Discrimination in HR Analytics
A fair Workflow.
Biased Data

Davide Zulato

December 9, 2022

1 Data

1.1 Simulating the Data

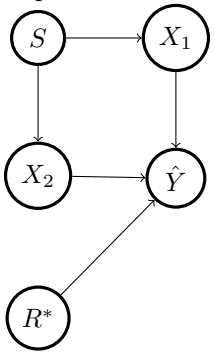
As suggested by Moritz Hardt unintended sources of unfairness can be introduced in data driven decision making through the training data because the reflect historical and human bias. Thus, the data generation process is a crucial step not only in the proposed workflow with simulated data but also in a real case study. the use of a simulated dataset with an intentionally introduced bias has the advantage of allowing knowledge of the sources of unfairness.In the context of human resources and recruitment, the measurement of the target variable is often subject to subjective decisions and is often obtained as a construct (e.g. from scores to binary outcome). For example, measuring the performance of salespeople through the number of vensites rather than through an analysis of customer reviews could lead to bias due to customer quality achieved and environmental conditions of the workplace [1].

If the training data reflect existing social biases against a minority, the algorithm is likely to incorporate these biases

The synthetic Dataset is composed of $n = 10000$ rows representing the candidates and $p = 12$ columns representing the *features*, measured variables for each individual. The p features are divided, consistent with the notation of literature, in independent features X and *Sensitive features* S (with S_a denoting the advantaged group and S_d for the disadvantaged group).The target variable is denoted as Y that in the case of binary classification takes value $Y = 0$ for the negative outcome (e.g. not assumption or churning from the company) and $Y = 1$ for the preferable outcome (e.g. intake or retention).

The next figure shows the logical graphical model for the variables

Figure 1: Relationship between variables, causal graph



The protected feature S is not involved directly in the definition of the target variable (Figure 1), but we obtain disparate impact through the *Proxies* and *Simpson's paradox*. The independent features not depending from the Sensitive feature S are notated with R^* (e.g. X_score).

Variable Simulation			
Variable Name	distribution	Formula for mean	link
S	$Binomial(\pi)$	$\pi = 0.2$	$identity$
Age	χ^2	$22 + \chi^2(0.5)$	$identity$
interview	$Poisson(\lambda)$	$\lambda = f(age, S, \eta)$	$identity$
GitHub account	$Binomial(\pi)$	$\pi = f(S, \eta)$	$logit$
Proxy	$Normal(\mu, 2)$	$\mu = f(S, \eta)$	$identity$
proxy2	$Beta(\alpha, \beta)$	$\alpha = f(proxy, age)$	$identity$
X score	$Normal(\mu, \sigma)$	$\mu = 100, \sigma = 5$	$identity$
score	$Poisson(\lambda)$	$\lambda = f(S)$	$identity$
Simpson score 1	$Normal(\mu, \sigma)$	$\mu = f(S)$	$identity$
Simpson score 2	$Normal(\mu, \sigma)$	$\mu = f(S)$	$identity$
Y	$Binomial(\pi)$	$\pi = f(.)$	$logit$
Group	$discrete$	$3 - levels$	$random$

1.2 Sensitive feature

The sensitive feature S is generated from a Binomial distribution with probability $\pi = 0.2$.

	S	P(S=s)
0	7928	0.74
1	2072	0.26

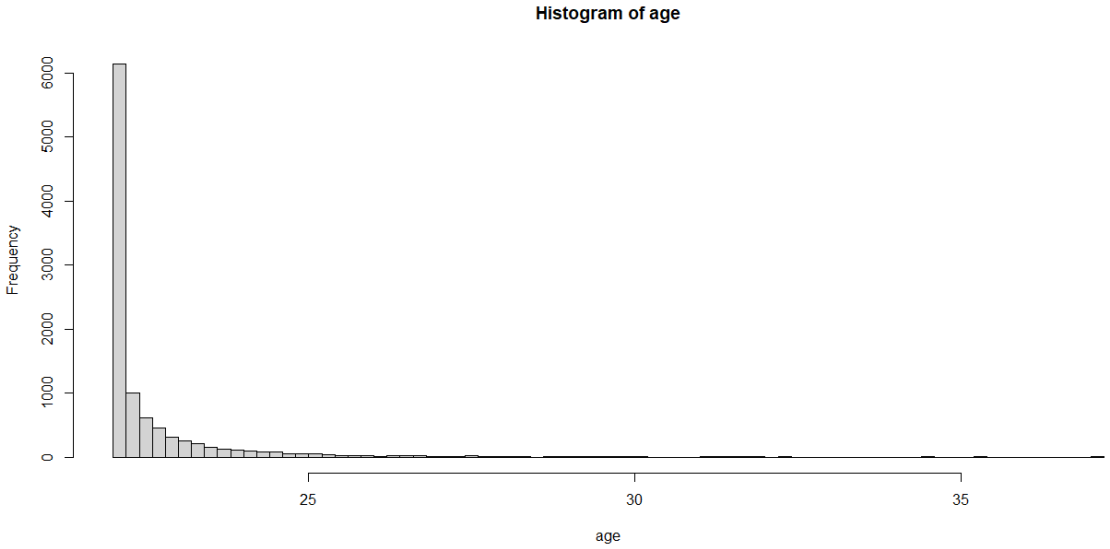
The idea is that most of individuals are from the advantaged group $S_a : S = 0$ in order to emulate a situation of sample size disparity with less data available about minorities. During this work we'll keep the notation S for the sensitive feature but we could imagine that S is the gender and S_a is female is the disadvantaged group. A scenario with 35% women is plausible in a stem context [1] (e.g. In Engineering and construction graduates in Italy we could find a similar distribution). In the simulation the sample size of the sensitive feature is kept low to exaggerate the sample size disparity, a source of unfairness presented by the literature. In a real-life scenario we could find bigger disparities, AIDS status for instance is considered a protected feature in Italian law and it affects roughly 2.2 per 100,000 individuals and it would be impractical to deal with it in this context.

1.3 Age

The age variable is obtained by summing a chi-square with 0.5 *df* to 22; $Age = 22 + \chi^2(0.5)$ the idea is to emulate a real-world situation of an HR dataset composed of individuals with recent bachelor's degrees. Thus, most of the individuals/candidates will be 22 years old and several outliers are present. In this context, the removal of outliers (of rows containing outliers) is not a viable solution since we are also interested in older individuals. In model development, the ideal approach might be to remove the age column or to imitate the mean value for the extreme values. The age variable is created independently of the other features.

[1] https://genderdata.worldbank.org/indicators/se-ter-grad-fe-zs/

Figure 2: Distribution of Age



The histogram of the variable *Age* (Figure 2)is strongly right-skewed and presents some outliers.

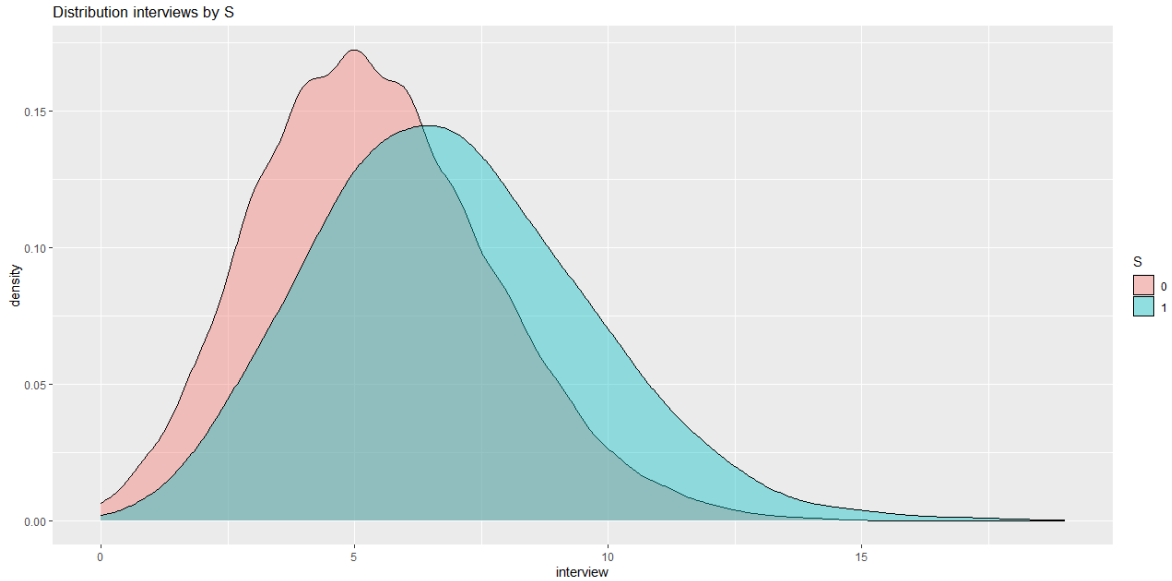
1.4 Interview

This feature, distributed as a Poisson, indicates the number of interviews taken by the candidate up to the time the dataset was recorded.

$\lambda = c + \beta_1 Age + \beta_2 S + \eta \rightarrow Interview \sim Po(\lambda)$
$\mu \sim N(0, 1)$

The mean and variance λ (rate) depends on the class and takes a larger value for S_d. The idea is that individuals belonging to the protected class are to some extent more active in the labor market. This variable is positively associated with the target variable Y. The mean is also a function of age (with $\beta > 0$) because we assume that an older individual has had more interviews.

Figure 3: Distribution of Interview by S



As we can observe from the distributions by S the mean of the group $S = 1$ is greater than the mean for the group $S = 0$. $\lambda = f(S, Age, \eta)$ where η is a noise normally distributed with mean $\mu = 0$ and variance $\sigma = 1$.

1.5 Github account

The GitHub account is intended as a proxy for the sensitive feature because it is constructed such that among Github account holders only 13% belongs to the disadvantaged class S_d, underrepresented respect to the advantaged class.

$Github\_account = c - \beta S + \eta$

Table 1: Github Account

		0	1
S	0	0.75	0.87
	1	0.25	0.13

The construction of this feature, similarly to the target variable, involves a linear relationship and *logit* transformation to obtain the probabilities. the obtained probabilities are then used as the probabilities of success π of a binomial distribution.

Table 2: Github Account

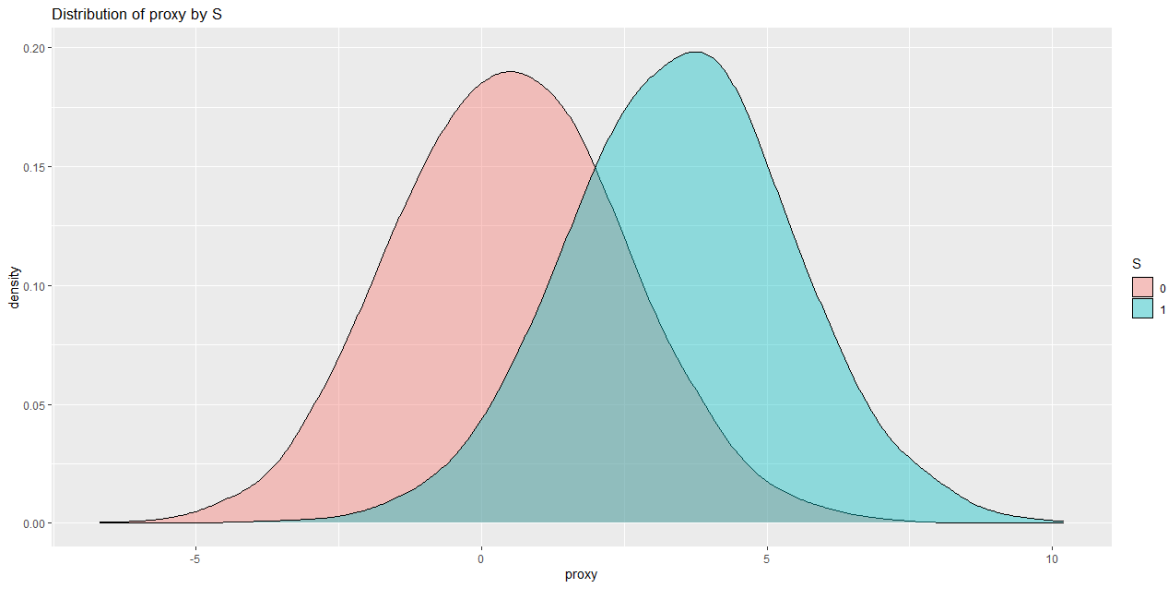
		0	1
S	0	0.61	0.39
	1	0.77	0.23

1.6 Proxy

This proxy is obtained as a linear with noise with the sensitive feature S used as function for the mean of a Gaussian distribution.

$\mu = \beta S + \eta \rightarrow Proxy \sim N(\mu, \sigma = 2)$

Figure 4: Distribution of Proxy

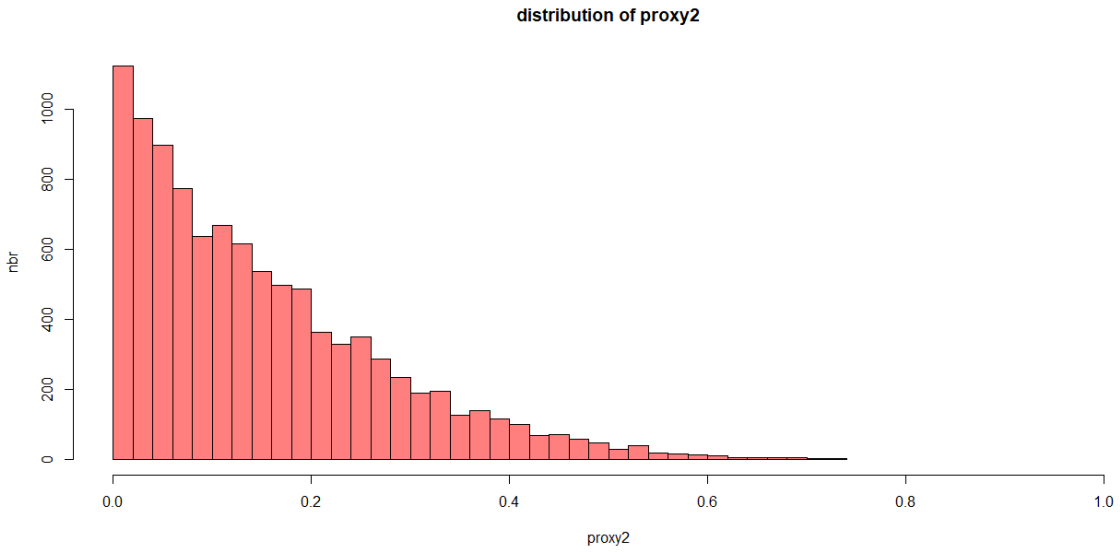


The correlation between the proxy feature is kept around $\rho = 0.5$ in order to avoid multicollinearity and the removing after a VIF $VIF = \frac{1}{1-R^2}$. The idea behind generating a Proxy feature is to continue to have Bias against a protected feature even after removing S. The Proxy feature, in order to generate disparate impact, must be negatively associated with the outcome if positive associated with the protected class (In this synthetic data the value $S = 1$ stands for the disadvantaged group S_d). The main difference between *proxy* and *Interview* is the relation with the target variable, positive for the interview (it makes sense that who participated more has a chance to get a positive outcome) and negative for the proxy.

1.7 Proxy2

The second proxy is built with a weak correlation with the target variable. This variable is distributed as a *beta* $proxy2 \sim Be(\alpha, \beta), \alpha, \beta > 0$ with the shape parameter α function of S, $Proxy$and Age ($\alpha = f(S, proxy, Age)$). The mean could be calculated as $E[Proxy2] = \frac{\alpha}{\alpha+\beta}$. In order to obtain a non-symmetrical distribution a non-centrality parameter $\lambda = 0.9$ is introduced.

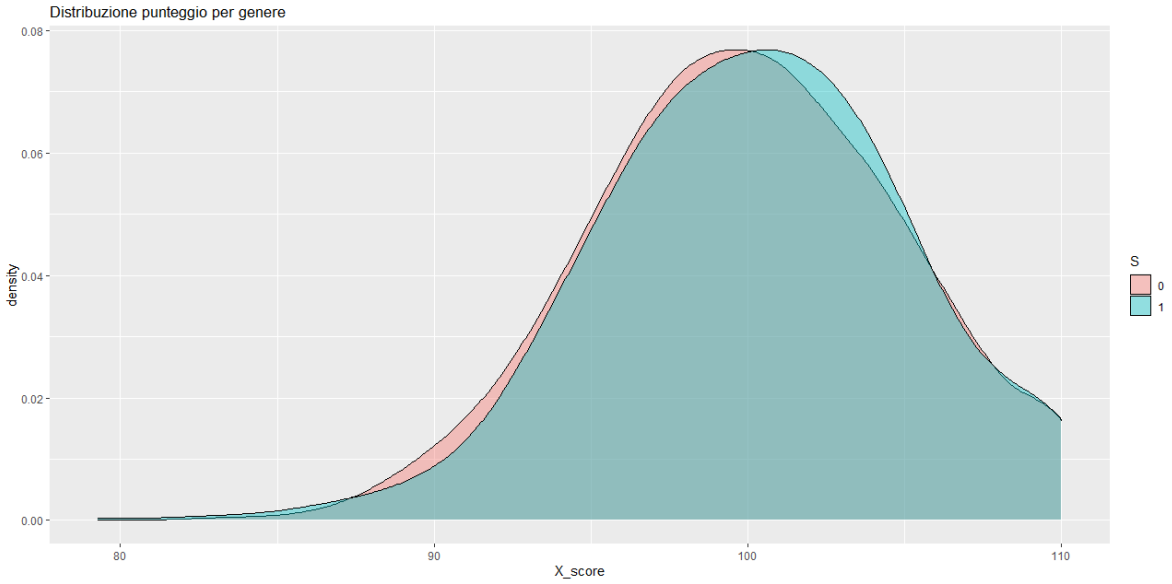
Figure 5: Proxy 2



1.8 X score

This feature is designed to represent the independent variable R^* in the causal graph in Figure 1. The $Xscore$

Figure 6: Distribution of X Score by S

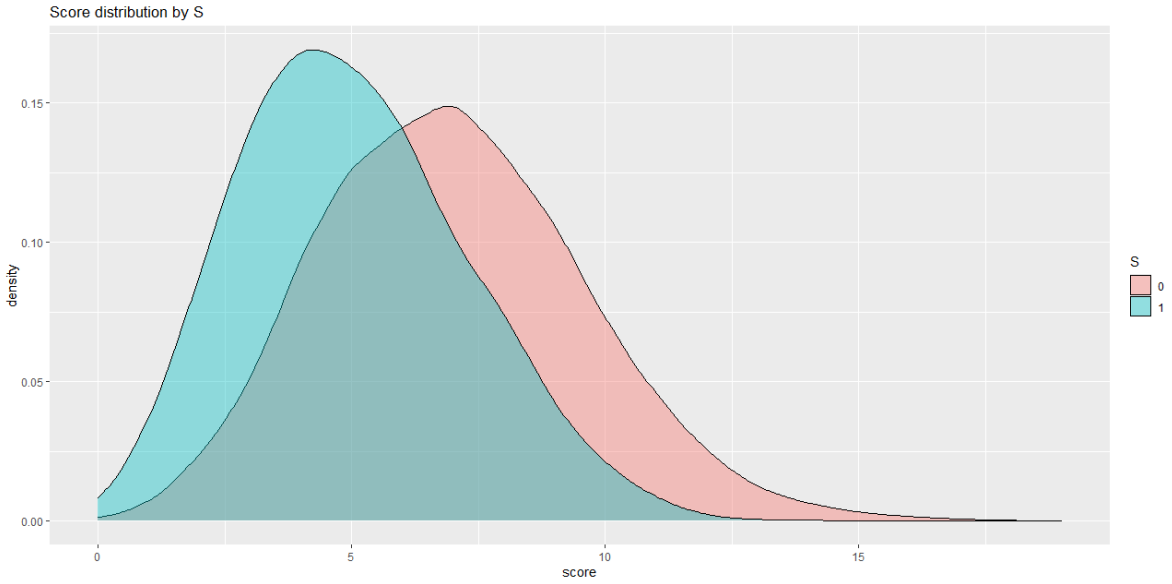


The t-test confirms the null-hypothesis of 0-difference between the means for the two groups S_a&S_d.

1.9 Score

This variable, which ideally represents a score assigned by the recruiter, is biased in the contexts of the protected class S_d: $(\lambda \mid S = S_d) < (\lambda \mid S = S_a)$. The underlying idea is to mimic a situation in the context of HRM where there is a human-implemented bias that could be perpetuated by data analysis. This variable is positively associated with the outcome Y.

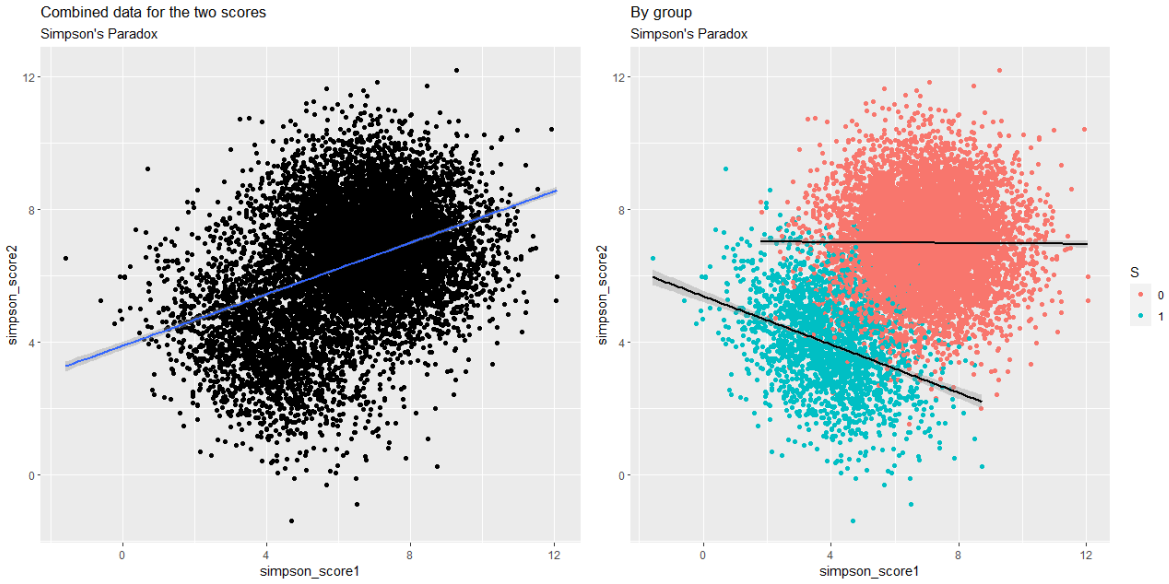
Figure 7: Distribution of Score by S



1.10 Simpson's Scores

The Simpsons Scores, involved in the generation of the target variable

Figure 8: Simson's Paradox in the synthetic Dat



The two Simpson's scores ($Simpsonscore1$ and $Simpsonscore1$) are built from a joint normal distribution with the following equations conditioned by protected feature, where u represents Simpsons score 1 and v represents the second Simpson's score [2].

$$(u, v|S = S_a) \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}2 & -0.1\\0 & 2\end{pmatrix}\right).$$

$$(u, v|S = S_d) \sim N\left(\begin{pmatrix}3\\3\end{pmatrix}, \begin{pmatrix}2 & -0.7\\-0.7 & 2\end{pmatrix}\right).$$

The correlation between the two variables involved is $\rho(v, u) = 0.380441$. Note that the global correlations between the vectors at global level has a positive sign while the covariances conditioning to the group $S = S_a$ are negative. The idea is that a linear regression analysis would suggest a significant positive association between the scores overall. However, if we look at the effect by group the picture is different.

1.11 Group

The *Group* is a randomly assigned treatment balanced and 3 treatment groups.

1.12 Model for the target variable

The target variable, a binary feature $Y \in [0, 1]$ is constructed as a function of all the features $Y = f(Age, interview, GitHub, Proxy, Proxy2, X\_score, Score, Simpson\_score1, Simpson\_score2)$ except for the sensitive feature S and the randomly assigned treatment group. Data has been centered and scaled to have regressors on the same scale before the construction of the variable Z and $inv-logit$ transformation.

$$\begin{aligned} Z = \beta * interview + \beta_1 * score + \beta_2 * Github.account - \beta_3 * age + \\ \beta_4 * X.score + \beta_5 * simpson.score1 + \beta_6 * simpson.score2 \\ +\beta_7 * simpson.score1 * simpson.score2 + \beta_8 * proxy + \beta_9 * proxy2 + \mu \end{aligned} \tag{1}$$

The variable Z obtained as a linear combination between regressors has been transformed via *inverse logit* (or *logistic function*;$pr = \frac{1}{1+exp(-z)}$) in order to obtain the probabilities linked to the outcome

[2] A practical example in HR management involves the relationship between Neuroticism(u) and salary(v) by educational level (s in our vocabulary) https://paulvanderlaken.com/

$\hat{Y}$ in an interval $[0, 1]$. This probability has been used as a π from a binomial distribution, obtaining the binary outcome Y. This Dataset is designed for a supervised learning task, the class labels of the target variable are given by the recruiter/HRM and we have to construct a predictor $\hat{Y} = f(S, X)$ that minimizes a loss function and that hopefully reduces the bias toward the sensitive feature.

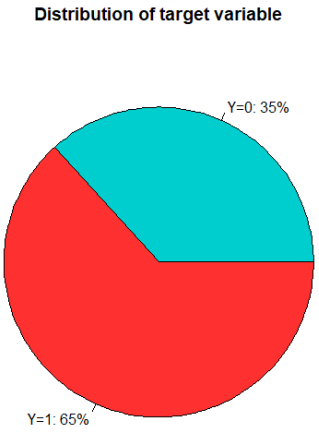
2 EDA

The Table 2 shows 4 rows from the simulated data

Table 3: Head of the Synthetic data frame

S	interview	Git_acc	proxy	proxy2	age	X_score	score	s_s1	s_s2	Y	Trt
0	6	1	-0.01	0.13	23	105	10	6.21	8.93	1	1
1	7	0	-1.23	0.12	22	100	4	2.74	4.90	0	2
0	4	0	0.23	0.00	22	94	4	7.37	6.34	1	2
0	2	0	-0.05	0.34	22	96	4	7.75	8.14	1	1

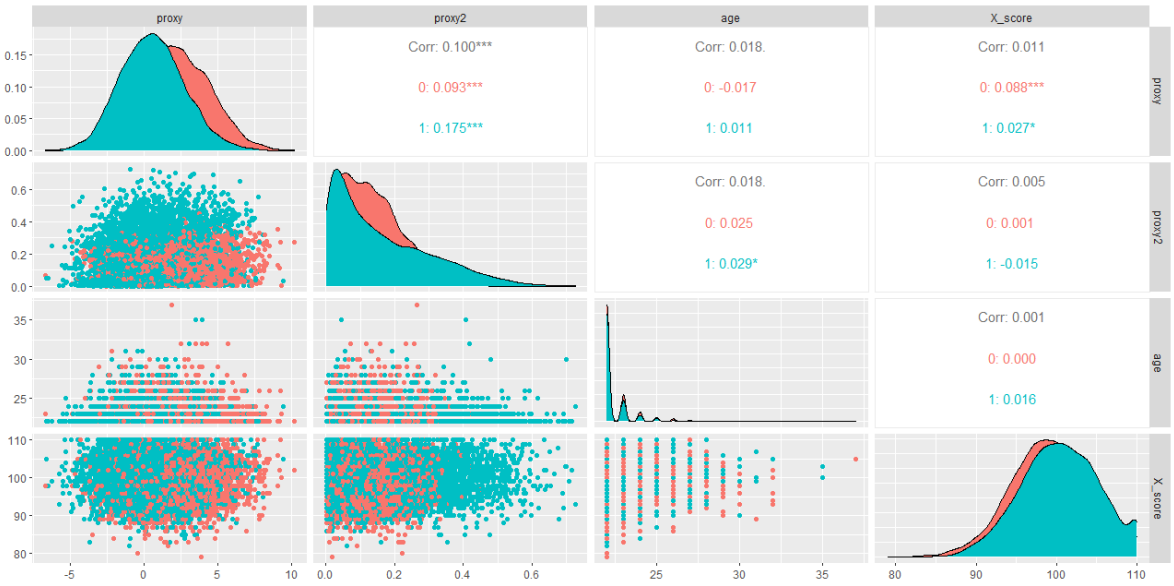
Figure 9: Distribution of the target variable



The Figure 9 Shows the overall distribution of the target variable, in a HR analytics context predictive algorithms are used in different steps and one can imagine many labels attributable to the binary response variable such as hiring, leaving the company, churning, retention etc...

In this paper, the positive outcome is denoted by $Y = 1$ and the non-preferable outcome by $Y = 1$. About 65% of the individuals in this dataset obtained a positive outcome but if we look at the distribution of Y conditional on S shown in Figure 2 the story changes.

Figure 10: EDA By Y



The exploratory analysis

Figure 11: EDA By Y

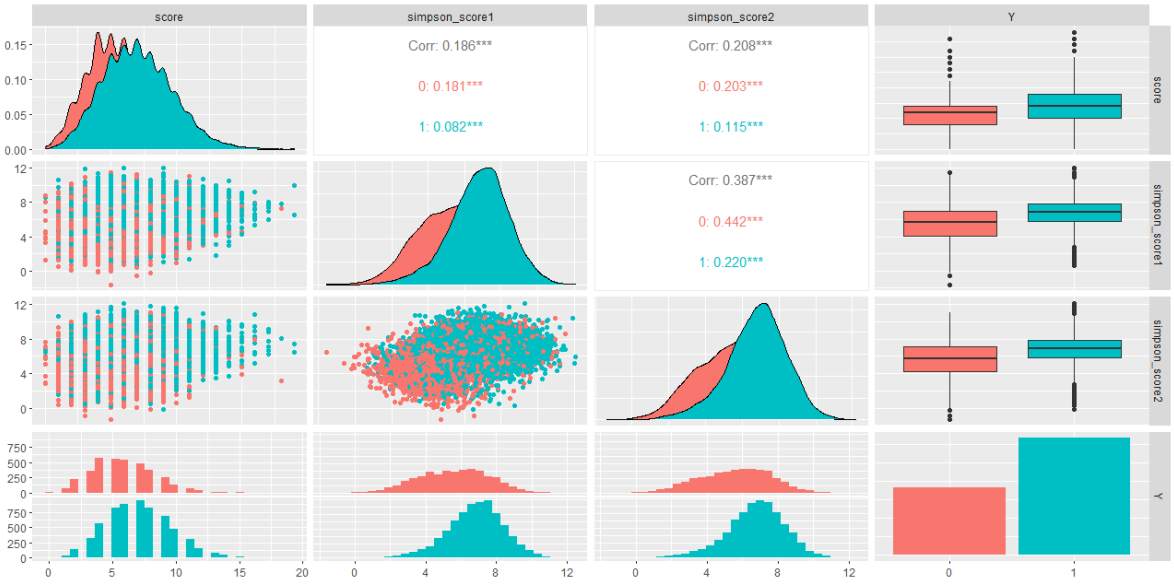
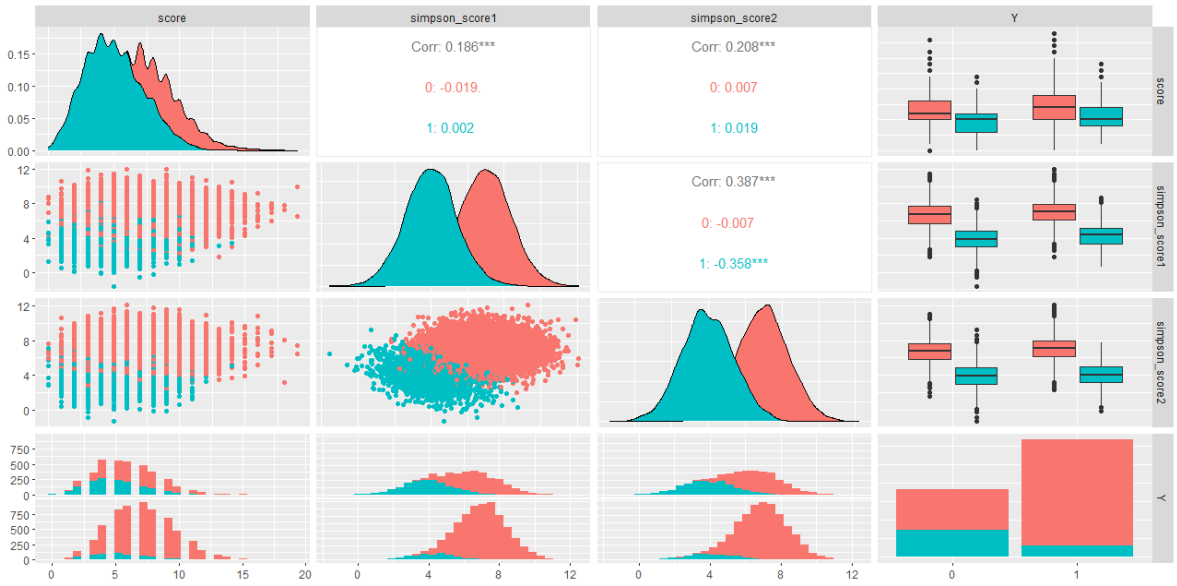


Figure 12: Discrete features and target variable

It is useful to look at the distribution of some interesting numeric features observed by S. Consistent with the mechanism that generated the data we observe that some variables see their distribution change significantly when observed conditionally with the sensitive feature S.

Figure 13: EDA By S

Figure 14: EDA By S



It is important to look at the linear correlations and VIF in order to detect a potential linear dependence between the features. An excessive values for the correlation would invalidate some models (like the linear model in case of multicollinearity).

Figure 15: Correlation plot

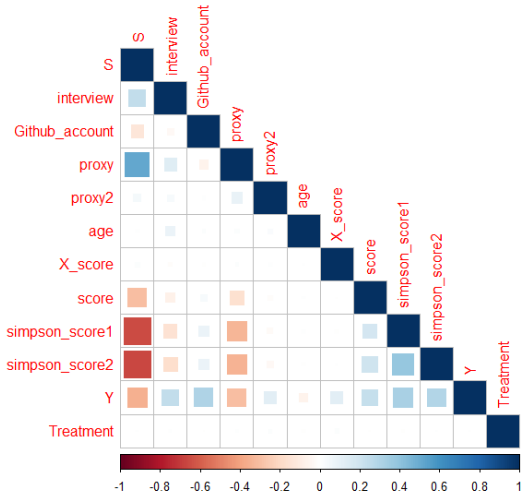
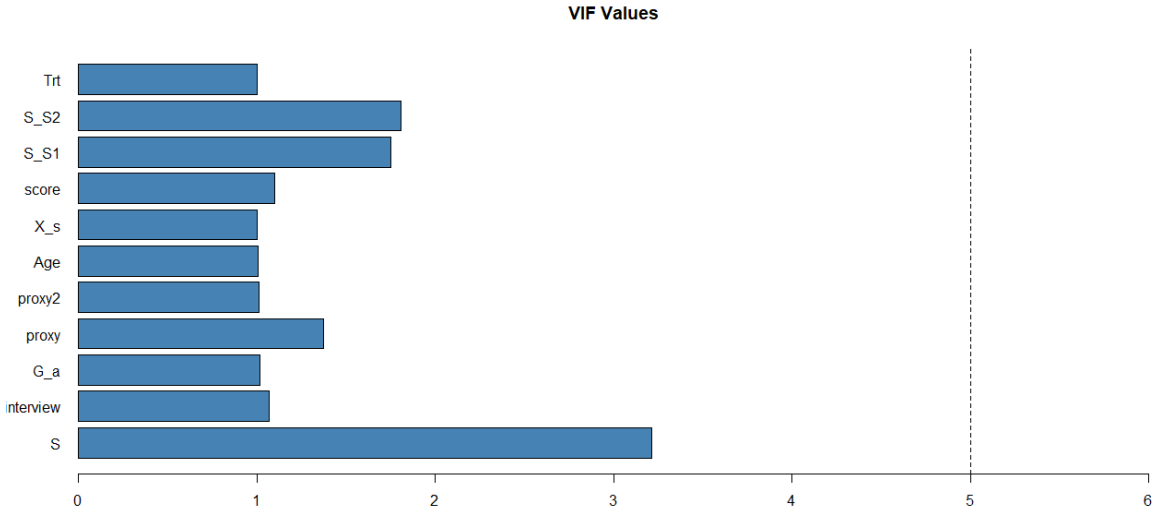


Table 4: Correlation Matrix; all features constrained as numeric

	S	int	G_a	prox	prox2	Age	X_s	score	S_S1	S_S2	Y
S	1.00	0.24	-0.14	0.52	0.05	0.01	0.02	-0.30	-0.65	-0.67	-0.35
int	0.24	1.00	-0.04	0.14	0.04	0.08	-0.01	-0.08	-0.16	-0.17	0.24
G_a	-0.14	-0.04	1.00	-0.07	0.01	0.01	-0.01	0.04	0.09	0.09	0.31
prox	0.52	0.14	-0.07	1.00	0.10	0.02	0.01	-0.16	-0.33	-0.34	-0.31
prox2	0.05	0.04	0.01	0.10	1.00	0.02	0.00	-0.02	-0.03	-0.03	0.13
Age	0.01	0.08	0.01	0.02	0.02	1.00	0.00	0.00	0.00	-0.00	-0.06
X_s	0.02	-0.01	-0.01	0.01	0.00	0.00	1.00	-0.00	-0.01	0.00	0.12
score	-0.30	-0.08	0.04	-0.16	-0.02	0.00	-0.00	1.00	0.19	0.21	0.23
S_S1	-0.65	-0.16	0.09	-0.33	-0.03	0.00	-0.01	0.19	1.00	0.39	0.32
S_S2	-0.67	-0.17	0.09	-0.34	-0.03	-0.00	0.00	0.21	0.39	1.00	0.30
Y	-0.35	0.24	0.31	-0.31	0.13	-0.06	0.12	0.23	0.32	0.30	1.00

Figure 16: VIF



A high value of VIF is interpretable as a symptom of collinearity (R^2 close to 1). A popular rule in the literature is to consider a variable collinear with another variable when VIF is greater than 5. This method is used to detect and possibly remove collinear features before model development. Looking at the Figure 2 , we are not led to remove features. THE VIF graph is to be observed along with the heat map showing Pearson correlation coefficients between variables in Figure 15.

3 Bias in data

Once we have identified (e.g. according to the legal framework) the sensitive feature S and the binary outcome Y with value 1 if favourable and 0 if not favourable, we can compute SPD and DI in order to measure the BIAS against a protected feature in our data. **Statistical Parity Difference** is defined as:

$$SPD = P(Y = 1|S = S_a) - P(Y = 1|S = S_d) = 0 \tag{2}$$

The Disparate Impact is defined as:

$$\frac{P(Y = 1|S = S_d)}{P(Y = 1|S = S_a)} \geq 0.8 \tag{3}$$

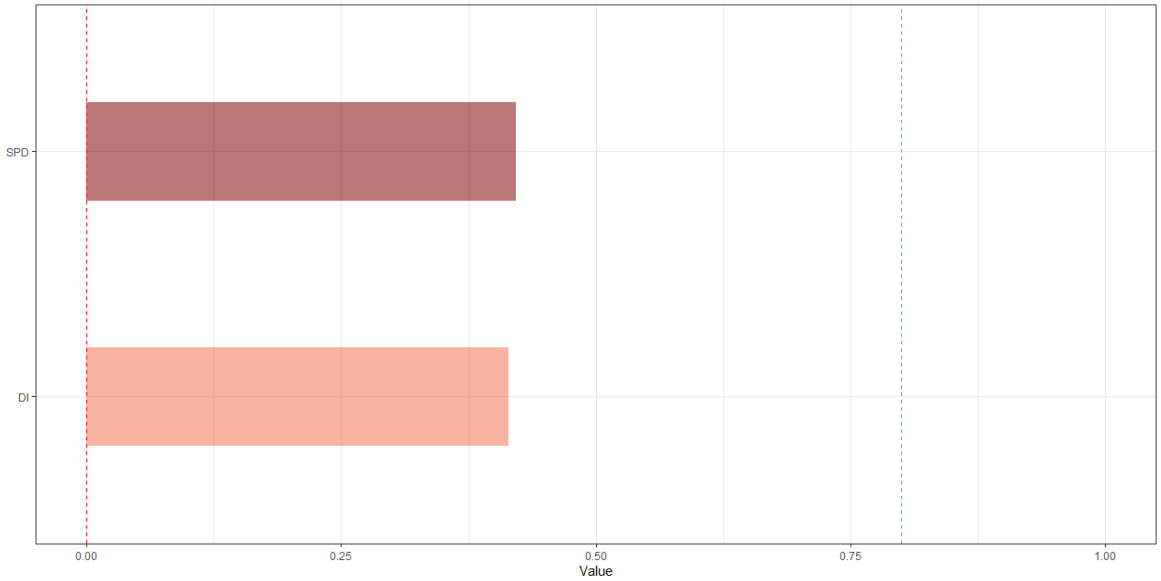
or equally

$$\frac{P(Y = 1|S = S_a)}{P(Y = 1|S = S_d)} \leq 1.25 \tag{4}$$

The disparate impact involves a threshold τ usually equal to 0.8. In other words, the probability that an individual from the group S_d would get $Y = 1$ should be at least 0.8 times the same probability for an individual belonging to the advantaged group S_a.

Statistical Parity difference & disparate impact	
SPD	DI
-0.4217392	0.4146449

Figure 17: SPD & DI



For convenience in the graphical representation, the modulus of the SPD was used since we assume that the direction of the bias is in favor of the advantaged class S_a. The red dashed line in Figure 17 represents the ideal value of SPD (as close as possible to 0) while the green dashed line represents the ideal value of DI in accordance with the 80% rule

References

[1] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.