Bias in AI Scientific Report

1st Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

***Abstract*—**

***Index Terms*—**

I. Project proposal

What I'd like to do for my course project is to analyse the human-centric German-Credit dataset for assessing credit risk.

The motivation for this project is to develop a fair machine learning ecosystem to detect, reduce, and eventually mitigate different types of bias that exist in the final outcome of the algorithm in various ways.

The concrete tasks I plan to do are as follows:

1) Perform ”cleaning”, binning, bucketing, and discrete-to-continuous feature transformations on the dataset.
2) Split the dataset into different demographic groups, by writing down the size of the groups, the average value for each (numeric) feature, the variance of each (numeric) feature, the mode for each categorical feature, and the three most frequent values for each categorical feature, each computed on the different demographic subgroups.
3) Observe any interesting differences between the different subgroups' statistics, and describe any bias observed and explain the reasons why this bias has happened from my point of view.
4) Naively split your dataset into training and testing sets by randomly sampling some of the data, for example, 70% train and 30% test.
5) Train your model and see how it generalises to the testing dataset. Explain my approach and findings
6) Subsample a new testing dataset in an unbiased way and representative of the task, for example, you may wish to ensure gender and age diversity
7) Retrain your model and see how it generalises to these new testing conditions.
8) Compare findings with the results in 3 and explain my approach.
9) If you have observed any sort of bias, please describe it and explain the reasons why this bias has happened from your point of view

The final work product of these tasks will be …

The particular context I'm thinking about is …

The technologies I plan to use in my implementation are as follows: the Python programming language, packages.

II. Project progress report

A. Data analysis

These tables need fixing.

	Age	cell3
size	1000	cell
average	35.54	cell
variance	cell8	cell

	Sex
size	1000
mode	35.54
most frequent	cell8
2nd-most frequent	cell8
3rd-most frequent	cell8

B. Conventional implementation

At this stage of the project, I have a biased dataset, and implement a conventional ML algorithm.

My chosen algorithm is …

The justification for selecting this algorithm is …

My approach is to split the dataset into training and testing sets by randomly sampling some of the data, with 70% for training and 30% for testing, followed by training the model on the training dataset, and then seeing how it generalises to the testing dataset.

My findings were …

My next approach was to subsample a new testing dataset in an unbiased way, by ensuring gender and age diversity by doing …, followed by retraining the model on the training dataset, and then seeing how it generalises to the testing dataset with new conditions.

My new findings were …, and compared to the previous results, they are ….

I observed there to be bias in … and …. From my point of view, the reason why this bias has happened is because ….

C. Fair machine learning implementation

In this step, I implement one of the fair ML methods of mitigating bias.

The solutions that were provided in the suggested project paper [1] are two methods of modifying data, called Combinatorial and Geometric repair. We choose only one algorithm to implement due to time constraints brought about by other academic commitments.

The proposed algorithm, combinatorial repair, is …

We test the performance of our trained model for the minority groups. Compared with the performance of our model over the majority group, …

We observe a reduction in algorithmic bias, which is …. We plot the results demonstrating this below:

We do get roughly the same results as in the project paper. We justify the reasons by …

References

[1] Michael Feldman. Computational fairness: Preventing machine-learned discrimination. 2015.