Analysing Bias in ML Tasks on the Census-Income Dataset

CSC4009 Assignment 1

Student Name: Adam Coyle
Student Number: 40178464
Student Email: acoyle22@qub.ac.uk

1: Task Selection, Chosen Technique and Model Training

For this assignment, it was decided to focus on the classification task for the census dataset. This decision was influenced by both previous experience with classification models and tasks, as well as the readily available online resources for applying classification techniques to the selected dataset. Classification was to be completed by using the Random Forests due to its ability to handle categorical features, of which there are many in the dataset, and its performance using a sufficiently large dataset. Selection of protected attributes to investigate for bias was based on trending issues in regard to bias: race and gender in particular, have been the focal points of many reports and articles in areas such as workplace diversity and job applications to name a few. We consider the male and female subgroups for the genders, and for the race group we consider whites and non-whites, since the small sample sizes for minority races would make a fair comparison difficult. Therefore, it was deemed more appropriate to consider them all as one group. In evaluating bias in the dataset, we will perform an analysis of both group fairness and the cause of unfairness. These choices are of particular interest within the context of the source of the dataset i.e. the economic conditions in the United States in the early 1990s. Therefore, there is an expectation that there will be some disparity between different subgroups in the dataset, and it will be interesting to observe how the trained model takes this into account (if at all).

The programming side of the assignment was completed in a Jupyter Notebook, due to its versatility for experimentation and testing. The assignment code relies heavily on the scikit-learn libraries to provide much of the heavy lifting (splitting the dataset, tuning model hyperparameters etc.) as well as the pandas library for aggregating the data.

The training pipeline consisted of pre-processing the provided training and testing datasets; samples containing missing data were removed to prevent the model being trained on potentially false data, and because the number of samples with missing data was small enough that it should not drastically affect model performance. We also apply a one-hot encoding to the many categorical features of the dataset to provide the model with only numerical data.

We used the Grid Search with 5-fold cross-validation framework provided by scikit-learn to train and test the model on different hyperparameters such as the number of estimators, maximum features, and criterion for the random forest classifier. Altogether, we evaluated 192 different permutations of the models hyperparameters and selected the model with the best parameters:

Criterion	Max Features	Min Samples Leaf	No. Estimators	No. Jobs
Gini	sqrt	10	100	-1

Table 1: The best parameters obtained from tuning the model's parameters.

The model is trained on the above hyperparameters using the provided training dataset. The two protected attributes are removed from the dataset for training, and similarly for evaluating the performance over the test data. This is the alternative to removing the attributes whilst the dataset still contains categorical data, and then appending it to the encoded dataset. Model performance is evaluated by calculating common metrics (accuracy, precision, recall etc.) from a confusion matrix. The raw numbers from the matrix (true positives, negatives etc.) are also tabulated.

2: Evaluation of Group Fairness

In order to evaluate the extent to which group fairness is adhered to/violated by the model, we must investigate the evaluation metrics for each group belonging to the protected attributes. In particular, we will focus on the raw numbers obtained from the confusion matrix, to evaluate whether or not the model predicts a certain class for one group in the same proportions as its counterpart. We will also compare the probability of outcome for the different groups alongside those of the overall dataset as a means of determining if the model favours one group over the other. Finally, we evaluate the disparity of each group, calculated by the difference between probability outcomes for the group and the overall dataset.

Race	Accuracy	Precision	Recall	F Score	Specificity	Error Rate	TP	TN	FP	FN
White	0.85	0.78	0.56	0.66	0.95	0.15	1900	9077	525	1468
Non-White	0.90	0.76	0.53	0.63	0.97	0.10	177	1701	57	155

Table 2: Evaluation Metrics of the model per race. Non-white races have been coalesced to mitigate the large imbalance between whites or other races.

Gender	Accuracy	Precision	Recall	F Score	Specificity	Error Rate	TP	TN	FP	FN
Male	0.82	0.77	0.58	0.66	0.92	0.18	1822	6475	529	1321
Female	0.93	0.83	0.46	0.58	0.99	0.07	255	4303	53	302

Table 3: Evaluation Metrics of the model per gender.

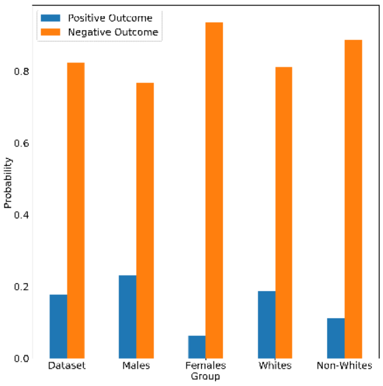

Figure 1: Probability of classification outcome for different groups.

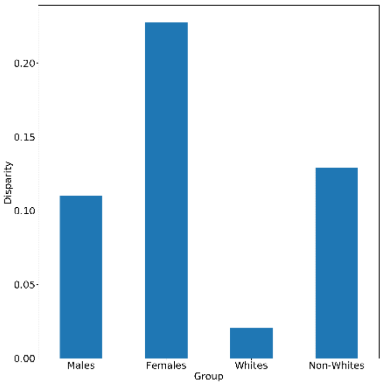

Figure 2: Disparity of the protected groups relative to the dataset's probability outcome.

Evidently, there are differences across all of the metrics: the accuracies for the male and white groups are lower than their counterparts by a reasonable margin, the recall of all groups is average, with a small difference between races, and a larger one between the genders. This metric is of particular interest when compared with the specificity, which remains quite high for all groups, and suggests that the model tends to favour a negative outcome in classification and infrequently predicts positive. The recall for the female group reveals that the model incorrectly labelled more than half of that group's positive samples, combined with the 12-point difference with males suggests a strong violation of group fairness for the genders, more so than the race groups, both of which have above 50% recall, and only a slight difference in their values. In contrast to the difference in recall rates, we observe the opposite trend wherein the male and white groups have lower specificity than their counterparts.

The comparable evaluation metrics for the race groups suggests that the model is much fairer for these groups but fails to achieve group fairness for the genders. For example, the ratio of false positives to false negatives for the race groups is about 1:3, whereas for the genders it is much more distorted. Males have a ratio of around 2:5 whereas females have a much larger value around 1:6. Consequently, this results in the female group having the best precision and specificity of all groups, as the model is not likely to make a positive prediction. These results can be confirmed by Figure 1. Figure 2 shows that the race group has a much lower cumulative disparity compared to the genders, where females have a considerably higher disparity than other groups. The relatively low disparity for whites could be explained by the fact that whites make up the majority of the dataset, so it is only natural that they would have similar outcomes.

3: Evaluation of the Cause of Unfairness

Unfairness in the model's outcome can be traced to two sources: the first is in the composition of the dataset and the balance of positive and negative samples. The second lies in the algorithm used by the model itself, and how it learns biases, either intentionally or unintentionally, from the training data. We will perform an exploratory analysis of the training data to identify any imbalances which might have promoted unfairness in the model's predictions. We will also take a close look at the learning algorithm for decision trees by visualizing one of the decision trees used in the ensemble, to determine whether or not the learning method is predisposed to unfair predictions.

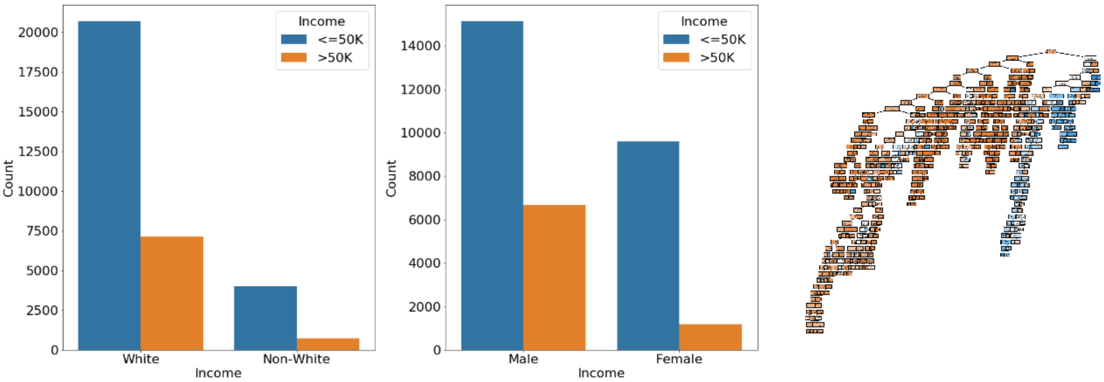

Figure 3: Left: Distribution of income classification for the protected groups. Right: a decision tree used in the random forest classifier.

The most striking observation from the above visualizations, is that a large majority of the samples are negative (i.e., most people make less than 50k), with only about 24% of the data being labelled positive across training and testing sets. However, the protected groups report different distributions, for example 30% of the male labels are positive, whilst females have only 10%. For the race group we observe whites with about 25% positive samples, which is comparable to the overall ratio, whilst non-whites have around 15% positive labels. Clearly there is greater skew in data for the gender group, since both subgroups show substantial deviations in terms of positive and negative samples in comparison to the overall dataset, whilst also having a larger imbalance than their racial counterpart (20% difference to 10%). Although the decision tree model is too small to interpret, visually it provides enough insight as to the classification process, that is, the model interprets many of the features as indicative of a negative sample. This could be attributed to the use of one-hot encodings for some features.

These distributions would explain why the model has a much higher rate of false negatives than false positives, as the data has a poor balance of positive and negative samples. It also might explain the noticeably poorer recall value for females than other groups, since that group has the worst skew in negative labels. What is interesting about the model decision tree, is that from a quick glance its ratio of positive to negative classifier conditions closely resembles the distribution of the training data that it uses. Therefore, it could be that the algorithm is biased due to the composition of the data, which could explain the higher-than-normal false negative rates. We can evaluate the effect of the training data on the model's performance by resampling the training data such that the label balance of the groups matches the overall trend in the dataset. For females, the bias becomes less noticeable while males remain relatively similar as expected, suggesting that the data is the primary cause of unfairness in the model.

Gender	Accuracy	Precision	Recall	F Score	Specificity	Error Rate	TP	TN	FP	FN
Male	0.81	0.79	0.54	0.64	0.94	0.19	1704	6561	443	1439
Female	0.91	0.61	0.65	0.63	0.95	0.09	364	4128	228	193

Table 4: Model performance on gender after resampling the training data to eliminate group deviations from overall labelling ratio.