Accuracy-Fairness Trade-off and Distribution Drift

[*]Ethics, Fairness and Explanation in AI (70076) @ Imperial College

1st AMIT VIKRANT DHAVALE
MSc AI
Imperial College London
London, UK
amit.dhavale24@imperial.ac.uk

I. Introduction

The report begins with a simple breakdown of the complex implementation details of model selection, as well as statements about methods adopted in empirical evaluations. The report is divided into 5 sections: **Introduction** presents an Introduction to the topic. **Task 1** performs model selection based on accuracy and fairness by varying the hyperparameters, **Task 2** employs the fairness mitigation technique of Rewighting for fairness improvement **Task 3** does feature selection and evaluates the chosen models in different states, **Conclusion** gives the conclusion for the above-mentioned analysis.

II. Task 1

In this task, Decision Tree Classifier (DTC) and Logistic Regression (LR) were selected to train and evaluate the models based on two key metrics: accuracy and fairness. Before training the models, all of the binary categorical attributes of the aif360 dataset were converted into their corresponding one-hot encoding forms. Accuracy was used to assess the overall performance of the models, while Equal Opportunity Difference (EOD) was chosen as a fairness metric to evaluate any disparities in outcomes for different groups. Equal Opportunity Difference (EOD) [1] focuses on fairness by evaluating the difference in true positive rates between groups, ensuring that the model provides equal opportunities across all subgroups. EOD helps identify and mitigate any disproportionate treatment between different groups, making it an essential metric for assessing and improving the fairness of machine learning models. **Figure 1** depicts the obtained mean performance for DTC.

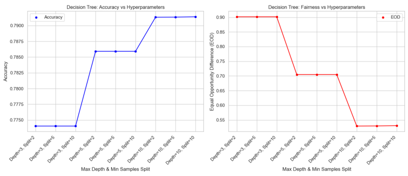

Fig. 1: Mean performance of Decision Tree Classifier (DTC)

From **Figure 1**, we observe a symmetric relationship between accuracy and EOD. Both metrics show a consistent insensitivity to variations in the minimum sample split. However, both metrics are significantly influenced by changes in the depth of the decision tree. Notably, as the tree depth increases, accuracy improves while EOD decreases, suggesting that deeper trees lead to better performance in terms of both accuracy and fairness. Next, **Figure 2** presents the mean performance obtained for Logistic Regression (LR).

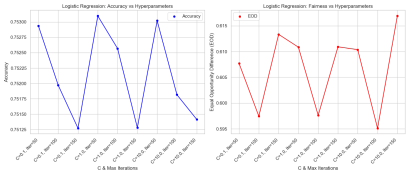

Fig. 2: Mean performance of Logistic Regression (LR)

From **Figure 2**, we observe a roughly cyclic trend where, as the number of iterations increases, both accuracy and EOD decrease. Additionally, minimal changes are observed in both metrics when varying the regularization parameter (C), indicating that regularization has a relatively small impact on the performance and fairness of the model. **Table I** depicts the final chosen hyperparameters for each model, based on their performance in terms of accuracy and fairness. **(Note: NA indicates that the hyperparameter does not apply to that model.)**

Model	Max Depth	Min Samples Split	Accuracy	EOD	Hyperparameters	C	Max Iter
DecisionTree	10.0	10.0	0.7914	0.5305	Depth=10, Split=10	NA	NA
DecisionTree	10.0	2.0	0.7913	0.5291	Depth=10, Split=2	NA	NA
LogisticRegression	NA	NA	0.7531	0.6109	C=1.0, Iter=50	1.0	50.0
LogisticRegression	NA	NA	0.7518	0.5951	C=10.0, Iter=100	10.0	100.0

TABLE I: Model Performance with Hyperparameters

Now, both the models are again trained on the training dataset with the newly chosen hyperparameters and evaluated on the test dataset, resulting in the performance which is illustrated in **Table II**.

Model	Test Accuracy	Test EOD
Decision Tree (Best Accuracy)	0.7920	0.5155
Logistic Regression (Best Accuracy)	0.7544	0.6225
Decision Tree (Best Fairness)	0.7919	0.5063
Logistic Regression (Best Fairness)	0.7526	0.5876

TABLE II: Test Accuracy and EOD for Each Model

III. Task 2

Now, a fairness-aware learning method, preprocessing with reweighting, was applied before performing the analysis presented in Task I. Pre-processing [2] is the act of modifying the observed dataset such that classifiers learned on the modified dataset do not display the biases inherent in our data. Thus, Preprocessing with reweighting [3] is a technique used to address bias in datasets and improve fairness in machine learning (ML) models by adjusting the weights of the training samples based on their group membership or other sensitive attributes, such as disability (sensitive attribute in the given coursework). By giving more weight to underrepresented groups and less weight to overrepresented ones, this method helps ensure that the model is not inclined toward one group over another. Thus, **Figure 3** depicts the obtained mean performance for DTC after Reweighting.

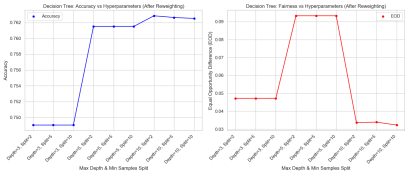

Fig. 3: Mean performance of Decision Tree Classifier (DTC) after Reweighting

From **Figure 3**, it can be observed that the accuracy increases significantly as the depth of the tree grows from 3 to 5. However, the increase in accuracy becomes relatively smaller as the depth continues to increase from 5 to 10. In contrast, the values of the Equal Opportunity Difference (EOD) display inconsistent behaviour, initially increasing with depth but subsequently decreasing after reaching a certain point. Next, **Figure 4** presents the mean performance obtained for Logistic Regression (LR) after Reweighting.

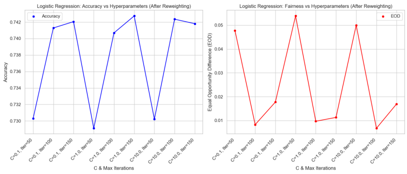

Fig. 4: Mean performance of Logistic Regression (LR) after Reweighting

From **Figure 4**, we observe a roughly cyclic trend where, as the number of iterations increases, accuracy increases and EOD decreases. Additionally, minimal changes are observed in both metrics when varying the regularization parameter (C), indicating that regularization has a relatively small impact on the performance and fairness of the model.

(Note: After applying reweighting, the EOD values significantly decreased for both models, indicating a notable improvement in the fairness of the trained models.)

Table III depicts the final chosen hyperparameters for each model, based on their performance in terms of accuracy and fairness after Reweighting. **(Note: NA indicates that the hyperparameter does not apply to that model.)**

Model	Max Depth	Min Samples Split	Accuracy	EOD	Hyperparameters	C	Max Iter
DecisionTree	10.0	2.0	0.7628	0.0337	Depth=10, Split=2	NA	NA
DecisionTree	10.0	10.0	0.7625	0.0323	Depth=10, Split=10	NA	NA
LogisticRegression	NA	NA	0.7427	0.0113	C=1.0, Iter=150	1.0	150.0
LogisticRegression	NA	NA	0.7423	0.0067	C=10.0, Iter=100	10.0	100.0

TABLE III: Model Performance with Hyperparameters after Reweighting

Now, both the models are again trained on the training dataset with the newly chosen hyperparameters and evaluated on the test dataset, resulting in the performance which is illustrated in **Table IV**.

Model	Test Accuracy	Test EOD
Decision Tree (Accuracy)	0.7605	0.0317
Logistic Regression (Best Accuracy)	0.7445	0.0062
Decision Tree (Fairness)	0.7601	0.0277
Logistic Regression (Fairness)	0.7392	0.0199

TABLE IV: Test Accuracy and EOD for Each Model after Reweighting

IV. Task 3

A. Literature Review

Feature selection [4] enhances ML models by removing irrelevant or redundant data, improving accuracy and efficiency. However, the AIF360 dataset contains 17 distinct features, making it impractical to perform an exhaustive grid search across all possible feature combinations (2^{17} **different combinations!**). To streamline the process and reduce computational cost, various feature selection techniques have been proposed to identify the most relevant features effectively. In [5], the authors discuss two feature selection techniques: **Forward Selection** and **Recursive Feature Elimination (RFE)**.

- **Forward Selection** follows a stepwise approach, beginning with an empty model. At each step, it iteratively adds the most impactful features until no further improvement occurs.
- **Recursive Feature Elimination (RFE)**, in contrast, starts with all features and systematically removes those with the lowest importance. After each iteration, the model is re-evaluated, and the least significant feature is discarded.

However, both methods focus mainly on optimizing performance metrics such as accuracy and do not take into account fairness when selecting features.

Similarly, in [6], the authors propose a **Decision Tree (DTC)** based feature selection method:

- This approach determines feature importance, prioritizing root nodes. Here, the features are removed iteratively until accuracy drops significantly.

Despite its effectiveness in identifying influential features, this method also fails to incorporate fairness considerations, making it unsuitable for applications where bias mitigation is essential. Finally, to address the aforementioned shortcomings, [7] proposes a fairness-aware feature selection approach.

This method first partitions the dataset based on the **sensitive attribute** and then applies multiple feature ranking techniques, such as **Forward Selection** and **Recursive Feature Elimination (RFE)**, separately for each demographic group. Once the feature rankings are obtained for each subgroup using these methods, the final feature ranking is determined by averaging the individual rankings across all demographic-specific datasets. This approach helps to reduce bias in feature selection.

B. Task 3-a

After assigning ranks to the distinct features in the dataset, a feature selection algorithm is developed, which is inspired by the **Forward Selection** method. This approach iteratively adds features to the model, starting with the highest-ranked feature (**Rank-1**). At each step, a new feature is added, and the model is re-evaluated based on the following metrics:

- **Accuracy**
- **Equalized Odds Difference (EOD)**
- **Fairness-Accuracy Tradeoff Score (FATS)**, a self-proposed metric defined as:

$$FATS = 0.5 \times \text{Accuracy} + 0.5 \times (1 - \text{EOD}) \qquad (1)$$

As shown in eq. (1), **Fairness-Accuracy Tradeoff Score (FATS)** balances predictive performance with fairness by considering both accuracy and the inverse of EOD. This ensures that feature selection considers not only model performance but also fairness in decision-making. Thus, if we consider FATS as the base evaluation metric, the sensitive attribute *DIS* becomes part of the selected feature set. However, if solely EOD is used as the evaluation metric, *DIS* does not get included in the selected subset of features. Table V presents the metric used for selecting features for each of the two chosen models.

Model	Selected Features	Avg EOD	Avg Accuracy	Metric
DTC	{RAC1P, GCL, SEX, AGEP, MAR, DEYE, SCHL, ESP, MIL, RELP, DIS, ANC, DEAR, CIT, DREM}	0.0351	0.7644	FATS
DTC	{RAC1P, GCL, SEX}	0.01597	0.5605	EOD
LR	{RAC1P, GCL, SEX, AGEP, MAR, DEYE, SCHL, ESP, MIL, RELP, DIS, ANC}	0.00538	0.7177	FATS
LR	{RAC1P, GCL, SEX}	0.00825	0.5519	EOD

TABLE V: Feature Selection Results for Decision Tree (DTC) and Logistic Regression (LR)

From Table V, it can be inferred that although fairness (as measured by EOD) can be improved by excluding the *DIS* feature, the overall FATS score is higher when *DIS* is included in both models. This indicates that the *DIS* feature is essential for improving model accuracy.

Therefore, for further analysis, the feature subset corresponding to the FATS metric is chosen for both the models.

C. Evaluation of Models on State Data (Excluding Florida)

In this section, both the Decision Tree (DTC) and Logistic Regression (LR) models after using Reweighting are trained on the selected subset of features from part (3-a) and are evaluated on data from various states, excluding Florida. The results are presented in terms of fairness (measured by EOD) and the fairness-accuracy tradeoff (FAST) metric.

Figure 5 shows the variation in fairness values (EOD) across different states for both models.

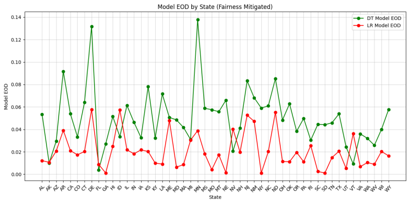

Fig. 5: Model EOD by State (Fairness Mitigated)

From Figure 5, it can be observed that the EOD values for the LR model are generally lower compared to the DTC model, indicating better fairness performance for LR across the states evaluated.

Similarly, Figure 6 illustrates the variation in FAST values across different states for both models.

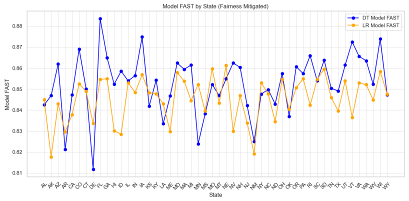

Fig. 6: Model FAST by State (Fairness Mitigated)

From Figure 6, it is evident that there is little difference in the FAST values between the LR and DTC models across the states. However, for the majority of the states, the FAST values for the DTC model are slightly higher than those of the LR model, indicating that DTC may have a slight advantage in balancing fairness and accuracy.

V. Conclusion

In conclusion, the choice of model—Decision Tree (DTC) or Logistic Regression (LR)—depends mainly on the employed evaluation criterion. If fairness is the primary concern, then LR can be considered as the preferred model due to its lower (EOD). However, if achieving a balance between good accuracy and fairness is important, then DTC provides a reasonable choice, as it offers a favourable tradeoff between model performance and fairness, as evidenced by the higher FAST values in most cases. Therefore, the decision between LR and DTC should be based on the specific priorities of the task, whether it is fairness, accuracy, or a balance of both. However, deploying this trained model across neighbouring states may not be ethically feasible because the dataset doesn't account for sector-specific employment (e.g., government, NGOs, or private sector). For example, a state may show high employment among disabled individuals, however in reality this employment may only be in the government or NGO sectors, while private sector employment may still be low. This could lead to inaccurate or biased predictions, as the model overlooks sector-specific employment. Therefore, deploying it without considering these details might not be ideal.

References

[1] Hardt, Moritz & Price, Eric & Srebro, Nathan. (2016). Equality of Opportunity in Supervised Learning. 10.48550/arXiv.1610.02413.

[2] Amal Tawakuli, Thomas Engel, Make your data fair: A survey of data preprocessing techniques that address biases in data towards fair AI, Journal of Engineering Research, 2024, , ISSN 2307-1877, https://doi.org/10.1016/j.jer.2024.06.016.

[3] Blow CH, Qian L, Gibson C, Obiomon P, Dong X. Comprehensive Validation on Reweighting Samples for Bias Mitigation via AIF360. Applied Sciences. 2024; 14(9):3826. https://doi.org/10.3390/app14093826

[4] Jie Cai, Jiawei Luo, Shulin Wang, Sheng Yang, Feature selection in machine learning: A new perspective, Neurocomputing, Volume 300, 2018, Pages 70-79, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2017.11.077.

[5] Reski Noviana, Enny Itje Sela . Performance Comparison Random Forest and Logistic Regression in Predicting Time Deposit Customers with Feature Selection. International Journal of Computer Applications. 186, 16 (Apr 2024), 33-38. DOI=10.5120/ijca2024923548

[6] Babagoli, Mehdi & Aghababa, Mohammad & Solouk, Vahid. (2019). Heuristic nonlinear regression strategy for detecting phishing websites. Soft Computing. 23. 10.1007/s00500-018-3084-2.

[7] Md Rahat Shahriar Zawad and Peter Washington. Evaluating Fair Feature Selection in Machine Learning for Healthcare. arXiv preprint arXiv:2403.19165, 2024.