CSC4009: FIP-ML – Assignment 1

Name: Jamie McKeown
Student Number: 40177467
Email: jmckeown35@qub.ac.uk

I have chosen to analyse bias within a *Classification* task using a *Gradient Boosting Machine* (GBM) model. Before training the model I inspected the full dataset. 76.07% of the participants made under 50K and 23.93% made over 50K. 3620 out of 48842 rows (7.41%) contained missing data. To handle the missing data I utilised a framework created by Jakobsen et al [1], specifically when to use data imputation. Complete case analysis (removing rows with missing data) is useful in settings with relatively insignificant amounts of missing data, however, in this case it is above the recommended 5% threshold. As I was unable to access the missing data, it was impossible to verify that there was no relationship between the 'missingness' of the data and any values (observed or missing) [2]. Consequently, the missing data could not be classified as MCAR (Missing Completely at Random). Therefore using complete case analysis in this situation would risk introducing bias results based on incomplete data.

The columns with missing values were occupation, working class and native country. When classifying the missing data as MAR (Missing at Random) or MCAR, I assumed data was likely to have a systemic relationship to the observed data because the most common missing fields were working class & occupation [2]. These fields directly relate to the measured attribute, whether a person makes above 50k. Unfortunately, in this scenario, it is not possible to prove whether or not the values are MNAR (Missing Not at Random) as we cannot determine the propensity of a value to be missing based off its value [2]. For example, the relationship between the working class of a person and the likelihood of them omitting it in the census. Thus I have made the assumption that the values missing are MAR, this will be considered when evaluating the model.

I used mode imputation for the missing values. While there is a greater risk of introducing bias, it is much less computationally expensive than imputation methods using KNN, MICE or Deep Learning [3]. To encode the categorical attribute, I performed label encoding for the measured attribute and one-hot encoding on the remaining categorical attributes to avoid bias being introduced in categories where there is no ordinal order. One exception to this was education, where I implemented ordinal encoding based off the level of education. For the remaining continuous attributes, I applied min max scaling. As the dataset is not large, I used original train/test split which had a 2:1 ratio.

My chosen protected attributes are *Sex* and *Race*. I decided to remove all the sensitive attributes during training, this aides with fairness through the concept of unawareness [5] and reduces the risk of proxy features as I theorise that there is a correlation between native-country and race as well as working class and occupation.

I used python based toolkits including NumPy, Pandas and Scikit-learn for handling the data as well as developing and testing the GBM model. I opted to use grid search to find the initial optimal hyperparameters that produced a GBM with the highest accuracy. This search produced a GBM with an accuracy of 82.5% using the following hyperparameters:

- learning_rate: 0.01
- max_depth: 7
- n_estimators: 500
- subsample: 0.7

When analysing the cause of unfairness, I used the Optuna library to aide analysis of the effects of hyperparameter settings on fairness. Particularly its ability to optimise hyperparameters based on an objective function, in which I used a TRE sampler to select hyperparameter values between an upper and lower bound. The library allowed me to plot parallel co-ordinate plots in which many (50 in this case) hyperparameter settings could be concurrently shown against the objective function.

1. Evaluation of Group Fairness

Group fairness within classification can be evaluated through model performance, demographic parity and equality of opportunity. To quantify performance, accuracy and a confusion matrices were derived for each group. Demographic parity states each group within the protected attribute should obtain the positive outcome at equal rates. This an appropriate definition of fairness when we are aware of historical biases which may have affected data quality [6], which will be further considered under cause of unfairness. To measure demographic parity, Disparate Impact (DI) [6, 10] was calculated between privileged groups (Male & White) and unprivileged groups. Demographic parity is enforced by several US government agencies where adverse impact of a group is determined using the 4/5ths rule [7] which states that if the selection rate is less than 80% when compared to the privileged group, it will adversely affect the unprivileged group [8]. Equality of Opportunity was measured using Equality of Opportunity (EqO) and Average Odds (AvO) differences [10]. EqO difference is the difference in True Positive (TP) rates and AvO difference is a measure of whether groups are equal in their differences of both False Positive (FP) and TP rates. **Note:** DI, equality of opportunity difference and average odds difference for race were calculated against the privileged group (White).

Prot. Attribute	**Sex**		**Race**				
Group	**Male**	**Female**	**White**	**Black**	**A.I.E.**	**A.P.I.**	**Other**
Accuracy	78.93%	89.88%	81.67%	90.50%	86.58%	78.68%	86.89%
True Negative Rate	65.39%	85.35%	70.38%	85.83%	83.22%	65.69%	79.51%
False Negative Rate	17.43%	6.80%	14.68%	7.23%	9.40%	16.67%	12.30%
False Positive Rate	3.64%	3.32%	3.65%	2.27%	4.03%	4.66%	0.82%
True Positive Rate	13.54%	4.54%	11.29%	4.68%	3.36%	12.99%	7.38%
Disparate Impact (DI)	0.4574			0.4648	0.4941	1.1810	0.5486
EqO Difference	-9.00%			-6.61%	-7.93%	1.70%	3.91%
AvO Difference	-4.66%			-4.00%	-3.78%	1.35%	-3.37%

Interestingly, overall model accuracy for females (89.88%) is greater than males (78.93%). However, the high accuracy of females is mainly comprised of the 85.35% True Negative (TN) rate. In contrast, the TP rate is 3x greater for males which is reflected in EqO difference of 9% and AvO difference of 4.66%. The DI is 0.4574, less than 0.8 and thus not adhering to the 4/5ths rule which suggests women face a significant adverse impact around their salary compared to men. Despite the model displaying bias against men in its overall accuracy, the more worrying unfairness lies in the demographic parity and substantial lack of equality of opportunity for women. The model does not adhere well to any metric for fairness for either sex, notably favouring men with regards to the positive outcome. For race, a similar pattern as sex emerged for the privileged group (White) compared to the unprivileged groups, with the exception of Asian-Pac-Islanders (A.P.I.). The privileged group experienced worse accuracy but greater positive outcome rates shown in the values of EqO and AvO. Furthermore, a massive gap exists in the DI values from the ideal 0.8, suggesting a lack of fairness. Noting the exception of A.P.I., the DI satisfies the 4/5ths rule implying fair demographic parity holds between A.P.I. and whites, with model satisfying group fairness in this regard. They have a slight advantage with equality of outcome compared to whites, however, these are the smallest differences across the all race groups. It's worth highlighting that A.P.I. group makes up less than 1% of the dataset. Black people face the biggest demographic parity with a DI of 0.4648 and face unfairness across positive outcomes with an AvO difference of 4.66%. Amer-Indian-Eskimo (A.I.E.) face the largest discrepancy in equality opportunity, 7.93%. The model does adhere to group fairness between whites and A.P.I., however, the other minority groups face significant bias with blacks, A.I.E.s facing adverse demographic disparity. Comparatively, within both race and sex, the privileged groups have the highest False Negative (FN) and highest TP rates. While the model has bias in misclassifying privileged groups, it may indicate a relationship between the FN and TP rates. Not only do the privileged groups have a significantly higher TP rate, the highest FN rate suggests a double-edged bias. On one hand, the model misclassifies privileged groups negatively more often, however, this suggests the actual number of people making >50K is higher than and not as fair as the model suggests. The DI across privileged and unprivileged in groups is similar, though sex faces more demographic parity on average. This is further highlighted in the next section when considering sex within race while analysing the cause of the unfairness.

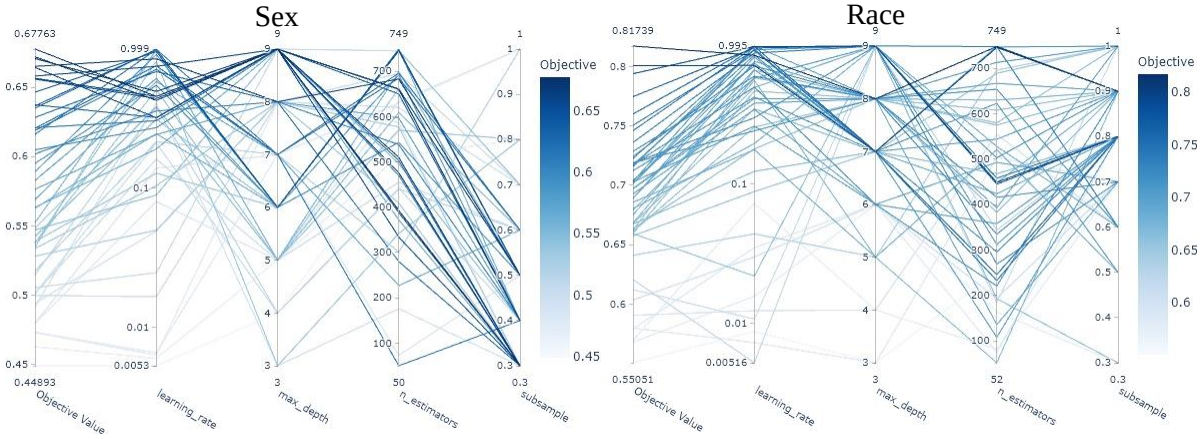
2. Cause of Unfairness

To understand the role of the dataset in causing unfairness, I measured the proportion of data made up by different groups within both protected attributes. Furthermore I analysed the distribution of those making over 50K within each group to see if there already exists unfairness within the data.

Prot. Attributes	**Sex**		**Race**									
Group	**M**	**F**	**White**		**Black**		**A.I.E.**		**A.P.I.**		**Other**	
Data Proportion	66.85%	33.15%	85.50%		9.59%		3.11%		0.96%		0.83%	
Proportion of Group >50K	30.38%	10.93%	25.40%		12.08%		11.70%		26.93%		12.32%	
			M	**F**	**M**	**F**	**M**	**F**	**M**	**F**	**M**	**F**
			31.55%	11.84%	18.26%	5.72%	14.04%	8.11%	33.93%	13.35%	15.54%	7.10%

One source of bias in the dataset is the unbalanced data and within sex and race there are clear disparities. As imbalanced data tends to be biased towards the majority class [9], this is likely a factor of the disparate impact favouring privileged groups. The tiny proportion of Asian-Pac-Islander people may have led to inaccurate modelling, hence why it appeared to be the privileged class. The table above shows the privileged groups earning >50K at higher rates than their counter parts. With 23.39% of all people making >50K, the fact that 30.38% of men compared to only 10.03% women make over >50K show a massive disparity across sex which undoubtedly causes unfairness in the model. This seem disparity is seen across whites and other races, though the disparity is greater for sex. A combination of the most or least privilege groups has a compound effect on unfairness. Disregarding the small, potentially inaccurate sample of A.P.I, white men are most likely to make >50K and women of colour are least likely. Black women have the smallest proportion (only 5.72%) being half as likely to make >50K compared to white women (11.84%). This could be a reflection of historical bias in US society, reflecting both racism and sexism. For example black people tend to live in lower socio-economic areas which strongly correlates to wealth [4] and there has been a well-documented gender pay gap.

To analyse the model's role on creating unfairness, I performed a study of the effects of different hyperparameters settings on disparate impact. I grouped the race groups into 'White' and 'Other'. The hyperparameter analysis aimed to maximise the P Percent score over 50 trials. The P Percent score is the minimum DI value between two groups and was used such that the objective value would not continue to increase and instead converge at 1, the ideal demographic parity.



The DI for Sex is maximised as learning rate and number of estimators increase and the subsample decreases. As they reach the upper bounds, DI appears to increase and thus the model is closer to fair demographic parity and group fairness. This is reflected in the race study, with the subsample being less restrictive to increasing fairness. Fascinatingly, the accuracy did not suffer large decreases, as the highest DI for Race & Sex produced training accuracies of 79.7% and 78.7%. However both TP and FP increased, even for the privileged group, suggesting a slight accuracy hit leads to more positive results at greater rate for unprivileged groups but still across the board. The DI for Sex improved from 0.4574 to 0.6776 and Race from 0.6116 to 0.8174, both significant improvements with the latter adhering to group fairness under demographic parity, suggesting that fairness can be greatly affected by hyperparameter choice within GBM models.

References

[1] J. Jakobsen, C. Gluud, J. Wetterslev and P. Winkel, "When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts", *BMC Medical Research Methodology*, 2017. Available: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-017-0442-1. [Accessed 10 February 2021].

[2] K. Grace-Martin, "How to Diagnose the Missing Data Mechanism - The Analysis Factor", *The Analysis Factor*, 2021. [Online]. Available: https://www.theanalysisfactor.com/missing-data-mechanism/. [Accessed: 10- Feb- 2021].

[3] W. Badr, "6 Different Ways to Compensate for Missing Data (Data Imputation with examples)", *Towards Data Science*, 2019. [Online]. Available: https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779. [Accessed: 11- Feb- 2021].

[4] D. Easton-Brooks and A. Davis, "Wealth, Traditional Socioeconomic Indicators, and the Achievement Debt", *The Journal of Negro Education*, vol. 76, no. 4, 2007. Available: https://www.jstor.org/stable/40037226?seq=1. [Accessed 15 February 2021].

[5] Z. Zhong, "A Tutorial on Fairness in Machine Learning", *Towards Data Science*, 2018. [Online]. Available: https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb. [Accessed: 11- Feb- 2021].

[6] V. Cortez, "How to define fairness to detect and prevent discriminatory outcomes in Machine Learning", *Towards Data Science*, 2019. [Online]. Available: https://towardsdatascience.com/how-to-define-fairness-to-detect-and-prevent-discriminatory-outcomes-in-machine-learning-ef23fd408ef2. [Accessed: 11- Feb- 2021].

[7] "Adverse Impact Analysis / Four-Fifths Rule", *Prevue HR*, 2009. [Online]. Available: https://www.prevuehr.com/resources/insights/adverse-impact-analysis-four-fifths-rule/. [Accessed: 13- Feb- 2021].

[8] N. Mondragon, "What is Adverse Impact? And Why Measuring It Matters", *HireVue*, 2018. [Online]. Available: https://www.hirevue.com/blog/hiring/what-is-adverse-impact-and-why-measuring-it-matters. [Accessed: 12- Feb- 2021].

[9] J. Brownlee, "A Gentle Introduction to Imbalanced Classification", *Machine Learning Mastery*, 2020. [Online]. Available: https://machinelearningmastery.com/what-is-imbalanced-classification. [Accessed: 13- Feb- 2021].

[10] A. Chidambaram, "Model Fairness & Transparency", *Medium*, 2020. [Online]. Available: https://medium.com/sfu-cspmp/model-transparency-fairness-552a747b444#:~:text=Fairness%20metrics%20considered%20are%20as,between%20unprivileged%20and%20privileged%20groups. [Accessed: 15- Feb- 2021].