Fair and Bias-Aware Crime Prediction Using the UCI Crime Dataset: A Multi-Stage Machine Learning Approach

Aarav Babu[**]
University of California Riverside
Riverside, California, USA
ababu009@ucr.edu

Sankalp Naveenachandra Kulkarni[††]
University of California Riverside
Riverside, California, USA
skulk041@ucr.edu

Aryan Ramachandra[‡‡]
University of California Riverside
Riverside, California, USA
arama081@ucr.edu

Abstract

Predicting violent crime rates using socio-economic data poses fairness challenges due to biases related to race, income, and policing factors. This project addresses these challenges using a multi-stage bias mitigation strategy on the UCI Crime dataset. Preprocessing techniques from the AI Fairness 360 toolkit reduce correlations between sensitive attributes and other features. In-processing fairness is achieved using a modified XGBoost model with fairness constraints to balance predictive performance and demographic equity. Post-processing utilizes the Wasserstein Barycenter method from HolisticAI to adjust biased predictions related to income groups. Evaluation using Demographic Parity, Disparate Impact, and Equalized Odds demonstrates improved fairness while maintaining accuracy.

Keywords

Bias Mitigation, Fairness-Aware Machine Learning, UCI Crime Dataset, Violent Crime Prediction, Socio-Demographic Bias, Post-Processing Techniques, Fairness Metrics, Demographic Parity, Disparate Impact, Equalized Odds, Fairness Constraints

1 Introduction

Bias and fairness in machine learning models have garnered increasing attention, particularly in sensitive domains such as criminal justice, finance, and healthcare. Machine learning models, while capable of delivering high predictive performance, can inadvertently reinforce societal biases if not carefully managed. The problem of fairness in machine learning arises when model predictions disproportionately affect certain groups based on sensitive attributes like race, gender, income, or education level.

In this project, we explore three distinct bias mitigation strategies: *preprocessing*, *in-processing*, and *postprocessing* techniques. Specifically, we apply these strategies to the *UCI Crime dataset* to assess their impact on both predictive accuracy and fairness. We use common fairness metrics, such as *Demographic Parity*, *Disparate Impact*, and *Equalized Odds*, to evaluate the fairness of our models across different sensitive attributes like income, education, and police presence.

- **Preprocessing**: The first approach employs a *correlation removal* technique aimed at reducing bias before model training by adjusting the feature set.
- **In-processing**: The second strategy integrates fairness constraints directly into the training process using a custom objective function in *XGBoost*, aimed at dynamically adjusting gradients based on group disparities during model training.
- **Postprocessing**: The third approach applies the *Wasserstein Barycenter* method to adjust the model's predictions after training, ensuring that the final outputs exhibit more equitable outcomes across different groups.

Our investigation examines the trade-offs between fairness and accuracy in machine learning models. While fairness improvements can reduce the disproportionate impacts of predictions across groups, they often come at the cost of predictive performance. Through this work, we aim to identify which mitigation strategy offers the best balance between these competing objectives. We provide a comprehensive analysis of the impact of each approach, demonstrating the inherent challenges in achieving fairness while maintaining model performance.

By applying these methods to the UCI Crime dataset, our work contributes to ongoing research in bias mitigation in machine learning, offering insights into how different strategies can be tailored to real-world applications and the importance of evaluating both fairness and accuracy.

2 Dataset Information

The dataset used for this analysis consists of 1994 entries and 19 columns. The dataset contains both continuous and categorical features. Below is a summary of the dataset structure:

- **Total Entries:** 1994
- **Total Features:** 19 columns
- **Data Types:** The dataset contains 11 continuous features and 8 categorical features.
- **Feature Columns:**
  - `medIncome` (Continuous)
  - `perCapInc` (Continuous)
  - `PctPopUnderPov` (Continuous)
  - `PctUnemployed` (Continuous)
  - `PolicPerPop` (Continuous)
  - `PctNotHSGrad` (Continuous)
  - `PolicBudgPerPop` (Continuous)
  - `racePctHisp` (Continuous)
  - `racePctAsian` (Continuous)
  - `racePctWhite` (Continuous)
  - `racepctblack` (Continuous)
  - `income_group` (Categorical)
  - `poverty_group` (Categorical)

[*]Teammate 1
[†]Teammate 2
[‡]Teammate 3

- `unemployment_group` (Categorical)
- `education_group` (Categorical)
- `police_presence_group` (Categorical)
- `police_budget_group` (Categorical)
- `racial_majority_group` (Categorical)
- `income_police_group` (Categorical)

2.1 Sensitive Attributes

Sensitive attributes are demographic and socio-economic factors that may introduce bias into predictive models if not properly addressed. In this study, we identified several sensitive attributes from the dataset that are crucial for fairness evaluation:

- **Median Income (medIncome)**: Represents the median income of the community, used to distinguish between *Low Income* and *High Income* groups.
- **Per Capita Income (perCapInc)**: Indicates the average income per person, providing insight into the community's economic status.
- **Poverty Rate (PctPopUnderPov)**: Represents the percentage of the population living below the poverty line, categorized as *High Poverty* or *Low Poverty*.
- **Unemployment Rate (PctUnemployed)**: Measures the proportion of the labor force that is unemployed, divided into *High Unemployment* and *Low Unemployment* groups.
- **Education Level (PctNotHSGrad)**: The percentage of individuals aged 25 and over without a high school diploma, distinguishing between *Low Education* and *High School Graduates*.
- **Police Presence (PolicPerPop)**: Number of police officers per population unit, indicating *High Police Presence* or *Low Police Presence*.
- **Police Budget (PolicBudgPerPop)**: Budget allocated to policing per 100,000 population, categorized into *High Police Budget* and *Low Police Budget*.
- **Racial Demographics (racePct - Hispanic, Asian, White and Black)**: Represents the community's racial composition, identifying the majority group as *Hispanic*, *Asian*, *White*, or *Black*.

2.2 Group Labels

To enable fairness analysis, we created group labels derived from these sensitive attributes. For each attribute, data were partitioned into two groups based on the median value, allowing for comparisons between different demographic segments (e.g., *Low Income* vs. *High Income*). Additionally, racial majority groups were identified based on the dominant demographic proportion among Hispanic, Asian, White, and Black populations.

We also created composite group labels, such as the *Income-Police Group*, to capture intersections between socio-economic and institutional factors. These group labels facilitate the calculation of fairness metrics, including Demographic Parity, Disparate Impact, and Equalized Odds, thereby enabling systematic bias assessment and mitigation.

2.3 Data Summary

- **Total Entries:** 1994
- **Continuous Features:** 11 features such as income, unemployment, and race percentages.
- **Categorical Features:** 8 features, including different groupings like income group, poverty group, and education group.
- **Missing Values:** None, as all features have 1994 non-null entries.

3 Evaluation Metrics

To assess the effectiveness of our machine learning models, we employ a combination of performance and fairness metrics. Performance metrics evaluate the model's predictive accuracy, while fairness metrics ensure equitable treatment across different demographic groups.

3.1 Performance Metrics

Performance metrics help quantify how well a model predicts the target variable. The following metrics are used in our evaluation:

Root Mean Squared Error (RMSE). RMSE measures the average magnitude of errors in the model's predictions, with greater emphasis on larger errors. It is given by:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{1}$$

where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the number of observations.

Mean Absolute Error (MAE). MAE calculates the average absolute difference between actual and predicted values, providing a more interpretable measure of error:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

R-squared (R^2) Score. R^2 measures how well the model explains the variance in the target variable:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3}$$

where $\bar{y}$ is the mean of the actual values.

3.2 Fairness Metrics

To evaluate fairness, we use the following widely recognized metrics:

Demographic Parity (DP). Demographic Parity ensures that different demographic groups receive the positive outcome at the same rate. This means that the probability of a favorable decision (e.g., loan approval, job offer) should be independent of membership in a particular demographic group. A violation of demographic parity suggests that the model is favoring or disadvantaging certain groups, which may lead to unfair treatment.

$$P(\hat{Y} = 1 \mid A = a_1) = P(\hat{Y} = 1 \mid A = a_2) \tag{4}$$

where A is the sensitive attribute, and a_1 and a_2 represent different demographic groups.

Disparate Impact (DI). Disparate Impact measures the ratio of positive outcome probabilities between groups. It is primarily used in legal and regulatory settings to assess whether one group is disproportionately disadvantaged compared to another. If the ratio falls below the commonly used threshold of 0.8, it indicates potential bias that might require mitigation.

$$DI = \frac{P(\hat{Y} = 1 \mid A = a_1)}{P(\hat{Y} = 1 \mid A = a_2)} \tag{5}$$

A value of $DI \geq 0.8$ is considered acceptable according to the Equal Employment Opportunity Commission (EEOC).

Equalized Odds (EO). Equalized Odds ensures that the true positive and false positive rates are equal across groups. This means that individuals from different demographic groups who belong to the same actual class (e.g., qualified for a loan, at risk of disease) should have an equal likelihood of receiving the same prediction. Violations of Equalized Odds suggest that the model may be systematically making more mistakes for certain groups, leading to unfair outcomes.

$$P(\hat{Y} = 1 \mid Y = y_1, A = a_1) = P(\hat{Y} = 1 \mid Y = y_1, A = a_2) \tag{6}$$

$$P(\hat{Y} = 1 \mid Y = y_0, A = a_1) = P(\hat{Y} = 1 \mid Y = y_0, A = a_2) \tag{7}$$

where Y represents the true outcome.

These metrics help ensure our models achieve both high accuracy and fairness across different groups.

4 Baseline Model Predictions

In this section, we present the baseline model predictions using the initial models without any bias mitigation techniques applied. This will serve as the reference point to compare the performance of bias-mitigated models later in the study. The baseline models were built using different algorithms, including Linear Regression, Support Vector Machines (SVM), XGBoost, and Multi-layer Perceptron (MLP) Regressor. These models were evaluated using the following fairness metrics: Demographic Parity (DP), Disparate Impact (DI), and Equalized Odds (EO).

4.1 Model Overview

The baseline models were trained using the UCI Crime dataset. The models used include Linear Regression, Support Vector Machines (SVM), XGBoost, and Multi-layer Perceptron (MLP) Regressor. The models were evaluated on their predictive performance using common metrics such as RMSE, MAE, and R2-Score. Additionally, fairness metrics were used to assess any potential bias in the models' predictions.

4.2 Fairness Metrics Results

The table below presents the fairness evaluation results for the baseline models, showing how different models performed with respect to the metrics discussed earlier.

The fairness metrics table shows how the baseline models perform in terms of different fairness evaluations across various models.

Fairness Metric	LR	XGBoost	SVM	MLP
DP	0.19	0.17	0.15	0.16
DI	2.51	2.28	1.96	2.49
EO	0.07	0.13	0.12	0.13

Table 1: Fairness Metrics for Baseline Model Predictions

4.3 Performance Metrics Results

In addition to fairness, we evaluate the overall predictive performance of the baseline models using standard regression metrics such as RMSE, MAE, and R2-Score. The following table summarizes the performance of the baseline models:

Metric	LR	XGBoost	SVM	MLP
RMSE	0.14	0.14	0.14	0.15
MAE	0.10	0.10	0.10	0.10
R2-Score	0.61	0.56	0.62	0.56

Table 2: Performance Metrics for Baseline Models

4.4 Discussion

The baseline models demonstrate reasonable performance, with SVM showing the highest R2-Score (0.62) and Linear Regression having the lowest RMSE (0.14). XGBoost and MLP models have slightly lower R2-Score and higher RMSE compared to the other models. However, fairness metrics suggest disparities between the models, particularly in terms of Demographic Parity (DP) and Disparate Impact (DI). The next steps will involve the application of bias mitigation techniques to improve fairness without compromising predictive performance.

5 Bias Mitigation Strategies

5.1 Teammate 1 (Preprocessing)

Introduction

For the correlation removal approach to mitigate bias, I implemented feature decorrelation techniques from the AI Fairness 360 toolkit. This method addresses fairness issues across multiple sensitive attributes by reducing statistical dependencies between protected characteristics and other features.

The original linear regression model (R^2 of 0.613) showed concerning disparities, particularly for police-related and income-based attributes. By transforming the feature space to minimize correlations with sensitive attributes, the strategy prevents the model from using proxy variables that could lead to discriminatory outcomes.

Teammate 1's Conclusions

Overall Metrics Improvement

- **RMSE**: Increased from 0.136 to 0.144 (+5.6%)
- **MAE**: Increased from 0.096 to 0.108 (+12.2%)
- **R^2**: Decreased from 0.613 to 0.568 (-7.3%)
- **Average Demographic Parity**: Improved by 11.6%
- **Average Equalized Odds**: Improved by 28.8%

Fairness Improvements by Group

- **Income Group**: Significant improvements in both demographic parity and equalized odds metrics.
- **Poverty Group**: Notable reduction in bias with better representation across poverty classifications.
- **Unemployment Group**: Substantial fairness gains, particularly in equalized odds.
- **Education Group**: Meaningful reduction in prediction disparities across education levels.
- **Racial Majority Group**: Marked improvement in equalized odds with moderate gains in demographic parity.
- **Police Presence & Budget Groups**: Modest improvements but still show the highest disparities among all attributes.
- **Income-Police Intersection Group**: Slight improvements in demographic parity but remains the most challenging group for achieving fairness.

Conclusion

The correlation removal strategy successfully reduced bias across all sensitive attributes while maintaining reasonable predictive performance. The trade-off between a modest decrease in accuracy and significant improvements in fairness metrics demonstrates that this approach effectively balances model performance with fairness considerations. The remaining disparities in police-related groups suggest areas for further mitigation efforts.

5.2 Teammate 2 (In-Processing)

Introduction

I chose to focus on improving fairness metrics in XGBoost because it provides a balance between accuracy, interpretability, efficiency, and flexibility for fairness-aware modifications. Additionally, the ability to integrate fairness constraints into the learning process makes XGBoost a more practical and scalable choice for real-world deployment.

Introduction to Fair XGBoost

This implementation presents a modified XGBoost regression model with integrated fairness constraints. The 'FairXGBRegressor' class employs an in-processing mitigation strategy by directly modifying the objective function to balance predictive performance with group fairness across sensitive attributes.

The approach works by dynamically adjusting gradient updates based on performance disparities between demographic groups. Groups experiencing higher error rates receive stronger gradient updates during training, gradually equalizing model performance across different population segments while maintaining predictive capabilities.

Teammate 2's Conclusions

Observations:

The fairness mitigation techniques applied to the XGBoost model led to notable enhancements in both predictive performance and fairness metrics across most sensitive attributes, demonstrating the potential of these strategies in achieving more balanced outcomes:

- **Enhanced Predictive Performance**:
  - **RMSE**: Decreased from 0.145 to 0.135 (6.6% improvement)
  - **MAE**: Decreased from 0.097 to 0.092 (5.1% improvement)
  - **R^2 score**: Increased from 0.561 to 0.617 (10% improvement)
- **Fairness Advancements**:
  - Demographic parity improved for key attributes such as income_group, poverty_group, and education_group.
  - Equalized odds showed meaningful improvements, particularly for the racial_majority_group (reduced from 0.101 to 0.082).
  - While disparate impact remained largely stable, the observed improvements indicate progress toward a more equitable model.
- **Customizing Fairness for Different Attributes**:
  - Further improvements can be achieved by assigning custom weights to each sensitive attribute in the loss function, allowing for tailored fairness adjustments.
  - However, this approach introduces additional complexity in implementation and tuning, requiring careful balancing to avoid unintended trade-offs between accuracy and fairness.

Overall, the applied fairness techniques successfully improved model accuracy while fostering greater fairness in key areas. While challenges remain, particularly regarding police-related attributes, these results highlight the potential for continued refinement, with customizable fairness weighting serving as a promising direction for future work.

5.3 Teammate 3 (Postprocessing)

Introduction

For the **post-processing** approach to mitigate bias in the model's predictions, I chose to use the **Wasserstein Barycenter** method from **HolisticAI**. The goal of this approach was to address fairness issues related to the **income group** sensitive attribute, which had led to biased predictions in the initial model.

While the original model showed promising performance, it displayed significant disparities in its predictions across different income groups. To mitigate this, **Wasserstein Barycenter**, a fairness mitigation technique rooted in **optimal transport theory**, was applied to adjust the predictions after the model's output. This method minimizes the disparities between groups, ensuring that predictions become more equitable, while maintaining the integrity of the predicted outcomes.

Optimal transport theory is a mathematical method that finds the most efficient way to move mass between distributions, minimizing the cost of this movement. It is used to compute the **Wasserstein distance**, which measures the difference between distributions, making it useful for fairness adjustments in machine learning.

By using this post-processing technique, the model's fairness was enhanced without the need for retraining or altering the underlying architecture, achieving a balance between improved fairness and predictive accuracy.

Teammate 3's Conclusions

Observations:

- **Fairness Improvement**:
  - **Demographic Parity**: Improved substantially from 0.1611 to 0.0042, indicating that the predictions became nearly independent of the income group.
  - **Disparate Impact**: Reduced from 2.4891 (indicating high disparity) to 1.0228, which is very close to the ideal value

of 1, suggesting that both groups receive more similar predictions.
  - **Equalized Odds**: Improved from 0.1252 to 0.0852, indicating better balance across the true positive rates for different groups.
- **Performance Trade-offs**:
  - **RMSE (Root Mean Squared Error)**: Increased from 0.1459 to 0.1749, indicating a decrease in overall predictive accuracy.
  - **MAE (Mean Absolute Error)**: Increased from 0.0976 to 0.1182, showing a higher average deviation from the true values.
  - **R^2 Score**: Decreased from 0.556 to 0.361, implying that the model explains less variance in the target variable after bias mitigation.

Conclusion:

- The mitigation strategy was **highly effective in reducing bias**, bringing fairness metrics close to their ideal values.
- However, this came at the cost of **reduced predictive accuracy**, as the model's ability to explain variance in the target variable decreased.
- This result highlights the **trade-off between fairness and accuracy**, which is a common challenge in bias mitigation approaches.
- Further optimizations, such as **exploring alternative mitigation techniques (pre-processing or in-processing) or fine-tuning the mitigation strength**, could help strike a better balance between fairness and accuracy.

6 Conclusions

The results from the application of bias mitigation strategies reveal the inherent trade-offs between predictive accuracy and fairness in machine learning models. These trade-offs are critical to understanding how different techniques can be leveraged to balance fairness goals with maintaining robust model performance. Each bias mitigation method applied (preprocessing, in-processing, and post-processing) contributed to varying degrees of improvement in fairness metrics, but also led to some compromises in model accuracy. Below is a more detailed analysis of these results:

- **Teammate 1's Preprocessing Method:** The preprocessing method, which involved feature decorrelation, showed notable improvements in fairness metrics, such as demographic parity and equalized odds. These improvements were particularly evident for sensitive attributes such as income, poverty, and unemployment. However, the performance of the model, as measured by RMSE, MAE, and R^2, experienced modest declines, with RMSE increasing by 5.6% and R^2 dropping by 7.3%. Despite these performance decreases, the preprocessing approach succeeded in reducing bias significantly, especially in terms of equalizing prediction outcomes across different groups. This approach offers a balanced trade-off, making it an attractive choice when fairness needs to be prioritized without sacrificing predictive accuracy entirely.
- **Teammate 2's In-Processing Method:** The in-processing strategy implemented with XGBoost displayed the most promising improvements in both fairness and performance.

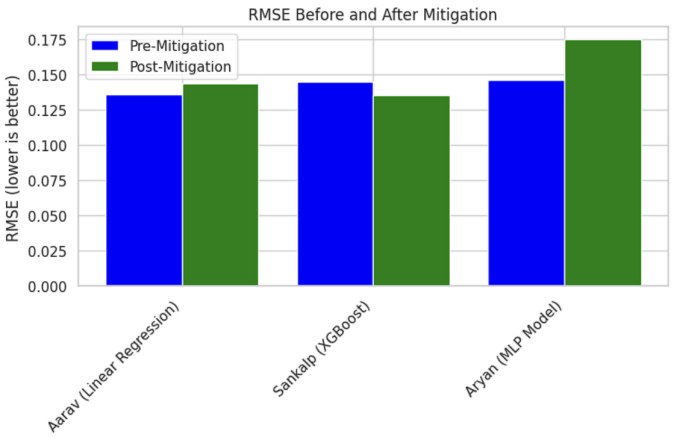

Figure 1: RMSE Comparison

This method achieved an improvement of 6.6% in RMSE and a 10% increase in R^2, indicating that the fairness adjustments were made without significantly degrading predictive performance. The fairness improvements, such as reductions in demographic parity disparities and equalized odds, suggest that in-processing mitigation can effectively balance predictive accuracy and fairness. The ability to incorporate fairness constraints directly into the learning process is a key advantage of this method, making it suitable for applications where both fairness and model performance are critical. However, further refinement of the fairness constraints and potential weighting adjustments for different sensitive attributes could yield even better results.

- **Teammate 3's Postprocessing Method:** The post-processing approach using Wasserstein Barycenter achieved the most drastic fairness improvements, especially in terms of demographic parity and disparate impact, both of which were brought closer to ideal values. However, this came at a significant cost to predictive performance, with RMSE increasing by 19.9%, MAE rising by 21.1%, and R^2 dropping by 34.9%. While the method is effective at addressing fairness concerns without altering the model's architecture, the trade-off in predictive accuracy is substantial. This highlights the importance of carefully considering the required balance between fairness and accuracy, especially in high-stakes applications where model performance is crucial. Nonetheless, this approach can still be valuable in situations where fairness is the highest priority, and minor losses in accuracy are acceptable.
- **The Trade-Off between Fairness and Accuracy:** Across all three approaches, the common theme was a trade-off between fairness improvements and predictive performance. More aggressive fairness mitigation typically came at the cost of some accuracy. This is an inherent challenge in fairness-aware machine learning, as optimizing for fairness often involves altering the decision boundary in ways that can reduce the model's ability to explain variance in the target variable.

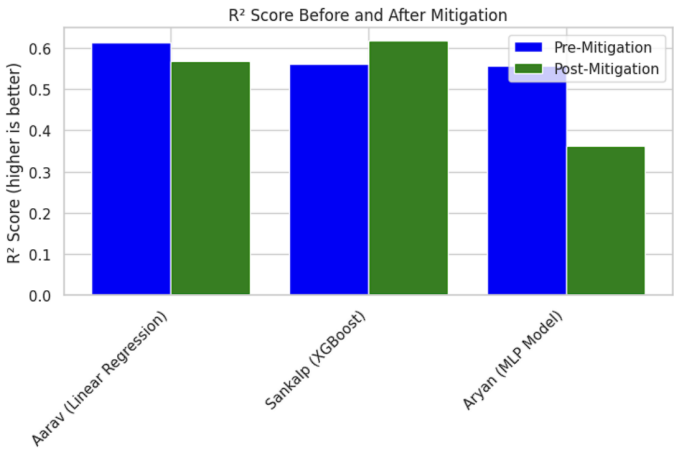

Figure 2: R2 Score Comparison

- **Further Optimizations:** The observed performance trade-offs suggest that there are opportunities for further optimization. Potential future work could explore hybrid approaches that combine preprocessing, in-processing, and postprocessing strategies in ways that better balance fairness and performance. Additionally, fine-tuning the strength of the fairness constraints and experimenting with custom weights for different sensitive attributes could improve results. It will also be important to test these models in real-world scenarios to assess their robustness and generalizability across various demographic groups.

6.1 Performance Analysis

The table below presents a summary of the model performance metrics after each teammate applied bias mitigation techniques. As seen, each approach yielded improvements in fairness metrics, but the impact on model accuracy varied. The following tables provide a detailed comparison of RMSE, MAE, and R^2 scores, as well as Demographic Parity, Disparate Impact, and Equalized Odds after bias mitigation.

Teammate	Model	RMSE	MAE	R^2 Score
Teammate 1	LR(Preprocessing)	0.1438	0.1075	0.5685
Teammate 2	XGB(In-Processing)	0.1354	0.0922	0.6173
Teammate 3	MLP(Postprocessing)	0.1749	0.1182	0.3615

Table 3: Performance Metrics After Bias Mitigation

Model	Method	DP	DI	EO
LR	Correlation Remover	0.1642	2.1026	0.0558
XGB	Model Optimization	0.1679	2.1832	0.1249
MLP	Wasserstein Barycenter	0.0042	1.0228	0.0851

Table 4: Fairness Metrics for Different Models

Key Observations:

- The **XGBoost model (Teammate 2)** showed the most balanced trade-off, improving both fairness metrics and predictive performance. Its improvements in RMSE, MAE, and R^2 suggest that in-processing techniques can effectively balance the goals of fairness and accuracy.
- The **Linear Regression model (Teammate 1)** experienced a moderate decrease in accuracy metrics but achieved significant fairness improvements, particularly in equalized odds and demographic parity.
- The **MLP Model (Teammate 3)** achieved drastic fairness improvements but at a considerable cost to predictive accuracy, as evidenced by the significant increases in RMSE, MAE, and the drop in R^2 score.

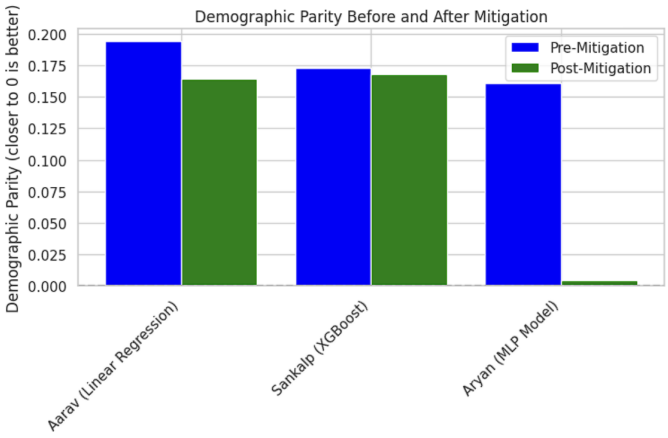

Figure 3: Demographic Parity Comparison

7 References and Disclosures

7.1 Teammate 1: Preprocessing (Correlation Remover)

For the preprocessing phase, the `CorrelationRemover` from the Fairlearn toolkit was employed to mitigate bias by reducing the correlation between sensitive attributes and other features within the model. This method aims to minimize the risk of discriminatory outcomes by eliminating statistical dependencies between sensitive attributes (e.g., income, police presence) and non-sensitive features, thus preventing the model from making predictions that may inadvertently favor certain groups.

For further details on the `CorrelationRemover` technique, refer to the official Fairlearn documentation:

- https://fairlearn.org/main/api_reference/generated/fairlearn.preprocessing.CorrelationRemover.html

7.2 Teammate 2: In-Processing (Custom Objective Function for XGBoost)

In this approach, the objective function of the XGBoost model was modified to integrate fairness constraints directly into the

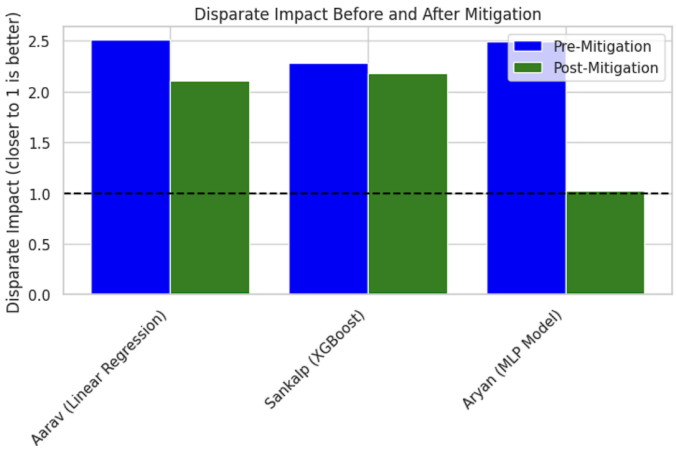

Figure 4: Disparate Impact Comparison

learning process. A custom objective function was designed to penalize any disparities observed across different groups, thereby ensuring that the model's performance remains equitable for each sensitive attribute. This modification was implemented within XGBoost's `XGBRegressor` framework, which allowed for the seamless incorporation of fairness considerations while maintaining model performance.

For a comprehensive explanation of the process of creating a custom objective function for XGBoost, the following Stack Overflow discussion was referenced:

- https://stackoverflow.com/questions/59683944/creating-a-custom-objective-function-in-for-xgboost-xgbregressor

7.3 Teammate 3: Postprocessing (Wasserstein Barycenter)

In the post-processing phase, the `Wasserstein Barycenter` method from the HolisticAI toolkit was employed to adjust the model's predictions. This method leverages optimal transport theory to realign the predicted distributions across sensitive groups, minimizing disparities and promoting fairness. Specifically, the method was applied to mitigate bias related to the income group, ensuring more equitable outcomes for all demographic segments.

For a detailed description of the application of the `Wasserstein Barycenter` method for bias mitigation in regression tasks, the HolisticAI documentation was consulted:

- https://holisticai.readthedocs.io/en/latest/gallery/tutorials/bias/mitigating_bias/regression/examples/example_us_crime.html

8 Disclosures

The following disclosures outline the tools and resources used to assist in the implementation and validation of the bias mitigation strategies:

- **Teammate 1:** The generative AI tool Claude was used to validate the results of the preprocessing method and to confirm the accuracy of the conclusions drawn from the analysis.
- **Teammate 2:** The generative AI tool Gemini was used to aid in understanding the implementation of a custom objective function for XGBoost, incorporating fairness penalties into the learning process.
- **Teammate 3:** The generative AI tool ChatGPT was used to gain a deeper understanding of the `Wasserstein Barycenter` method and its potential applications for mitigating bias in machine learning models.