Student Number: 244800

Contents

1. Introduction

As the usage of artificial intelligence systems and applications in our daily lives grow, ensuring fairness in the design and engineering of such systems has become increasingly important. As AI system are used to make vital decisions in a variety of delicate contexts, it is critical to guarantee that these decisions do not reflect biased behaviour toward specific groups. In machine learning, data bias is a type of error in which some parts of a dataset are given more weight and/or representation than others [5]. Since the algorithm does not analyse all of the information in the data, it has a tendency to learn the wrong signals in a systematic manner. The relevant relationship between data inputs and targeted outputs may be missed by an algorithm due to model bias. Results of a biased dataset include a skewed conclusion, low accuracy levels, and analytical errors which do not effectively represent a model's use case. As machine learning models use training data to learn and do its job, it is important for this data to be representative of real world. Bias not only harms those who are discriminated against, but it also limits the potential of AI for business and society by instilling distrust and causing inaccurate outcomes.

In this report, we will be looking at and studying machine learning models by using and analyzing different methods with the goal of improving accuracy and fairness and finding the best model. We will look at and compare models on the criteria of most accurate and most fair.

2. Experiment Setup

We will be making use of the AI Fairness 360 toolkit, which is an extensible open-source library containing techniques developed by the research community to help detect and mitigate bias in machine learning models throughout the AI application lifecycle [1]. The aim of the toolkit is to encourage a better grasp of fairness metrics and mitigating strategies.

2.1. Dataset

We have chosen two public datasets, the Adult dataset and the German dataset. The Adult dataset, also referred to as "Census Income", presents a binary classification problem in which one tries to predict whether an individual receives a salary greater than 50K [7]. [1] carried out data processing, which resulted in 32,561 instances and 44 attributes, and is regarded as highly discriminative towards gender.

The German Credit dataset [3] contains information on whether or not a creditor granted a loan application access to one, as well as information about the applicant. The information includes relevant data about an applicant's credit history, savings, and work, as well as demographic information such as age, sex, and marital status. Creditors can use data like credit history, money, and employment to reliably forecast whether or not an application would return their debts; however, data like age and sex should not be utilised to determine whether or not an applicant should be issued a loan.

Table 1 contains information about the datasets that are used in the experiment. Each dataset has a single protected attribute that subdivides into privileged and unprivileged groups.

Various metrics have been defined for determining whether a trained machine learning model contains ethical bias. The fairness metric compares the rate at which a marginalised group receives a cer-

Dataset	Protected Attribute		Class Label	
	Privileged	**Unprivileged**	**Favorable**	**Unfavorable**
Adult [9]	Sex-Male	Sex-Female	High Income	Low Income
German Credit [3]	Sex-Male	Sex-Female	Good Credit	Bad Credit

Table 1: Dataset

tain outcome or result to the rate at which a privileged group receives the same outcome or result. Existing publications have proposed a number of fairness criteria, with the majority of them focusing on fairness classification. A set of commonly used fairness metrics are equal opportunity and equalized odds [4] which have been widely used to assess discrimination based on protected attributes. Our aim is to lessen the disparity in treatment each individual receives from the model. To measure such differences for group fairness, we will use the Equal Opportunity Difference metric [2]. The purpose of the group EOD is to determine if those who should qualify for an opportunity are equally likely to do so regardless of their membership in a particular group.

- **Equal Opportunity Difference (EOD):** Difference of True Positive Rates (TPR) for unprivileged and privileged groups

$$EOD = TPR_U - TPR_P$$

2.2. Methodology

A machine learning process consists of collecting and pre-processing data, selecting features, training the model and obtaining the metrics. This project focused on different models to detect and address the fairness issue in metrics step through 5-fold cross validation by varying the trade-off hyper-parameters to help analyze the effect of proposed solutions. We use several classification algorithms to collect and analyze their performances on the data. These algorithms are:

- **Logistic Regression:** Logistic Regression Classifier implemented for creating baseline results using Scikit-Learn library [8]. We used 'l2' regularization which helps to prevent over-fitting.

- **Support Vector Machine:** Support Vector Machine Classifier implemented for creating baseline results using Scikit-Learn library [8].

- **Reweighing:** We will also apply reweighing [6], a pre-processing bias mitigation algorithm which assigns weights to the training data as follows:

$$W(s,c) = \frac{|X(s)| \cdot |X(c)|}{|X(s,c)| \cdot |X|}$$

  where s is the value of sensitive attribute and c is the value of binary class label.

- **Adversarial Debiasing:** An in-processing strategy for learning a classifier to improve predicted accuracy while reducing an adversary's ability to deduce the protected feature from the predictions [10].

3. Experiments & Results

We start the experiment by defining where the bias is in the features of the dataset. As described in section 2.1, we will be focusing on the gender feature for both the Adult and German datasets. We split the dataset into train and test with 70/30% ratio to estimate the performance of machine learning algorithms. Standard scaler is applied to normalize the input dataset after the splitting process.

We use the classification algorithms described in section 2.2 and begin by determining each model's baseline performance, then choose the best performing model for further optimization via hyper-parameter tuning. The aim of this stage is to analyse whether or not better generalisation could correspond to fairer models. We utilise the scikit-learn

GridSearchCV() function, which loops through predefined hyperparameters and uses cross validation to evaluate the model for each combination. As a result of the process, we will be able to select the model with the best accuracy across all combinations of hyperparameters. Cross-validation is used to evaluate the algorithms' effectiveness. The cross-validation iterates through the folds, using one of the K folds as the validation set and the remaining folds as the training set at each iteration. This procedure is repeated until all of the folds have been used as validation sets. In our experiment, we made use of 5-fold cross validation, which can be seen in figure 1. The dataset is used to produce five data sets of similar size (folds). Each algorithm is evaluated based on its average performance after being trained on data from four folds and then tested on the fifth, a process that is repeated five times to ensure that each fold is tested exactly once. The accuracy results are saved, and the combination with the highest accuracy is selected as the best performing.

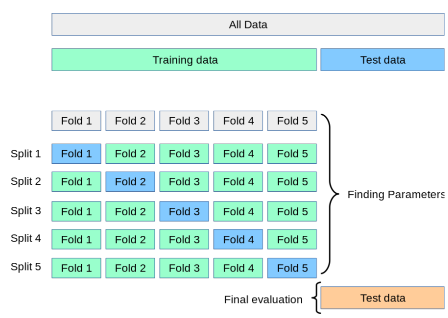

Figure 1: 5-fold Cross Validation

Second task is comparing the models using an algorithmic fairness method. Using the models selected in the previous task, we apply reweighing, as explained in section 2.2, to the data and carry out the same experiment as before by hyperparameter tuning utilizing 5-cross validation and analyzing the impact on accuracy and fairness metrics.

For the last task, we make use of *Race* attribute, a non-binary sensitive feature to analyze the algorithmic fairness for methods. The race attribute consists of White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black. To proceed with the analysis, we must first change these traits in such a way that they will allow us to do so. We start by defining where the bias is in the dataset, *privileged groups = 1* and *unprivileged groups = 0*. We also need to map thw protected attributes, *White = 1* and *Non-white = 0*. After defining the bias in the dataset, we continue by mapping each race to privileged and unprivileged groups, which consists of *White = 1.0* and *Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black = 0.0*. This process allows us to carry out fairness analysis on methods on the non-binary sensitive features.

We also conducted a random oversampling experiment, which involves increasing the number of minority class instances or samples by producing new cases or repeating some instances. We can conclude from the data that occurrences belonging to the underprivileged group make up a smaller percentage. As the fairness problem can be observed as under-representation of the discriminated group, we will perform oversampling based on the value of the sensitive attribute, which results in the dataset having the same number of instances for both the privileged and unprivileged groups.

3.1. Results

The fairness metric EOD, defined in section 2.1, is used to determine how biased the data is on the sensitive attributes during the study, whereas the accuracy is used to analyse the overall prediction performance. For a fair model, free of bias, difference metrics should be close to or be 0.

Table 2 shows results of the three tasks which have been carried out in the experiment. Comparing the baseline performance of Logistic Regression and Support Vector Machine with, we can see that they have a high accuracy rate of 80% and 78% respectively for the adult dataset. However, despite having high accuracy rates, the models are both extremely biased as the EOD metric for both of them are extremely low.

With the help of GridSearch, described in section 3, we were able to search through predefined parameteres and utilise cross validation to evaluate the model for each combination. Through several experiments using C to perform hyperparameter tuning, we were able to see different accuracy scores across different C values. For LR models, with a low C value, we were able to see an increase in EOD value which indicates a model with better fairness metric. However, the accuracy rate was lower than the base model, dropping to 76%. A lower C value indicates that the model is giving complexity more weight at the expense of fitting the data. As a result, a high hyperparameter value C implies that training data is more essential and more accurately reflects real-world data, whereas in our case a low value shows the exact opposite. This was also the case for SVM, where the accuracy dropped to 78% when compared to the baseline model, while the EOD fairness score was 0.033, which is an acceptable number in terms of fairness. Looking at the models after applying the mitigation bias method, reweighing, we can see that the accuracy rate for both models increased compared to hyperparameter tuning to 79.06% and 79% respectively.

Looking at the German dataset, we can see that hyperparameter tuning has resulted in a decrease in accuracy for LR model, 72.66% to 67.33%, while the SVM model has seen an increase, from 67.67% to 70%. However, as a result of the same process, the fairness metric has seen an improvement in both models, 0.033 and 0.7 respectively.

For the last task, after modifying the non-binary sensitive attributes of the adult dataset in order to allow for the analysis, we have decided to carry out adversarial debiasing to reduce model bias. Adversarial bias works by first predicting the target using the pre-processing techniques carried out on the training data, and then attempting to predict the sensitive attribute using the previous stage's predictions. The model withouth debiasing resulted in an accuracy rate of 80.56% and an equal opportunity difference of -0.1650. After carrying out debiasing, the accuracy rate saw a very slight decrease, to 80.43%, while the equal opportunity difference increased significantly, to 0.0008.

We have also performed oversampling on the data to compensate for the imbalance that is present in the data. Through random oversampling, the results suggests that this approach managed to improve fairness of the predictive model, while it saw an accuracy rate of 73.68% accross 5 folds. Comparing with the accuracy results as seen in table 2, we can say that it is in par with the other fairness treatment methods.

4. Conclusion

Detecting unfairness in AI is difficult but not impossible. In this paper, we have looked at different ways of detecting the issue of fairness in machine learning methods on two different datasets. As baseline algorithms we used logistic regression and support vector machines. One way of addressing the issue of fairness explored in this paper is tuning of hyperparameters in which we have looked at the performance of the models with varied parameter value across 5 folds. We also analyzed the effectiveness of bias mitigation approaches on the data. We focused on data dimension, where we sampled the training dataset to balance the size of each category when one was under-represented, which results in reducing the fairness issue using logistic regression.

Although we employed logistic regression and support vector machines in this experiment, we may use alternative classification models, such as deep neural network models, for future investigation. In order to generalise the outcomes of the experiment to all available metrics, we can also cover other metrics and definitions of fairness for evaluation bias. It might also be worthwhile to look into how human biases might be translated into machine learning biases, as in our experiment, we investigated machine learning biases that have already been examined and how they can be handled using mitigation techniques.

Model	Dataset	Treatment	Accuracy (%)	EOD
LR	Adult	None	80.41%	-0.434
		Hyperparameter Tuning	76.37%	0.044
		Reweighing	79.06%	0.035
	German	None	72.66%	-0.052
		Hyperparameter Tuning	67.33%	0.033
		Reweighing	72.85%	0.066
SVM	Adult	None	80.64%	-0.466
		Hyperparameter Tuning	78.74%	0.033
		Reweighing	79%	0.016
	German	None	67.67%	-0.327
		Hyperparameter Tuning	70%	0.07
		Reweighing	70.71%	0.006

Table 2: Comparison of results across the models

References

[1] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. T. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang. AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *CoRR*, abs/1810.01943, 2018. 1

[2] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. T. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang. AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *CoRR*, abs/1810.01943, 2018. 2

[3] D. Dua and C. Graff. UCI machine learning repository, 2017. 1, 2

[4] M. Hardt, E. Price, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. 2

[5] T. International. Seven types of data bias in machine learning. 2021. 1

[6] F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012. This paper is an extended version of the papers [3, 13, 14]. 2

[7] R. Kohavi et al. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, pages 202–207, 1996. 1

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Van-

derplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 2

[9] UCI. UCI adult dataset, 1994. 2

[10] B. H. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 335–340, New York, NY, USA, 2018. Association for Computing Machinery. 2