Towards a Framework for the Global Assessment of Sensitive Attribute Bias Within Binary Classification Algorithms

Adrian Byrne[1,2], Ivan Caffrey[2] and Quan Le[1]

[1] *CeADAR, NexusUCD, University College Dublin, Belfield Office Park, Unit 9, Clonskeagh, Dublin 4, Ireland*
[2] *Idiro Analytics, Clarendon House, 39 Clarendon Street, Dublin 2, Ireland*

Abstract

This paper proposes a framework for undertaking bias monitoring with respect to binary classification algorithms. We present reproducible methods using a reproducible dataset rich in sensitive attribute information that can help identify both problematic bias and problematic proxies. Our demonstration uses the IPUMS International 10% random sample of the 2016 Irish Census survey courtesy of the Central Statistics Office in Ireland and centres on whether sex or ethnicity or both and their interaction significantly contribute to the prediction of our owner or renter target controlling for our social indicator of occupation and age. We develop our bias monitoring framework around this demonstration. Our focus is global as we are interested in monitoring the overall model performance as well as the contribution of each feature rather than locally picking out individual row instances. We deploy explainable AI functions (ELI5, RFECV and SHAP) on both clear/transparent (logistic regression) and opaque/black-box (random forest classifier) models in order to assess the level of inferential agreement on the underlying dataset despite the algorithms having different predictive capabilities. Our proposed framework can be extended to any classification or regression task as it is designed to be model agnostic so long as there is access to a structured, tabular dataset. The framework is designed to be as fair as possible to practitioners whilst also providing robust bias detection that citizens can have confidence in.

Keywords

Bias monitoring, classification algorithms, sensitive attributes, explainable AI

1. Introduction

Binary classification algorithms predict whether something is a 'one' or a 'zero' based on information fed to the algorithm. Applications for such classification algorithms are numerous and non-trivial, e.g. hiring or not, promotion or not, admission or not, treatment or not, lending or not, detention or not. The non-trivial nature of these classification algorithms coupled with their increasing use and application, as technology advances, has prompted policymakers to propose legislating against potential harms to people as a direct result of these binary classification algorithms.[1]

Within the legislative text of the EU's proposed Artificial Intelligence Act (AIA), it is stated that

"In order to protect the right of others from the discrimination that might result from the bias in AI systems, the providers should be able to process also special categories of personal data, as a matter of substantial public interest, in order to ensure the bias monitoring, detection and correction in relation to high-risk AI systems."

Bias monitoring is explicitly covered within Article 10 (data and data governance) of the proposed Act and the non-trivial examples outlined above are all high-risk AI system examples according to Annex III of the Act. With the Act due to be enacted in 2024, bias monitoring of algorithms (including binary classification) in high-risk sectors is soon to be a necessary consideration for providers and users of such systems for automated

EWAF'23: European Workshop on Algorithmic Fairness, June 07–09, 2023, Winterthur, Switzerland
EMAIL: Adrian.byrne@ucd.ie (A. 1)
ORCID: 0000-0003-4887-2572 (A. 1)


[1] For the EU, see https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52021PC0206
For the US, https://oag.dc.gov/release/ag-racine-introduces-legislation-stop

decision-making purposes. This paper proposes a framework for undertaking bias monitoring with respect to binary classification algorithms.

Machine learning (ML) models that use historical data for predictions can inherit biases from the past which can lead to increasing discrimination [1]. These ML systems can perpetuate unfairness due to human bias existing in the training data [2]. Operators of these systems seeking to reduce the risk of harmful outcomes for consumers (and their businesses), should promote the use of bias mitigation proposals thereby creating a path towards algorithmic fairness [3]. However, it is still the case that bias and fairness are application dependent and there is no universally agreed upon method for resolving such issues, i.e. there exists a lack of standardisation [4]. Caton and Haas (2020)[5] elaborate on this heterogeneity in their survey of research literature on fairness in ML. Further to the lack of consensus on bias and fairness metrics, the authors identify a growing knowledge gap between state of the art ML system advancement and the ”man-on-the-street” implying that it is getting harder for the general public to understand these systems. The authors also suggest that practitioners require assistance in the identification of protected variables and groups as well as their proxies. Caton and Haas conclude by suggesting it may be preferable to fix a bias issue by improving the underlying data sample, i.e. collect more data which better represents minority or protected groups, than try to fix a discriminatory ML model.

Returning to the lack of standardisation, there are an increasing number of international standards being published on artificial intelligence (AI). Recent published examples include ISO/IEC 22989 (2022)[6]; AI concepts and terminology, ISO/IEC 23053 (2022)[7]; Framework for AI Systems Using ML, and ISO/IEC 23894 (2023)[8]; Guidance on AI risk management. Bias is also addressed via ISO/IEC TR 24027 (2021)[9]; Bias in AI systems and AI aided decision making as well as the forthcoming ISO/IEC CD TS 12791 [10]; Treatment of unwanted bias in classification and regression machine learning tasks. A common theme across all these standards is ongoing verification and validation in relation to AI/ML systems and the ISO/IEC/IEEE 29119 series of software testing standards is cited in an attempt to define an internationally-agreed set of standards for software testing that can be used by any organisation when performing any form of software testing [11]. Therefore, there are and will continue to be international standards on AI/ML systems and associated biases in an attempt to design ”best practice”. These standards may interact with forthcoming legislation on AI and it is anticipated that this paper will complement the drive for better standards in AI/ML bias detection.

Our proposed framework draws upon the work of O’Neil (2022)[12], Flores et al (2016)[13], and Skeem and Lowenkamp (2016)[14] when testing for bias in both the degree and functional form of prediction as a function of sensitive attributes (i.e. personal data). This includes inspecting the effect of interactions between sensitive attributes and key features associated with targets of interest. Our framework enables the identification of both problematic bias and problematic proxies.[2] We define ‘problematic’ as sensitive attribute(s) being statistically (i.e. non-ignorably) important despite controlling for the most important features and importance is determined via post-model [agnostic] explainable AI (XAI) techniques namely Permutation Importance [15], Recursive Feature Elimination with Cross Validation [16], SHapley Additive exPlanations [17] and Absolute Value Test Statistics [18]. We take the opportunity in this paper to test these different XAI techniques on our chosen dataset to check for (in)consistencies across their respective outputs. Problematic bias and proxies are assessed in the same way using modelling and post-modelling methods. This framework is designed for structured, tabular data and requires access to sensitive attribute data which is consistent with the idea that fair algorithms require sensitive attributes to achieve certain fairness goals [2], which is akin to the idea of “fairness through awareness”.

2. Data

In order to illustrate our proposed bias monitoring framework, we utilise the IPUMS International 10% random sample of the 2016 Irish Census survey courtesy of the Central Statistics Office (CSO) in Ireland [19]. We chose this dataset as it contains a rich source of sensitive attribute data and is freely available for bona fide

[2] In this paper, proxy is taken to mean association with sensitive attribute(s) rather than association with the unobserved variable(s) that one may attempt to measure by proxy.

researchers. The sample constitutes 491,122 person records. To avoid household duplication, we focus on the primary person in each household, i.e. the primary survey responder in each household labelled person number one. This emphasis reduces our sample size to 181,448 (37% of the total sample). Our binary classification target is owner or renter. We chose this target as a plausible indicator of wealth. Our key feature is a binary social indicator of occupation, i.e. non-manual or higher versus manual or lower. We chose this feature as a plausible indicator of being a desirable borrower from a lender's perspective. We also include age between 30 and 64 as we hypothesise this feature would be associated with both our chosen target (owner or renter) and our chosen key feature (social indicator of occupation). Including age only between 30 and 64 further reduces our sample size down to 121,530. Our chosen sensitive attribute features are sex (male or female) and ethnicity (Irish or non-Irish). Within our chosen dataset, owner or renter and ethnicity variables contain missing data; 6,995 and 2,151 cases respectively with 1,233 missing for both variables and 918 missing for ethnicity only. This means the complete case dataset used in this paper contained 113,617 cases.

3. Methods

Kang et al (2022)[20] note that most studies on group fairness focus on a single sensitive attribute which they find unrealistic and call for more studies to include multiple sensitive attributes when assessing fair outcomes. Given our paper includes multiple sensitive attributes (i.e. sex and ethnicity), we have the opportunity to investigate both main and interaction effects, rather than a single attribute treated in isolation. With our framework, we propose to look at the interaction of any statistically important sensitive attribute with other sensitive attribute(s) (to possibly uncover vulnerable subgroups) as well as the most important feature(s) (to possibly uncover moderating/amplifying effects). This approach aligns with the concept of 'intersectionality'. Intersectionality identifies multiple factors (such as sex and ethnicity) of advantages and disadvantages across time and space. Kimberlé Crenshaw,[3] law professor and social theorist, first coined the term intersectionality in her 1989 paper "Demarginalizing The Intersection Of Race And Sex: A Black Feminist Critique Of Antidiscrimination Doctrine, Feminist Theory And Antiracist Politics." Crenshaw pointed to the 1976 case DeGraffenreid v. General Motors,[4] in which the plaintiffs alleged hiring practices that specifically discriminated against black women and that could not be described as either racial discrimination or sex discrimination alone. In this paper, we are interested in the interactions of these sensitive attributes in pursuit of bias detection within binary classification algorithms as these interactions may hide and exacerbate discrimination if left unchecked.

Prior to the modelling stage, all variables are assessed for serial correlation via correlation matrices and variance inflation factor (VIF) analysis. This is to identify any multicollinearity issue which may compromise our ability to robustly explain the results we observe. If any serial correlation is detected among the features, we can either select the feature (among the serially correlated set) with the strongest correlation with the target or undertake a dimension reduction exercise, such as principal components analysis, and select the resulting summarised feature if the data allows for a clear, explainable summary of the observable features. Also, initial descriptive statistics will help inform us if there is any missing data that may require further analysis to determine if the missing data can be considered either (completely) missing at random or not. Further to undertaking these checks, we explore within-feature target proportion distributions to determine the level of class imbalance for each feature as well as pairs of features via their interaction.

Before presenting our main findings using the IPUMS CSO dataset, we first present our general, statistical, whole model approach for assessing bias. In doing so, our starting assumption is that there is a structured, tabular dataset which contains a target, features and sensitive attributes and that there is a model for this dataset. Our approach is designed to flag any concerning areas, from a bias perspective, that merit further consideration. Therefore, our generalised proposed framework for tackling structured, tabular data for binary classification algorithms (which can be extended to cover regression algorithms too) with respect to automated decision making (especially in high-risk areas as defined by the EU AIA) is as follows:

1. Determine the dataset to be used, e.g. target, key features and sensitive attributes. Feature importance techniques can be utilised to remove redundant features if needs be.

[3] https://www.law.columbia.edu/faculty/kimberle-w-crenshaw
[4] https://www.ywboston.org/2017/03/what-is-intersectionality-and-what-does-it-have-to-do-with-me/

2. Descriptive statistics with application of the class imbalance inspection between target, key features and sensitive attributes as well as serial correlation/multicollinearity and missing data checks
3. Run model with key features on target to establish the base model results
4. Run model with key features and sensitive attributes on target to identify any problematic bias (as determined by post-model [agnostic] techniques)
5. Run model with target and sensitive attributes on key features to identify any problematic proxies (as determined by post-model [agnostic] techniques)

4. Results[5]

As our target is binary, we deployed both random forest classifier and logistic regression algorithms in Python and analysed the following models:

Model 1: Owner or renter ~ social occupation indicator
Model 2: Owner or renter ~ social occupation indicator + age
Model 3: Owner or renter ~ social occupation indicator + age + sex
Model 4: Owner or renter ~ social occupation indicator + age + ethnicity
Model 5: Owner or renter ~ social occupation indicator + age + sex + ethnicity
Model 6: Missing owner/renter ~ social occupation indicator + age + sex + ethnicity
Model 7: Social occupation indicator ~ Owner or renter + age + sex + ethnicity
Model 8: Owner or renter ~ social occupation indicator + age + sex + ethnicity + (sex*ethnicity)
Model 9: Owner or renter ~ social occupation indicator + age + sex + ethnicity + (social occupation indicator *ethnicity)
Model 10: Owner or renter ~ social occupation indicator + age (Irish only sample)
Model 11: Owner or renter ~ social occupation indicator + age (non-Irish only sample)
Model 12: Owner or renter ~ social occupation indicator + age (Irish male only sample)
Model 13: Owner or renter ~ social occupation indicator + age (non-Irish female only sample)

5. Concluding remarks

This purpose of this paper was to present reproducible methods using a reproducible dataset rich in sensitive attribute information that can help identify both problematic bias and problematic proxies. Our demonstration centred on whether sex or ethnicity or both and their interaction significantly contributed to the prediction of our chosen target controlling for social indicator of occupation and age. We developed our bias monitoring framework around this demonstration. Our focus was global as we were interested in monitoring the overall model performance and the contribution of each feature rather than locally picking out individual row instances.[6] We deployed explainable AI functions (ELI5, RFECV and SHAP) on both clear/transparent (logistic regression) and opaque/black-box (random forest classifier) models in order to assess the level of inferential agreement on the underlying dataset despite the algorithms having different predictive capabilities.[7] We conclude our demonstration by highlighting a concern from a bias monitoring perspective, i.e. when sensitive attribute features (and their interactions) are shown to be statistically important despite controlling for key associated features in relation to targets of interest, we argue that there exists a non-ignorable situation that requires further analysis and rectification before being deemed legally compliant and certified unbiased (from the perspective of the forthcoming EU AIA).

One key limitation of this paper is that we did not explore newer, more complex black-box models whereby the accuracy achieved may have been far superior to the least complex binary classification algorithm, i.e. logistic regression, thereby diminishing the influence of the latter algorithm's output on this paper's findings. However, our results were consistent across models and suitably strong enough for us to suggest that more complex models would not negate the concerning problematic bias around the ethnicity feature thanks to the

[5] It is not possible for us to present our results in this short paper due to space constraints. Full results will be presented at EWAF'23.
[6] Such local explainability was beyond the scope of our paper.
[7] We were able to make this comparison as model accuracy scores proved to be very similar.

underlying dataset we used. A second key limitation was our narrow use of our chosen dataset. The paper focused on specific sensitive attribute features (i.e. sex and ethnicity) and a particular age range (30-64). This may limit the generalisability of the proposed framework to other sensitive attributes and/or age ranges. This study was also performed on one dataset which limits the generalisability of insights developed. We consider this paper to be an introduction to our proposed framework and we plan to address the two key aforementioned limitations in a forthcoming research paper. Our future work suggestions are in line with Bhaumik and Dey (2022)[21] who call for extending the framework to other classification/regression models as well as launching pilots in industry. We agree with the authors who say industry pilots would "provide insights, which are essential to extend the current scope of the audit framework beyond technical aspects to include organizational and process related aspects."

In conclusion, we anticipate our proposed framework can work with any model (as it is designed to be model agnostic) in conjunction with an accompanying structured, tabular dataset. The framework is designed to be as fair as possible to practitioners whilst also providing robust bias detection that citizens can have confidence in. The emphasis is on raising awareness of any potentially concerning, problematic biases and/or proxies. Failure to get that right and proposed solutions become meaningless for practitioners and the general public.

6. References

[1] Wisniewski, J., & Biecek, P.: fairmodels: a Flexible Tool for Bias Detection, Visualization, and Mitigation in Binary Classification Models. R Journal, 14(1) (2022).

[2] https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb, last accessed 2023/04/14.

[3] https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/, last accessed 2023/04/14.

[4] https://towardsdatascience.com/bias-detection-in-machine-learning-models-using-amazon-sagemaker-clarify-d96482692611, last accessed 2023/04/14.

[5] Caton, S., & Haas, C.: Fairness in machine learning: A survey. arXiv preprint arXiv:2010.04053 (2020).

[6] ISO/IEC 22989: Information technology — Artificial intelligence — Artificial intelligence concepts and terminology (2022).

[7] ISO/IEC 23053: Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML) (2022).

[8] ISO/IEC 23894: Information technology — Artificial intelligence — Guidance on risk management (2023).

[9] ISO/IEC TR 24027: Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making (2021).

[10] ISO/IEC CD TS 12791, Information technology — Artificial intelligence — Treatment of unwanted bias in classification and regression machine learning tasks (working draft last accessed 2023/04/14).

[11] ISO/IEC/IEEE 29119-1: Software and systems engineering — Software testing — Part 1: Concepts and definitions (2013).

[12] O'Neil, C.: Explainable Fairness: A Framework. ORCAA presentation (2022).

[13] Flores, A. W., Bechtel, K., & Lowenkamp, C. T.: False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. Fed. Probation, 80, 38 (2016).

[14] Skeem, J. L., & Lowenkamp, C. T.: Risk, race, and recidivism: Predictive bias and disparate impact. Criminology, 54(4), 680-712 (2016).

[15] https://eli5.readthedocs.io/en/latest/blackbox/permutation_importance.html, last accessed 2023/04/14.

[16] https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFECV.html, last accessed 2023/04/14.

[17] https://shap.readthedocs.io/en/latest/, last accessed 2023/04/14.

[18] Altman, D. G., & Bland, J. M.: How to obtain the P value from a confidence interval. Bmj, 343 (2011).

[19] Minnesota Population Center. Integrated Public Use Microdata Series, International: Version 7.3 [Ireland Census 2016, CSO]. Minneapolis, MN: IPUMS, 2020. https://doi.org/10.18128/D020.V7.3 (2020).

[20] Kang, J., Xie, T., Wu, X., Maciejewski, R., & Tong, H.: Infofair: Information-theoretic intersectional fairness. In 2022 IEEE International Conference on Big Data (Big Data) (pp. 1455-1464). IEEE (2022).

[21] Bhaumik, D., & Dey, D.: An Audit Framework for Technical Assessment of Binary Classifiers. arXiv preprint arXiv:2211.09500 (2022).