Bias in AI Scientific Report

I. Project proposal

In this course project, we will be analysing the human-centric German-Credit dataset for assessing credit risk, which is one of the suggested projects[1]. We intend to implement one of the two methods of modifying data, called Combinatorial repair and Geometric repair, proposed by [3]. The motivation for this project is to assess the paper's claim that the repair performs favourably in terms of training classifiers that are both accurate and unbiased.

We believe the intention suits this submodule as bias in artificial intelligence because we will be developing a fair machine learning ecosystem to detect, reduce, and eventually mitigate different types of bias that may exist in the final outcome of the algorithm in various ways. The concrete tasks we plan to do are as follows:

1) perform any 'cleaning', binning, or 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 demographic subgroup;
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 our point of view;
4) choose a conventional algorithm to implement on the biased dataset, and describe the algorithm and the justification for selection;
5) naively split the dataset into training and testing sets, then train the model and see how it generalises to the testing dataset, as well as explain our approach and findings;
6) subsample a new testing dataset in an unbiased way, then retrain the model and see how it generalises to these new testing conditions;
7) compare our findings with the previous results and explain our approach;
8) describe any sort of bias we observed, and explain the reasons why this bias has happened from our point of view;
9) implement one of Combinatorial repair or Geometric repair, and describe the algorithm;
10) test the performance of the trained model for the repaired datasets, compare it with the performance of the model over the unrepaired dataset, and describe any reduction in algorithmic bias we observe;
11) state whether we get roughly the same results as the project paper, and, if not, reconsider our code or justify the reasons; and,
12) describe from our point of view any sort of reduction in accuracy we 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 we observe after the fair machine learning implementation.

As we formulate this project, the particular context we are thinking about is applying for credit loans. The dataset might exhibit a greater proportion of good risk for applicants from specific demographic groups (such as age groups) over another. Biases such as this, termed *disparate impact*, should be avoided by classifiers, as it is unethical and prohibited by law [1].

The technologies we plan to use in our implementation are the Python programming language, the Pandas package for data analysis, and the Scikit-learn package for the implementation of the artificial intelligence (AI) algorithm.

II. Project progress report

A. Data analysis

Out of 'cleaning', binning, or discrete-to-continuous feature transformations we did on the dataset, we did 'cleaning' and binning. We did 'cleaning' by dropping the *Checking account* and *Saving accounts* features because they had instances with missing values. And, we did binning by replacing the *Age* feature, which has integer values, with the *Age_Group* feature, which has categorical values. The four categorical values of *Age_Group* and the range of integer values, inclusive, of *Age* 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 subgroups' statistics are presented in Tables I, II, and III.

The interesting differences between the different subgroups' statistics we observe are as follows: there is more than twice as many *male* than *female*; and, the sizes of the *Young* and *Young Adults* subgroups are each larger than the *Senior* and *Elder* subgroups combined.

[1] kaggle.com/janiobachmann/german-credit-analysis-a-risk-perspective

TABLE I
TABLE OF THE SIZE OF EACH DEMOGRAPHIC SUBGROUP.

	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 EACH DEMOGRAPHIC SUBGROUP.

		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

By using the demographic parity group fairness metric [4], we observe there to be bias in both the *Sex* and *Age_Group* demographic groups: a greater proportion of *female* than *male* was assigned (credit) *Risk* as *bad*; while, a greater proportion of *Young* was assigned (credit) *Risk* as *bad* compared to other age groups. The probability of *bad* predictions for each subgroup is presented in Table IV.

From our point of view, the reason why this bias happened is because there is lower representation of minority subgroups in the dataset, such as *female* or *Elder*, compared to majority subgroups, such as *male* or *Young*.

B. Conventional implementation

At this stage of the project, we have a biased dataset. We implement a conventional machine learning (ML) algorithm that is widely used, to solve the classification problem of predicting an instance as having *good* or *bad* credit risk.

The algorithm we choose to implement is the support-vector machine (SVM) [2]. An SVM is a discriminative classification algorithm which finds the curve (in two dimensions) or manifold (in multiple dimensions) that divides classes of data from each other with the maximum margin. In two dimensions, the margin is the width of the separating line or curve, which can be the perpendicular distance to the nearest data point when plotted on a graph. The justification for selecting this algorithm is that it is the same conventional ML model used by the suggested paper [3].

Our first approach was to do the following:

1) naively split the dataset into training and testing sets by randomly sampling some of the data, with $\frac{2}{3}$ for training and $\frac{1}{3}$ for testing;
2) perform feature scaling on numerical data by standardisation, so the values have zero-mean and unit-variance;
3) encode categorical features using a one-hot scheme, and target labels with values 1 for *good* and 0 for *bad*;
4) exhaustively consider all parameter combinations by grid search to find the best for the SVM with L_2 regularisation;
5) train the model on the training dataset and make sure the model is not over-fitting by cross-validation; and,
6) see how the model generalises to the testing dataset, by calculating an accuracy score.

TABLE III
TABLE OF THE THREE MOST FREQUENT VALUES FOR EACH CATEGORICAL FEATURE FOR EACH DEMOGRAPHIC SUBGROUP.

		Housing	Purpose	Risk
male	mode	own (517)	car (243)	good (499)
	2nd freq.	free (89)	radio/TV (195)	bad (191)
	3rd freq.	rent (84)	furniture/equipment (107)	-
female	mode	own (196)	car (94)	good(201)
	2nd freq.	rent (95)	radio/TV (85)	bad (109)
	3rd freq.	free (19)	furniture/equipment (74)	-
Young	mode	own (248)	radio/TV (117)	good(234)
	2nd freq.	rent (113)	car (102)	bad (137)
	3rd freq.	free (10)	furniture/equipment (84)	-
Young Adults	mode	own (278)	car (128)	good(264)
	2nd freq.	free (39)	radio/TV (93)	bad (91)
	3rd freq.	rent (38)	furniture/equipment (58)	-
Senior	mode	own (143)	car (79)	good (150)
	2nd freq.	free (40)	radio/TV (51)	bad (53)
	3rd freq.	rent (20)	furniture/equipment (36)	-
Elder	mode	own (44)	car (28)	good (52)
	2nd freq.	free (19)	radio/TV (19)	bad (19)
	3rd freq.	rent (8)	business (9)	-

TABLE IV
TABLE OF THE PROBABILITY OF 'RISK' ASSIGNED AS 'BAD' FOR EACH DEMOGRAPHIC SUBGROUP DURING DATA ANALYSIS.

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

Our findings from the first approach were as follows: the best parameter combination for the classifier was a regularisation parameter C of value 0.75 and a polynomial kernel function of degree 4; the mean accuracy on the training data was 80.2%; the mean accuracy after cross-validation was 72.4%; and, the accuracy on the testing data was 69.2%.

Our second approach was to subsample a new testing dataset in an unbiased way, followed by retraining the model on the training dataset as before, and then seeing how it generalises to the testing dataset with new conditions. We chose to ensure gender diversity by employing uniform sampling [5] based on the expected probabilities to meet demographic parity. By doing so, we made the number of *female* (deprived community) with *good* (positive class) labels (DP), the number of *female* with *bad* (negative class) labels (DN), the number of *male* (favoured community) with *good* labels (FP), and the number of *male* with *bad* labels (FN) all equal. We did this by drawing 109 samples uniformly from each of DP, DN, FP, and FN, respectively — 109 was the largest size we could use to get an equal number of samples. When splitting this subsampled dataset into training and testing sets, we applied stratified sampling to preserve the expected probabilities.

Our findings from the second approach were as follows: the best parameter combination for the classifier was a C of value 0.75 and a sigmoid kernel function; the mean accuracy on the training data was 74.8%; the mean accuracy after cross-validation was 62.1%; and, the accuracy on the testing data was 55.5%.

Compared to the previous results, these new results have worse accuracy scores at each stage. This suggests there was discrimination against the *female* demographic subgroup in the first approach. From our point of view, the reason why this bias has happened is because, in the first approach, the classifier used the *Sex* feature to help classify the target, *Risk*.

C. Fair machine learning implementation

In this step, we implement one of the fair ML methods of mitigating bias. We chose only one algorithm to implement: Geometric repair. The justification for this choice is that this method of modifying data seems easier to implement than Combinatorial repair, and we are limited on time due to other course commitments.

The Geometric repair algorithm is as follows:

1) Given a biased dataset D, with unprotected features Y and stratifying (protected) features S, make all stratified groups (all possible combinations of protected features), and let these be the values of S.
2) Store the size of each stratified group; if a group has size 0, ignore it.
3) Pick a number of quantiles, with the maximum number equal to the size of the smallest stratified group, so that there will be at least one entry per quantile.
4) For each Y feature with orderable values, over its unique values, for each stratified group, find the median value at the 1^{st} quantile.
5) For each Y feature, find the median value of the median values, preferring the smaller item in the case of even-length lists, and call this the target value for all values of the feature in the 1^{st} quantile of each stratified group.
6) For each Y feature, for each original value in the 1^{st} quantile, update the original value to the repair value,

$$rv = ((1 - \lambda) * original) + (\lambda * target).$$

7) Proceed similarly for each remaining quantile.

Using this algorithm, we generate 11 datasets, for $\lambda = 0.0, 0.1, 0.2, \ldots, 1.0$, where $\lambda = 0.0$ is no repair and $\lambda = 1.0$ is a full repair. The Y columns affected by the repair are *Job*, *Credit amount*, and *Duration*. Then, we retrain and test the model for each dataset, by the first approach. We plot results for the accuracy and Disparate Impact (DI) score in Figures 1 and 2, where

$$DI = \frac{Pr\,[C = + \mid X = x]}{Pr\,[C = + \mid X = \bar{x}]}$$

as defined in [3]. The lower the DI score is, the fairer the algorithm.

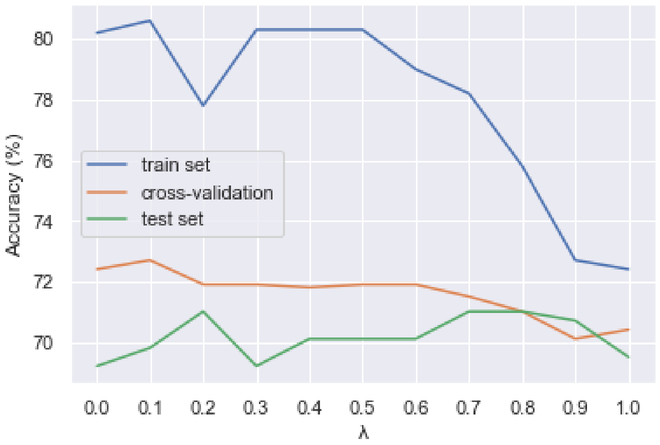

Fig. 1. Plots of accuracy against values of λ.

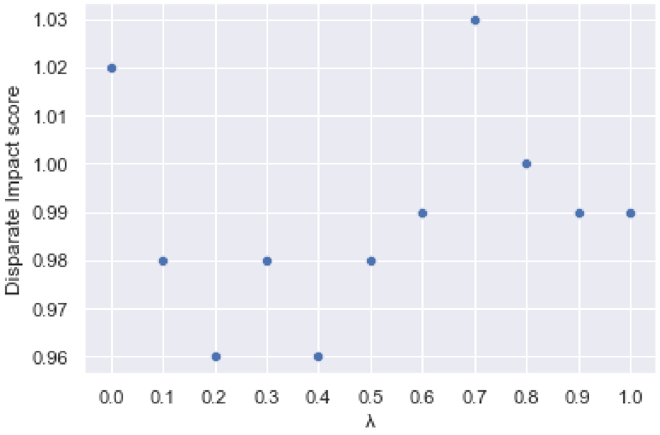

Fig. 2. Plot of Disparate Impact score against values of λ.

Comparing the performance of the model on the biased dataset ($\lambda = 0$) against the model on the repaired datasets ($\lambda > 0$), we see that accuracy generally decreases as λ increases. Also, the DI score is less for every repaired dataset except for $\lambda = 0.7$, which may be an outlier. We observe from these results that there is a reduction in algorithm bias in exchange for a reduction in accuracy, and therefore confirm the same observation and claim made by [3].

References

[1] Solon Barocas and Andrew D. Selbst. Big data's disparate impact. *California Law Review*, 104(3):671–732, 2016.

[2] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, September 1995.

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

[4] Pratik Gajane. On formalizing fairness in prediction with machine learning. *CoRR*, abs/1710.03184, 2017.

[5] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, December 2011.