Investigating Bias in AI models, and implementing ways to mitigate it

Section 1

Task 1.1

The sensitive attributes were determined to be ‘Gender’ and ‘BAMEyn’ determined by the laws specified by the UK equality act. The reason being that these features, are protected to prevent discrimination. Hence, this gives us a privileged population consisting of ‘Male’, ‘Non-Bame’ and an unprivileged population consisting of ‘BAME’ and ‘Female’. This is proven by the exploratory Data analysis performed and represented in Tables below.

Task 1.2

Figure 1 : Tables Outlining Outcome for Different groups as a % of total applicants

Gender	Shortlisted	Interviewed	Offer
Male	48.7%	34%	11.2%
Female	24.7%	13%	1.2%

BAMEyn	Shortlisted	Interviewed	Offer
BAME	15.7%	10%	0.8%
Non-BAME	43.4%	26%	7.1%

Figure 2 : Tables outlining the Number of Applicants

Sensitive Feature	Number of Applicants
Male	78
Female	202
Bame	159
Non-Bame	121

Fig 1 and 2 representing the percentages of applicants receiving interviews(Shortlisted), and to put that into perspective the number of applicants are represented in Fig 2. This clearly signifies that women are a lot less likely to get an offer in comparison to men. This is also strengthened by the fact that there were more women applicants, despite less women receiving offers than men.

Task 1.3

The method used to determine disparity is the 80 % test more commonly known as the disparate impact test. This test compares two groups: the proportions of unprivileged and privileged applicants who obtain a positive output.

Figure 3 : Interview Shortlist Disparity for Different Groups

	Privileged Group	Unprivileged Group	Privileged Group	Unprivileged Group
Sensitive Attribute	Male	Female	Non-Bame	Bame
Number of Applicants	78	202	159	121
Number Invited for Interview	38	50	69	19
Percentage of Interview	48.7%	24.8%	43.3%	15.7%
Ratio	0.51		0.36	

Figure 4 : Offer Percentage Disparity for Different Groups

	Privileged Group	Unprivileged Group	Privileged Group	Unprivileged Group
Sensitive Attribute	Male	Female	Non-Bame	Bame
Number of Applicants	78	202	159	121
Number of Offers	18	10	20	8
Percentage of Offers	23.1%	5%	12.6%	6.6%
Ratio	0.22		0.52	

From these results it is clear that all ratios are significantly below the 0.80 threshold value, as mentioned by the Disparate test. Hence, confirming our suspicion that there is a clear disparate violation. Also, from these results it is clear that for our Target Variable Offer, the Gender attribute plays a more significant role in disparity in comparison to the attribute BAME.

Task 1.4
It is determined that a chi-squared test is appropriate in our case because of two reasons Firstly being a "goodness of fit" statistic, it measures how well the observed distribution of data fits with the distribution that is expected if the variables are independent.Hence, determining if getting an Interview is dependent upon our sensitive variables, also because our variables are categorical.

Null hypothesis : Assuming that the distribution of data is due to chance
Alternative hypothesis : Null hypothesis is wrong
If p-value < 0.05 : Reject Null

Figure 5 : Table for Calculating Chi-Squared Values

	Shortlisted	Not-Shortlisted	Chi-Squared Value	P-Value	Reject or Accept Null Hypothesis
BAME	50	152	15.0	0.00011	Reject
Non-BAME	38	40			
Male	69	90	24.45	0.00001	Reject
Female	19	102			

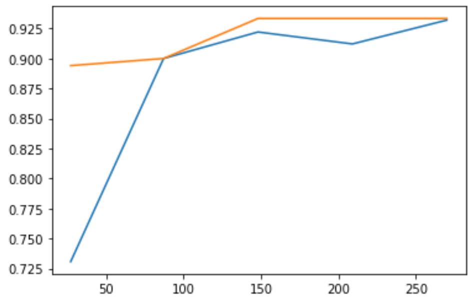
As the P-value is smaller than the Threshold of 0.05, it means that it is highly unlikely that the distribution of the data is due to chance, hence rejecting the Null hypothesis. This confirms that our predictions are dependent on our sensitive attributes ‘Gender’ and ‘BAME’.

This provides us statistical evidence that the dataset is biased.

Task 1.5
Code implementation under Markdown 1.5. In the code from scipy library chi2_contingency function is imported to calculate the p-value, degrees of freedom and the expected value. This p-value is then compared to alpha which is set as the threshold, and if the p-value is lower than alpha H0 is rejected. From the implementation in the code the p-values obtained are equivalent to calculation in Task 1.4, hence successfully implementing our proof in code.

Section 2
Task 2.1
Before training our Naive model, some data preprocessing is performed to ensure best possible results, for this firstly ‘Applicant Code’ was dropped from our dataset, as it is not useful. Also, ‘JoinNY’ and ‘AcceptNY’ columns are dropped, as those features can not predict our target variable ‘Offer’. Also, a candidate joining or accepting the job doesn’t affect the probability of receiving an offer. The null values were also filled with 0, these null values were present because some applicants did not get as far as others, hence have no values corresponding to those columns(eg having null in accept because you were not offered a job). Also, all values were standardly scaled, which is the same as all features being encoded to 0 and 1. This was done to ensure that attributes such as ‘BAMEyn’ with range 1 and 2 values is not weighed more than the attribute such as gender with range 0-1. The data was then split into 70 % training and 30 % testing data. In the train test split function the parameter straify was set

to True in order to ensure that in both the training and testing datasets, the proportion of applicants who received a job offer and those who did not are the same. Our Naive Model is determined to be Logistic Regression, due to the ability to mould the loss function with regularisation terms for removing bias[1], and also it is easier to implement in the library used which is Pytorch. For implementation, inbuilt pytorch functions such as backward and forward are used.The disparity index is initially 0.1 for my regular classification model, and the accuracy is 0.9034.The model's performance on the dataset is tested by the learning curve in Figure 6. From Fig 6 it is clear that the difference between testing and training accuracy is very minute. As a result of the minimal variance, no overfitting occurs. It is also obvious from the excellent accuracy that there was no underfitting.. For this reason the loss function has not been modified to add a L1 or L2 regularisation term. Finally, to mitigate bias Gender is chosen as the sensitive feature, this is because from our calculations above it is clear that Gender causes greater disparity than BAME, hence being more dominant.



Task 2 Results

From Fig 7 it is clear that the accuracy of the model is very high with 90%, and also the AUC score is very high. However looking at the f1 scores of both genders it is found that although the f1 score for Men is 0.95, f1 score for women is only 0.50. Hence, this Naive classifier is extremely biased towards Men having greater likelihood of getting the Offer. Furthermore, after investigating the predictions, it is clear that the model is clearly being influenced by gender. From this it is clear that Women are the underprivileged group.

Task 3 Implementation

For adversarial bias, the pytorch model was implemented by creating another instance of the Logistic Regression Class, the point of the adversarial model is that it takes in the predicted values from the Regular Naive Classifier, which in this case would be getting an offer or not. The adversarial then uses these values and tries to predict our sensitive attribute which is Gender. If the sensitive attribute is correctly predicted from our target variable, the regular classifier is penalised for this prediction. Hence what this does is it reduces the accuracy of the regular classifier, by maximising the loss function of adversarial. These models are pretrained, with the regular pretrained in Task 1 and adversary pretrained on predictions of the regular(biased) classifier. Finally to implement the complete model, the adversary and regular classifier are run together for epoch iterations, and based on the weights of the adversary at each iteration the gradient of the classifier is altered. This is done by the function in paper[1]. The overall model is based on two criterions, to measure their overall performance as a classifier. The first criterion being the Balanced Error Rate and accuracy, being the standard metrics. For the second criteria is the measure of the model's fairness using the Zernel Fairness defined as: $Zemel\ Fairness = prob(C = +|S = \overline{s}) - prob(C = +|S = s)$. The lower the Zernel Fairness score it signifies a lower level of discrimination in the model.

Task 4 Results

Training a new model on this training set yielded an accuracy of 86% and a loss of 0.114 for the regular classifier, with a 1.112 loss for the adversary classifier.This does signify a decrease in predictive performance of the model, but also the model fairness improved with a Zernel Fairness being now of 0.12 from 0.15, suggesting that there was only a 12% greater chance of Male getting an offer in comparison to women. Although this does not signify an equal opportunity for two demographics, it has still moved in the right direction. Although the decrease in accuracy is not desirable, the accuracy decrease is small and there is an increase in fairness. Finally, even though the notion of fairness is not satisfied, it can be concluded that with a bit more experience there is a possibility that with preprocessing techniques such as massaging the bias can be removed further

Task 5 Conclusion

From the results that are collected in this report, it can be shown that Adversarial debiasing is an effective method that limits discriminartion, whilst maintaining a high predictive accuracy.

Figure 7

Accuracy	0.903
AUC	0.84
Disparity	0.1
Weight	0.11

[1] https://dl.acm.org/doi/pdf/10.1145/3278721.3278779
[2] Rich Zemel, Toni Pitassi Yu Wu, Kevin Swersky, and Cynthia Dwork.
Learning Fair Representations. The International Machine Learning
Society, 2013.
[3]https://github.com/python-engineer/pytorchTutorial/blob/master/08_logistic_regression.py
[4]Geeks for Geeks
[5]