Task 1: **Theo Hinton-Hallows Bias In AI Coursework**

Q1) This dataset has 2 sensitive attributes. As defined by the 2010 equality act, Sex, and Race are both protected characteristics, so 'Gender', and 'BAMEyn' are sensitive attributes. Within the 'Gender' attribute, gender 1 (male) is the privileged group and gender 2 (female) is the unprivileged group. I found this by calculating the proportion of men who were given an offer (~0.231), and the proportion of women who were given an offer (~0.0495) – Men are more likely to be given offers than women. Within the 'BAMEyn' attribute, I found that group 2 (people who are not Black, Asian, or Minority Ethnic) is the privileged group and group 1 (people who are BAME) is the unprivileged group. I found this by calculating the proportion of non-BAME people who were given an offer (~0.126), and the proportion of BAME people who were given an offer (~0.0661) – People who are not BAME are more likely to be given offers.

Q2) (fig 1)

	Male	Female	BAME	Not BAME
Total	78	202	121	159
Total given offers	18	10	8	20
Total not given offers	60	192	113	139
Proportion given offers	0.231	0.0495	0.0661	0.126
Proportion not given offers	0.769	0.950	0.934	0.874
Total shortlisted for interview	38	50	19	69
Total not shortlisted for interview	40	152	102	90
Proportion shortlisted for interview	0.487	0.248	0.157	0.434
Proportion not shortlisted for interview	0.513	0.752	0.843	0.566

Q3) Disparity analysis:

Observed frequencies (and Expected frequencies: [row total * column total / overall total] in brackets) {and normalised square residuals: $([O-E]^2)/E$ in curly brackets}:

Gender/invited to interview (fig 2):

	Male	Female	Total
Shortlisted for interview	38 (24.51) {7.42}	50 (63.49) {2.87}	88
Not shortlisted for interview	40 (53.49) {3.40}	152 (138.51) {1.31}	192
Total	78	202	280

BAME/invited to interview (fig 3):

	BAME	Not BAME	Total
Shortlisted for interview	19 (38.03) {9.52}	69 (49.97) {7.25}	88
Not shortlisted for interview	102 (82.97) {4.36}	90 (109.03) {3.32}	192
Total	121	159	280

Gender/Offered Job (fig 4):

	Male	Female	Total
Offered job	18 (7.8) {13.34}	10 (20.2) {5.15}	28
Not Offered job	60 (70.2) {1.48}	192 (181.8) {0.57}	252
Total	78	202	280

BAME/Offered Job (fig 5):

	BAME	Not BAME	Total
Offered job	8 (12.1) {1.39}	20 (15.9) {1.06}	28
Not Offered job	113 (108.9) {0.15}	139 (143.1) {0.12}	252
Total	121	159	280

Q4) **Gender Bias test:** (read data from fig 2)

H_0: Gender and getting shortlisted for an interview are independent – Women are just as likely to get shortlisted as men

H_1: The likelihood of getting shortlisted is not the same for men and women

5% significance level

$$X^2 := \sum(O_i - E_i)^2 \,/\, E_i$$

$$X^2 \sim \chi_1^2$$

(X^2 is defined as the sum of all the normalised square residuals given a simple random sample of an unbiased population (assuming H_0) and it is distributed by the Chi-Squared distribution with 1 degree of freedom)

Observed result:

$\hat{X}^2$= 7.42 + 2.87 + 3.40 + 1.31 = 15

p-value = $P\left(X^2 \geq \hat{X}^2 \middle| H_0\right) = 1.075 \times 10^{-4} < 0.05$

Therefore, there is sufficient evidence to reject H_0. Therefore, there is sufficient evidence to suggest that the dataset is biased towards men in the shortlisting process.

Racial Bias test: (read data from fig 3)

H_0: Race and getting shortlisted for an interview are independent – BAME people are just as likely to get shortlisted as non-BAME people

H_1: The likelihood of getting shortlisted is not the same for BAME people and Non-BAME people

5% significance level

$$X^2 := \sum(O_i - E_i)^2 \,/\, E_i$$

$$X^2 \sim \chi_1^2$$

Observed result:

$\hat{X}^2$= 9.52 + 7.25 + 4.36 + 3.32 = 24.45

p-value = $P\left(X^2 \geq \hat{X}^2 \middle| H_0\right) = 7.626 \times 10^{-7} < 0.05$

Therefore, there is sufficient evidence to reject H_0. Therefore, there is sufficient evidence to suggest that the dataset is biased towards non-BAME people in the shortlisting process.

Q5) Same process in q3 and q4 implemented in jupyter notebook. I used SciPy [1] to access the CDF of the chi-squared distribution in python. I also used Pandas [2] to extract the data from excel. Leads to same conclusions as above although numbers are more accurate due to fewer approximations.

Task 2:

Q1) For the "regular classification algorithm" I created a custom implementation of a basic Logistic Regression algorithm. I did this using Scikit-learns [3] tools for custom machine learning algorithms, so I could use some of their other functions such as train_test_split easily. I also used Numpy [4] for some data manipulation. I had to create a custom algorithm so I would be able to easily add in the adversarial algorithm later. My model is trained by implementing gradient descent on a set of weights that are inputs to the prediction function. I split the dataset into a 70% training set and a 30% testing set and trained the model using the training data. I then evaluated the model on the test set.

Q2) The initial algorithm had an accuracy of ~91.6% on the test set. The confusion matrix is as follows:

	Offer actually received	Offer not actually received
Predicted offer received	4	2
Predicted offer not received	5	73

The model yielded the following scores: precision 0.66, negative predictive value 0.94, recall 0.44, selectivity 0.973, F1 score 0.53, and ROC AUC score 0.98. This model is heavily biased towards men, as the probability of a man being predicted an offer was 0.22, whereas that probability for women was 0 (these figures are likely slightly inaccurate due to the small sample size of the test set). Similarly, it is biased towards non-BAME people as the probability of a BAME person being predicted an offer was 0.049 whereas the same probability for a non-BAME person was 0.093.

Q3) I Implemented an adversarial Logistic Regression algorithm for each of the sensitive attributes in the dataset. I based my implementation on the method of bias mitigation set out by Zhang et al. [5] in their report. I chose to follow the equality of opportunity [6] notion of fairness which equalises the probability of predicting an offer given an offer is actually received and the value of the sensitive attribute. I chose this as it makes sure new offer predictions aren't just chosen randomly to satisfy the fairness check.

Q4) The algorithm with the adversarial component had an accuracy of 77.4% on the test set. The confusion matrix is as follows:

	Offer actually received	Offer not actually received
Predicted offer received	9	19
Predicted offer not received	0	56

The model yielded the following scores: precision 0.32, negative predictive value 1, recall 1, selectivity 0.746, F1 score 0.49, and ROC AUC score 0.93. The scores are likely lower than the original model as by removing the biases the only indicators the model has on whether to give an offer is whether or not the person was interviewed. As a result, the model has learnt to give an offer to everyone who was interviewed and not give an offer to anyone who was not interviewed. The odds for all sensitive groups of predicting an offer given an offer was actually received was 1 so equalised opportunity was achieved.

Q5) Overall the adversarial algorithm was successful in mitigating bias, as it achieved equality of opportunity. The odds for all sensitive groups of predicting an offer given an offer was actually received was 1. Whereas the original algorithm yielded 0.571 for men, 0 for women, 1 for BAME people, and 0.375 for non-BAME people – certainly not equalised opportunity. However, in the process of doing this, the algorithm was made much less accurate, as it simply learned to give offers to anyone who was interviewed, and not give offers to anyone who wasn't interviewed – it did this as beyond the sensitive attributes there were not many useful indicators of whether to give an offer left in the dataset. As a result, the model is no longer very useful.

Bibliography

[1] P. Virtanen et al., "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods,* vol. 17, pp. 261-272, 2020.

[2] The pandas development team, *pandas-dev/pandas: Pandas,* Zenodo, 2020.

[3] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research,* vol. 12, pp. 2825 - 2830, 2011.

[4] C. R. Harris, "Array programming with NumPy," *Nature,* vol. 585, no. 7825, pp. 357-362, 2020.

[5] B. H. Zhang, B. Lemoine and M. Mitchell, "Mitigating Unwanted Biases with Adversarial Learning," *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society,* pp. 335-340, 2018.

[6] M. Hardt, E. Price and N. Srebro, *Equality of Opportunity in Supervised Learning,* arXiv, 2016.