CSC4009: FIP-ML – Assignment 1

**Name**: Jamie McKeown

**Student** **Number**: 40177467

**Email**: jmckeown35@qub.ac.uk

I have chosen to analyse the bias within a **Classification** task using a simple 3 layer neural network based off [1].

Firstly, I evaluated the amount of data points within the dataset that contained missing values and this totalled to 3620 rows (7.41% of the original total). This is a significant portion of the full dataset. Complete case analysis is useful in settings where insignificant amounts data, however this is above 5% which is above a threshold recommended by a 2017 paper on missing data within clinical trials [1]. If the data qualified as MCAR (Missing Completely at Random), I would have opted for complete case analysis as well. However, I could not verify if there was no relationship between the missingness of the data and any values observed or missing [2] thus I did not want risk biased results based on incomplete data. I came to the conclusion that between MAR (missing at random) vs MCAR, the data was more likely to have a systemic relationship to the observed data as the most common fields missing were workclass & occupation which directly relate to the measured attribute, whether a person makes above 50k. The next most common was country of origin which is a sensitive attribute, while there would a relationship to race. Unfortunately it is difficult to work whether or not the values are MNAR (Missing Not at Random) as we cannot determine the propensity of a value to missing based of its value, i.e. someone in lower education leaving an education field blank. Thus I have made the assumption that the values missing are MAR, this will be considered when evaluating the model.

At the risk of including bias within the dataset, I have decided to make use of Imputation using mode for missing values, while there is a risk of introducing bias, it is easier and computationally much less expensive than other methods such imputation using KNN, MICE or Deep Learning [3]. To encode the categorical values (not including sensitive data) I have performed label encoding for the classification task variable and one-hot encoding on the other variables as no order this ya boi wanted no correlation between them. As the dataset is not very large, I decided to keep the original training/test split of 2:1.

I also applied a minmax scalar to the continuous variables. I decided to remove all of the sensitive variables before training the model. As I believe including some would correlate to others that I may have chosen as my my main two and would limit a variety of choice in analysing bias. I.e if I chose marital-status this relates very closely to relationship. I chose to use the Gradient Boost Machine model. I then ran a grid search across several parameters to tune the model the highest accuracy I could get.

[1] <https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-017-0442-1>

[2] <https://www.theanalysisfactor.com/missing-data-mechanism/>

[3] <https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779>

<https://machinelearningmastery.com/imbalanced-classification-with-the-adult-income-dataset/>

<https://www.stxnext.com/blog/getting-started-machine-learning-python/>

<https://godatadriven.com/blog/towards-fairness-in-ml-with-adversarial-networks/>

<https://machinelearningmastery.com/evaluate-gradient-boosting-models-xgboost-python/>

<https://machinelearningmastery.com/one-hot-encoding-for-categorical-data/>

<https://www.analyticsvidhya.com/blog/2016/02/complete-guide-parameter-tuning-gradient-boosting-gbm-python/>

<https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74>

**1. Evaluation of Group Fairness**

To measure the GDM models adherence to group fairness I chose to use a few main metrics. Firstly, a confusion matrix was derived for each group across the two protected features, ‘Sex’ and ‘Race’, enabling analysis of model performance between groups. Demographic Parity between groups was calculated as it states each group within the protected class should receive the positive outcome at equal rates (noting the positive outcome in this case is ‘>50K’) and is an appropriate definition of fairness when we are aware of historical biases which may have affected the data [1] in this case sexism and racism in America [refer me daddy]. Equality of Opportunity/Equalised odds/Predictive rate parity… was used to measure

Eoin – should I include the raw numbas

<https://towardsdatascience.com/how-to-define-fairness-to-detect-and-prevent-discriminatory-outcomes-in-machine-learning-ef23fd408ef2>

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Sex** | | **Race** | | | | |
|  | **Male** | **Female** | **White** | **Black** | **A.I.E.** | **A.P.I.** | **Other** |
| Test 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% |
| Demographic Parity | 0.457 | |  | 0.4648 | 0.4941 | 0.2813 | 0.5486 |
| Eq. of Opportunity | 0.801 | | 0.7813 | 0.9182 | 0.5486 | 0.4954 |

**2. Cause of Unfairness**

Skews in data

See open article dickead

Visualise model/Black box

Assumptions within the model?

Hyperparameter differences?