Human Interaction and Emerging Technologies (IHIET 2023), Vol. 111, 2023, 749–755
https://doi.org/10.54941/ahfe1004082

Proposition of an Utilitarianism and Fair Objective Function Building Method Based on Values and Socio-Economic Consequences for Data-Driven Decisions

Christian Goglin

Data & AI Dpt., ICD Business School 12 rue Alexandre Parodi, Paris 75010, France

ABSTRACT

In this short article, we propose a method to setup the objective function of machine learning binary classifier used in data driven decision. The goal is to take fair decisions aligned with an ethical value system and based on the long-term consequences of prediction errors for all stakeholders. The proposed method is based on human in the loop with an ethical committee to define the appropriate setup of the objective function, depending on the context of the decision. The setup parameters are of three categories: the fairness criteria, the ethical values and the weights associated to socio-economic long-term consequences of prediction errors.

Keywords: Binary classifier, Group fairness, Socio-economic consequences, Ethical values, Ethical committee

INTRODUCTION

In data-driven decisions based on a binary classification machine learning model, the algorithm learns the model from the data by optimizing a function measuring the empirical error, named the objective function. Much work has been done to make these decision models fairer by optimizing group fairness. The multiple group fairness criteria proposed (independence, separation, sufficiency...) aim to implement static affirmative actions to improve group fairness (Kozodoi *et al*., 2022). But these mitigation actions based on fairness processor, are made at the expense of the performance of the model (for example measured by the accuracy) and therefore of its initial objective. Also, these methods do not take into account the perspective of the long-term socio-economic consequences of the predictions errors for all stakeholders, and resulting from the model being made fairer by affirmative actions (Liu *et al*., 2018). In the classical example Compas[1], revealed in the ProPublica article[2], if we reason about the consequences of prediction errors, the antagonism between predictive performance (accuracy) and fairness, leads to the

[1]COMPAS = Correctional Offender Management Profiling for Alternative Sanctions

[2]https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

© 2023. Published by AHFE Open Access. All rights reserved.

following question: are we willing to increase the recidivism rate of defendants to improve fairness between ethnic groups? This trade-off may or may not be shocking depending on one's value system. Indeed, we are in the context of a moral dilemma in which two values are in competition: the integrity of individuals and the equity between groups. Then this question may arise if the social consequences of less discrimination between groups result in more errors of justice, which is a dramatic effect. Thus, not only the values come into play, but also the relative importance of the consequences of predictions impacting on both prisoners and future victims, and specifically the erroneous predictions.

To address this issue, we propose a method consisting in generalizing analysis outcomes from two cases, that of bank credit granting and that of predictive justice.

Credit Granting Case

The banking credit granting process is based on a credit score, measuring the expected default risk of the borrower. Nowadays, this score is often computed using a machine learning model trained on a huge historical dataset.

Regarding the fairness, and more precisely the group fairness, scholars propose numerous criteria and processors to mitigate the fairness issue. The mitigation is performed by optimizing the chosen fairness criteria. Let's consider two groups in the population, defined on a protected attribute x_a, for instance the gender, with $x_a = 0$ for female and $x_a = 1$ for male.

As explained by Kozodoi *et al.* (2022), the separation criteria (SEP) suits well the specific context of the credit granting. This criterion means the classifier don't make more mistakes (prediction errors) for one group than for the other. If we consider the confusion matrix resulting of the model evaluation, for both groups, the signification of this constraint is:

$$FPR_{\{xa=0\}} = FPR_{\{xa=1\}} \; and \; FNR_{\{xa=0\}} = FNR_{\{xa=1\}}$$

The separation criteria, defined as follow, must be minimized:

$$\text{SEP} = (1/2)|(FPR_{\{xa=0\}} - FPR_{\{xa=1\}}) + (FNR_{\{xa=0\}} = FNR_{\{xa=1\}})|$$

With

FPR = False Positive Rate = the classifier predicts no default but the borrower default

FNR = False negative rate = the classifier predicts a default, but the borrower would not default

SEP is part of our first criteria, associated to the fairness ethical value, formally, we write:

$$\text{FAIRNESS\_CRITERION} = V_0.\alpha_0.\text{SEP}$$

With V_0 the fairness value subjective relative importance and α_0 the subjective relative severity of the fairness risk, both are scalar $\epsilon[0, 1]$. Subjective means the result of human judgement.

Also, in this use case, we have two stakeholders, the bank, and the borrower. Let's then consider the whole group of borrowers (female and male)

one hand, and the bank on the other hand, to examine the consequences of prediction errors.
Stakeholder 1, Bank:

- Case *FPR*: the consequence is economic, it's a loss due to the default of the borrower.
- Case *FNR*: the consequence is economic, it's a shortfall in income due to the credit not granted (no cumulated interest income)

For each harmful consequence, we identify the related ethical value, chosen in the EU guidance document "Design and Ethics of Use Approaches for Artificial Intelligence"[3]

Associated value to *FPR* and *FNR* is 'social well-being' corresponding to following danger: the financial stability of the bank.

(Verbraken *et al.*, 2014) propose an indicator, the expected maximum profit (EMP), built on FPR and FNR for this specific context that best suit bank objective than Accuracy [(TP+TN)/(TP+TN+FP+FN)] since it takes in account the real expectation: the profit and lost.

Note: consequences are asymmetric. In average, *FPR consequences* severity > *FNR consequences* severity, even if this assertion closely depends on interest rates.

This note means we can weight the consequences related to FPR and FNR differently.
Stakeholder 2, borrower:

- Case *FPR*: the default has important economic consequences for the borrower and first, a risk of over indebtedness. There are also some psychologic consequences like ill-being due to over-indebted since people experience anxiety linked to material difficulties such as the seizure of furniture, electricity, or telephone cuts, or even eviction from their home. Dependence on family and institutions becomes extreme, especially if the over-indebted person is unable to have a home or a bank card. This anxiety can then lead to depression and suicidal thoughts.
- Case *FNR*: the consequence is a risk of banking exclusion with economic and social consequences since the person is not able to realize its professional projects (starting a business or financing professional training) and may ask help to his family and friends. There are also some psychologic consequences like ill-being due to difficulties in obtaining a job and integrating in society (Gloukoviezoff, 2009).

Again, consequences are asymmetric with *FPR* consequence severity > *FNR* consequence severity, meaning we could weight those consequence differently.

Associated value to *FPR* and *FNR* is the same: 'individual well-being' corresponding to following danger: personal bankruptcy or difficulties to integrate in society and mental health.

[3] https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-by-design-and-ethics-of-use-approaches-for-artificial-intelligence_he_en.pdf

If at the beginning of this research, our first idea was to define an economic indicator measuring the consequences of *FPR* and *FNR* for borrower, to have an equivalent to the EMP indicator and be able to mix both indicators, we dismiss this idea since for borrowers, consequences are not only economic but also psychosocial.

After this brief analysis, we can define our secund criterion, related to socioeconomic consequences of classifier prediction errors for stakeholders:

$$\text{SOCIO\_ECO\_CRITERION} = (V_1.\alpha_1 + V_2.\alpha_2).FPR + (V_3.\alpha_3 + V_4.\alpha_4).FNR$$

V_i and $\alpha_i$are scalar $\epsilon[0, 1]$ with $i = 1..4$

V_1 and V_3 are the subjective severity related to social well-being value and α_1, α_3 the subjective severity of the social well-being risk for the bank regarding *FPR* and *FNR*.

V_2 and V_4 are the subjective severity related to individual well-being value and α_2, α_4 the subjective relative severity of the individual well-being risk for the borrower regarding *FPR* and *FNR*.

This secund indicator replaces the Accuracy, the goal is to minimize the prediction errors but, taking in account the subjective severity of values at stake (V_i) and the relative severity of socioeconomic consequences at stake for all stakeholders (α_i).

The final objective function to minimize mix two criteria, the first one is related to the fairness while the secund one is related to the socioeconomic consequences:

$$\text{CRITERION} = \text{FAIRNESS\_CRITERION} + \text{SOCIO\_ECO\_CRITERION}$$

Prédictive Justice Case

This second case relates to the predictive justice, when a score, based on machine learning, predicts the likelihood of committing a future crime by a defendant.

Concerning the first criterion related to the group fairness, we choose the independence (IND) criteria, meaning the model must not disadvantage one group. Here, groups are defined based on a protected ethnic attribute x_a, with $x_a = 0$ for Afro-American and $x_a = 1$ for White defendants. IND criterion means:

$$PR_{\{xa = 0\}} = PR_{\{xa = 1\}}$$

With

PR = Positive Rate = Proportion of defendants for which the classifier predicts they will not commit a future crime

The independent criterion, defined as follow, must be minimized:

$$\text{IND} = |(PR_{\{xa = 0\}} - PR_{\{xa = 1\}})|$$

IND is part of our first criteria, associated to the fairness value, formally, we write:

$$\text{FAIRNESS\_CRITERION} = V_0.\alpha_0.\text{SEP}$$

In this case, we have two stakeholders, the defendant, and the people, meaning anybody that can be victim of the defendant in case of reoffending. Let's examine the socioeconomic consequences of prediction errors:
Stakeholder 1, Defendant:

- Case *FPR*: the defendant is released but commit a crime, the main consequence, for him, is a new period of imprisonment.
- Case *FNR*: the defendant is not released while he would not commit a crime, the main consequence, for him, is deprivation of freedom.

Associated value to *FPR* and *FNR* is 'freedom' corresponding to following danger: the imprisonment.

Consequences are asymmetric, this time for an ethical reason (defendant free will) rather than the severity of the consequences: *FNR consequences* > *FPR consequences* because with *FNR*, the harmful consequences are not depending on the defendant whereas with *FPR*, the defendant is guilty.
Stakeholder 2, The people:

- Case *FPR*: the defendant is released but commit a crime, the consequence for the victim may be severe and even fatal with the death.
- Case *FNR*: the defendant is not released while he would not commit a crime, the consequence for the people is null.

Associated value to *FPR* 'individual well-being' corresponding to following danger: to be victim of a crime.

Consequences are obviously asymmetric, *FPR consequence severity* > *FNR consequence severity*.

We can now define our secund criterion, related to socioeconomic consequences of classifier prediction errors for stakeholders.

$$\text{SOCIO\_ECO\_CRITERION} = (V_1.\alpha_1 + V_2.\alpha_2).FPR + V_3.\alpha_3.\text{FNR}$$

(NB: α_4 is null, see upper)
The final objective function to minimize is again as follow:

$$\text{CRITERION} = \text{FAIRNESS\_CRITERION} + \text{SOCIO\_ECO\_CRITERION}$$

This second case is useful for two reasons:

1) This time, the socioeconomic consequences of prediction errors are more psychosocial than economic.
   This point supports the idea to have dedicated weight to measure the consequences severity rather than a quantitative measure like the Expected Maximum Profit, because psychosocial consequences cannot be reduced to a monetary amount. To determine the severity weights, we propose a qualitative analysis, resulting from discussions between stakeholders, lawyer and ethicist expert, and a deliberation, rather than a mathematical model that cannot embed the long-term consequences and the complexity of the psychosocial dimension.
2) The "Compas" case, mentioned in the introduction, is a real illustration of predictive justice. In this case, the software company solution (based

on machine learning), compute a score, predicting the likelihood a released defendant would commit a future crime. First, "*a disproportionate number of black defendants were 'false positives': they were classified by COMPAS as high risk but subsequently not charged with another crime*." as explain (Courtland, 2018) but what is more interesting for us is the illustration of a dilemma where two values are in competition: the fairness and the safety (Individual well-being) as explained by Müller (2020). In Table 1 below, we can see that human judges take fair justice decisions since the error rate (60%) are equal between groups (White and Afro-American) but the total accuracy (60%) of their decisions is lower than the total accuracy of decisions computed by robot judge (65%) which in turn are less fair (error rates disadvantage Afro-American).

This point supports the idea to quantify the relative importance of values at stake in the decision where values derive from the harmful consequences of the prediction errors.

Table 1. Tension between Accuracy and fairness in predictive justice (Müller, 2020).

	Human Judge (fair)	**Robot Judge (accurate)**
Whites	60%	72%
Afro-Americans	60%	58%
Accuracy	60%	65%

Généralization

In the previous cases, we 've chosen the same analysis approach. We have defined a two-criteria objective function with group fairness as first criterion and prediction errors (FPR and FNR) as proxy of socio-economic consequences. We have identified a system of ethical values, specific to the decision-making context, chosen in the ethical values and principles proposed in the EU guidance document "Design and Ethics of Use Approaches for Artificial Intelligence". This value-based method is in line with other approach more general used to design electronic systems with ethical values, like 'Ethics by design' and 'Value-sensitive design' (Stahl *et al.*, 2023).

We also have measured the socio-economic harmful of prediction errors (FP and FN) with scalar weights (α_i with $i = 1..4$) because some social consequences cannot be reduced to monetary equivalent amount, like the shame sentient resulting of the banking exclusion or the freedom privation due to unfair extended imprisonment.

We think both the ethical values at stake, and the weights, could be setup by an *ad hoc* committee, composed by the stakeholders (banker and borrower representants in the first case), plus ethicist and lawyer. We recommend a decision-making process in two steps with at first, an open discussion between ethical committee members, followed by a vote. The first step would allow to identify ethical values related to prediction errors for each stakeholder and to discuss the relative importance of socio-economic harmful consequences. The discussion would also focus on the appropriate group-fairness criterion to use in the case context. The vote would then consist

of filling in a questionnaire where ethical values and weights would be asked using a numerical scale (Osgood scale for instance). Ethical values and weights would then be averaged.

The outcome of the ethical committee deliberation process would be the objective function to minimize and implement in the classifier model.

CONCLUSION

Nowadays, data driven decisions are more and more frequently used to decide resources allocation (credit granting), to accept or reject application or to assist public authorities (social benefits, justice, police…). Artificial Intelligence systems are often machine learning based binary classifier, fueled by empirical data. It's now well-known that numerous bias exist in the training Dataset leading to unfair decision between groups. Criteria proposed by scholars help to mitigate this risk, but other dimensions are not addressed by the literature, the ethical values dilemma as well as the importance of the long-term consequences of the prediction errors at stake in the decision and depending on a specific context. The method presented here aims to answer this gap. Since this research is ongoing, this method still need to be tested with empirical data for comparison.

REFERENCES

Courtland, R. (2018) 'As machine learning infiltrates society, scientists grapple with how to make algorithms fair.', *Nature*, 558(7710), pp. 357–360.

Gloukoviezoff, G. (2009) 'L'exclusion bancaire : de quoi parle-t-on ? Une perspective française', *Vie sciences de l'entreprise*, N° 182(2), pp. 9–20.

Kozodoi, N., Jacob, J. and Lessmann, S. (2022) 'Fairness in Credit Scoring: Assessment, Implementation and Profit Implications', *European Journal of Operational Research*, 97(3), pp. 1083–1094.

Liu, L. T. *et al.* (2018) 'Delayed Impact of Fair Machine Learning'. arXiv. Available at: https://arxiv.org/abs/1803.04383 (Accessed: 14 March 2023).

Müller, V. C. (2020) 'Ethics of Artificial Intelligence and Robotics', in E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy*. Winter 2020. Metaphysics Research Lab, Stanford University. Available at: https://plato.stanford.edu/archives/win2020/entries/ethics-ai/ (Accessed: 17 April 2021).

Stahl, B. C., Schroeder, D. and Rodrigues, R. (2023) *Ethics of Artificial Intelligence: Case Studies and Options for Addressing Ethical Challenges*. Cham: Springer International Publishing (SpringerBriefs in Research and Innovation Governance). Available at: https://doi.org/10.1007/978-3-031-17040-9.

Verbraken, T. *et al.* (2014) 'Development and application of consumer credit scoring models using profit-based classification measures', *European Journal of Operational Research*, 238(2), pp. 505–513. Available at: https://doi.org/10.1016/j.ejor.2014.04.001.