Bias in Machine Learning for Income Prediction

WRJX35

***Abstract*—Binary classification is a machine learning problem aiming to correctly identify class labels of data instances. These algorithms all rely on the provision of a data set from which they can learn to what extent each feature of a data instance influences its class label. As a result, the performance of a model is directly dependent of the quality of the data supplied to the algorithm used to train that model and as such, any bias existing in the dataset will also influence the predictions made by the model.**

Keywords—Algorithmic Bias, Binary Classification, Machine Learning

I. Project Proposal

The aim for this project is to create a classifier capable of accurately identifying whether an individual earns above or below a specified threshold income whilst minimising bias towards a sensitive attribute.

This classifier will be created using the UCI Adult Dataset [1] collected from 1994 US census data. This dataset is an ideal candidate for creating such a model as it is human centric in nature and contains the sensitive features of *age, gender* and *race* and an appropriate *income* class whereby each instance is classified as earning greater than \$50k or less than/equal to \$50k. Whilst the initial assumption as to how to make a model non-discriminative with respect to the sensitive features would be to simply omit them (a solution known as fairness through unawareness), this does not deal with the influence that the sensitive feature may have on other attributes. For example, it is generally accepted that there is a gender pay-gap, whereby on average men are paid more than women. According to a report by payscale.com [2], women earn \$0.82 for every \$1 earned by men. The reason for this disparity is multifaceted and is influenced by factors such as the difference in the distribution of men and women in different employment sectors and not solely upon discrimination against women (although this does have an impact). Therefore, simply removing the gender column from the dataset will not remove the underlying bias within the dataset.

To tackle the bias, a more complex approach is necessary. In this project, a model called *Classification with No Discrimination (CND)* proposed by Kamiran and Calders [3] will be implemented. This model is used to minimise the potential bias within a dataset with respect to a sensitive feature by first implementing a naïve implementation of a machine learning algorithm. The resultant model is then used to predict the class for every instance in the in the training set. The training set is then split into two subsets, *candidates for promotion (CP)* and *candidates for demotion (CD). CP* are instances where a member of the unprivileged demographic group did not receive a positive classification† and *CD* are instances where a member of the unprivileged demographic group did receive a positive classification. The *CP* group are then ordered by decreasing probability and the *CD* group by increasing probability. The first *n* entries in each group then have their class label changed.

$$n = \frac{(S_s \times S_{\bar{s}+}) - (S_{\bar{s}} \times S_{s+})}{S_s + S_{\bar{s}}}$$

† In the case of this classifier, a positive classification would be one such that the income is defined as being greater than \$50k.

Where:

- S_s is the number of entries in the under-privileged class
- $S_{\bar{s}}$ is the number of entries in the privileged class
- S_{s+} is the number of entries in the under-privileged class receiving a positive prediction
- $S_{\bar{s}+}$ is the number of entries in the privileged class receiving a positive prediction

This results in a new dataset which, when used to train a model, results in a classifier which discriminates minimally with regards to the selected sensitive feature.

The overall models will be evaluated based on two distinct criteria. The first criterion will be a measure of their overall performance as a classifier, specifically their *accuracy* and *Balanced Error Rate (BER)*. These are standard metrics used when measuring the performance of a classification model on unseen data.

The second criterion will be a measure of the models' fairness. This will be ranked using *Zemel Fairness* [4] and *Disparate Impact* as defined in [5].

$$Zemel\ Fairness = prob(C = +|S = \bar{s}) - prob(C = +|S = s)$$

$$Disparate\ Impact = \frac{prob(C = +|S = s)}{prob(C = +|S = \bar{s})}$$

Where:

- C is the prediction returned by the classifier
- $+$ represents a prediction of having an *income* greater than \$50k
- $S = s$ represents the under-privileged class
- $S = \bar{s}$ represents the privileged class

Note that a lower Zemel score and a higher Disparate Impact score signify a lower level of discrimination in the model.

Throughout this project, all code will be written in *python*, using an *IPython Notebook* and making use of modules including *pandas* for data cleaning and preparation, *matplotlib* and *seaborn* for preliminary analysis and visualisation and *scikit-learn* for creating, training and analysing the performance of the models.

II. Data Cleaning & Preliminary Analysis

Before creating a model from the dataset, it first needs to be cleaned so that it can be understood by machine learning algorithms. The first step in this process was to abstract the *marital-status* feature such that the values *married-civ-spouse* and *married-AF-spouse* were grouped into one class called *married* and *married-spouse-absent* was joined to the *separated* class. This feature alongside the *occupation* and *relationship* features were then one-hot encoded due to their un-orderable nature. The *race* feature was abstracted into white and non-white values and the *native-country* feature into United-States and ex-pat. In both cases the values were encoded as 1s and 0s respectively. The *gender* feature was

mapped to 1 for male instances and 0 for female instances. This means that in the two discrete sensitive classes (*race* and *gender*), the privileged class was encoded as a 1 value. The class labels were also converted into binary values, with 0 representing instances with an *income* below or equal to $50k, and 1 representing instances with an *income* above $50k.

For the sake of investigating bias in this project, *gender* was selected as the sensitive attribute due to its high correlation of 0.214628 with the *income* class, suggesting that men are more likely to receive a positive prediction. On splitting the dataset by this feature, the first thing that becomes apparent is the under-representation of women. Of the 48,842 instances collected, 32,650 represent men whilst just 16,192 represent women, which is not reflective of the real gender distribution of the United States. This is an example of representation bias which could lead to the model performing poorly on the female demographic.

The two subsets can then be broken down and compared by calculating the *mean* and *variance* of each continuous feature, and the three most common values for each of the discrete data.

A. Continuous Data

Feature	***Mean***		***Variance***	
	Male	***Female***	***Male***	***Female***
age	39.5	36.9	179.9	199.9
fnlwgt	$1.92\text{x}10^5$	$1.85\text{x}10^5$	$1.14\text{x}10^{10}$	$1.07\text{x}10^{10}$
education-num	10.1	10.0	7.08	5.66
capital-gain	1326.2	580.7	$7.0\text{x}10^7$	$2.60\text{x}10^7$
capital-loss	100.4	61.5	$1.85\text{x}10^5$	$1.15\text{x}10^5$
hours-per-week	42.4	36.4	146.9	142.8

Among the continuous features, the most notable disparity between the two demographics were in the *capital-gain* and *capital-loss* features, with men having the larger *mean* value for both features by a considerable amount. The rest of the features seemed to be relatively similar in distribution.

B. Discrete Data

Male			
Feature	***1st***	***2nd***	***3rd***
workclass	Private	Self-Employed	Local-gov
education	HS-Grad	Some-college	Bachelors
marital-status	married-civ-spouse	Never Married	Divorced
occupation	Craft-repair	Exec-managerial	Prof-specialty
relationship	Husband	Not-in-family	Own-child
race	White	Black	Asian-Pac-Islander
native-country	United-States	Mexico	Philippines

Female			
Feature	***1st***	***2nd***	***3rd***
workclass	Private	Local-gov	State-gov
education	HS-Grad	Some-college	Bachelors
marital-status	Never-married	Divorced	Married-civ-spouse
occupation	Adm-clerical	Other-service	Prof-specialty
relationship	Not-in-family	Unmarried	Own-child
race	White	Black	Asian-Pac-Islander
native-country	United-States	Mexico	Philippines

Of the discrete features, there was a notable difference in the *workclass* feature, where there were significantly more men in the self-employed bracket and in the *occupation* class, women appeared to favour administrative and clerical roles as opposed to crafts/trades like men.

C. Class Labels

As a further part of the analysis, it was also necessary to check the distribution of the class labels within the dataset. It was discovered that 30.4% of men had an *income* greater than $50k as opposed to just 11.0% of women. This level of disparity was expected given the previously mentioned correlation between the *gender* feature and the *income* class label.

It is clear that the distribution of the data is not the same in both demographics, suggesting that there are some underlying patterns that may influence the predictions made by the model within these subgroups. Some of these patterns do explain for the disparity between the number of men and women earning over $50k, such as the significantly higher mean value of *capital-gain* in men and the higher number of men in executive/managerial roles, which usually pay higher salaries. Whilst their impact on the *income* class is logical, their relationship to the sensitive class will impact the models' performance on the under-privileged demographic.

III. Naïve Model

The first step in implementing a fair classifier is to implement a naïve model with no bias mitigating measures in place, in order to establish a baseline with which to compare the performance and fairness of the improved model.

In this specific case a Random Forest Classifier will be used. In classification, Random Forest uses an ensemble of Decision Trees in order to predict class labels by taking the majority vote of predictions made by the individual trees. This type of algorithm applies well to the project as it generally achieves very high accuracy and also provides a measure of the probability each instance belongs to each class, determined from the percentage of the ensemble voting each way, which is necessary in order to implement *Classification with No Discrimination* later in the project.

Before training the model, the encoded dataset was split into a training and testing set whereby the training set accounted for 70% of the data and the test set for 30%. This was done using *scikit-learn*'s *train_test_split* function which randomly assigns the instances from the dataset into each subset.

Training the model on the training set, provided initial *accuracy* of 85.7% and a *BER* of 0.225 (3d.p.) when tested using the test set. This suggests the model to have a strong predictive power. However, in measuring the model's

discrimination between the two demographic groups, it can be seen that the model has a *Zemel Fairness* score of 0.177, meaning a there is 17.7% increased chance that a male will be predicted to have an income greater than $50k than a female. The *Disparate Impact* score for this model was 0.313.

IV. Representative Model

As mentioned in SECTION II, there is a significant under-representation of women in the dataset. To get an idea of how the model would generalise to a dataset more representative of the true gender distribution in the united states, oversampling techniques were used to increase the size of the female demographic such that it accurately reflected the 49.1:50.9 male to female ratio of the United States population. The dataset was then again split into training and test sets of 70% and 30% respectively, this time stratifying by *gender* in order to ensure the gender distribution was maintained in both sets.

The performance of the model trained and tested on this representative dataset actually yielded an improved performance, with *accuracy* increasing to 90.4% and *BER* decreasing to 0.173 (3d.p.). The fairness of the model also improves, if only slightly, with the *Zemel Fairness* score reducing to 0.158 and the *Disparate Impact* score increasing to 0.384. This suggests that the oversampling and gender-based stratification on the dataset have helped to remove (at least partially) representation bias.

V. Fair Model

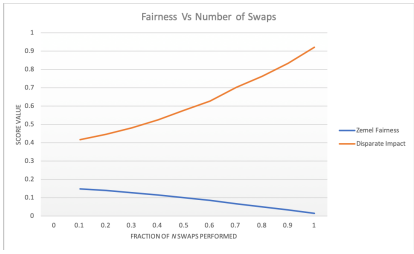
Whilst the removal of representation bias from the dataset is a step in the right direction with regards to creating a fair predictive model, the oversampling methods have not dealt with the bias caused by the underlying patterns within the dataset, i.e. the effect of the different distributions of the other features within the two demographic groups.

Kamiran and Calders data massaging procedure, as part of *Classification with No Discrimination,* was performed on the training set used to train the representative model, using the model to calculate the probabilities associated with each data instance in the set. The full *n* swaps were performed, ensuring that the resultant modified training set contained minimal discrimination. Note that the testing set remained unmodified such that the evaluation of the *Classification with No discrimination* approach will be performed on biased data, ensuring a fair comparison. Breaking down the now modified training set showed that the correlation between the gender feature had reduced to 0.000288, with the percentage of both men and women with a positive class label equal to 20.5% (3 s.f.).

Training a new model on this training set yielded an *accuracy* of 86.2% and a *BER* of 0.227. Whilst this signifies a decrease in the predictive performance of the model, the over fairness of the model of the improved significantly, with a new *Zemel Fairness* score of just 0.015, meaning that there is only a 1.5% increased chance of a male being predicted to have an *income* greater than $50k than a female. The *Disparate Impact* score also rose to 0.921, suggesting an equality of opportunity had almost been reached between the two demographics. Although the decrease in predictive performance is obviously undesirable, due to the fact that the decrease in *accuracy* is relatively small, the trade-off appears acceptable in order to have a significantly fairer model.

Similar results can be seen in a later paper [6] by the same authors, which show the significant decrease in discrimination *CND* results in on the same dataset as used in this project. Although they use different classification algorithms to predict the probabilities, the underlying principles remains the same.

The overall effect of the data massaging procedure can be visualised by repeating the procedure using different fractions of the calculated *n* value as the number of swaps to be made in the training set and plotting the effect that this has on the *Zemel Fairness* and *Disparate Impact* scores of the classifier trained from this set.



The same can be done with the *accuracy* and *BER* values to visualise the decrease in predictive performance resultant from the massaging procedure.

VI. Limitations

One issue with the *Classification with No Discrimination* approach is that it directly alters the class labels in the training set, such that the modified set is not truly representative of real-world instances. As such, a fair model created using this approach can expect to display a decrease in performance as observed in SECTION V. In a later paper [6] Kamiran and Calders address the 'intrusive' nature of this approach and propose two new solutions that do not alter the class labels of the data instances but apply weights or resample the data based on the predictions of a classifier trained on the data that does not take any bias mitigating measures (similar to the process in *CND*). Generally, these methods maintain a higher level of accuracy (due to their less intrusive nature) but not do not reduce discrimination to the same extent as *CND*. As such, these methods are still bound by the same accuracy-discrimination trade off as *CND*. All these methods of bias mitigation are viable options and should be chosen depending on whether accuracy or discrimination is more important in the given application. In the implementation in this project, the *CND* was chosen due to the main focus being on reduction of discrimination.

VII. Conclusion

From the results collected within this project, it can be shown that *Classification with No Discrimination* is an effective method for creating a machine learning model that limits discrimination within a sensitive attribute, whilst maintaining a high, albeit reduced, level of predictive performance.

REFERENCES

[1] Kohavi, R. and Becker, B. (1996). UCI Machine Learning Repository [https://archive.ics.uci.edu/ml/datasets/adult]. Irvine, CA: University of California, School of Information and Computer Science.

[2] PayScale. 2021. *Racial and Gender Pay Gap Statistics for 2021 | PayScale*. [online] Available at: <https://www.payscale.com/data/gender-pay-gap> [Accessed 14 April 2021].

[3] F. Kamiran and T. Calders, "Classifying without discriminating", *2009 2nd International Conference on Computer, Control and Communication*, 2009.

[4] Rich Zemel, Toni Pitassi Yu Wu, Kevin Swersky, and Cynthia Dwork. Learning Fair Representations. The International Machine Learning Society, 2013.

[5] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (*FAT '19*). Association for Computing Machinery, New York, NY, USA, 329–338. DOI:https://doi.org/10.1145/3287560.3287589

[6] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination", *Knowledge and Information Systems*, vol. 33, no. 1, pp. 1-33, 2011.