Multi-Perspective Explanation of Data Bias in AI: A Case Study

Tosin Adewumi
Machine Learning Group, Luleå University of Technology, Sweden
tosin.adewumi@ltu.se

Abstract

Explanations are largely lacking in some Machine Learning (ML) systems but having explanations is very helpful because they clarify events. In this case study, we use 7 bias metrics in the AI Fairness 360 (AIF360) library to explain bias in the German Credit Dataset (GCD) in a credit scoring scenario from multiple perspective. Some of the metrics applied are applicable in Natural Language Processing (NLP). Investigations reveal that bias exists in the dataset for the Sensitive Attribute (SA) *age*. As a contribution, we show that having multiple perspectives (through multiple metrics) of bias gives a clearer assessment compared to a single one. We highlight some of the mitigation algorithms that are available for handling the bias. We publicly release our source code[1]

1 Introduction

Explanability is the ability to use information (i.e. explanans) to clarify the causes or reasoning for an event (i.e. explanandum), contributing a better understanding of the event, as a result (Strevens, 2011; Burkart and Huber, 2021). It is a very useful concept in Machine Learning (ML) systems, just as it is in other human endeavors. However, it is lacking to a great extent in many deep learning systems (Heuillet et al., 2021; Dhar et al., 2023). In critical application domains such as healthcare, forensics, criminal justice, and credit scoring, among others, this becomes all the more important, especially with the observed challenge of bias (Hassani, 2021). One important thing to note about explanations is that they should comply with ethical standards or regulations, such as the EU's General Data Protection Regulation (GDPR), and for explanations to be useful, they should be easy to understand by the stakeholders.

Bias may be defined as systematic error arising from prejudices, which may be based on a certain Sensitive Attribute (SA), such as age (Antoniak and Mimno, 2021; Mehrabi et al., 2021; Alkhaled et al., 2023). It has strong relation to unfairness. A couple of examples of metrics for estimating bias are Mean Difference (MD) (Thissen et al., 1986) and Disparate Impact Ratio (DIR) (Feldman et al., 2015). These and more are discussed in the next section.

In this case study, we research what effect the sensitive attribute *age* has on data. We explain bias in the popular German Credit Dataset (GCD) (Hofmann, 1994) using the tool AI Fairness 360 (AIF360) (Bellamy et al., 2019). We provide multiple (7) perspectives with the following bias metrics: Consistency, DIR, MD, Number of Positives (NoP), Number of Negatives (NoN), Smoothed Empirical Differential Fairness (SEDF) (Foulds et al., 2020), and Base rate. **Our contribution** is that we show having multiple perspectives of bias gives a clearer and more insightful assessment compared to a single metric.

The rest of this paper is structured as follows. In Section 2, we discuss the method used in this work. We present and discuss the results in Section 3, including the explanations. In Section 4, we discuss some of the related work. We conclude this work in Sedction 5.

2 Method

We used the GCD and the AIF360 library. These are discussed briefly in the following subsections. We split the dataset in the ratio 7:3 for training and test sets. For SA, we use only age and distribute the age brackets according to Kamiran and Calders (2009), where the privileged group are people equal to or older than 25 and the unprivileged group are those younger than 25. For evaluation, we use only the training set, since this is what models are typically trained on and we use the seven bias metrics identified in the following subsection. They are based on the *BinaryLabelDatasetMetric* class and

[1] colab.research.google.com/drive/1ID5xji4Q9YwyKHjfoQ9jFXpLBHHsL8iv#scrollTo=ghzJIojSvAdk

we provide explanations through *MetricJSONExplainer*.

2.1 German Credit Dataset (GCD)

The dataset categorizes people by a set of attributes as bad or good credit risks. It has 1,000 samples and 20 features. The age distribution in the dataset for the privileged and unprivileged groups (for the cut-off age of 25) is given in Figure 1. It shows the privileged group has more than 800 people.

2.2 AI Fairness 360 (AIF360)

The tool is a library of bias mitigation algorithms and over 70 bias metrics.[2] It can be used to examine and mitigate bias and discrimination in ML models. Some of the state-of-the-art (SotA) mitigation algorithms include Fair Data Adaptation (FDA) (Plečko et al., 2021), Grid Search Reduction (GSR) (Agarwal et al., 2019), and Rich Subgroup Fairness (RSF) (Kearns et al., 2018), among others.

2.3 Metrics

The following bias metrics are used in this study.

1. Consistency. It compares the class of prediction of a given data point to its k-nearest neighbors (Zemel et al., 2013).

2. Disparate Impact Ratio (DIR). The ratio of predicted favorable outcomes for the group that is unprivileged to the privileged one (Feldman et al., 2015).

3. Mean Difference (MD). It compares the privileged and unprivileged groups by subtracting the former's percentage of favorable results from the latter (Thissen et al., 1986). It may also be called *Statistical Parity*.

4. Number of Positives. This computes the number of positives in the total instances of the data[2].

5. Number of Negatives. This also computes the number of negatives in the total instances of the data[2].

6. Smoothed Empirical Differential Fairness (SEDF). It is the estimation of the posterior predictive distribution of a Dirichlet-multinomial model, where Differential Fairness (DF) extends the 80% rule in the U.S. Code of Federal Regulations (Commission et al., 1970). It does so to protect multi-dimensional intersectional categories (Foulds et al., 2020).

7. Base rate. This is conceptualized from stereotypes (Cao and Banaji, 2016; Locksley et al., 1980, 1982). Neglect of it will result in predictions that deviate from what is statistically likely.

[2] ai-fairness-360.org

3 Results and Discussion

The results of the seven metrics show there's bias in the original dataset for the sensitive attribute *age* when the privileged and unprivileged groups are compared.

Consistency The Consistency result of 0.681143 is not equal to 1, the preferred and fair value, hence, this implies that similar inputs are not treated similarly. The following Listing shows the *JSON*-formatted explanation for the prediction.


```
{
  "metric": "Consistency",
  "message": "Consistency (Zemel, et
      al. 2013): [0.68114286]",
  "description": "Individual fairness
      metric from Zemel, Rich, et al.
      \"Learning fair
      representations.\", ICML 2013.
      Measures how similar the labels
      are for similar instances.",
  "ideal": "The ideal value of this
      metric is 1.0"
}
```


DIR The DIR = 0.8338. This means the privileged group has favourable outcomes. A value of 1 would have indicated no disparate impact while a value above 1 would have indicated the unprivileged group has favourable outcomes. The following Listing shows the *JSON*-formatted explanation for the prediction.


```
{
  "metric": "Disparate Impact",
  "message": "Disparate impact
      (probability of favorable outcome
      for unprivileged instances /
      probability of favorable outcome
      for privileged instances):
      0.8338481338481339",
  "numPositivePredictionsUnprivileged":
      66.0,
  "numUnprivileged": 111.0,
  "numPositivePredictionsPrivileged":
      420.0,
  "numPrivileged": 589.0,
```

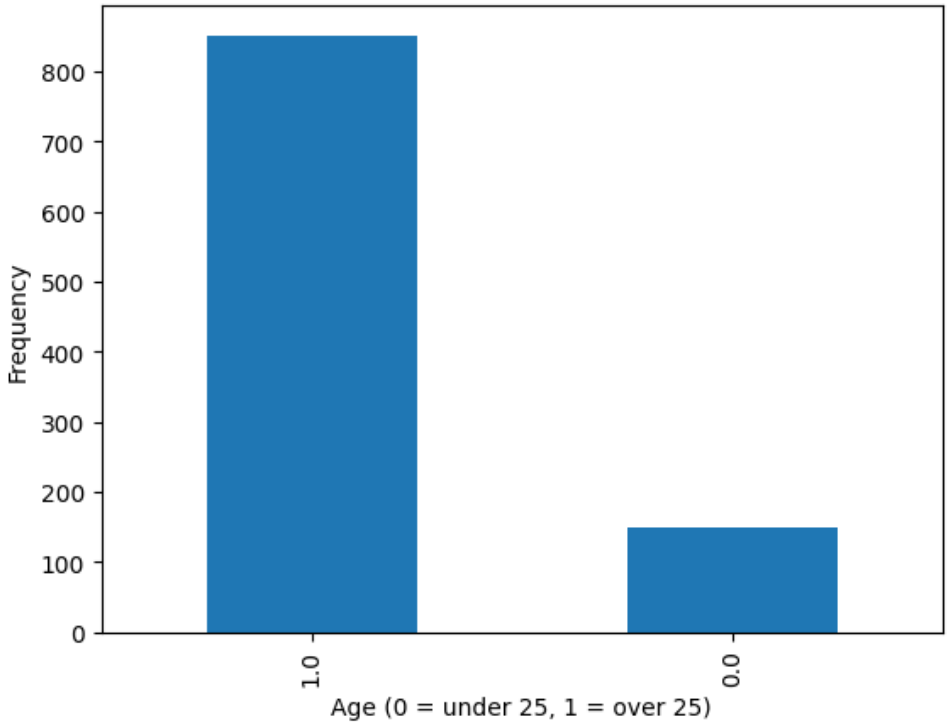

Figure 1: Age distribution in the GCD

```
  "description": "Computed as the ratio
      of rate of favorable outcome for
      the unprivileged group to that of
      the privileged group.",
  "ideal": "The ideal value of this
      metric is 1.0 A value < 1 implies
      higher benefit for the privileged
      group and a value >1 implies a
      higher benefit for the
      unprivileged group."
}
```

Mean Difference The MD result = -0.1185. This means the unprivileged group has less favourable outcomes compared to the privileged group. The score implies the privileged group was getting 11.85% more positive outcomes. A value of 0 would mean there's no difference in the outcomes for the groups. The Listing below shows the *JSON*-formatted explanation for the prediction.

```
{
  "metric": "Mean Difference",
  "message": "Mean difference (mean
      label value on unprivileged
      instances - mean label value on
      privileged instances):
      -0.11847841049878394",
  "numPositivesUnprivileged": 66.0,
  "numInstancesUnprivileged": 111.0,
  "numPositivesPrivileged": 420.0,
  "numInstancesPrivileged": 589.0,
  "description": "Computed as the
      difference of the rate of
      favorable outcomes received by
      the unprivileged group to the
      privileged group.",
  "ideal": "The ideal value of this
      metric is 0.0"
}
```

Number of Positives The Number of Positives identified in the data is 486. The Listing below shows the *JSON*-formatted explanation for the prediction.

```
{
  "metric": "Number Of Positives",
  "message": "Number of
      positive-outcome instances:
      486.0",
  "numPositives": 486.0,
  "description": "Computed as the
      number of positive instances for
      the given (privileged or
      unprivileged) group.",
  "ideal": "The ideal value of this
      metric lies in the total number
      of positive instances made
      available"
}
```

Number of Negatives The Number of Negatives identified in the data is 214. The Listing below shows the *JSON*-formatted explanation for the prediction.

```
{
  "metric": "Number Of Negatives",
```

```
  "message": "Number of
      negative-outcome instances:
      214.0",
  "numNegatives": 214.0,
  "description": "Computed as the
      number of negative instances for
      the given (privileged or
      unprivileged) group.",
  "ideal": "The ideal value of this
      metric lies in the total number
      of negative instances made
      available"
}
```


The other results for SEDF and Base rate are 0.346483 and 0.694286, respectively, which both indicate bias. We observe that combining the explanation of the various metrics, for which are shown, give us better appreciation of the bias in the data. This is a preferred first step in order to mitigate the bias in the data before utilizing it for training.

4 Related Work

The challenge of bias in data is one that has persisted for years and cuts across many fields of AI or research, including Natural Language Processing (NLP) (Blodgett et al., 2020; Adewumi et al., 2024). As a result, there have been efforts at detecting and mitigating bias in Machine Learning (ML) for many years, which has led to the introduction of metrics for quantifying bias in data or models (Locksley et al., 1980; Cao and Banaji, 2016). Such efforts include the work of Zemel et al. (2013), who introduced the Consistency metric, Feldman et al. (2015) with DIR, and (Thissen et al., 1986) with MD.

Although many metrics for quantifying bias and fairness exist, each one on its own may not be sufficient to paint the true picture of affairs, hence the reason why some researchers create new metrics that address the short coming of an older one. This is the case with the SEDF, which builds on the original Differential Fairness (Foulds et al., 2020).

5 Conclusion

In this work, we presented a case study of explaining bias using the AIF360 in the original GCD. Seven metrics were experimented with, some of which are applicable in NLP, such as DIR. We observed that bias exists in the data for the SA *age*. Being able to measure bias, as a first step, gives the opportunity to mitigate such bias. The AIF360 library provides multiple bias mitigation algorithms, such as FDA, GSR, and RSF. Future work may include exploring more bias metrics and the available mitigation strategies to ascertain which is most effective in a given scenario.

6 Limitations

Being a case study, this work is limited to the specific case of the GCD dataset. The library is also limited to the AIF360.

Acknowledgements

This work is supported by the European Commission-funded project "Humane AI: Toward AI Systems That Augment and Empower Humans by Understanding Us, Our Society and the World Around Us." The author wishes to thank *Thilakarathne Dilhan* and our colleagues at the *Responsible Artificial Intelligence Group at Umeå University* for their contributions.

References

Tosin Adewumi, Lama Alkhaled, Namrata Gurung, Goya van Boven, and Irene Pagliai. 2024. Fairness and bias in multimodal ai: A survey. *arXiv preprint arXiv:2406.19097*.

Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. 2019. Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning*, pages 120–129. PMLR.

Lama Alkhaled, Tosin Adewumi, and Sana Sabah Sabry. 2023. Bipol: A novel multi-axes bias evaluation metric with explainability for nlp. *Natural Language Processing Journal*, 4:100030.

Maria Antoniak and David Mimno. 2021. Bad seeds: Evaluating lexical methods for bias measurement. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1889–1904.

R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, S. Nagar, K. Natesan Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang. 2019. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4:1–4:15.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is

power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Nadia Burkart and Marco F Huber. 2021. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70:245–317.

Jack Cao and Mahzarin R Banaji. 2016. The base rate principle and the fairness principle in social judgment. *Proceedings of the National Academy of Sciences*, 113(27):7475–7480.

Equal Employment Opportunity Commission et al. 1970. Guidelines on employee selection procedures. *Federal Register*, 35(149):12333–12336.

Tribikram Dhar, Nilanjan Dey, Surekha Borra, and R Simon Sherratt. 2023. Challenges of deep learning in medical image analysis—improving explainability and trust. *IEEE Transactions on Technology and Society*, 4(1):68–75.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268.

James R Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. 2020. An intersectional definition of fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 1918–1921. IEEE.

Bertrand K Hassani. 2021. Societal bias reinforcement through machine learning: a credit scoring perspective. *AI and Ethics*, 1(3):239–247.

Alexandre Heuillet, Fabien Couthouis, and Natalia Díaz-Rodríguez. 2021. Explainability in deep reinforcement learning. *Knowledge-Based Systems*, 214:106685.

Hans Hofmann. 1994. Statlog (German Credit Data). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5NC77.

Faisal Kamiran and Toon Calders. 2009. Classifying without discriminating. In *2009 2nd international conference on computer, control and communication*, pages 1–6. IEEE.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International conference on machine learning*, pages 2564–2572. PMLR.

Anne Locksley, Eugene Borgida, Nancy Brekke, and Christine Hepburn. 1980. Sex stereotypes and social judgment. *Journal of Personality and Social psychology*, 39(5):821.

Anne Locksley, Christine Hepburn, and Vilma Ortiz. 1982. Social stereotypes and judgments of individuals: An instance of the base-rate fallacy. *Journal of experimental social psychology*, 18(1):23–42.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35.

Drago Plečko, Nicolas Bennett, and Nicolai Meinshausen. 2021. fairadapt: Causal reasoning for fair data pre-processing. *arXiv preprint arXiv:2110.10200*.

Michael Strevens. 2011. *Depth: An account of scientific explanation*. Harvard University Press.

David Thissen, Lynne Steinberg, and Meg Gerrard. 1986. Beyond group-mean differences: The concept of item bias. *Psychological bulletin*, 99(1):118.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International conference on machine learning*, pages 325–333. PMLR.