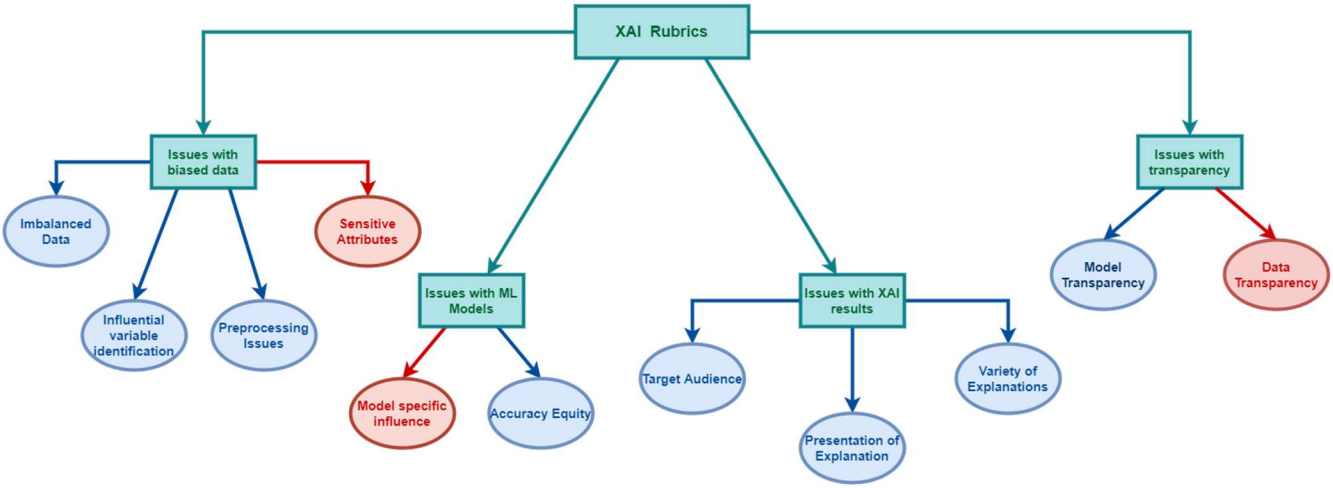
# An Analysis of the Implications of Fairness Resolution on Multiple Fairness Metrics

Rohan Sanjay Pawar
*Georgia Institute of Technology*
*rohan.pawar@gatech.edu*

Aditya Salian
*Georgia Institute of Technology*
*asalian@gatech.edu*

Tianyi Liu
*Georgia Institute of Technology*
*tliu422@gatech.edu*




## Abstract

The introduction of fairness in Artificial Intelligence is an active field of research to explain the interpretability and transparency of AI models, insensitive and life-changing decisions of candidates in cases like risk-based sentencing and loan applications. However, there exist multiple definitions of fairness, adding a potential subjective element to machine learning models. However, due to the subjective definition of fairness, accounting for one type of fairness might have implications on other types' definitions on fairness. With our project, we work on multiple biased datasets to benchmark the performance of models before and after some fairness constraints are resolved and analyze the impact of these fairness models in terms of fairness definitions. The benchmark rubrics identified in the scope of this result can aid improve the explainable AI frameworks to better analyze their post-processing steps.

## 1. INTRODUCTION

### 1.1 Problem Motivation

With the growing ubiquity of machine learning algorithms in our everyday lives, there is no denying that these algorithms play a significant part in the decision-making process of almost every individual [1]. This also implies that due to these algorithms playing a crucial part in this decision-making process, machine learning algorithms must be unbiased in their predictions to avoid being unfair to data points belonging to a certain category that does not contribute to the prediction values. In our research towards fairness in machine learning, we observed that there is no formal definition of fairness. Furthermore, due to this lack of clarity in defining fairness, a lot of debate precedes on concluding whether or not a particular dataset is "fair".

Artificial intelligence has also been previously used to make predictions in subjective topics like criminal risk assessment algorithms and loan applications to eliminate human bias and derive data-driven recommendations. In this context, although it is clear to the human understanding that attributes like age, race and gender are not indicative of whether or not a person is a potential risk to society. Machine learning algorithms must be able to explain the statistical correlations based on which a certain model makes predictions. Current explainable AI (XAI) tools like AI360, FairLearn, and LIME address these concerns but they too, come with their set of limitations. These tools cannot detect issues with data preprocessing, sensitive attributes like race and gender, and the choice of a model for a particular dataset. In this project, using Microsoft's Fairlearn, we have studied two topics- firstly, how satisfying one definition of fairness (equalized odds) with Fairlearn affects other fairness definitions, like predictive parity and overall accuracy equality, and the overall accuracy of the model on the dataset; and secondly, comparing the results of using Fairlearn's post-processing for solving equalized odds affects a Logistic Regression model and a Random Forest Model. Furthermore, we have also attempted to work on developing metrics for evaluating the transparency and fairness of models in a more objective manner [1]. Furthermore, to get an exhaustive understanding of explainable AI, we compare Fairlearn [2] against IBM's XAI-360 framework [3] for mitigating all the fairness definition violations. This is critical in understanding the core working of these two fairness frameworks in terms

of the effects that might be propagated by mitigating a given fairness definition violation.

## 1.2 Current Approaches

In our research for previous work in explainable AI, we discovered that current explainable AI research generally falls under two categories of strategies. First, researchers have resorted to using machine learning models that are naturally easier for humans to comprehend and interpret [4]. As machine learning models became deeper and more complex, researchers gradually shifted focus towards developing tools to help explain the decision-making from more complex models [4]. Current approaches mainly focus on identifying features that are more important to the decision-making process of the model [5]. Some of these approaches have the limitation that they only weigh the importance of features, without regard to whether that feature positively or negatively contributes to the model prediction result. Other Explainability paradigms such as LIME [6], are only designed to explain a single prediction and don't provide general insight to a model as a whole.

Previous explainable AI research has demonstrated some notable limitations to current explainable AI [5]. One of the main limitations of these approaches is that they may not focus on the sensitive attributes that are important to the fairness integrity of the model. In addition, these approaches have the limitation that they cannot comment on the choice of machine learning model and can only comment on the output. Another weakness of these approaches is that they cannot detect a model that was inherently trained on dirty or poorly collected data, such as data with a selection bias or those with data pre-processing issues.

Furthermore, to our knowledge, we have not seen any currently existing work that comments on how addressing a fairness constraint breaches other fairness definitions, and how sensitive different machine learning algorithms are towards.

## 1.3 Our Key Ideas

Given that we have not seen any currently existing work in this direction, our goal is to explore how training with fairness constraints to address a particular fairness metric will impact other fairness metric definitions. More specifically, we want to explore if addressing a particular fairness metric by training with fairness constraints designed to satisfy these metrics will inadvertently breach an alternative fairness metric or affect it significantly in any way, adversely or otherwise. In the rest of this paper, we attempt to conduct an analysis that will help us better understand the relationship between fairness resolution and fairness metrics.

Our contributions are as follows: 1) We conduct exploratory data analysis on three different datasets known for potential issues with sensitive attributes and train baseline models for each dataset. We also re-train each classifier using two different XAI tools, Microsoft's Fairlearn and IBM's XAI-360, to optimize on a specific fairness metric. 2) Our main contribution, we conduct additional analysis to analyze how the new classifiers perform under not only the original specified fairness metric, but also alternative fairness metrics.

# 2. RELATED WORK

## 2.1 Current Literature

### 2.1.1 Defining Fairness in Machine Learning

Fairness in machine learning is defined in terms of whether or not a protected attribute exists which divides the population into privileged and disadvantaged subjects across the population [5]. The subjective nature of fairness has led to several categories of fairness. Classification parity assumes a classifier to be fair if it generates a positive classification for the privileged and unprivileged classes with an equal probability.

Calibration is another definition of fairness that assumes a classifier to be fair if, for any predicted probability, subjects from the privileged and unprivileged have the same likelihood for positive classification. Similarly, conditional statistical parity proposes that a classifier needs to classify the privileged and unprivileged subjects with an equal likelihood over a specific set of factors. The key difference between calibration and conditional statistical parity is that statistical parity does not consider the fraction of positive predictions over the number of all predictions for any probability.

To summarize, definitions of fairness are categorized into two classes- individual and group fairness. Individual fairness attributes include fairness through awareness and unawareness and counterfactual fairness focus on similar outcomes for individuals belonging to the same group. Whereas group fairness metrics like demographic parity, calibration, and statistical parity similarly emphasize different groups.

### 2.1.2 Fairness Prediction with Disparate Impact

Recidivism prediction instruments (RPI's) are gaining a lot of traction in the current judicial systems for their ability to assess the likelihood of recidivism for each individual. But recently, questions were raised with regards to the fairness of these RPI's [7]. Numerous researchers [7] have tried to address these questions by assessing the classifiers provided in the COMPAS dataset against numerous fairness criteria. It is crucial to realize that all the fairness criteria cannot be satisfied by the COMPAS dataset due to the inherent varied distribution of recidivism across various groups. It is also necessary to analyze the impact of the RPI when it doesn't satisfy the base error rate balance criteria [7].

Another important aspect while evaluating the fairness of COMPAS is the novelty of ProPublica's analysis. A key term that has come up a lot in the research is disparate impact, which alludes to settings wherein a penalty policy can lead to disproportionate and unintended impacts on a particular

group [7]. This notion is critical in realizing that any RPI which is fair on a given set of fairness criteria can still lead to disparate impact for a different set of fairness criteria.

To summarize, fairness is a social and ethical concept and it cannot be statistically determined by some measures. Given the inherent nature of COMPAS, the varied distribution of the data across various groups makes it impossible to satisfy all the fairness criteria and it is also inherent that a disparate impact will arise in datasets similar to COMPAS, regardless of the fairness criteria used [7].

### 2.1.3 Framework for Evaluating Explainable AI

Another feature commonly explored alongside the fairness in Machine Learning models is Explainable AI (XAI). As the machine learning models become more and more complex by the day, it is crucial to understand the inner workings of the models in an intuitive manner. XAI frameworks provide an extensive array of tools to elucidate model behavior.

Although XAI can help interpret model behavior, a valid concern raised by researchers is the impact of XAI on biased or unfair models and how it can lead to fairwashing [8] among the users. One way of analyzing the current XAI frameworks is to identify a holistic set of rubrics [5] that can help examine these frameworks. Once this is identified, several XAI frameworks like Lime [1], IBM's AI Explainability 360 (XAI 360) [3], random forests, and a simple logistic regression model were trained on the COMPAS dataset and compared against these rubrics to understand how do these models perform in terms of the fairness rubrics.

The fairness rubric identified over here lays the foundation for the researchers to benchmark new XAI frameworks to make the framework more holistic in terms of fairness apart from the explainability.

## 2.2 Key Idea Inspiration

In previous sections, we have seen many different types of fairness metrics, as well as different ways to train for and attempt to resolve a breach in a single specific metric. Our intuition is that, since we are only attempting model retraining to resolve one fairness metric at a time, there may be some trade-off in the evaluation of a different fairness metric. Some natural but important questions for us to ask are: "Can we satisfy multiple fairness metrics simultaneously?" and "Are we likely to break a different fairness metric while fixing another?".

# 3. PROBLEM DEFINITION

In this project, our goal was to conduct an exploratory data analysis (EDA) for breach of fairness constraints of three datasets- COMPAS, Loan Prediction, and Adult Dataset. To verify whether the classifier trained on this dataset was fair or not, we trained two separate fairness unaware classifiers- a Logistic Regression classifier and a Random Forest Classifier on these datasets and used the predictions from the datasets against the ground truth to evaluate whether the dataset caused classifiers to be unfair towards a certain unprotected attribute. We worked with three definitions of fairness- overall accuracy equality, predictive parity, and equalized odds in all our datasets.

The next step was to conduct an EDA to check for the aforementioned fairness definition breaches on the datasets by comparing the predictions with the ground truth. After conducting this EDA, worked towards solving the fairness constraint which was the most prominent- which was equalized odds in all three cases, using Microsoft's Fairlearn library. We then proceeded to compare the predictions of the fairness aware classifier with the ground truth to again evaluate the breach of fairness definitions to analyze how addressing one fairness definition affects the other definitions of fairness. This experiment was performed on both classifiers- Logistic Regression and Random Forest to benchmark the difference in the behavior of models after a fairness constraint is solved.

To extend our understanding of various explainable AI tools, we tried to mitigate the aforementioned violations of fairness definitions with IBM's XAI-360 tool. The experiments performed were similar to Fairlearn, wherein we take in the original model and check XAI-360's performance in terms of the fairness definitions that were violated and how effective was XAI-360 to mitigate them.

# 4. MAIN RESEARCH ALGORITHMS

This section lays down the foundation for the working of Fairlearn and the XAI 360 framework. This research utilizes the post-processing libraries of these frameworks to mitigate the fairness definitions violation and it is important to understand how these frameworks handle these violations. The following subsections provide a holistic view of FairLearn's

### 4.1 Fairlearn

Microsoft's fairlearn framework has support for post-processing algorithms that alters the output of the original model to mitigate fairness definitions violations. Fairlearn uses either demographic parity or the equalized odds constraint [2] in the post-processing steps to make the model fair. The fairlearn framework uses a wrapper model around the original model for mapping the predict method in the original model to the predict_proba method to achieve better estimates [9].

The first step of the framework is to find the threshold values to divide the data for all the groups in the sensitive attributes that are provided as part of the input. The original classifier generates a huge number of threshold rules and for each rule, fairlearn generates the FNR and TPR of that group [9]. Fairlearn then plots the ROC [2] for all these rules which results in a probabilistic classifier. The interesting observation here is that fairlearn produces various

| Dataset | Model | Overall Accuracy Equality (Protected \| Unprotected) | | Predictive Parity (Protected \| Unprotected) | | Equalized Odds (Protected [FPR, FNR]\| Unprotected [FPR, FNR]) | | Model Accuracy (in %) | |
|---|---|---|---|---|---|---|---|---|---|
| | | Before | After | Before | After | Before | After | Before | After |
| COMPAS | Logistic Regression | 0.70 \| 0.58 | 0.73 \| 0.63 | 0.56 \| 0.70 | 0.50 \| 0.67 | 0.20, 0.58 \| 0.34, 0.34 | 0.37, 0.38 \| 0.33, 0.43 | 65.9 | 61.88 |
| | Random Forest | 0.84 \| 0.91 | 0.83 \| 0.91 | 0.6 \| 0.69 | 0.58 \| 0.70 | 0.21, 0.48 \| 0.35, 0.34 | 0.26, 0.40 \| 0.342, 0.35 | 66.93 | 66.82 |
| Loan Prediction | Logistic Regression | 0.70 \| 0.58 | 0.74 \| 0.63 | 0.94 \| 0.93 | 1 \| 1 | 0.88, 0.07 \| 1, 0.07 | 0.91, 0 \| 1, 0 | 68.08 | 71.63 |
| | Random Forest | 0.84 \| 0.91 | 0.83 \| 0.91 | 0.95 \| 1 | 1 \| 0.93 | 0.41, 0.04 \| 0.22, 0 | 0.41, 0.07 \| 0.22, 0 | 85.48 | 84.4 |
| Adult Dataset | Logistic Regression | 0.89 \| 0.74 | 0.90 \| 0.74 | 0.30 \| 0.28 | 0.22 \| 0.19 | 0.04, 0.72 \| 0.03, 0.70 | 0.01, 0.81 \| 0.01, 0.78 | 79.59 | 79.96 |
| | Random Forest | 0.93 \| 0.82 | 0.93 \| 0.82 | 0.61 \| 0.64 | 0.62 \| 0.66 | 0.10, 0.36 \| 0.03, 0.39 | 0.10, 0.34 \| 0.03 \| 0.38 | 85.73 | 85.73 |

probabilistic classifiers that collaboratively predict the label for the given input. The aim of fairlearn here is to resolve equalized odds and that is achieved by extracting the convex hull [2], [9] for the ROC for all the groups in the sensitive attribute. The ROC curves for the two groups in the COMPAS dataset are shown in *Figure 2*.

In the case of equalized odds, wherein the FPR and TPR for all the corresponding groups are the closest all the while minimizing the error rate, and hence the problem is decomposed to finding the error overlap of the convex hulls of the ROC curves of all the groups.

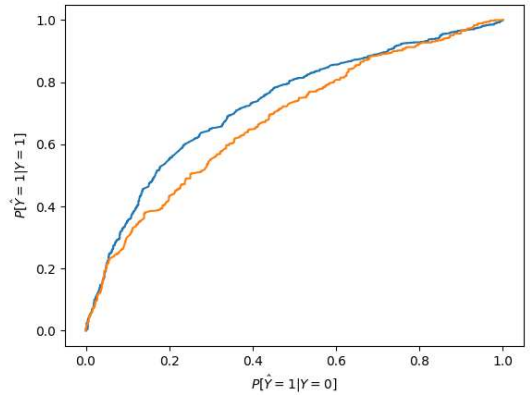



**Fig. 2** ROC curves for race category in COMPAS dataset

### 4.2 XAI-360

Similar to Fairlearn, XAI 360 also supports libraries for post-processing steps on existing models to make them fair [10]. XAI-360 leverages calibrated equalized odds to calibrate classifier score outputs in order to determine the probabilities with which each label is updated to make the classifier fair.

The post-processing model fit function takes in the parameters as the ground truth and predicted scores and it computes the parameters for equalizing the generalized odds while preserving the calibration [3], [10]. Similarly, the predict function of XAI-360 perturbs the predicted scores to derive new labels to satisfy the equalized odds constraints while preserving the calibration [3].

## 5. EXPERIMENTAL STUDY

### 5.1 Datasets

For our project, we worked with three datasets- ProPublica COMPAS dataset [11], Loan Prediction Dataset [12], and Adult Dataset [13]. COMPAS classifier dataset consists of recidivism scores of over 10000 defendants with their outcomes within 2 years of decision and consists of 137 attributes. ProPublica's COMPAS assigns higher recidivism scores to the African-American race type and lower recidivism scores to the Caucasian race type. For this dataset, we take the race attribute as our sensitive attribute. The loan Prediction dataset is a much smaller dataset from Kaggle which has 612 records with 12 attributes that predict whether a loan application will be granted a loan based on attributes like their marital status, income, etc. This dataset has also been accused of being biased towards the male gender in approving loans. For this dataset, we take the gender attribute as our sensitive attribute. The Adult Income dataset is a larger dataset consisting of 48k entries that try to predict if an adult has an income of over or under USD $50k based on 13 attributes such as education level, age, marital status, race, and gender. This dataset is heavily skewed with Caucasians more heavily represented than Black individuals, and also heavily skewed with more males than females. For this dataset, we take gender as our sensitive attribute.

### 5.2 Evaluation Metrics

Our experiment consists of four evaluation metrics- overall accuracy of the model, equalized odds fairness, predictive parity fairness, and overall accuracy equality fairness of the classifier trained on the dataset. **Overall Accuracy Equality** fairness is assessed by comparing the accuracy of the model on privileged and unprivileged attribute classes of predicting the ground-truth label, like getting a loan approved, reoffending, or having an income over $50,000. **Predictive Parity** fairness is checked by checking the probability for both groups for getting a positive predictive label like reoffending, getting the loan approved, or having income over $50,000. **Equalized Odds** fairness metric requires the model to have similar probabilities of having false positive and false negative rates for both groups.

### 5.3 Baselines and Results

For our experimental setup, we first trained two unaware classifiers- a random forest classifier and a logistic regression classifier, which gave us predictions for the test data. Using the ground truth and predictions of both these models, we

**Table 1** Experimental Results for XAI frameworks

worked towards evaluating the fairness metrics of both these models. The baselines for our project in addressing fairness were the naive fairness unaware Logistic Regression and Random Forest Classifier. The Explainable AI tools serving as our baselines are Fairlearn and XAI360. *Table 1*. summarizes our results where red highlighted cells denote fairness constraint breaches after solving equalized odds, and the green highlighted cells indicate the increase in accuracy of our fairness aware model. "Protected" denotes the dominant class of the sensitive attribute i.e. Caucasian, male, whereas "Unprotected" denotes the non-dominant class of the sensitive attribute i.e. Black, female.

### 5.3.1 Results on COMPAS Dataset

For the COMPAS Dataset, the Random Forest classifier proved to be much more robust than the Logistic Regression classifier as it had an accuracy of 66.93% which dipped to 66.82% after solving the equalized odds fairness constraint while the Logistic Regression classifier had an accuracy of 65.90% which dipped to 61.88%. For Random Forest, the overall accuracy equality is similar for both classes at 68.76% for Caucasians and 65.66% for the African-American class. The fairness constraint was not broken after achieving equalized odds fairness as the accuracy for the Caucasians group slightly dipped to 68.62% and for the African-American group, it was 65.56%. The predictive parity fairness for Random Forest was unfair and we noticed that solving equalized odds exacerbated this fairness constraint as the predictive parity for the Caucasian group decreased from 60.16% to 58.54%, while that for the African-American group slightly increased from 69.54% to 69.63%, thus increasing the difference between the probabilities for both groups.

The Logistic Regression classifier had low overall accuracy and it dropped significantly after solving equalized odds from 65.90% to 61.88%. The overall accuracy equality was maintained after solving the equalized odds fairness metric although the probabilities for Caucasian and African-American groups dropped from 0.6568 and 0.6605 to 0.6260 and 0.6138 respectively. Predictive parity fairness constraint was not satisfied before solving equalized odds, and the disparity between the two groups increased as the predictive parity for Caucasian and African American groups changed from 0.56 and 0.70 to 0.50 and 0.67 respectively.

The equalized odds for random forest classifier improved significantly as the false positive rate (FPR) and false-negative rate (FNR) for the Caucasian race got closer from 0.212 and 0.476 to 0.257 and 0.405 respectively, and the FPR and FNR for the African-American group also got closer from 0.346 and 0.340 to 0.342 and 0.345 respectively. For logistic regression, we observed a drastic improvement in the similarity of FPR and FNR for the Caucasian race as they changed from 0.20 and 0.575 to 0.372 and 0.376 respectively. This shift was drastic and this could have contributed to the 5% drop in accuracy for the classifier. However, this did cause the FPR and FNR for the African-American group to breach the constraint again as they dropped from being almost equal at 0.34 and 0.338, respectively to observably different at 0.331 and 0.431.

### 5.3.2 Results on Loan Prediction Dataset

For the Loan Prediction Dataset, we trained both models on 422 records and tested them on 140 records out of which we had 117 records for Males and 24 records for females. This data imbalance affected the accuracy of the Logistic Regression classifier but Random Forest proved to be much more robust as it had an accuracy of 85.81% while the Logistic Regression classifier had an accuracy of 68.08%.

The overall accuracy equality for biased towards females for the RF classifier, achieving over 91% accuracy for the female class, and 84% for the male class. The bias against males increased after achieving equalized odds fairness as the female accuracy remained the same but male accuracy dipped to 82.9%. The Logistic Regression classifier performed much worse and was biased against females, achieving an accuracy of 58% for females, and 70% for males. This performance significantly improved after achieving equalized odds by applying the Fairlearn Threshold Optimizer Postprocessing and increased the accuracy of the model to 62.5% for females and 73.5% for males, while still being slightly biased towards males.

The predictive parity fairness for Random Forest had similar probabilities for both genders and there was a slight decrease in this fairness after applying post processing as the probabilities for males and females changed from 0.95 and 1 respectively, to 1 and 0.93. Predictive parity was maintained for logistic regression both before and after achieving equalized odds as the probabilities for males and females increased from 0.94 and 0.93 respectively, to 1 for both.

The equalized odds for the random forest classifier slightly improved as the false positive rate (FPR) and false-negative rate (FNR) for the male gender got closer from 0.41 and 0.04 to 0.41 and 0.07 respectively, while that for the female gender remained constant at 0.22 and 0 respectively. For logistic regression, this false-negative rate for females decreased from 0.067 to 0 while the false positive rate remained the same at 1. For males, the false positive rate for males increased from 0.88 to 0.91while the FNR decreased from 0.06 to 0.

### 5.3.3 Results on Adult Income Dataset

For the Adult Income Dataset, we trained both models on 32561 records out of which we had 21790 records for Males and 10771 records for females. After training, our random forest classifier had an accuracy of 85.74% while the Logistic Regression classifier had an accuracy of 79.60%.

The overall accuracy equality for bias towards females for the RF classifier, achieving over 82% accuracy for the female class, and 93% for the male class. The bias against females

increased after achieving equalized odds fairness as the female accuracy remained the same but male accuracy increased to 82.9%. The Logistic Regression classifier performed slightly worse and was biased against females, achieving an accuracy of 74% for females, and 89% for males, which remained the same after applying Fairlearn Threshold Optimizer Postprocessing.

The predictive parity fairness for Random Forest had similar probabilities for both genders and there was a slight decrease in this fairness after applying post-processing as the probabilities for males and females changed from 0.61 and 0.64 respectively, to 0.62 and 0.66. Predictive parity for logistic regression both before and after achieving equalized odds as the probabilities for males and females decreased from 0.30 and 0.28 respectively, to 0.22 and 0.19 respectively.

The equalized odds for the random forest classifier slightly improved as the false positive rate (FPR) and false-negative rate (FNR) for the male gender slightly improved from 0.10 and 0.36 to 0.10 and 0.34 respectively, while that for the female gender also slightly improved at 0.03 and 0.39 to 0.03 and 0.38 respectively. For logistic regression, this false-negative for both males and females showed a significant increase.

### 5.4 Summary and Analysis of Results

Our results can be seen above in Table 1. From our analysis, we observe that when attempting to optimize the baseline models for equalized odds, in general the predictive parity and overall accuracy equality of the models suffered slightly as a result. This evidence suggests that resolving models based on one fairness metric has a slight adverse effect on other fairness metrics.

## 6. Future Scope

The research tries to analyze various aspects of explainable AI and its capabilities in mitigating the fairness definition violations. The research scope can be extended to provide a more comprehensive overview of these fairness metrics as defined in the XAI rubrics in the previous sections of this research. Furthermore, Another interesting domain could be to leverage our tool to enhance XAI frameworks to address some of the other fairness metrics that are violated after post-processing.

## 7. Conclusion

Our research tries to analyze various XAI frameworks on biased datasets based on the XAI rubrics for a holistic framework. Furthermore, we conduct the exploratory data analysis on these biased datasets to understand the fairness definitions. This research benchmarks the XAI frameworks on the highlighted rubrics and further checks the overall effect of these XAI tools after mitigating the violated fairness metrics. Our motivation is to realize the effects of these post-processing steps on other fairness definitions and this research tries to reason how these XAI frameworks mitigate the fairness violation and how these frameworks can adapt their post-processing algorithm to minimize the adverse effects of these steps.

## REFERENCES


[1] S. Tolan, "Fair and Unbiased Algorithmic Decision Making: Current State and Future Challenges," *ArXiv190104730 Cs Stat*, Jan. 2019, Accessed: Dec. 14, 2021. [Online]. Available: http://arxiv.org/abs/1901.04730

[2] "fairlearn.postprocessing package — Fairlearn 0.7.0 documentation." https://fairlearn.org/v0.7.0/api_reference/fairlearn.postprocessing.html (accessed Dec. 14, 2021).

[3] "AI Explainability 360." https://aix360.mybluemix.net/ aix360.mybluemix.net (accessed Dec. 14, 2021).

[4] K. Alikhademi, B. Richardson, E. Drobina, and J. E. Gilbert, "Can Explainable AI Explain Unfairness? A Framework for Evaluating Explainable AI," *ArXiv210607483 Cs*, Jun. 2021, Accessed: Dec. 14, 2021. [Online]. Available: http://arxiv.org/abs/2106.07483

[5] K. Alikhademi, E. Drobina, D. Prioleau, B. Richardson, D. Purves, and J. E. Gilbert, "A review of predictive policing from the perspective of fairness," *Artif. Intell. Law*, Apr. 2021, doi: 10.1007/s105 06-021-09286-4.

[6] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," *ArXiv160204938 Cs Stat*, Aug. 2016, Accessed: Dec. 14, 2021. [Online]. Available: http://arxiv.org/abs/1602.04938

[7] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *ArXiv161007524 Cs Stat*, Oct. 2016, Accessed: Dec. 14, 2021. [Online]. Available: http://arxiv.org/abs/1610.07524

[8] U. Aivodji, H. Arai, O. Fortineau, S. Gambs, S. Hara, and A. Tapp, "Fairwashing: the risk of rationalization," in *Proceedings of the 36th International Conference on Machine Learning*, May 2019, pp. 161–170. Accessed: Dec. 14, 2021. [Online]. Available: https://proceedings.mlr.press/v97/aivodji19a.html

[9] M. Hardt, E. Price, and N. Srebro, "Equality of Opportunity in Supervised Learning," *ArXiv161002413 Cs*, Oct. 2016, Accessed: Dec. 14, 2021. [Online]. Available: http://arxiv.org/abs/1610.02413

[10] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On Fairness and Calibration," in *Advances in Neural Information Processing Systems*, 2017, vol. 30. Accessed: Dec. 14, 2021. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/b8b9c74ac526fffbeb2d39ab038d1cd7-Abstract.html

[11] "How We Analyzed the COMPAS Recidivism Algorithm — ProPublica." https://www.propublica.org

/article/how-we-analyzed-the-compas-recidivism-algorithm (accessed Dec. 14, 2021).
[12] “Loan Data Set.” https://kaggle.com/burak3ergun/loan -data-set (accessed Dec. 14, 2021).
[13] “UCI Machine Learning Repository: Adult Data Set.” https://archive.ics.uci.edu/ml/datasets/Adult (accessed Dec. 14, 2021).