

How Differential Robustness Creates Disparate Impact: A European Case Study

Charles Wan¹, Leid Zejnilović² and Susana Lavado²

¹Rotterdam School of Management, Erasmus University

²Nova School of Business and Economics, Universidade NOVA de Lisboa

Abstract

We formalize a notion of differential robustness. An algorithm is differentially robust to an event if the event has disparate impact on the performance of an algorithm for different groups in the population. We illustrate it with a case study of the real world deployment of a predictive algorithm in a European public employment service.

Keywords

differential robustness, causal mechanism, disparate impact, distribution shift, concept drift, algorithmic fairness

1. Introduction

An algorithm may be *differentially robust* to likely distribution shifts as a result of *variance in causal mechanisms*. This happens when there are distinct causal mechanisms at work for different groups. Even if causal inference is used to model the relationship between features and label, the distinct causal mechanisms at work for different groups are not invariant [1] and may be more or less susceptible to changes in background conditions. Possible or likely changes in the world will then have disparate impact on the algorithm's performance, leading to less accurate interventions for some group(s). Thus, algorithmic fairness is not simply a criterion that obtains under certain distributional assumptions. One should be able to either anticipate possible and likely changes in the real world or evaluate their effects on the model post hoc in a timely manner. In our paper we develop a notion of **differential robustness** and elucidate it with a case study of the real world deployment of a predictive algorithm in a European public employment service prior to and during the COVID-19 pandemic.

2. Differential Robustness

2.1. Definition

We formalize the notion of **differential robustness** for an algorithm as follows:

EWAF'23: European Workshop on Algorithmic Fairness, June 7–9, 2023, Winterthur, Switzerland


✉ wan@rsm.nl (C. Wan); leid.zejnilovic@novasbe.pt (L. Zejnilović); susana.lavado@novasbe.pt (S. Lavado)

🌐 <https://wan-charles.github.io/> (C. Wan); <http://www.zejnilovic.com/> (L. Zejnilović)

🆔 0000-0001-8284-1353 (C. Wan); 0000-0002-4209-4637 (L. Zejnilović); 0000-0002-1088-6357 (S. Lavado)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

1. Given training data set $\{(\mathbf{x}_i, y_i)\}_i \sim \mathcal{D}$ with $\mathbf{x}_i \in \mathbb{R}^{d_{in}}$, $y_i \in \{0, 1\}$, and group membership $g_i \in \{1, 2\}$ and hypothesis class \mathcal{H} , train a predictor $h : \mathbb{R}^{d_{in}} \mapsto \{0, 1\}$ that minimizes empirical risk. Individuals are assigned interventions T^0 and T^1 respectively based on model predictions.
2. Assume that there is an ordering of pairs of outcome and intervention with respect to welfare: $(y_i = 0, T^0) = (y_i = 1, T^1) > (y_i = 0, T^1) \gg (y_i = 1, T^0)$. That is, assigning intervention T^0 to individuals who actually belong to the positive class generates the greatest harm.
3. Suppose $\mathcal{D} \xrightarrow{\text{shift}} \mathcal{D}'$ for test time with $\{(\mathbf{x}'_j, y'_j)\}_j \sim \mathcal{D}'$. If $P(y' = 1|h(\mathbf{x}') = 0, g = 2) - P(y = 1|h(\mathbf{x}) = 0, g = 2) > P(y' = 1|h(\mathbf{x}') = 0, g = 1) - P(y = 1|h(\mathbf{x}) = 0, g = 1)$, then h is more robust to this particular distribution shift for group 1 than for group 2 with respect to welfare. In other words, the false negative rate for group 2 increases relative to group 1.
4. If the above holds for all *likely* shifts $s : \mathcal{D} \xrightarrow{\text{shift}} \mathcal{D}'$, then h is more robust to distribution shifts for group 1 than for group 2 with respect to welfare.

2.2. Toy Example

This can be illustrated with a toy example. Suppose there are two groups: $\{x = 0, x = 1\}_{g=1}$ and $\{x = 2, x = 3\}_{g=2}$. Given background conditions B the causal mechanism C is as follows: $x = 0$ or $x = 2 \rightarrow y = 0$; $x = 1$ or $x = 3 \rightarrow y = 1$. An intervention improves welfare if and only if $y = 1$. Suppose an algorithm h is able to predict y perfectly from x and an intervention is assigned if $h(x) = 1$. Now imagine a change in background conditions $B \xrightarrow{\text{change}} B'$ that induces a change in the causal mechanism $C \xrightarrow{\text{change}} C'$, with $C' : x = 0 \rightarrow y = 0$; $x = 1, x = 2$ or $x = 3 \rightarrow y = 1$. The predictor h is *differentially robust* to such a change since under h the false negative rate for group 2 increases relative to group 1.

3. Case Study

We elucidate the notion of **differential robustness** with a case study from the European Union. The context is a public employment service where an XGBoost-trained model was deployed to help counselors assess whether unemployed individuals are at risk of long-term unemployment (LTU), defined as being unemployed for a year or longer. We ran a pilot study from October 2019 to June 2020 and subsequently collected data on the employment outcomes of the unemployed individuals. During the period of the pilot study the COVID-19 pandemic hit and considerably changed the causal structure of unemployment with service workers in the tourism industry being among the most affected [2, 3]. The model was trained on historical data and, therefore, captured historical patterns of causal relationships.

While it is arguable whether a shock such as the COVID-19 pandemic *a priori* represents a *likely* change, the case nonetheless demonstrates how events in the world, expected and

unexpected, can change causal mechanisms and shift data distributions in such a way that the model's performance is differentially affected for different groups. Indeed, our analyses show that as a result of the COVID-19 pandemic the model's false negative rate increases in particular for service workers in the tourism industry. This means that as a group they were more vulnerable to not being allocated the needed interventions conditional on having received a negative prediction from the model. The model – and the policy designed around its predictions – is differentially robust to a shock such as the COVID-19 pandemic for different groups in the population.

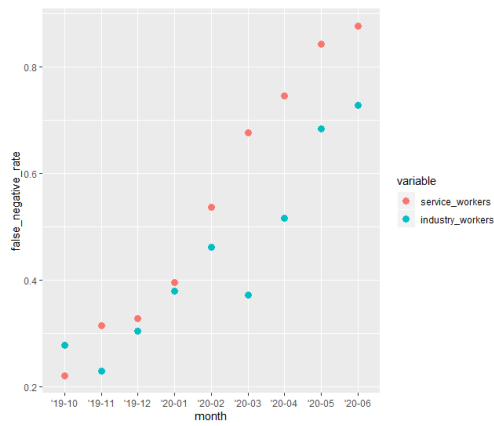
Table 1

The algorithm was differentially robust to the COVID-19 shock for service workers and industry workers as well as for non-EU/EEA nationals and Portuguese nationals.

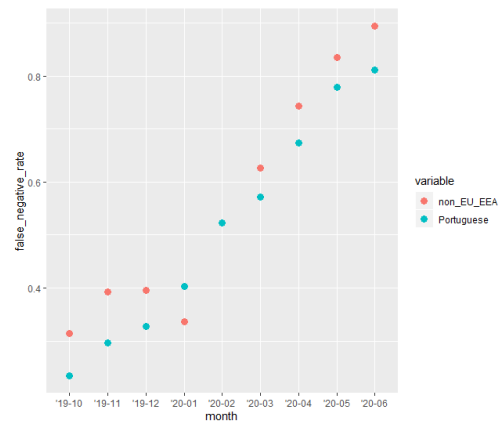
	service workers	industry workers	non-EU/EEA	Portuguese	all
$P(y' = 1 h(\mathbf{x}') = 0)$	0.781	0.474	0.790	0.696	0.713
$P(y = 1 h(\mathbf{x}) = 0)$	0.381	0.343	0.397	0.375	0.375
difference	0.400	0.131	0.393	0.321	0.338

Figure 1 shows the evolution of the false negative rates for service vs industry workers and non-EU/EEA vs Portuguese nationals before and during the first few months of the COVID-19 pandemic. We observe that the false negative rates for service and industry workers track each other fairly closely before the pandemic but the gap between them expands drastically in favor of industry workers after the onset of the pandemic in Portugal. The gap between the false negative rates for non-EU/EEA and Portuguese nationals is more volatile before the onset of the pandemic but settles into a stable pattern in favor of Portuguese nationals after the onset of the pandemic.

Figure 1: False negative rates before and during COVID-19

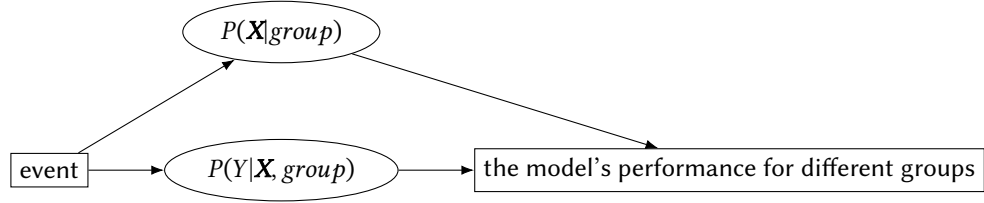


(a) False negative rates for service vs industry workers



(b) False negative rates for non-EU/EEA vs Portuguese nationals

Figure 2: Causal graph. An event can cause both covariate shift and concept drift. When there are distinct causal mechanisms at work for different groups, concept drift can differentially affect the model's performance.



This leaves open the question of whether the observed differences in the false negative rate are due to covariate shift, where $P(\mathbf{X})$ changes, or concept drift, where $P(Y|\mathbf{X})$ changes [4]. We are interested in the effects of a shock on the causal mechanism itself, i.e. $P(Y|\mathbf{X})$. This is because covariate shift in such a short period of time is likely due to sampling and not any change in the real distribution of features for a particular group. We can isolate the effects of concept drift by running the regressions below with `risk_score` as a control variable, where `risk_score` denotes the raw probability score output by XGBoost. This stratifies the unemployed individuals into bins of equal assessed risk given the features \mathbf{X} and thereby (to a large extent if not completely) controls for covariate shift as a source of variation in the false negative rate. The dummy variable `covid` is 1 if the candidate was registered after the start of the COVID-19 pandemic. The regressions are run over the subset of observations where the algorithm gives a negative prediction of LTU.

$$\text{FN} = \beta_0 + \beta_1 \text{service_workers} + \beta_2 \text{covid} + \beta_3 \text{service_workers} \times \text{covid} + \beta_4 \text{risk_score} + \varepsilon$$

$$\text{FN} = \gamma_0 + \gamma_1 \text{industry_workers} + \gamma_2 \text{covid} + \gamma_3 \text{industry_workers} \times \text{covid} + \gamma_4 \text{risk_score} + \eta$$

$$\text{FN} = \delta_0 + \delta_1 \text{non_EU_EEA} + \delta_2 \text{covid} + \delta_3 \text{non_EU_EEA} \times \text{covid} + \delta_4 \text{risk_score} + \xi$$

$$\text{FN} = \zeta_0 + \zeta_1 \text{Portuguese} + \zeta_2 \text{covid} + \zeta_3 \text{Portuguese} \times \text{covid} + \zeta_4 \text{risk_score} + \sigma$$

The regression results in Tables 2 and 3 show that the coefficients for `service_workers × covid` and `non_EU_EEA × covid` are positive and statistically significant. We can interpret this as evidence that concept drift caused by COVID-19 led to an increase in the false negative rate for service workers (non-EU/EEA nationals) relative to other groups. Conversely, the coefficients for `industry_workers × covid` and `Portuguese × covid` are negative and statistically significant. Concept drift caused by COVID-19 led to a decrease in the false negative rate for industry workers (Portuguese nationals) relative to other groups.

Table 2
LPM regression results

	<i>Dependent variable:</i>			
	FN	FN	FN	FN
	(1)	(2)	(3)	(4)
covid	0.322*** (0.006)	0.365*** (0.005)	0.337*** (0.005)	0.408*** (0.014)
service_workers	0.018 (0.009)			
industry_workers		−0.030* (0.015)		
non_EU_EEA			0.046* (0.022)	
Portuguese				−0.025 (0.013)
service_workers × covid	0.085*** (0.012)			
industry_workers × covid		−0.178*** (0.018)		
non_EU_EEA × covid			0.071** (0.025)	
Portuguese × covid				−0.082*** (0.015)
risk_score	0.739*** (0.048)	0.670*** (0.048)	0.769*** (0.049)	0.839*** (0.049)
Observations	32,610	32,610	32,610	32,610
R^2	0.1242	0.129	0.122	0.1239
Adjusted R^2	0.1241	0.1289	0.1219	0.1238
<i>Note:</i>		*p<0.05; **p<0.01; ***p<0.001		

4. Conclusion

Our work shows that to ensure algorithmic fairness it is not sufficient for criteria of statistical fairness to obtain. If the model is especially or uniquely vulnerable to performance degradation for a particular group in possible scenarios, fairness will be elusive. Possible and likely changes in the world should be considered and their differential impact on the algorithm's performance for different groups carefully evaluated. This requires the judicious incorporation of domain knowledge into the decision-making process. Alternatively, there should be continual post hoc evaluation of an algorithm's differential robustness, for which the analysis in our paper could serve as a model of performance monitoring.

Table 3
Logit regression results

	<i>Dependent variable:</i>			
	logit(FN) (1)	logit(FN) (2)	logit(FN) (3)	logit(FN) (4)
covid	1.355*** (0.028)	1.564*** (0.026)	1.425*** (0.025)	1.790*** (0.065)
service_workers	0.080 (0.042)			
industry_workers		−0.129 (0.066)		
non_EU_EEA			0.207* (0.099)	
Portuguese				−0.116* (0.058)
service_workers × covid	0.448*** (0.057)			
industry_workers × covid		−0.789*** (0.082)		
non_EU_EEA × covid			0.393*** (0.116)	
Portuguese × covid				−0.420*** (0.070)
risk_score	3.449*** (0.226)	3.152*** (0.225)	3.563*** (0.228)	3.879*** (0.231)
Observations	32,610	32,610	32,610	32,610
Log-likelihood	−20201.36	−20122.92	−20247.13	−20211.38
<i>Note:</i>		*p<0.05; **p<0.01; ***p<0.001		

References

- [1] J. Woodward, Causation with a human face: Normative theory and descriptive psychology, Oxford University Press, 2021.
- [2] A. S. Lopes, P. Carreira, Covid-19 impact on job losses in portugal: who are the hardest-hit?, International Journal of Manpower (2021).
- [3] A. S. Lopes, A. Sargento, P. Carreira, Vulnerability to covid-19 unemployment in the portuguese tourism and hospitality industry, International Journal of Contemporary Hospitality Management (2021).
- [4] C. Huyen, Designing Machine Learning Systems, " O'Reilly Media, Inc.", 2022.