Towards Truthworthy Classifiers: Bias Reduction in Tree-Based Classifiers via Pruning

Track 1: Multi-Objective Optimization for AI Trustworthiness in Tabular Data Classification

Team ARM

The team

Rishi Kalra

- BSc Physics @ UCL
- MSc Data Science & Machine Learning @ UCL

Murtaza Dhanerawala

- BSc Physics @ Warwick
- MSc Data Science @ UCL

Avanija Menon

- BSc Astrophysics @ UCL
- MSc Artificial Intelligence @ Imperial

Temi Oluwole

- MSci Physics @ UCL

Problem Statement

Types of bias:

1. Statistical Bias:
2. **Algorithmic Bias:**
3. Data Bias: training data may not represent the population
4. **Societal Bias: Social bias can lead to bias in the data collected**

1. Maintain high performance across multiple tasks
2. Minimise bias and maximise fairness
3. Use the COMPAS dataset

> **“The expansion of AI into many aspects of public life requires extending our view to consider AI within the larger social system in which it operates.” Reva Schwartz - NIST 2022**

Objectives

1. Use HolisticAI's existing framework to optimise for multiple objectives in accuracy, bias and explainability
2. Attempt to integrate this multi-objective optimisation in the training process

Solution Overview

1. Enhance the existing optimisation process

a. Add more Holistic AI metrics to the objective function
b. Bayesian Optimisation Method + Hyperparameter Search

2. Add fairness optimisation and explainability to the decision tree / random forest training process via pruning

a. Pre-Pruning: Prune leaves as the tree is built / during split selection
b. Post-Pruning: Prune leaves after the tree is built
c. Custom fairness objective that incorporates multiple metrics

Our Metrics

Fairness Deviation	(|Average Odds Difference| + |Treatment Inequality Deviation|) / 2
Fairness Penalty	Fairness Weight x log2(Fairness Deviation)

TPR = True Positive Rate

TNR = True Negative Rate

FPR = False Positive Rate

FNR = False Negative Rate

Explanation

Average Odds Difference

|TNR(Group 1) - TNR(Group2) + |TPR(Group1) - TPR(Group2)|

Idea

Promote equality of **correct** predicted outcomes

Treatment Inequality Deviation

|FNR(Group 1) - FNR(Group2) + |FPR(Group1) - FPR(Group2)|

Idea

Promote equality of **incorrect** predicted outcomes

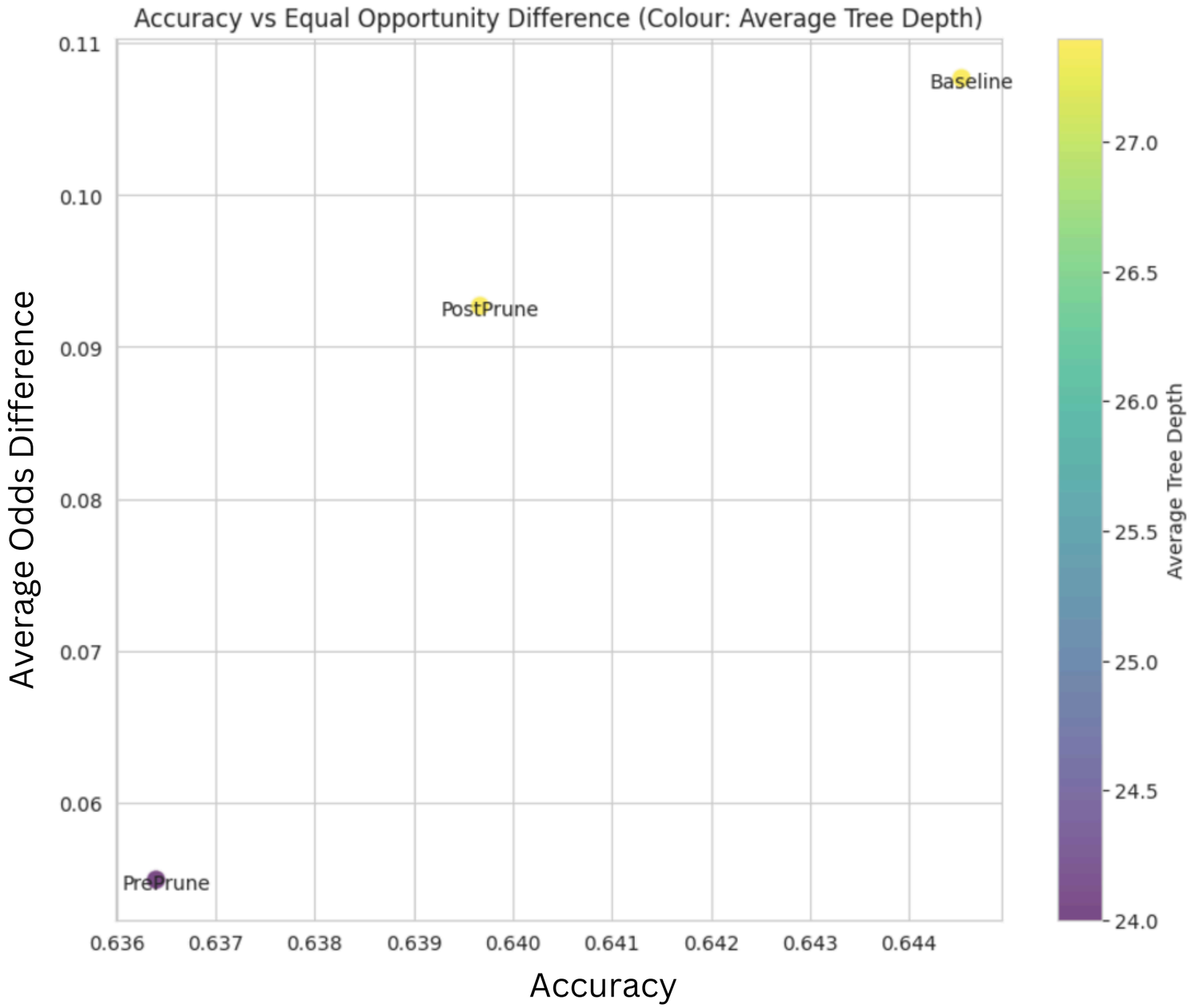
System design

	Pre-Pruning	Post-Pruning
Training	RF Training ↓ Split Selections: **Gain - Fairness Penalty** ↓ Build Trees ↓	RF params Fairness weighting ↓ **Split Based on gain as usual** ↓ Prune based on fairness penalty and accuracy, with set threshold
Evaluation Metrics	**Performance:** Accuracy	**Bias:** Treatment Inequality Average Odds Differece	**Explainability** Average Tree Depth

Results 1


Comparison of Metrics Across Different Methods
Holistic AI Baseline
Optuna on All Metrics
Values
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Accuracy
Equal Opportunity Diff
Average Odds Diff
Metrics

Results 2



- PrePruning has less depth + much smaller EOD for small difference in accuracy

Business value: Governance + Compliance

“Unacceptable risk AI systems are systems considered a threat to people and will be banned.” “Social scoring: classifying people based on behaviour, socio-economic status or personal characteristics” EU union 2023

Improve explainability by pruning: we can more easily identify if socio-economic status or personal characterstics are used in decision-making

“Local Law 144 of 2021 regarding automated employment decision tools (“AEDT”) prohibits employers and employment agencies from using an automated employment decision tool unless the tool has been subject to a bias audit” NYC GOVT 2021

Example: LL144 requires calculating bias metrics like disparate impact of automated employment decision tools --> optimising this in training would help with compliance

Business value: Sustainability compliance

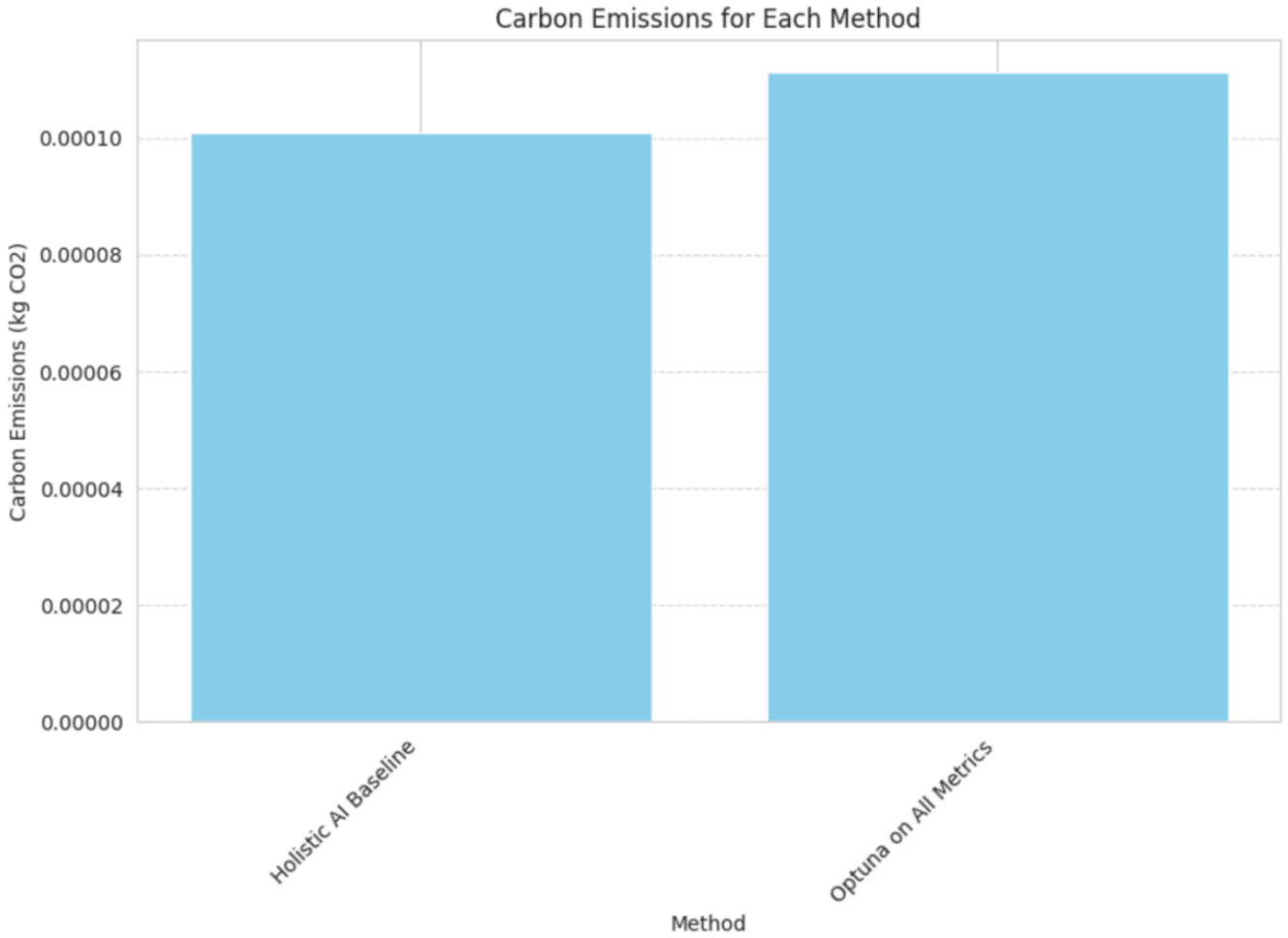
By measuring the carbon emissions of our models with Code Carbon making it easier for companies which will have to comply with the Corporate Sustainability Reporting Directive (CSRD) when reporting their impact on the environment.

“There is a Dickensian quality to the use of AI when it comes to our environment: It can make our planet better, and it can make our planet worse,” said US Senator Ed Markey. “Our AI Environmental Impacts Act would set clear standards and voluntary reporting guidelines to measure AI’s impact on our environment. The development of the next generation of AI tools cannot come at the expense of the health of our planet.”

"AI adoption has driven higher energy use, prompting leaders to reevaluate sustainability goals. Developing energy-efficient models and leveraging sustainable IT practices not only reduces environmental impact but enhances cost-efficiency, aligning innovation with long-term resilience and organizational success" Christina Shim, Chief sustainability officer, IBM

Sustainability Metrics

By measuring the carbon emissions of our models with Code Carbon making it easier for companies which will have to comply with the Corporate Sustainability Reporting Directive (CSRD) when reporting their impact on the environment.



We can get better results using an Optuna hyperparameter optimisation, but this is more carbon intensive

So training one model using the pre or post pruning method is more efficient .

Academic Value + Novelty

Pruning is often used for reducing overfitting + improving explainability

Recent work uses hyperparameter optimization to balance fairness metrics with accuracy [1] and pruning in medical contexts [2]

Our contributions include introducing our pre and post pruning methods with a custom fairness penalty

[1] A. Cruz, Pedro Saleiro, Catarina Bel'em, C. Soares and P. Bizarro. "Promoting Fairness through Hyperparameter Optimization." 2021 IEEE International Conference on Data Mining (ICDM) (2021): 1036-1041. https://doi.org/10.1109/ICDM51629.2021.00119.
[2] Wu, Yawen, Dewen Zeng, Xiaowei Xu, Yiyu Shi and Jingtong Hu. “FairPrune: Achieving Fairness Through Pruning for Dermatological Disease Diagnosis.” International Conference on Medical Image Computing and Computer-Assisted Intervention (2022).

Risk considerations

- **Performance:** Balance fairness and performance through weighting
- **Fairness:** Incorporate the fairness penalty within pruning
- **Sustainability:** Training one model is less carbon intensive compared to training many and finding the optimal
- **Robustness:** Protecting characteristics may help reduce reliance on these features
- **Privacy:** Our algorithm ensures that protected characteristics are not input to the model but rather used to steer the systemic bias towards its lower limit
- **Explainability:** Pruning implicitly optimises explainability by reducing tree depth and making splits more interpretable

Future work & limitations

Technical

- We postulate that: the systemic bias with respect to each protected characteristic has a non-zero lower limit - analogous to Bayesian priors.
  - A probabilistic method to determine the lowest possible bias in a dataset would lead to:
    - 1. Knowledge about how fair AIs can be in the limit
    - 2. A process for judging training data based on its inherent bias
- Add pairwise systemic bias scores to reduce bias in cases where there are >1 underprivileged groups
  - e.g. {”African-American”, “Hispanic-American, “Asian-American”}
- Target robustness by adding adversarial networks to make the model more secure to changing distributions
- Encode data within preprocessing to increase privacy

Future work & limitations

Sustainability

- Use a more objective carbon emission calculator - since it depends a lot on computer hardware

Thank You For Listening!