Matthew Elliott
40153557
melliott21@qub.ac.uk

CSC4009
Analysing Bias in ML Tasks on the Census-Income Dataset

Task: Classification
ML model: Random Forrest
Analysis: group fairness and cause of unfairness

The python library Pandas was used to process the data. Sklearn was used to create a Random Forrest model. MS Excel was responsible for creating graphs Fig 1 – 4 and Tableau for graphs Fig 5 – 6.

Out of the multiple protected attributes available, sex and race were chosen. Income was classified into two categories: earning less than or equal 50K and over 50K USD. People earning over 50K USD were considered to be more privileged than their counterpart.

For sex, male was considered to be the privileged and female unprivileged. The race attribute has 5 categories of race. The white category was taken as the privileged group. Since the white race accounted for 86% of the data, all the other non-white races, which was majority black, were combined together to form a non-white unprivileged group. There was insufficient data to analyse on a per-race basis.

The data was pre-processed to remove any entries with null values as well as the irrelevant final weight feature. Income of greater than 50K was mapped to 1 and less than or equal to was mapped to 0 to allow averages to be created.

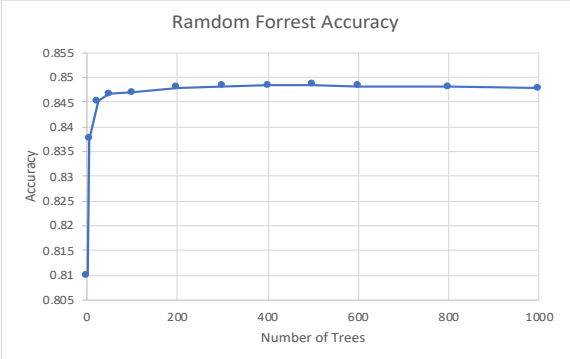

Fig 1, Accuracy of Random Forrest Model in predicting income with varying number of trees

The Accuracy of Fig 1 was assessed using 5-fold cross validation. The optimum number of trees to use was 200. Using more than 200 had no real accuracy gain while having a detrimental effect on computational costs.

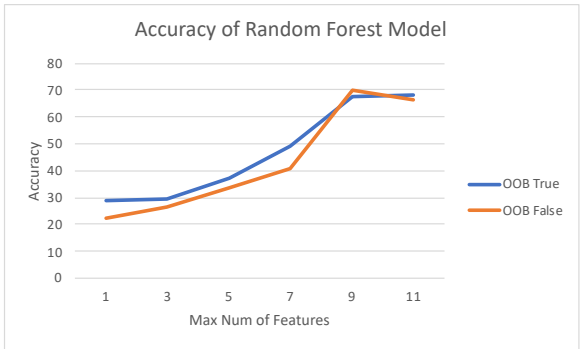

Fig 2, Assessing the performance of a Random Forrest Model under different hyperparameters.

The accuracy of the model remained at 0.845 ± 0.002 for all combinations of hyperparameters. The only difference observed was computational performance. OOB estimation was kept at the default value of false. Max features was kept to the default value of the square root of the number of features.

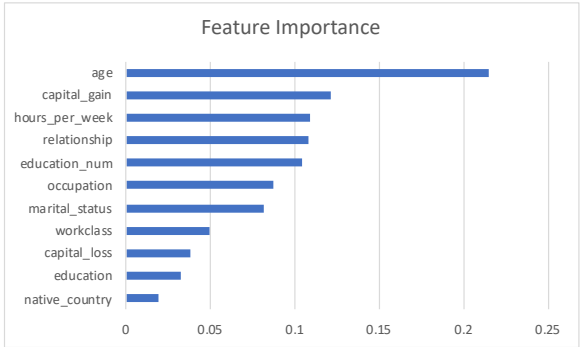

Fig 3, Feature importance used for predicting an individual's income with protected attributes

The most important features of the model used to predict income were observed and noted for help with model analysis in later sections. The data was obtained from the "*feature_importances_*" value.

Assessing group fairness

No features apart from race and sex were excluded from the model depending on which protected group was being analysed.

To measure unfairness of a particular feature, for each privileged - unprivileged pair, the number of predictions of individuals earning over 50K was compared to the actual number of individuals that earnt over 50K. This gives us a relative bias, regardless of sample size of the privileged and unprivileged groups as to how likely they are to be predicted to earn over 50K.

$$Bias\ Score = \frac{Num\ predicted\ to\ earn\ over\ 50K}{Actual\ num\ earning\ over\ 50K}$$

A score of above 1 would mean the model has bias in favour of predicting a group earning over 50K. Likewise, a score of below 1 would show the model was bias against predicting a group was earning more than 50K per year.

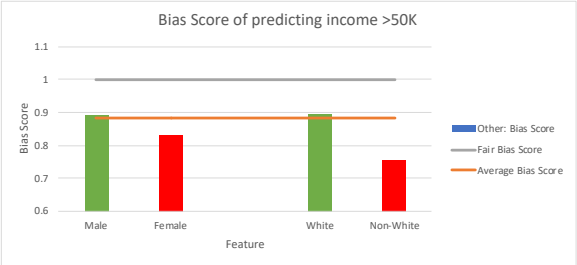

Fig 4, Chart to show exclusive bias of gender and race on the ML model's prediction of group income with green-red privileged-unprivileged groups.

From Fig 4, looking at the bias score, it is clear that there is a bias in terms of the model's prediction of group fairness. For both pairs of analysis, the privileged groups, male and white, had an above average bias score and far exceeded the unprivileged group. Interestingly, while there was roughly equal bias between the 2 privileged groups, male and white, there was a large difference between the 2 unprivileged groups, female and non-white. The non-white group had a larger bias against predicting an individual earnt above 50K than the other underprivileged group, female.

The model had more bias against the unprivileged race group, non-white, than the sex equivalent, Female.

It is important to note that the Average bias score (0.882) was well below the fair bias score of 1. This implies that even though the model had a greater bias of predicting the privileged groups to be in the over 50K income bracket compared to their unprivileged counterpart, all groups experienced an unfair bias, predicting a lower rate of earning above 50K that the test data used.

Cause of unfairness

As seen from Fig 3, the age attribute (importance value of 0.215) is significantly greater than any other attribute. Capital gain came a clear second at 0.122.

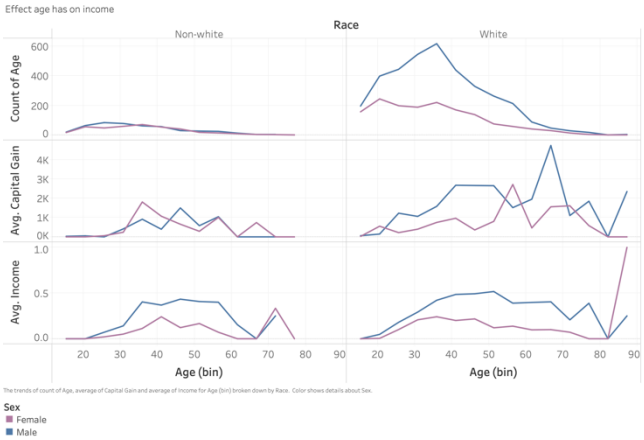

Fig 5 Charts showing how the distribution of sampled participants correlated to a bias of predicting a higher income for privileged groups

Fig 3 showed that age and capital gain were the two most important groups in predicting the income of a person. Since the age was by far the most important factor, the tiles are all in terms of age distribution so we can further analyse the consequences of age on predictions made by the model. Capital income, the second most important feature is also considered.

As you can see from Fig 4, the distribution of ages was unfair in terms of race with whites having an average age of 38.4 vs 37.5 for non-white. A larger difference is experienced when analysing the genders with the average age of a male being 38.9 and female being 37.0.

The second row highlights the average capital gain per age bracket for whites is roughly twice the respective non-white values. For the white race, we can see that males have almost twice the capital gain as females per age bracket. A pattern that is not present in the non-white group.

The third row shows average income for females is typically about half that of males of the same race. Between white and non-white races, the white race earns slightly more than their non-white counterpart respective of age and gender.

To assume that the model is wrong for predicting income on age is incorrect. However, the model fairness could be possibly be improved my reweighting age down to the same weight as the second most important feature (see Fig 3). This could improve the race and even more so, due to the larger difference, sex fairness between the privileged and unprivileged groups.

Could there be discrimination in the data itself?

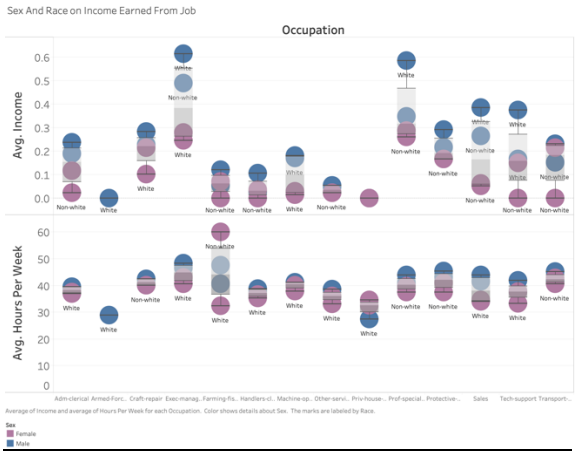

Fig 6, Grapy to show the average income of an individual when compared to their sex and race.

Fig 6 shows that there is still bias when looking at how much the different groups earn in terms role and average hours per week.

Looking at Fig 6, the largest divergence between protected features come between sex where the male, independent of race, will earn almost twice as much as the Female. Between races, the white race will earn more than the non-white races the majority of the time, regardless of sex. This shows there is clear discrimination in the data, with the unprivileged groups for race and sex earning less then their counterpart, almost without exception, for the same role.

The number of hours each week only greatly varies in terms of farming in terms of race. This suggests there is a bias towards the non-white and female groups.