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 [].

What I intend to implement is one of the two methods of modifying data that [1] proposes, called Combinatorial and Geometric repair.

The motivation for this project is to replicate the paper's claim that the repair performs favourably in terms of training classifiers that are both accurate and unbiased.

The reasons I believe that the implementation suits this submodule as bias in artificial intelligence is that it lets me develop a fair machine learning ecosystem to detect, reduce, and eventually mitigate different types of bias (such as the "80% rule" of disparate impact) 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, and compute and report information for each feature for each group.
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) Describe my chosen conventional algorithm to implement on the biased dataset, as well as describe the justification for selection.
5) Naively split the dataset into training and testing sets.
6) Train the model and see how it generalises to the testing dataset, as well as explain my approach and findings.
7) Subsample a new testing dataset in an unbiased way, then retrain the model and see how it generalises to these new testing conditions.
8) Compare my findings with the previous results and explain my approach.
9) Describe any sort of bias I observed, and explain the reasons why this bias has happened from my point of view.
10) Implement one of Combinatorial or Geometric repair, and describe the algorithm.
11) Test the performance of the trained model for the minority groups, and compare it with the performance of the model over the majority group.
12) Describe any reduction in algorithmic bias I observe.
13) State whether I get roughly the same results as the project paper, and, if not, reconsider my code or justify the reasons.
14) Describe from my point of view any sort of reduction in accuracy I observe.

The final work product of these tasks will be appropriate and proper plots and graphs that demonstrate any reduction in algorithmic bias and any sort of reduction in accuracy I observe after the fair machine learning implementation.

As I formulate this project, the particular context I'm thinking about is applying for credit loans.

The technologies I plan to use in my implementation are as follows: the Python programming language, the Pandas package for data analysis, and the Scikit-learn package for the implementation of the AI algorithm.

II. Project progress report

A. Data analysis

In terms of "cleaning", binning, bucketing, and discrete-to-continuous feature transformations I did on the dataset, I did only bucketing: I converted the column 'Age' with integer values to the column 'Age_Group' with four categorical values. The categorical value (age group) and range of integers (age range), inclusive, that they correspond to are as follows: **Young,** 19-29; **Young Adults,** 30-40; **Senior,** 41-55; **Elder,** 55+.

Then, we 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. The two different demographic groups are age group and sex, and their different demographic subgroups, respectively, are young, young adults, senior, or elder, and male or female. The different subgrouos' statistics are presented in Tables I, II, and III.

TABLE I
TABLE OF THE SIZE OF THE GROUPS.

	male	female	Young	Young Adults	Senior	Elder
size	690	310	371	355	203	71

TABLE II
TABLE OF THE AVERAGE VALUE AND VARIANCE OF EACH (NUMERIC) FEATURE FOR THE DIFFERENT DEMOGRAPHIC SUBGROUPS.

		Job	Credit amount	Duration
male	average	1.9	3448.0	21.6
	variance	0.4	8412806.3	154.7
female	average	1.8	2877.8	19.4
	variance	0.5	6776346.3	122.1
Young	average	1.8	3089.0	20.8
	variance	0.3	7261837.7	142.6
Young Adults	average	2.0	3375.5	21.5
	variance	0.4	7646336.1	139.2
Senior	average	1.9	3366.4	20.2
	variance	0.4	7986564.4	146.1
Elder	average	1.8	3430.4	20.5
	variance	0.7	13329819.2	192.5

The interesting differences between the different subgroups' statistics that I observe are as follows: there are more than twice as many males than females; the sizes of the 'Young' and 'Young Adults' subgroups are each larger than the 'Senior' and 'Elder' subgroups combined; and, 'Young Adults' are the only subgroup where the mode of the 'Checking account' feature is 'moderate', and not 'little'.

I have observed bias in both the gender and age demographic groups: a greater proportion of females than males were assigned a 'bad' credit risk; and, a greater proportion of 19-29 year-olds (the 'Young' subgroup) were assigned a 'bad' credit risk compared to other age groups. This data is presented in Table IV.

From my point of view, the reasons why this happened are because there are fewer females than males in the data, and the size of the 'Young' subgroup is greater than any of the other age groups, so the data is not representative.

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.

TABLE III

TABLE OF THE THREE MOST FREQUENT VALUES FOR EACH CATEGORICAL FEATURE FOR THE DIFFERENT DEMOGRAPHIC SUBGROUPS.

		Housing	Saving accounts	Checking account	Purpose	Risk
male	mode	own (517)	little (409)	little (186)	car (243)	good (499)
	2nd most frequent	free (89)	moderate (71)	moderate (183)	radio/TV (195)	bad (191)
	3rd most frequent	rent (84)	quite rich (47)	rich (43)	furniture/equipment (107)	-
female	mode	own (196)	little (194)	little (88)	car (94)	good(201)
	2nd most frequent	rent (95)	moderate (32)	moderate (86)	radio/TV (85)	bad (109)
	3rd most frequent	free (19)	rich (19)	rich (20)	furniture/equipment (74)	-
Young	mode	own (248)	little (242)	little (115)	radio/TV (117)	good(234)
	2nd most frequent	rent (113)	moderate (42)	moderate (112)	car (102)	bad (137)
	3rd most frequent	free (10)	quite rich (19)	rich (24)	furniture/equipment (84)	-
Young Adults	mode	own (278)	little (201)	moderate (100)	car (128)	good(264)
	2nd most frequent	free (39)	moderate (41)	little (81)	radio/TV (93)	bad (91)
	3rd most frequent	rent (38)	quite rich (24)	rich (18)	furniture/equipment (58)	-
Senior	mode	own (143)	little (117)	little (57)	car (79)	good (150)
	2nd most frequent	free (40)	moderate (16)	moderate (39)	radio/TV (51)	bad (53)
	3rd most frequent	rent (20)	quite rich (15)	rich (15)	furniture/equipment (36)	-
Elder	mode	own (44)	little (43)	little (21)	car (28)	good (52)
	2nd most frequent	free (19)	rich (5)	moderate (18)	radio/TV (19)	bad (19)
	3rd most frequent	rent (8)	quite rich (5)	rich (6)	business (9)	-

TABLE IV

TABLE OF THE PERCENTAGE OF EACH DEMOGRAPHIC SUBGROUP WHERE 'RISK' IS 'BAD'.

	male	female	Young	Young Adults	Senior	Elder
%bad	27.7	35.2	36.9	25.6	26.1	26.8