University of Bologna

Analysis of Fairness-Accuracy Tradeoff

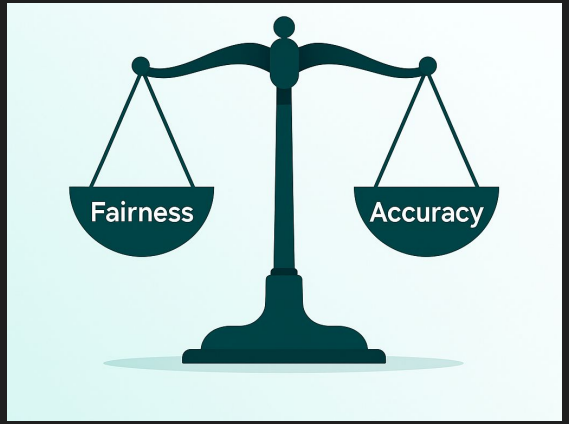
Professor: Michele Lombardi

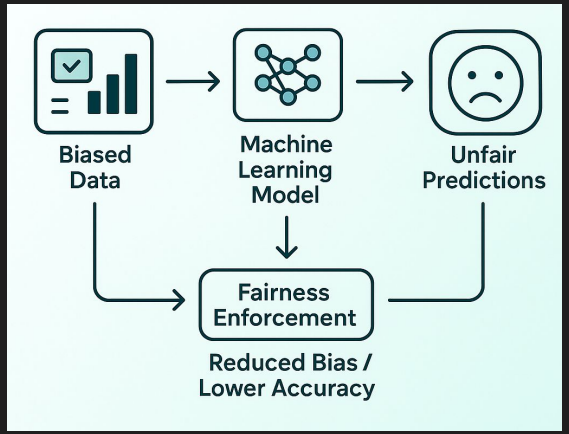
TA: Luca Giuliani

Mehregan Nazarmohsenifakori

Why Fairness–Accuracy Tradeoff Matters?

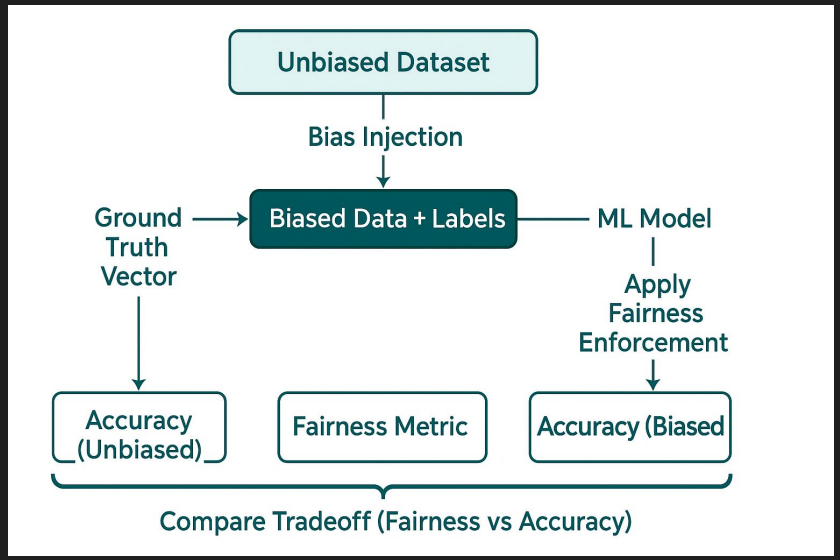
- Machine Learning models are widely used in **sensitive decisions** (e.g., hiring, healthcare, justice)
- These models often learn from **historically biased data**
- **Fairness interventions** can reduce bias, but **may lower predictive accuracy**
- This creates a **tradeoff**: fairness vs performance
- Understanding and managing this tradeoff is essential for **responsible and trustworthy AI**





Project Objective & Setup

- **Goal**: Study how **fairness enforcement** affects both **model accuracy** and **fairness metrics**
- **Setup**: Train on **biased data**, evaluate using:
  - **Accuracy w.r.t. biased ground truth**
  - **Accuracy w.r.t. hidden unbiased vector**
  - **Fairness gap** (e.g., **Demographic Parity**)



- Built a **synthetic data generator:**
  - Introduces controlled bias
  - Keeps access to **true unbiased labels**
- Enables a **controlled, ground-truth-aware analysis** of the fairness–accuracy tradeoff

Synthetic Dataset Construction

- Developed a **custom data generator** for full control over bias and label noise
- Supports **both binary and continuous sensitive attributes** via configuration
- Features (x0, x1) are derived from unbiased latent features (u0, u1)
- Bias injected via three mechanisms:
  - ➕ **Additive** (shifted mean)
  - ✖️ **Multiplicative** (rescaled features)
  - 🌀 **Nonlinear** (bias depends on sign/magnitude)
- Generates two outputs:
  - **uy = unbiased label (ground truth)**
  - **y = biased label (used for model training)**

Why Not Use Continuous ?

- Project goal: **evaluate fairness interventions** commonly used in real-world applications
- Most standard fairness metrics (e.g., Demographic Parity, Equalized Odds) are:
  - ✳️ Defined for **binary sensitive attributes**
  - ❌ Not directly applicable to continuous attributes without preprocessing
- Continuous settings require:
  - Discretization or thresholds (adds noise & complexity)
  - Different evaluation strategies (e.g., fairness over intervals or distributional metrics)
- To maintain **clarity, comparability, and focus**, this study centers on **binary group bias**

Baseline Models

- **Models Compared**:
  - **Logistic Regression** *(liblinear solver)*
  - **Decision Tree** *(max_depth = 5)*
  - **Random Forest** *(n_estimators = 100, max_depth = 5)*

- **Metrics Used**:
  - ✅ **True Accuracy**: prediction vs **unbiased labels**
  - ✅ **Observed Accuracy**: prediction vs **biased labels**
  - ✅ **Demographic Parity Gap (DPG)**: $|P(\hat{y}=1 \mid \text{group } 0) - P(\hat{y}=1 \mid \text{group } 1)|$

- **Logistic Regression** provided the **best balance** between accuracy and fairness across all bias scenarios. It was selected as the **main model** for further fairness experiments.

Training & Evaluation Strategy

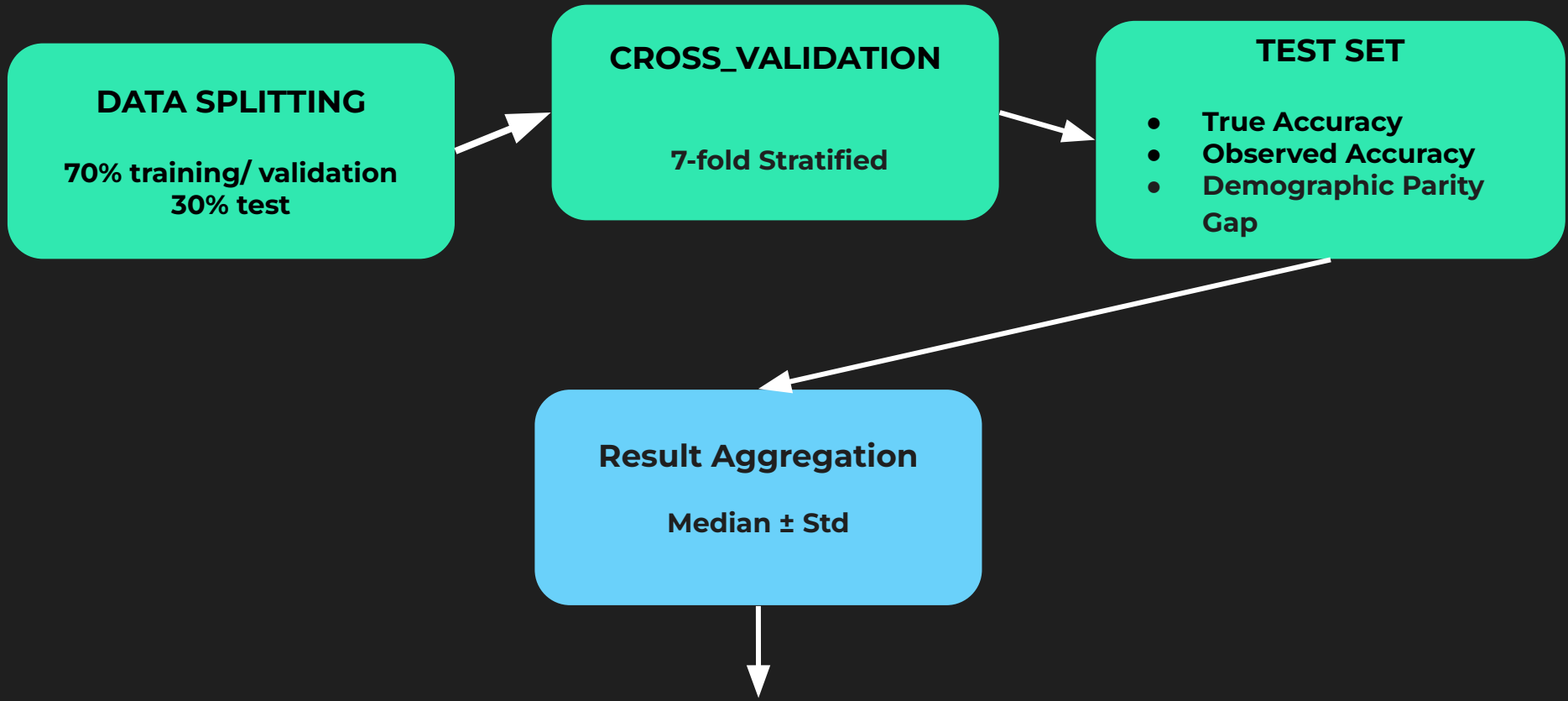
- **Data Splitting**:
  - Dataset is first split **70% training/validation, 30% test**
  - Test set is **fixed and untouched during training**
  - Only the **training/validation split** is used in **cross-validation**

- **Cross-Validation**:
  - Performed **7-fold Stratified Cross-Validation** on training/validation set
  - Models trained on 6 folds, evaluated on **fixed test set**
  - Ensures balanced distribution of biased labels in each fold

- **Result Aggregation:**
  - For each model:
    - Metrics are computed for each fold (7 total) on **test set**
    - Final results reported as: **Median ± Standard Deviation** across folds

- This approach provides **robust, variance-aware estimates** of both fairness and performance — minimizing the effect of outlier splits.

Training & Evaluation Pipeline



Fairness Enforcement Techniques

Goal: To **reduce bias** in predictions between sensitive groups while maintaining as much accuracy as possible.

Pre-processing

- **Re-weighting:**

Adjusts sample weights to balance group distributions.

- **Label Massaging:**

Flips labels for some samples to equalize outcomes

In-processing

- **Exponentiated Gradient:**

Trains model with fairness constraint using ε thresholds.

- **Grid Search Reduction:**

Finds best model satisfying fairness-accuracy tradeoff.

Post-processing

- **Threshold Optimization:**

Adjusts prediction thresholds per group to balance outcomes.

- These five methods cover **all three fairness enforcement stages** — allowing direct comparison across data-level, model-level, and output-level interventions.

Bias Scenarios

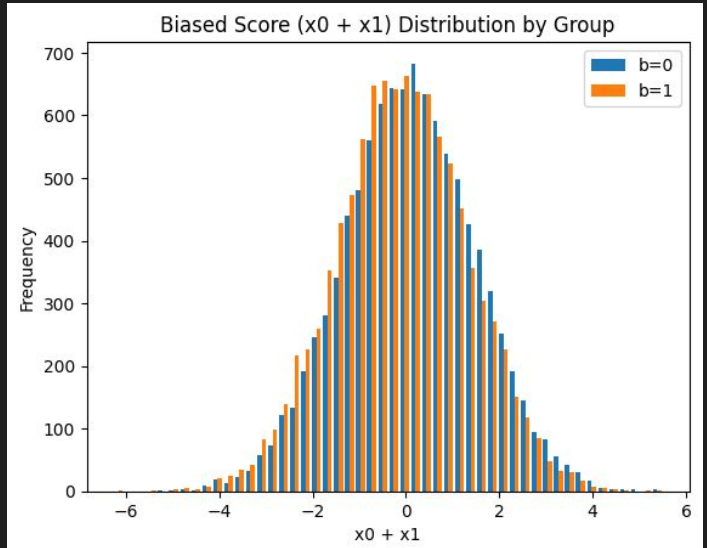
Scenario	Bias Type
Mild Additive	Additive (feature-level)
Severe Additive	Additive (feature-level)
Mild Multiplicative	Multiplicative (feature-level)
Nonlinear Bias	Nonlinear (feature-level)
Interaction Bias	Additive + Interaction (feature & label)

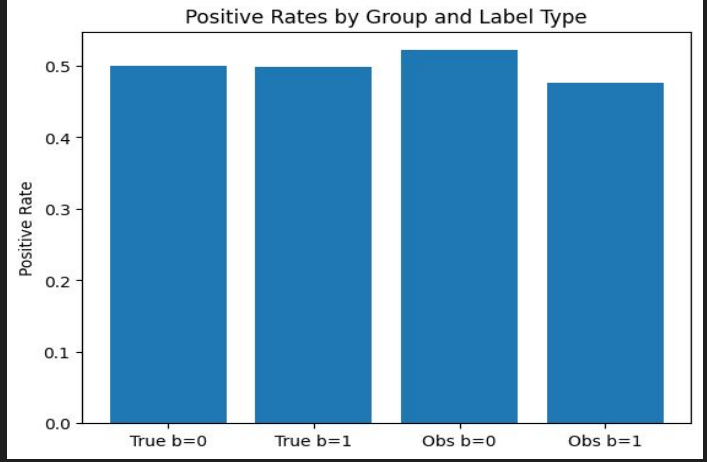
Scenario 1 – Mild Additive Bias

- **Description:**
  - Protected Attribute: Binary (b ∈ {0,1})
  - Bias Type: Additive
  - Bias Strength: ±0.5
  - Mechanism:

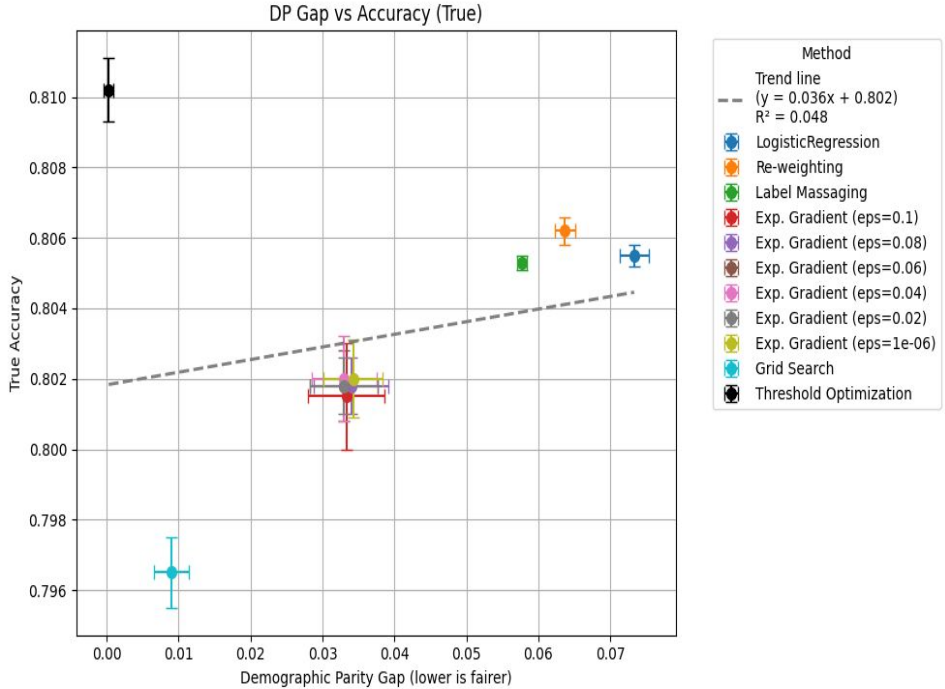
```
x = u + bias_strength * (b == 0 ? +1 : -1)
```

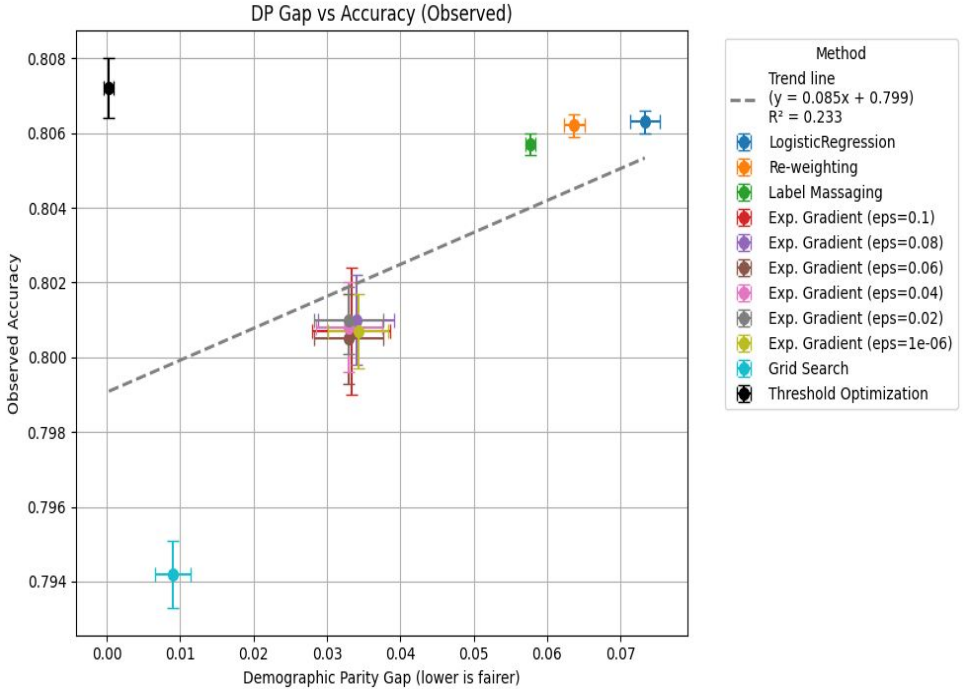
- **Score Distribution** shows that group b=0 receives slightly higher feature values, confirming the additive bias in x0 + x1.
- **Observed label rates** are skewed: group b=0 has more positives, even though the true labels were balanced — revealing the impact of bias on outcomes.





Results of Mild Additive Bias





Fairness–Accuracy Trade-off Analysis: Mild Additive Bias

- **Logistic Regression** is the strongest **baseline**, balancing high accuracy and moderate fairness.
- **Threshold Optimization** delivers the **best overall result** — it improves both **accuracy** and **DPG** significantly.

- **Re-weighting** achieves slight fairness gains **without sacrificing accuracy**.

- **Exponentiated Gradient** reduces bias reliably **but costs some accuracy**.

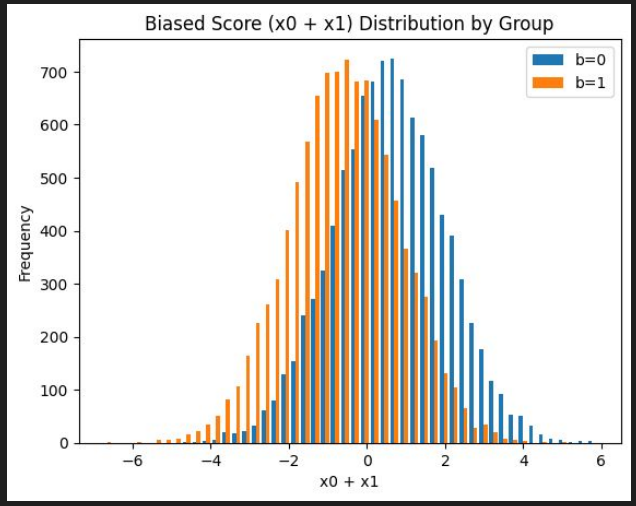
- **Grid Search** gives the **very low DPG**, but with a **clear drop in accuracy.**

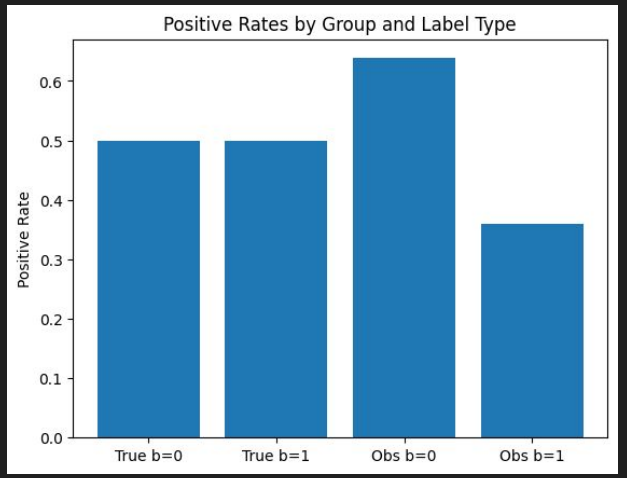
- Trade-off strength:
  - **True Accuracy vs. Fairness**: weak trade-off ($R^2 = 0.048$)
  - **Observed Accuracy vs. Fairness**: mild trade-off ($R^2 = 0.233$)

- **Threshold Optimization** clearly stands out by enhancing both fairness and predictive performance.

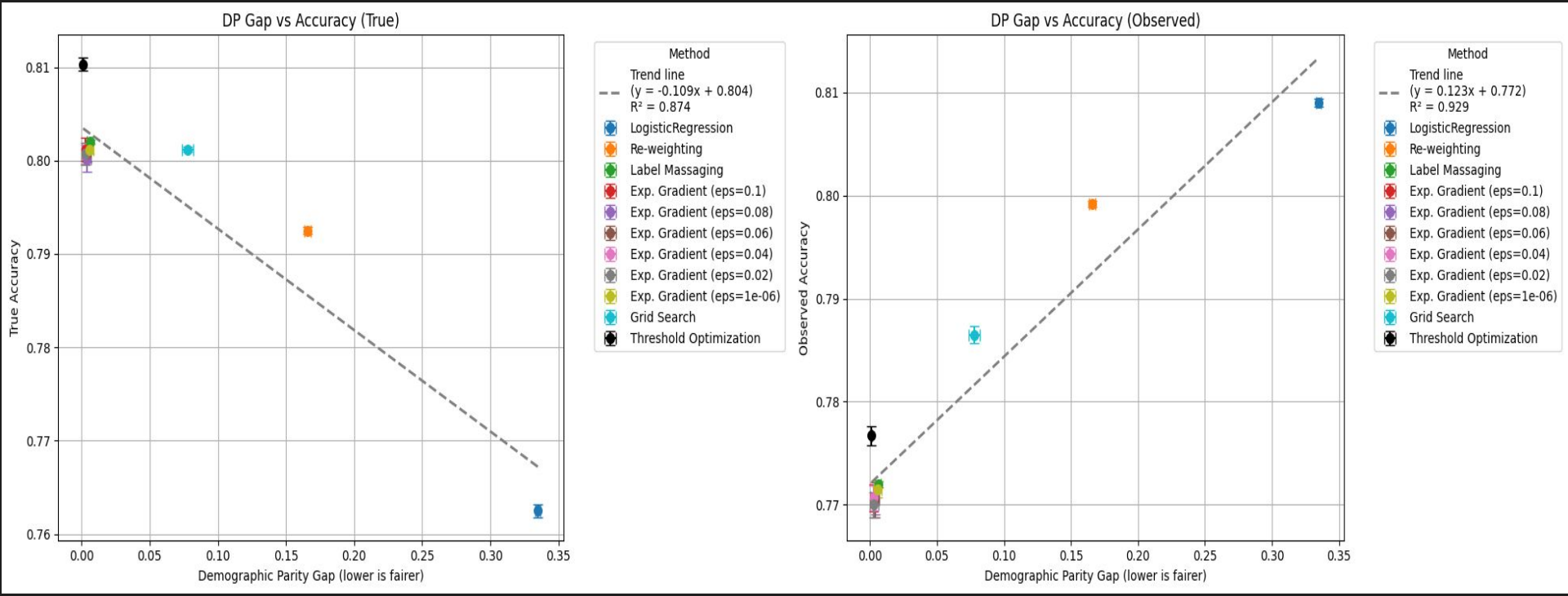
Scenario 2 – Severe Additive Bias

- **Description:**
  - Protected Attribute: Binary (b ∈ {0,1})
  - Bias Type: Additive
  - Bias Strength: ± 2
  - Mechanism: Like the previous scenario
- **Feature Distribution Plot**: The histogram shows a **strong separation** between b=0 and b=1 in the biased score (x0 + x1), caused by a **larger additive shift** than in the mild case.
- **Label Distribution Plot**: The positive rate for group b=0 is significantly higher than b=1 under **observed labels**, reflecting a **more severe label bias**.





Results of Severe Additive Bias



Fairness–Accuracy Trade-off Analysis: Severe Additive Bias

- **Logistic Regression** overfits to biased labels (high observed accuracy) but suffers in fairness and true accuracy.
- **Threshold Optimization** outperforms all others with, **Highest True Accuracy, Lowest DP Gap and Best overall fairness–accuracy trade-off.**
- **Label Massaging** and **Exponentiated Gradient** offer near-fair results (~0.006 DPG) with only **minimal drop in accuracy.**
- **Grid Search** is moderately fair but less accurate.
- **Re-weighting** preserves observed accuracy but **fails to correct bias.**
- **Trade-off strength:**
  - **R^2 = 0.874 (True)** → **Fairer models generalize better**
  - **R^2 = 0.929 (Observed)** → **Biased accuracy improves as fairness drops**

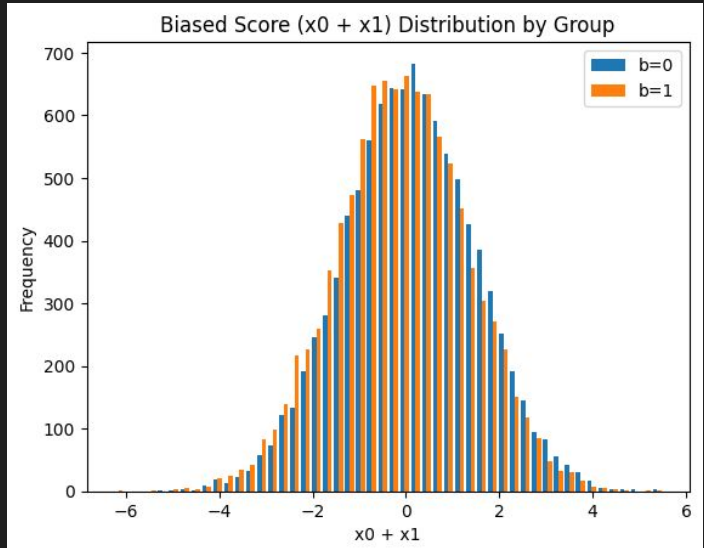
Scenario 3 – Mild Multiplicative Bias

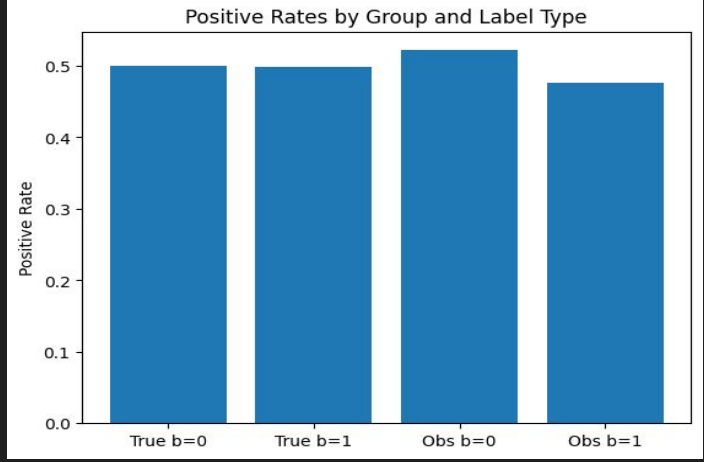
- **Description:**
  - Protected Attribute: Binary (b ∈ {0,1})
  - Bias Type: Multiplicative
  - Bias Strength: ±0.3
  - Mechanism:

```
x = u × (1 + bias_strength × (b == 0 ? +1 : -1))
```

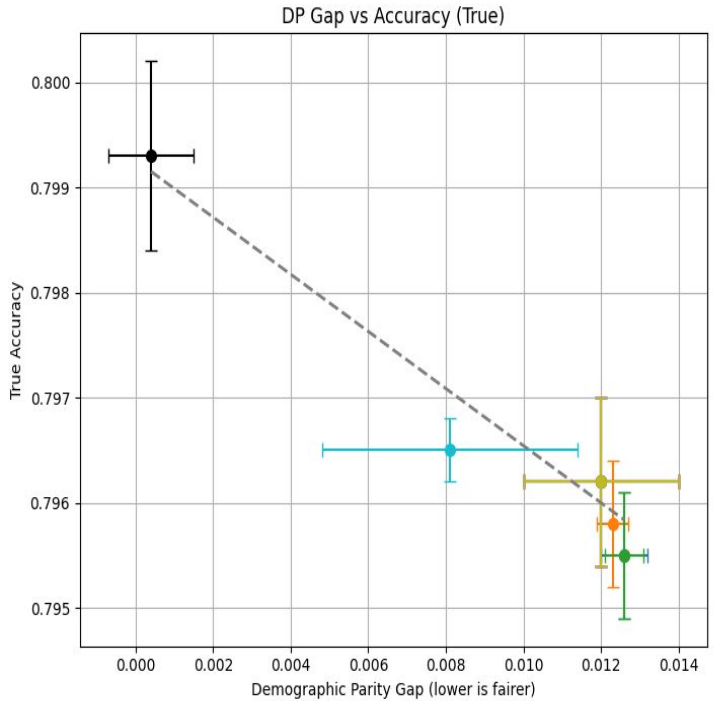
- **Score Distribution Plot**: Group b=0 scores are slightly stretched outward (higher values), while b=1 scores are compressed.

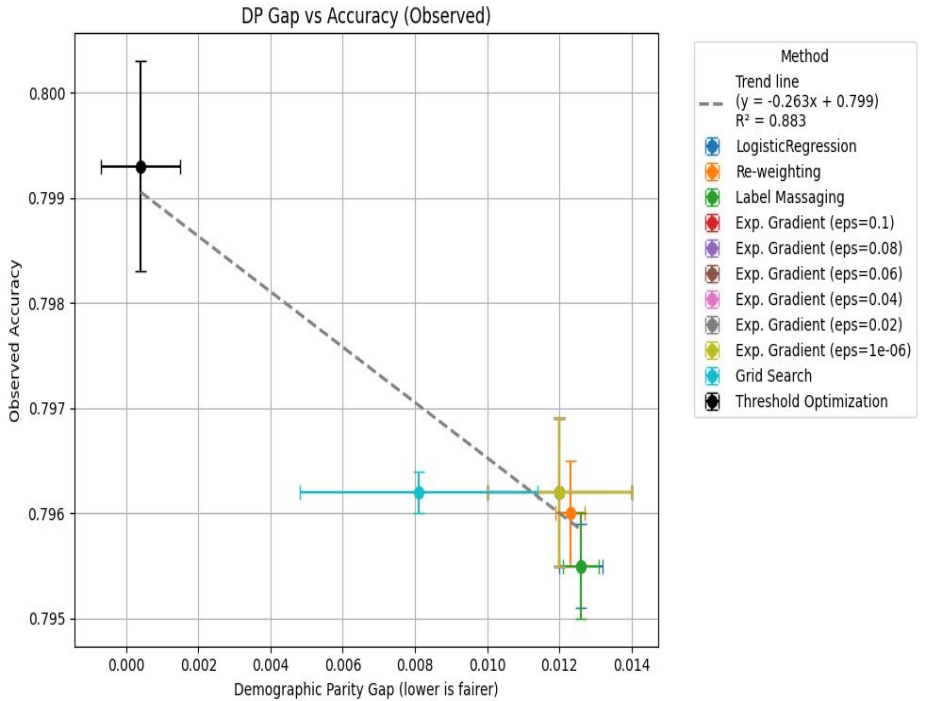
- **Positive Rate Plot:** Both groups still have **similar positive label rates,** meaning this bias doesn't heavily distort the label balance like the additive biases did.





Results of Mild Multiplicative Bias





Fairness–Accuracy Trade-off Analysis: Mild Multiplicative Bias

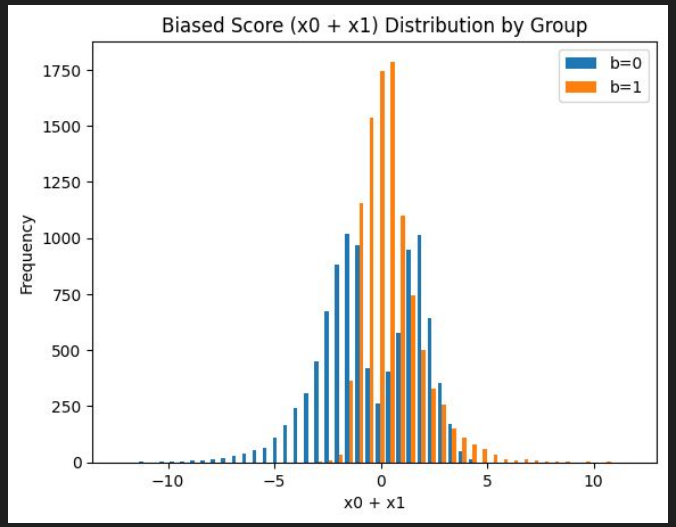
- **Logistic Regression** performs strongly even without fairness intervention — showing high true accuracy and low demographic parity gap.
- **Threshold Optimization** again dominates, eliminating bias almost entirely while also slightly boosting accuracy.
- **Grid Search** and **Exponentiated Gradient** offer solid fairness gains with almost no accuracy drop, while **pre-processing** methods remain stable but less impactful.
- **Trade-off is minimal:** accuracy improves slightly as fairness increases — bias is mild enough that most methods work well.
- **R^2 = 0.923 (True)** → As models become fairer, true accuracy increases.
- **R^2 = 0.883 (Observed)** → Bias correction slightly improves observed performance too.
- **In low-bias settings, fairness and accuracy are not conflicting — they often improve together.**

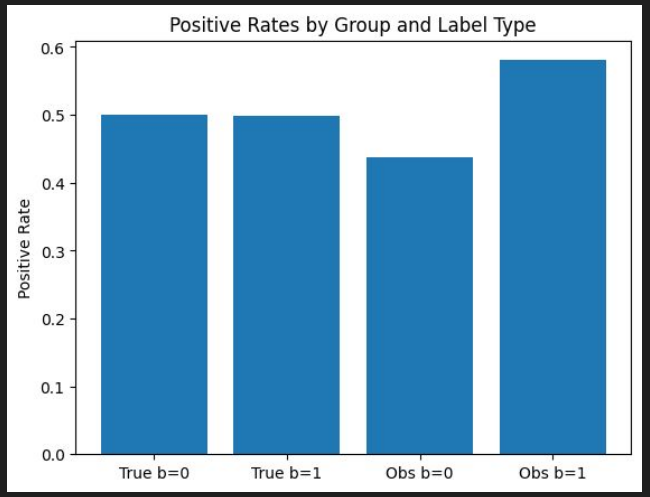
Scenario 4 – Nonlinear Bias

- **Description:**
  - Protected Attribute: Binary ($b \in \{0,1\}$)
  - Bias Type: Nonlinear (Sign-based transformation)
  - Bias Strength: ± 1
  - Mechanism:

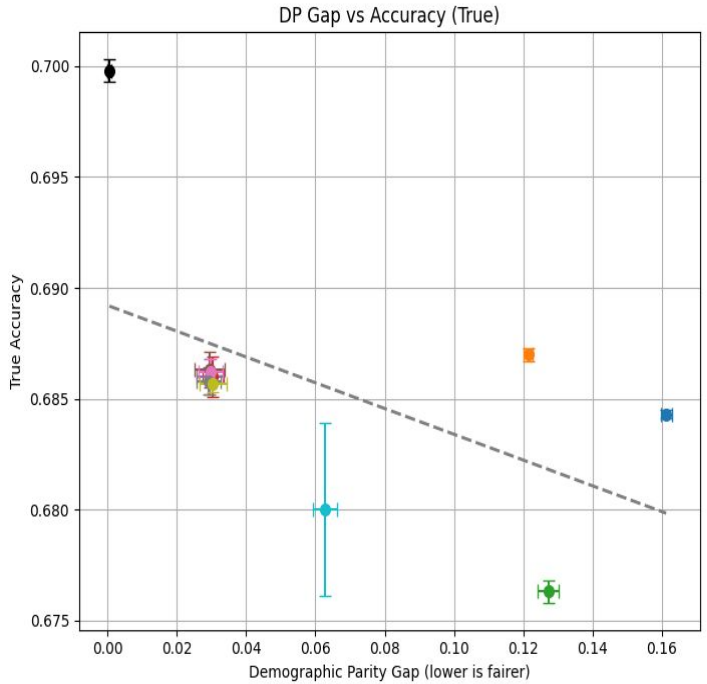
$$x = u + \text{Bias\_strength} \times (b == 0 \ ? \ +1 : -1) \times \text{sign}(u)$$

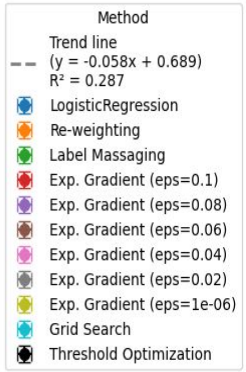
- **Score Distribution Plot**: Group $b = 0$ scores are skewed outward, while $b = 1$ scores are compressed due to the sign-based nonlinear transformation.
- **Positive Rate Plot:** Group $b = 1$ receives significantly more positive labels, exposing how the nonlinear bias disrupts label balance.

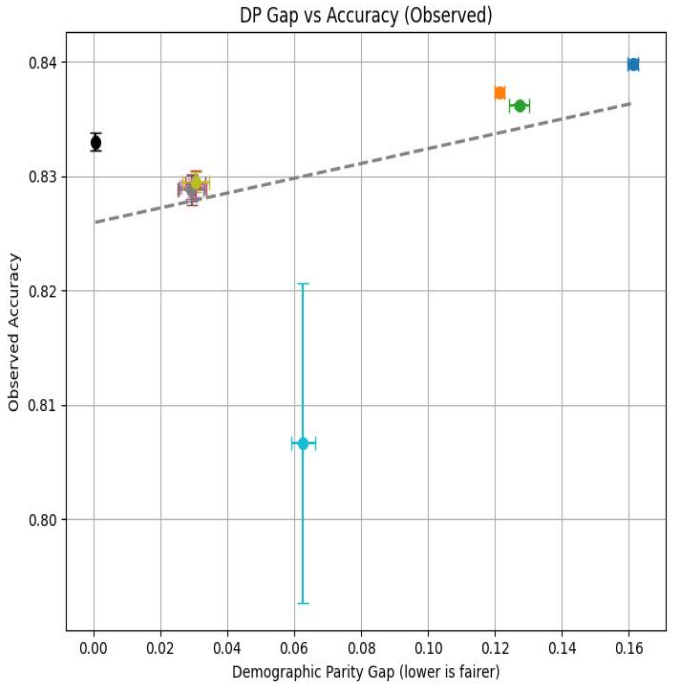


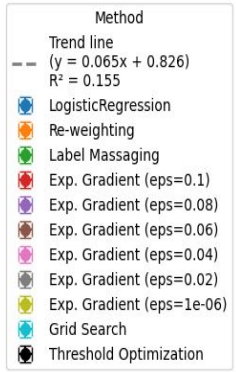


Results of Nonlinear Bias









Fairness–Accuracy Trade-off Analysis: Nonlinear Bias

- **Logistic Regression** shows **high observed accuracy** but **overfits to biased patterns,** with low true accuracy and a large DP gap
- **Threshold Optimization** again dominates — delivering the **highest true accuracy and near-zero bias. Best overall fairness–accuracy balance in this complex scenario.**
- **Exponentiated Gradient** offers **consistent fairness improvements** with only **a small trade-off in true accuracy. Grid Search** performs moderately across all metrics — **more fair than baseline**, but clearly outperformed by TO and EG.
- **Re-weighting** and **Label Massaging** reduce bias slightly, but **don't significantly improve fairness or accuracy.**
- **Trade-off is moderate**: fairness improvements help true accuracy, but **complex distortion makes gains harder** than in mild-bias cases:
  - **R^2 = 0.287 (True)** → Some correlation: **fairer models generalize slightly better.**
  - **R^2 = 0.155 (Observed)** → Weak pattern: observed accuracy **does not reflect fairness** well.
- In **nonlinear bias, post-processing methods shine**, while others may struggle to disentangle real patterns from distortion.

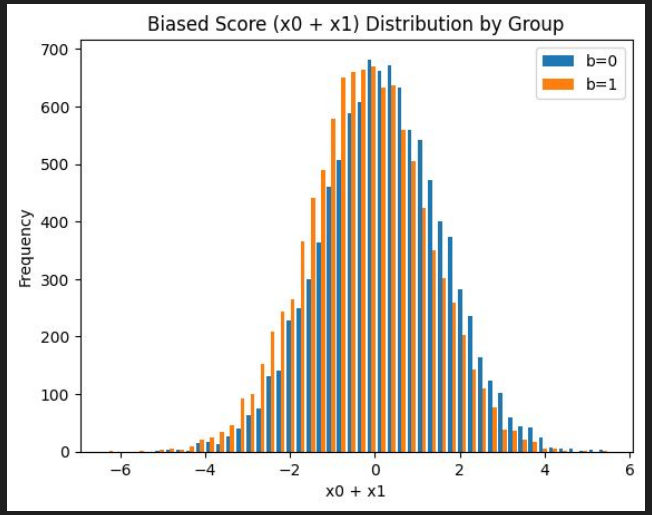
Scenario 5 – Interaction Bias

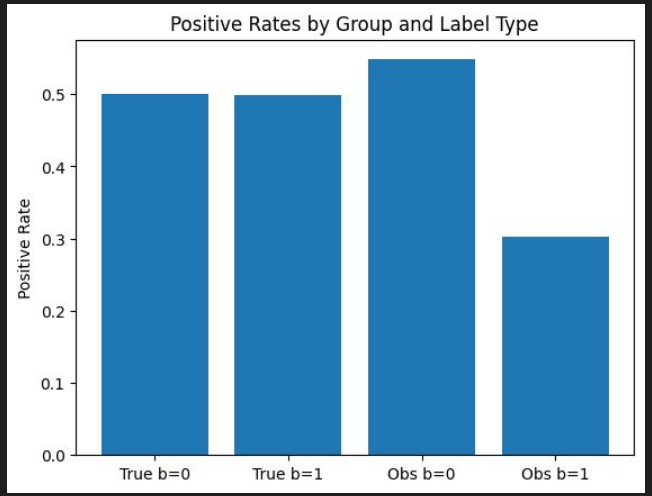
- **Description:**
  - Protected Attribute: Binary ($b \in \{0,1\}$)
  - Bias Type: Additive on features + Interaction on labels
  - Bias Strength: ± 1
  - Mechanism:

$$\text{Labels: } y = (x_0 + x_1 + 1.0 \cdot b \cdot x_0 \cdot x_1 + \text{noise} \geq 0)$$

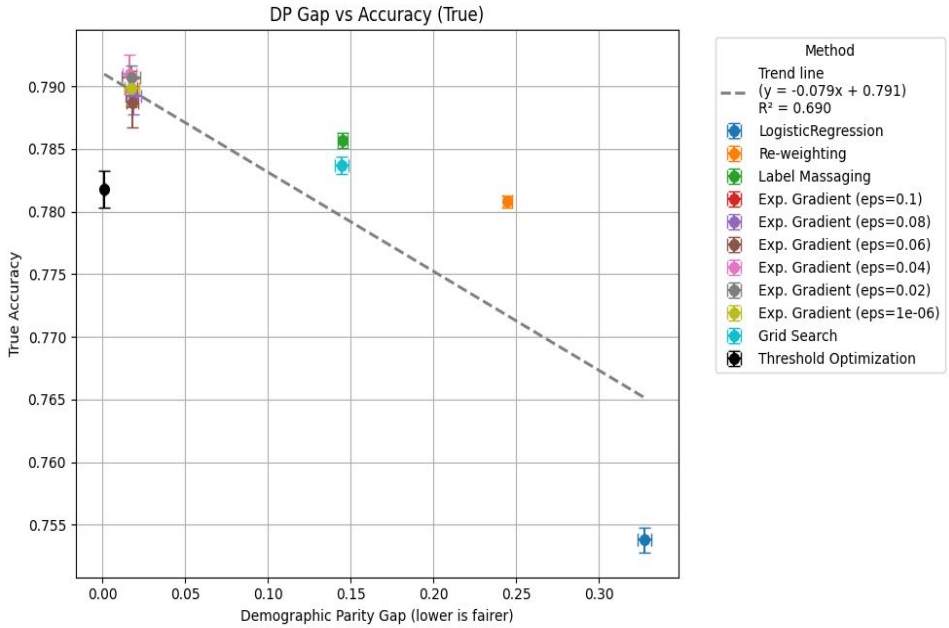
- **Score Distribution Plot**: The feature distributions for both groups are **visually similar**, suggesting **no obvious feature-based bias** — similar to mild additive bias.

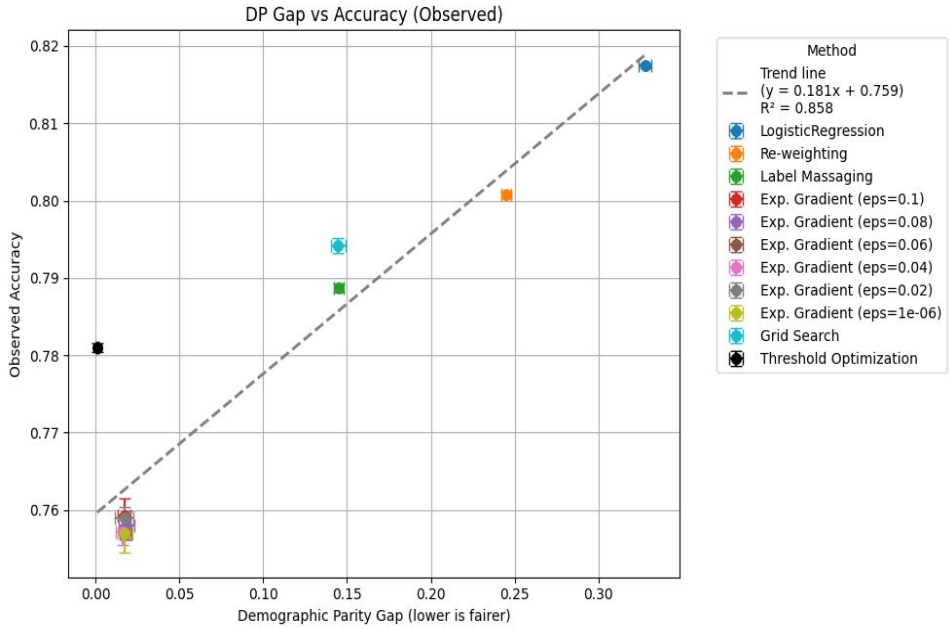
- **Positive Rate Plot:** A **sharp disparity emerges: group b=1 receives far fewer positive labels** due to the **interaction term in the label function.**





Results of Interaction Bias





Fairness–Accuracy Trade-off Analysis: Interaction Bias

- **Logistic Regression** shows **high observed accuracy** but **severe bias** — large DP gap and low true accuracy.
- **Threshold Optimization** achieves **near-zero bias** and solid accuracy — **best when fairness is critical.**
- **Exponentiated Gradient** offers the **best balance of fairness and true accuracy**, outperforming most other methods
- **Pre-processing methods** (Label Massaging, Re-weighting) improve fairness slightly but are **less effective**. Grid Search show moderate fairness improvement but can't match the top methods in this complex scenario.
- **Trade-off is strong**:
  - **True accuracy improves as fairness increases ($R^2 = 0.69$)**
  - **Observed accuracy drops due to label distortion ($R^2 = 0.858$)**
- **Advanced methods are essential** — interaction bias is complex and can't be solved with simple fixes.

Fairness–Accuracy Trade-off: Overall Conclusion

- **Threshold Optimization** consistently delivered **best fairness** (near-zero DP gap) and **top true accuracy** — especially strong in complex bias scenarios.
- **Exponentiated Gradient** offered the **ideal balance trade-off** between fairness and accuracy — ideal when modifying the training process is allowed.
- **Grid Search** showed **moderate fairness gains**, useful in mild/moderate bias but less robust in complex scenarios.
- **Pre-processing methods** (Re-weighting, Massaging) were **simple and effective in light bias,** but **failed under severe distortions.**
- **Observed accuracy can be misleading:** models that look good on biased labels often **perform poorly on true labels** — so it's important to **evaluate fairness carefully.**

Fairness–Accuracy Trade-off: Overall Conclusion

- **Bias severity shapes strategy:**
    - Mild bias → many methods work well
    - Complex bias → **post-processing/in-processing needed**

- **Key Insight:**

  Fairness isn't free, but it's achievable.
  By aligning **bias type with the right technique**, we can build models that are **both equitable and effective.**

Limitations

- **Synthetic Bias Only:**

  The study was limited to artificially generated bias scenarios — real-world bias may behave differently.

- **Single Fairness Metric (DPG):**

  Only Demographic Parity Gap was used; other metrics (e.g., Equal Opportunity) may offer deeper insight.

- **Binary Sensitive Attribute:**

  All experiments assumed a binary protected group; real-world data often includes multiple or continuous attributes.

Future Work

- **Apply to Real-World Datasets:** Test fairness enforcement methods on real datasets to assess generalization beyond synthetic settings.
- **Explore Alternative Fairness Metrics:** Include Equal Opportunity, Equalized Odds, and calibration for a more complete evaluation.
- **Vary Bias Type & Strength:** Investigate more diverse and extreme bias patterns by tweaking bias mechanisms and strengths.
- **Try Additional Fairness Methods:** Evaluate other fairness techniques (e.g., adversarial debiasing, fairness constraints in deep learning).

Thank you!

Contact me

mehrega.nazarmohseni@studio.unibo.it