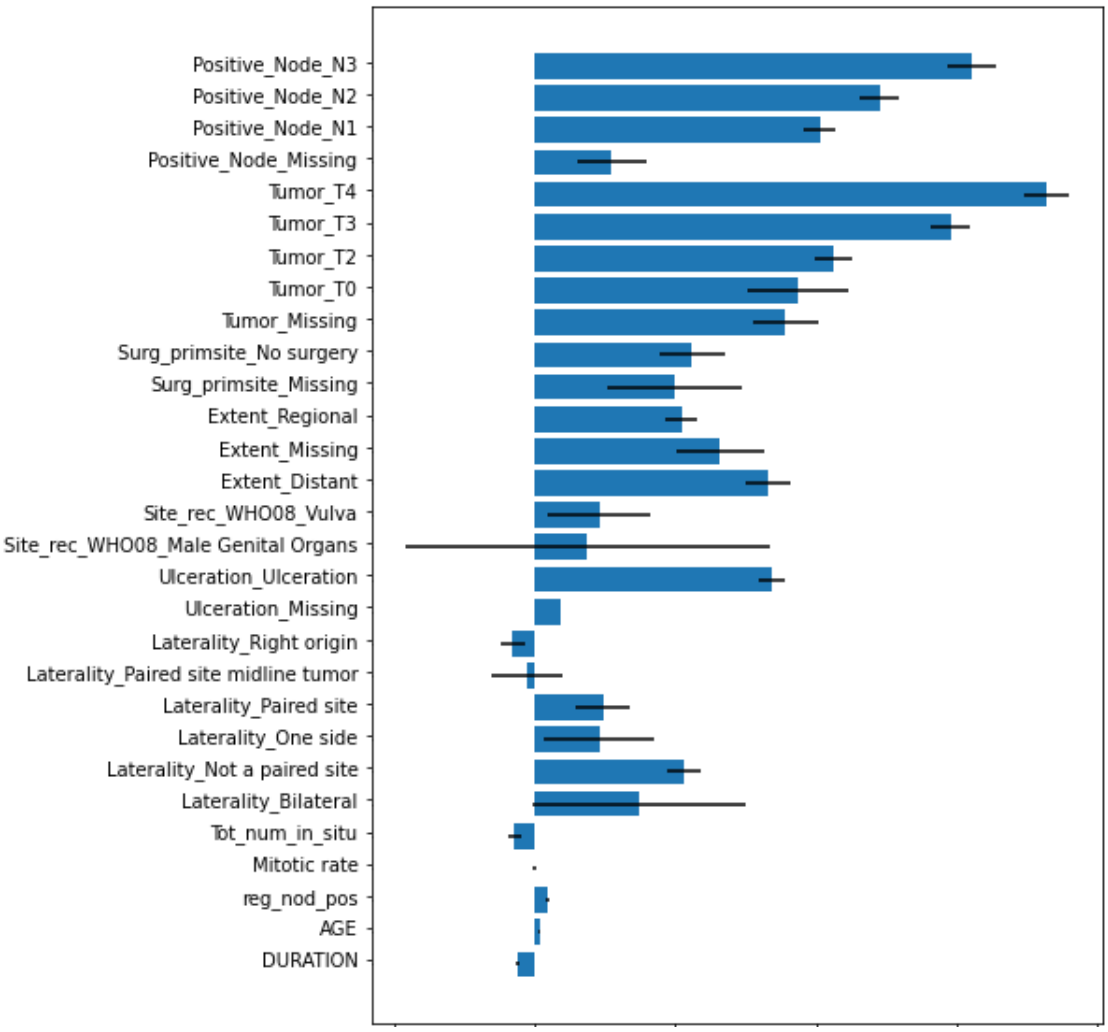


Figure 30: Coefficients and their standard errors (model with transformed variables)

Performance evaluation Looking at the ROC curve in figure 31, the model performances have once again downgraded: we went from an AUC of 0.8778 (model without the protected variables) to an AUC of 0.8534, which still indicates good predictive performance.

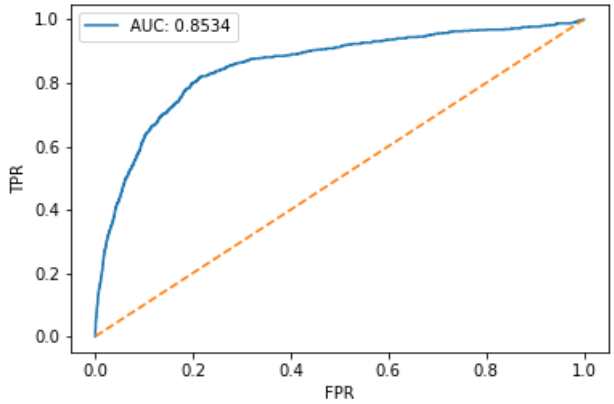


Figure 31: ROC curve for the model with transformed variables

The model using transformed variables still exhibits good predictive performance, although slightly worse than in the previous case.

Fairness evaluation

- Globally, the AR, TPR and FPR are higher than for the model without protected variables. The AR and FPR are 0.1 point higher and the FPR 0.7 point higher.

Without protected variables		Transformed variables	
(%)	Global	(%)	Global
AR	99.20	AR	99.30
TPR	99.29	TPR	99.39
FPR	86.26	FPR	86.92

Table 47: Global fairness metrics

- By sex
  - The acceptance rates are now only 0.03 point apart.
  - The true positive rates are only 0.04 point apart.
  - The false positive rates are further apart and have changed signs.

  We have approximately reached statistical parity when looking at the Sex variable: both groups are treated fairly by the model. We have also approximately reached equal opportunity. As the FPR are now 9.02 points apart, we are further away from the definition of equalized odds, with now the Female category being disadvantaged by the model.

Without protected variables				Transformed variables			
(%)	Female	Male	Difference	Sex	Female	Male	Difference
AR	99.46	98.97	0.49	AR	99.32	99.29	0.03
TPR	99.52	99.10	0.42	TPR	99.41	99.37	0.04
FPR	88.76	84.97	3.79	FPR	80.75	89.77	-9.02

Table 48: Fairness metrics by sex

- By origin
  - The acceptances rates are now very close to each other. The order is the same except for Black and Asian that switched places (the two lowest acceptance rates). The variance in acceptances rates is of 1.82e−50.
  - The true positive rates are closer to each other than in the case without protected variables. The order is the same except for the American Indian/AK Native that went from second highest TPR to second lowest. The variance in true positive rates is of 1.23e−50.
  - The false positive rates are further apart than before, with a variance of 0.12. The order has been changed: Hispanic had the highest FPR but now has the fourth highest. Black and American Indian/AK Native respectively had the fifth and fourth highest but now have the same highest. Missing still has the lowest.

  We have approximately reached statistical parity and equal opportunity when looking at the Origin variable, as we have very close acceptance and true positive rates. For equalized odds, we are further away from fairness as the false positive rates are further apart.

Without protected variables

(%)	White	Hispanic	Black	Asian or Pacific Islander	American Indian /AK Native	Missing	Var
AR	99.23	98.76	93.85	96.68	99.10	100.00	4.37e−4
TPR	99.31	98.88	95.32	97.05	99.54	100.00	2.71e−4
FPR	87.07	89.66	45.45	70.00	66.67	0.00	9.28e−2

Transformed variables

(%)	White	Hispanic	Black	Asian or Pacific Islander	American Indian /AK Native	Missing	Var
AR	99.30	98.93	98.88	98.62	99.04	99.96	1.82e−5
TPR	99.39	99.18	98.85	99.16	99.02	99.96	1.23e−5
FPR	87.45	76.92	100.00	63.64	100.00	0.00	1.18e−1

Table 49: Fairness metrics by origin

- By marital status
  - The acceptance rates are now almost equal, with a variance of 1.83e−50.
  - The same goes for the true positive rates, with a variance of 1.70e−50.
  - The false positive rates are a lot further apart.

  We have approximately reached statistical parity and equal opportunity. For equalized odds, we are further away.

Without protected variables

(%)	Married	Missing	Single	Widowed	Divorced	Separated	Unmarried	Var
AR	99.01	99.89	99.06	97.74	98.14	97.75	97.48	6.85e−5
TPR	99.12	99.91	99.15	98.13	98.29	97.96	97.44	6.23e−5
FPR	85.43	92.75	87.96	78.72	87.32	85.71	100.00	3.78e−3

With transformed variables

(%)	Married	Missing	Single	Widowed	Divorced	Separated	Unmarried	Var
AR	99.12	99.77	99.06	99.19	99.07	98.40	98.47	1.83e−5
TPR	99.23	99.79	99.18	99.42	99.21	98.66	98.47	1.70e−5
FPR	86.13	92.86	84.40	86.96	90.12	75.00	0.00	9.31e−2

Table 50: Fairness metrics by marital status

The indirect discrimination method has proven its success: we have approximately – because of the focus on correlation instead of higher order dependence – reached statistical parity. For one of the other metrics, the model is fairer, but for the last one we are further away from fairness.

5.6 Conclusion on the methods

Just like in the simulated case, we compared the performance and fairness of the logistic regression model using three different types of explanatory variables: all variables, only non-sensitive variables, and transformed non-sensitive variables. Visual results by sensitive variables can be found in appendices H, I and J.

The model using all variables has the best results in terms of AUC, with only non-sensitive variables we have a decrease in AUC and finally with transformed non-sensitive variables the AUC is the lowest. This all shows the trade-off between performance and fairness.

In terms of fairness, the model using all variables treats unfairly the different protected groups. By Sex, Male is disadvantaged under all three definitions of fairness. By Origin, Black is the most disadvantaged group under the statistical parity and equal opportunity definitions. For equalized odds, we cannot conclude. By Marital status, Separated is the most disadvantaged category under the statistical parity definition and Widowed under the two others.

When we remove the protected variables, thus avoiding direct discrimination, the results depend on the relationships the protected variables have with the other variables. By sex, fairness improves under all definitions, but Male is still disadvantaged. By Origin, fairness improves under the statistical parity and equal opportunity definitions, with Black still being the most disadvantaged group. Under the equalized odds definition, one involved metric improves but the other does not. By Marital status, fairness improves under all definitions. For statistical parity, Unmarried is now the most disadvantaged group instead of Separated. For the other definitions, Widowed is still the most disadvantaged group.

Finally, applying the change of basis method transformed the non-sensitive variables and we now have a null correlation with the sensitive variables. Applying the model, we have the expected results: we are now very close to respecting statistical parity. Looking at the two other definitions, we are closer to equal opportunity. Looking at the metrics for equalized odds, the fairness is worse, so we can conclude that all fairness definitions are not compatible with each other.

Conclusion

The goal of this thesis was to mitigate indirect discrimination when using Machine Learning models for insurance mortality data. We have provided the reader with an innovative method based on mathematical concepts of linear algebra to achieve this goal. Looking first at a simulated case allowed us to grasp the intricate effects of discrimination mechanisms, and gave us insights as to how to tackle the real use case.

We have made a few approximations, as reminded throughout the thesis. Firstly, we have chosen to focus on only one definition of fairness, statistical parity. Although it is the most popular in research articles, we are not certain that it will be the one chosen by regulators and policy makers in their control for fairness. Secondly, statistical parity is an independence condition, and we have approximated dependence with its first-order component, correlation. Our method is therefore a first step towards fairness, but might not be sufficient if variables have intricate and complex relationships with each other. Although, as we have mentioned, if regulators decide to use such a metric to control fairness, they will need to take into account acceptability threshold around the goal value of the metric, because of its statistical nature. Lastly, with our transformation of the non-sensitive variables, we have interpretability issues as the new vectors are a combination of the old ones. Explaining to an insured or even to other less technical employees of the insurance company, why premiums vary may be tricky as they are based on the mortality rates that come from the use of the transformed variables.

Further developments around this subject should thus focus on expanding the method to all forms of dependence instead of correlation only, and even looking at ways of adapting the method to other fairness definitions. Other improvements may include adaptation to regression problems, which are also encountered by insurers.

Bibliography

A Reminder on confidence intervals

As fairness metrics are statistical, confidence intervals have to be taken into account. A legal decision revolving around a metric cannot overlook this aspect and if a value is given as a fairness standard, there has to be some interval within which is it acceptable to be.

Suppose we have a sample $(X_1, \ldots, X_n)$ following a normal distribution $\mathcal{N}(\mu, \sigma^2)0$. The goal is to provide a confidence interval for $\mu 0$. With $\hat{\mu}$ and $\hat{\sigma}$ the unbiased mean and standard deviation estimators:

$$\hat{\mu} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

Then $T = \sqrt{n}\frac{\hat{\mu}-\mu}{\hat{\sigma}}$ follows a Student law with (n-1) degrees of freedom. Choosing the quantile t such that $-t < T < t$, we can write

$$\hat{\mu} - \frac{t\hat{\sigma}}{\sqrt{n}} < \mu < \hat{\mu} + \frac{t\hat{\sigma}}{\sqrt{n}}$$

Proof.

$$\begin{aligned}\mathbb{P}(-t < T < t) &= \mathbb{P}(T < t) - \mathbb{P}(T < -t)\\ &\mathbb{P}T < t - \mathbb{P}(T > t)\\ &= 2\mathbb{P}(T < t) - 1\end{aligned}$$

So

$$\mathbb{P}(-t_{\alpha/2} < T < t_{\alpha/2}) = 2(1 - \alpha/2) - 1 = 1 - \alpha$$

So the confidence interval is

$$-t_{\alpha/2} < T < t_{\alpha/2} \iff \hat{\mu} - t_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}} < \mu < \hat{\mu} + t_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}$$

□

The formula can be approximated thanks to the asymptotic distribution of $\hat{\mu}$ which is normally distributed: $\hat{\mu} \sim \mathcal{N}(\mu, \sigma^2/n)0$. Therefore,

$$\hat{\mu} - |z_{\alpha/2}|\frac{\hat{\sigma}}{\sqrt{n}} < \mu < \hat{\mu} + |z_{\alpha/2}|\frac{\hat{\sigma}}{\sqrt{n}}$$

with $z_{\alpha/2}$ the standard normal quantile of level $\alpha/20$. For $\alpha = 5\%, |z_{2.5\%}| = 1.960$.

B Reminder on the logistic regression

This section gives a quick reminder about the logistic regression [12]. A binary logistic model is used to model the probability of a certain class: $Y \in \{0,1\}0$. The model assumes that the outcomes conditioned on the explanatory variables follow a Bernoulli distribution:

$$Y_j | X_j^{(1)} = x_j^{(1)}, \dots, X_j^{(n)} = x_j^{(n)} \sim \mathcal{B}(p_j)$$

with $p_j = \mathbb{P}(Y_j = 1 | X_j^{(1)} = x_j^{(1)}, \dots, X_j^{(n)} = x_j^{(n)})$ and that the *logit* of this probability is modelled as a linear combination of the n explanatory variables:

$$\begin{aligned} logit(\mathbb{P}(Y_j = 1 | X_j^{(1)} = x_j^{(1)}, \dots, X_j^{(n)} = x_j^{(n)})) &= b_0 + b_1 x_j^{(1)} + \dots + b_n x_j^{(n)} \\ &= \mathbf{B} \cdot \boldsymbol{x_j} \end{aligned}$$

with $\mathbf{B} = (b_0, \dots, b_n)$ and $\boldsymbol{x_j} = (1, x_j^{(1)}, \dots, x_j^{(n)})^T$
As a reminder,

$$logit(p) = ln(\frac{p}{1-p}), p \in (0,1)$$

and

$$logit^{-1}(x) = \frac{1}{1+e^{-x}}, x \in \mathbb{R}$$

The vector $\mathbf{B}$ is estimated by maximizing the log-likelihood. The likelihood of an independently and identically distributed sample of size J is

$$\begin{aligned} L &= \mathbb{P}(Y_1 = y_1, \dots, Y_J = y_J | X^{(1)} = x^{(1)}, \dots, X^{(n)} = x^{(n)}) \\ &= \prod_{j=1}^{J} \mathbb{P}(Y_j = y_j | X_j^{(1)} = x_j^{(1)}, \dots, X_j^{(n)} = x_j^{(n)}) \\ &= \prod_{j=1}^{J} \frac{1}{1 + e^{-(b_0 + b_1 x_j^{(1)} + \dots + b_n x_j^{(n)})}} \end{aligned}$$

We can then find the $\hat{\boldsymbol{B}}$ that maximizes log(L) and classify a new individual j by applying Bayes' rule:

$$\begin{aligned} Y_j = 1 &\iff \mathbb{P}(Y_j = 1 | \boldsymbol{X_j} = \boldsymbol{x_j}) > \mathbb{P}(Y_j = 0 | \boldsymbol{X_j} = \boldsymbol{x_j}) \\ &\iff \mathbb{P}(Y_j = 1 | \boldsymbol{X_j} = \boldsymbol{x_j}) > \frac{1}{2} \\ &\iff \frac{1}{1 + e^{-\hat{\boldsymbol{B}} \cdot \boldsymbol{x_j}}} > \frac{1}{2} \\ &\iff \hat{\boldsymbol{B}} \cdot \boldsymbol{x_j} > 0 \end{aligned}$$

So

$$Y_j = \mathbb{1}_{\hat{\boldsymbol{B}} \cdot \boldsymbol{x_j} > 0}$$

C The Gram-Schmidt process

The Gram-Schmidt process [28] is a method for orthonormalizing a set of vectors in an inner product space. The process takes a finite and linearly independent set of vectors $(u_1, \ldots, u_n)$ and produces a set of orthogonal vectors $(v_1, \ldots, v_n)0$. The process works as follows:

$$
\begin{aligned}
v_1 &= u_1 \\
v_2 &= u_2 - \frac{\langle u_2, v_1 \rangle}{\langle v_1, v_1 \rangle} v_1 \\
&\ldots \\
v_n &= u_n - \sum_{k=1}^{n-1} \frac{\langle u_n, v_k \rangle}{\langle v_k, v_k \rangle} v_k
\end{aligned}
$$

A first idea was therefore to use this process. The issue with this method is that we obtain a transformation of all variables, both sensitive and non-sensitive, except for the first one as $v_1 = u_1$, and they are all orthogonal to each other, which is not what we are looking for.

D Variable description (pseudo database)

`AGE`: Age of individual in time interval
Numerical (int)
18-80

`Age_dx`: Age of individual at diagnosis
Numerical (int)
18-80

`AYA_site_rec20`: Type of melanoma
Categorical
Nodular melanoma, Superficial spreading/low cumulative sun damage melanoma, Other malignant

`DURATION`: Time since diagnosis (in years)
Numerical (int)
$\mathbb{N}$

`Death_skin`: Individual died of skin melanoma in time interval
Numerical (int)
$\{0, 1\}$

`EXPO`: Exposure of individual in time interval
Numerical (float)
$[0, 1]$

`Extent`: How far the tumor has spread at diagnosis
Categorical
In situ, Localized, Regional, Distant, Missing

`Laterality` Side of the body tumor originated on
Categorical
Right origin, Left origin, Bilateral, ...

`Mar_stat` Marital status of individual at diagnosis
Categorical
Married, Single, Divorced, ...

`Metastasis`: Metastatic state at diagnosis
Categorical
M0, M1

`Mitotic rate`: Number of mitoses per mm^2 at diagnosis
Numerical (int)
0-11

`Positive_Node`: Spread to regional lymph nodes at diagnosis
Categorical
N0, N1, N2, N3, Missing

`Origin`: Origin of individual
Categorical
Hispanic, White, Asian, ...

`reg_nod_ex`: Number of regional lymph nodes that were removed and examined
Numerical (int)
$\mathbb{N}$

`reg_nod_pos`: Number of regional lymph nodes that contain metastases at diagnosis
Numerical (int)
$\mathbb{N}$

`Sex`: Sex of the patient
Categorical
Male, Female

`Site_rec_WHO08`: Origin site of primary tumor
Categorical
Pleura, Vulva, Male Genital Organs

`Stage`: Cancer stage at diagnosis
Categorical
I, II, III, IV

`Surg_LN`: Surgery of lymph nodes
Categorical
Surgery, No surgery

`Surg_oth`: Surgery of other sites
Categorical
Surgery, No surgery

`Surg_primsite`: Surgery of primary site
Categorical
Surgery, No surgery

`Tot_num_benin`: Total number of benign tumors at diagnosis
Numerical (int)
$\mathbb{N}$

`Tot_num_in_situ`: Total number of in situ tumors at diagnosis
Numerical (int)
$\mathbb{N}$

`Tumor`: Tumor size at diagnosis
Categorical
T0, T1, T2, T3, T4

`Ulceration`: Presence of ulceration (break on the skin) at diagnosis
Categorical
No ulceration, Ulceration, Missing

`YEAR`: Year of time interval
Numerical (int)
2004-2018

`Yr_dx`: Year of diagnosis
Numerical (int)
2004-2018

E Coefficients of the model with all selected variables

Variable	Coefficient	Standard error	P-value
const	-7.1992	0.118	0.000
DURATION	-0.0618	0.007	0.000
AGE	0.0211	0.002	0.000
Sex_Female	-0.3976	0.043	0.000
reg_nod_pos	0.0369	0.007	0.000
Mitotic rate	0.0043	0.006	0.494
Tot_num_in_situ	-0.0540	0.023	0.020
Laterality_Bilateral	0.7446	0.331	0.025
Laterality_Not a paired site	0.5161	0.060	0.000
Laterality_One side	0.1908	0.200	0.339
Laterality_Paired site	0.3998	0.093	0.000
Laterality_Paired site midline tumor	-0.0055	0.126	0.965
Laterality_Right origin	-0.0595	0.044	0.175
Ulceration_Missing	0.0898	0.099	0.362
Ulceration_Ulceration	0.7611	0.046	0.000
Site_rec_WHO08_Male Genital Organs	0.1275	0.561	0.820
Site_rec_WHO08_Vulva	0.6528	0.176	0.000
Origin_American Indian/AK Native	0.3888	0.292	0.183
Origin_Asian or Pacific Islander	0.1445	0.156	0.355
Origin_Black	0.3856	0.168	0.021
Origin_Hispanic	0.1710	0.108	0.114
Origin_Missing	-3.0668	1.001	0.002
Mar_stat_Divorced	0.3712	0.064	0.000
Mar_stat_Missing	-0.4660	0.071	0.000
Mar_stat_Separated	0.2985	0.225	0.185
Mar_stat_Single	0.2562	0.055	0.000
Mar_stat_Unmarried	0.0923	0.426	0.828
Mar_stat_Widowed	0.3620	0.087	0.000
Extent_Distant	0.8813	0.077	0.000
Extent_Missing	0.5145	0.167	0.002
Extent_Regional	0.5354	0.055	0.000
Surg_primsite_Missing	0.3650	0.251	0.146
Surg_primsite_No surgery	0.6385	0.126	0.000
Tumor_Missing	1.0028	0.122	0.000
Tumor_T0	0.6995	0.183	0.000
Tumor_T2	1.0626	0.066	0.000
Tumor_T3	1.4361	0.071	0.000
Tumor_T4	1.6748	0.078	0.000
Positive_Node_Missing	0.3923	0.133	0.003
Positive_Node_N1	0.8841	0.055	0.000
Positive_Node_N2	1.0769	0.066	0.000
Positive_Node_N3	1.4402	0.086	0.000

Table 51: Coefficients of the model with selected variables

F Coefficients of the model without protected variables

Variable	Coefficient	Standard error	P-value
const	-7.4933	0.108	0.000
DURATION	-0.0662	0.007	0.000
AGE	0.0225	0.002	0.000
reg_nod_pos	0.0384	0.007	0.000
Mitotic rate	-0.0019	0.006	0.761
Tot_num_in_situ	-0.0680	0.023	0.004
Laterality_Bilateral	0.4795	0.369	0.194
Laterality_Not a paired site	0.5809	0.059	0.000
Laterality_One side	0.3412	0.187	0.067
Laterality_Paired site	0.3723	0.093	0.000
Laterality_Paired site midline tumor	0.0578	0.124	0.642
Laterality_Right origin	-0.0745	0.044	0.088
Ulceration_Missing	0.1616	0.095	0.090
Ulceration_Ulceration	0.7454	0.046	0.000
Site_rec_WHO08_Male Genital Organs	-0.1704	0.624	0.785
Site_rec_WHO08_Vulva	0.3408	0.176	0.053
Extent_Distant	0.9361	0.078	0.000
Extent_Missing	0.6202	0.157	0.000
Extent_Regional	0.6152	0.056	0.000
Surg_primsite_Missing	0.2659	0.247	0.281
Surg_primsite_No surgery	0.5068	0.122	0.000
Tumor_Missing	1.0455	0.118	0.000
Tumor_T0	0.7275	0.182	0.000
Tumor_T2	1.1378	0.066	0.000
Tumor_T3	1.5450	0.071	0.000
Tumor_T4	1.8266	0.078	0.000
Positive_Node_Missing	0.3997	0.124	0.001
Positive_Node_N1	0.9418	0.055	0.000
Positive_Node_N2	1.1609	0.066	0.000
Positive_Node_N3	1.4787	0.087	0.000

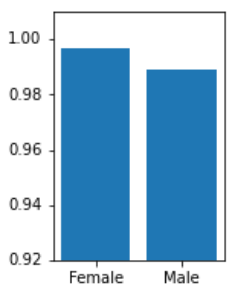
Table 52: Coefficients of the model without protected variables

G Coefficients of the model with transformed variables

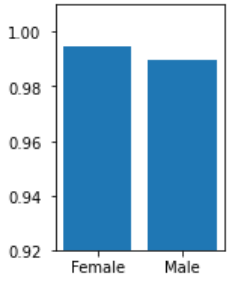
Variable	Coefficient	Standard error	P-value
const	-5.7105	0.031	0.000
DURATION	-0.0617	0.007	0.000
AGE	0.0169	0.002	0.000
reg_nod_pos	0.0425	0.007	0.000
Mitotic rate	0.0006	0.006	0.930
Tot_num_in_situ	-0.0732	0.024	0.002
Laterality_Bilateral	0.3711	0.379	0.327
Laterality_Not a paired site	0.5309	0.061	0.000
Laterality_One side	0.2291	0.194	0.237
Laterality_Paired site	0.2427	0.095	0.011
Laterality_Paired site midline tumor	-0.0276	0.128	0.829
Laterality_Right origin	-0.0791	0.043	0.067
Ulceration_Missing	0.0913	0.097	0.348
Ulceration_Ulceration	0.8420	0.047	0.000
Site_rec_WHO08_Male Genital Organs	0.1864	0.648	0.774
Site_rec_WHO08_Vulva	0.2293	0.182	0.207
Extent_Distant	0.8314	0.079	0.000
Extent_Missing	0.6594	0.155	0.000
Extent_Regional	0.5229	0.056	0.000
Surg_primsite_Missing	0.4959	0.237	0.037
Surg_primsite_No surgery	0.5580	0.117	0.000
Tumor_Missing	0.8921	0.118	0.000
Tumor_T0	0.9362	0.178	0.000
Tumor_T2	1.0608	0.066	0.000
Tumor_T3	1.4789	0.071	0.000
Tumor_T4	1.8200	0.078	0.000
Positive_Node_Missing	0.2745	0.122	0.024
Positive_Node_N1	1.0127	0.056	0.000
Positive_Node_N2	1.2251	0.067	0.000
Positive_Node_N3	1.5559	0.087	0.000

Table 53: Coefficients of the model with transformed variables

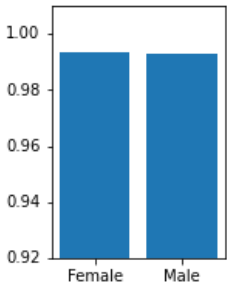
H Fairness metrics by sex



(a) All variables

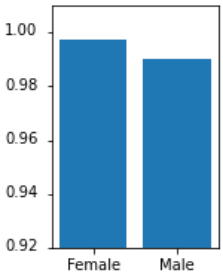


(b) Without protected

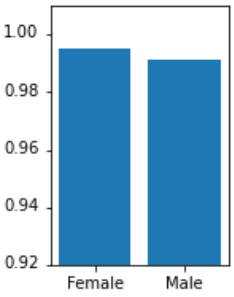


(c) Transformed variables

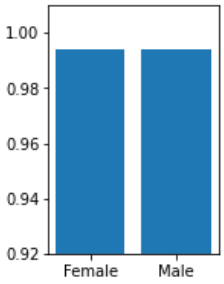
Figure 32: Acceptance rates $\frac{TP+FP}{TP+TN+FP+FN}$ by sex



(a) All variables

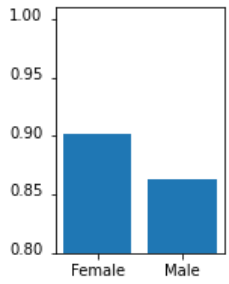


(b) Without protected

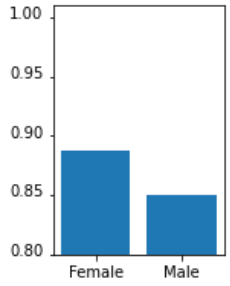


(c) Transformed variables

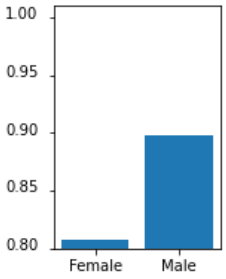
Figure 33: True positive rates $\frac{TP}{TP+FN}$ by sex



(a) All variables



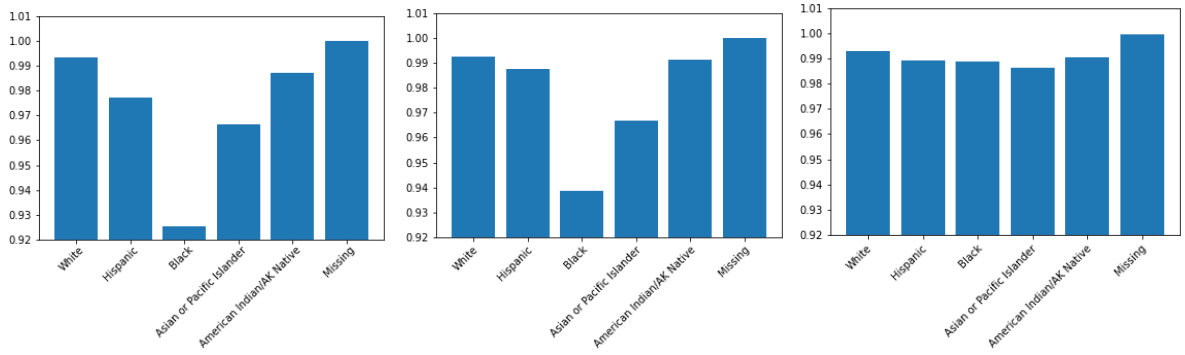
(b) Without protected



(c) Transformed variables

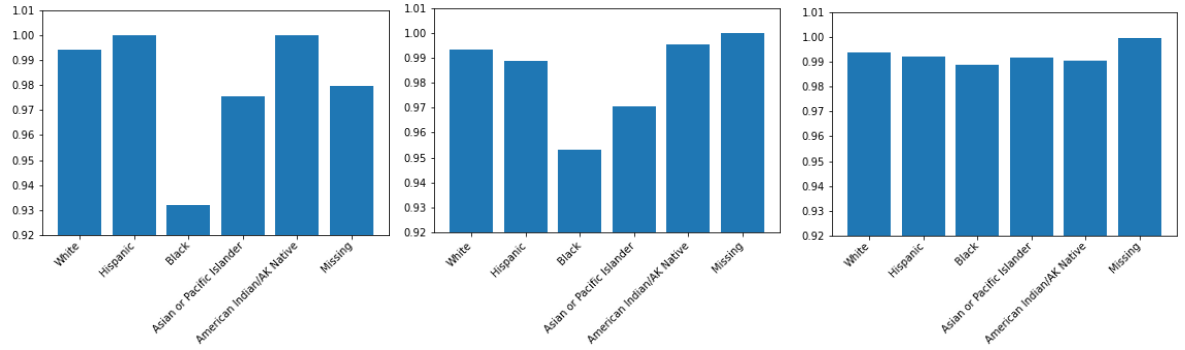
Figure 34: False positive rates $\frac{FP}{FP+TN}$ by sex

I Fairness metrics by origin



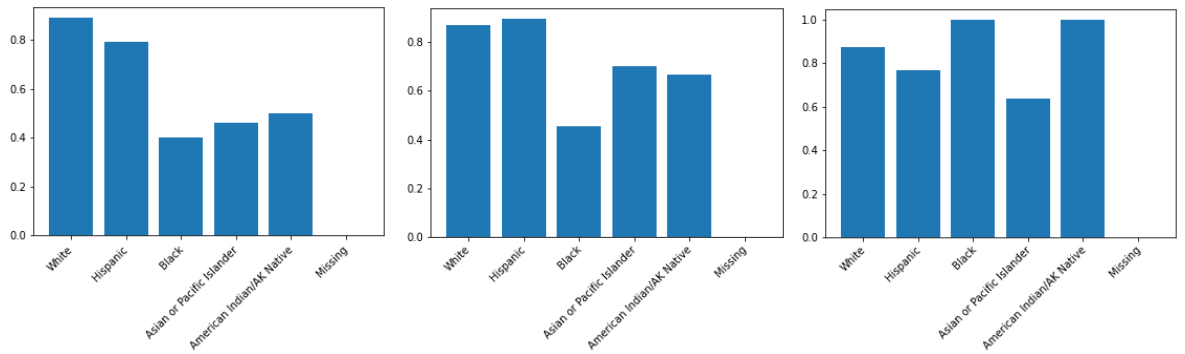
(a) All variables (b) Without protected (c) Transformed variables

Figure 35: Acceptance rates $\frac{\text{TP+FP}}{\text{TP+TN+FP+FN}}$ by origin



(a) All variables (b) Without protected (c) Transformed variables

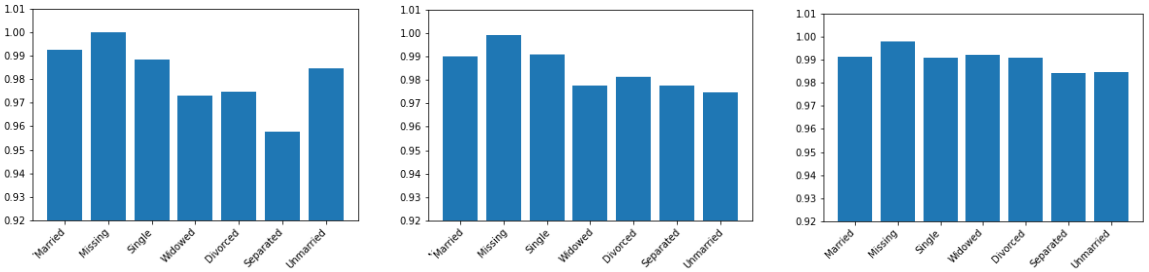
Figure 36: True positive rates $\frac{\text{TP}}{\text{TP+FN}}$ by origin



(a) All variables (b) Without protected (c) Transformed variables

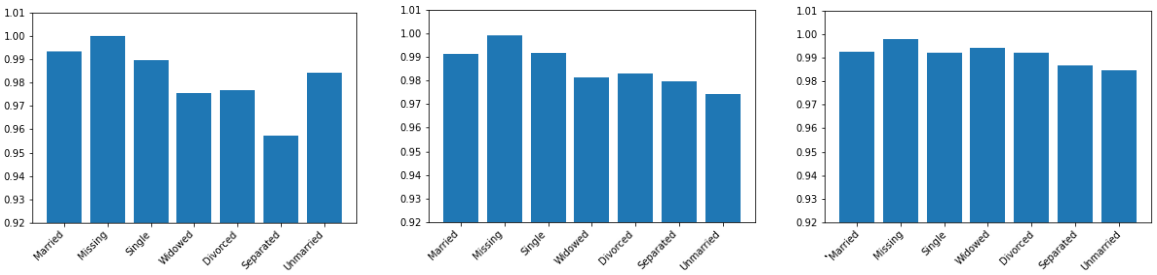
Figure 37: False positive rates $\frac{\text{FP}}{\text{FP+TN}}$ by origin

J Fairness metrics by marital status



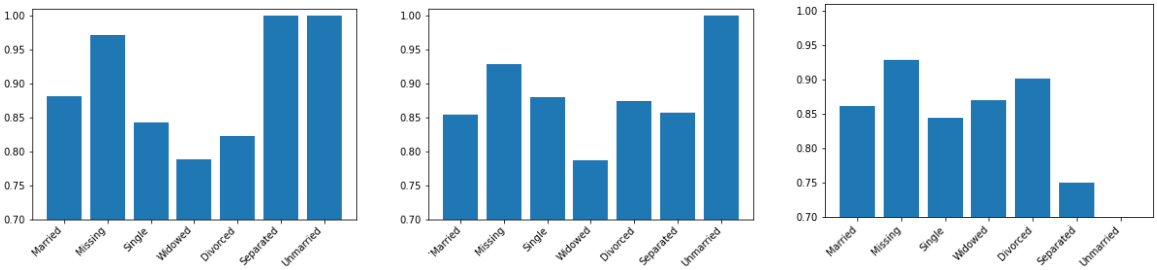
(a) All variables (b) Without protected (c) Transformed variables

Figure 38: Acceptance rates $\frac{\text{TP+FP}}{\text{TP+TN+FP+FN}}$ by marital status



(a) All variables (b) Without protected (c) Transformed variables

Figure 39: True positive rates $\frac{\text{TP}}{\text{TP+FN}}$ by marital status



(a) All variables (b) Without protected (c) Transformed variables

Figure 40: False positive rates $\frac{\text{FP}}{\text{FP+TN}}$ by marital status