



Equality AI Fair Pre-processing Machine Learning Recipe

Are you concerned that data and algorithmic biases lead to machine learning (ML) models that treat individuals unfavorably based on characteristics such as race, gender, or political orientation?

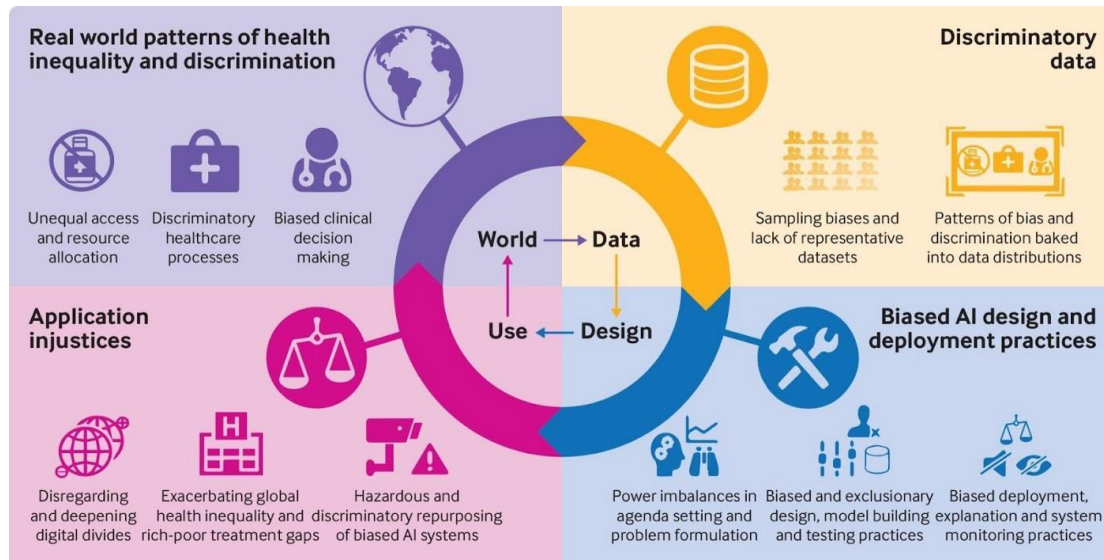


Figure 1: An infographic showing the long-term consequences of bias finding its way into our data in one example, the health sector (image from the British Medical Journal).

Fairness-based ML offers a potential solution by incorporating **bias mitigation methods** and **fairness metrics** into ML Operations (MLOps). We have conducted extensive literature review and [theoretical analysis](#) on a large number of fairness metrics and mitigation methods to create our [responsible_mlops](#) open source ML software framework for fitting a fair pre-processing ML model.

We are excited to share the results of our first use case! Through hospital admission, we will illustrate a workflow recipe with pre-processing based bias mitigation for predicting hospital admissions. Here we will not be showing any of the code but will walk through how we ran our case study, with explanations of conceptually executing each step or series of functions in our `responsible_mlops`. Functions will be denoted by trailing parentheses.

To follow along with the code, go clone our [responsible_mlops](#) GitHub repository and open the script [example_dataset_NHAMCS.R](#).

Recipe Instructions

1. Define Research Question
2. Connect to Source Data
3. Select Fairness Metric and Mitigation Strategy
4. Data Preparation
5. Fit Prediction Model
6. Compute Model Results and Fairness Score

Ingredients

- Your research question (or run our use case)
- Source data (or use our sample data)
- Fairness metric
- Bias Mitigation method
- Integrated development environment (IDE), e.g. R studio
- R programming language, Python (coming soon)

7. Run Mitigation
 8. Compute Model Results and Fairness Score after Mitigation
 9. Compare Model Results and Fairness Score Before and After Mitigation
- Access to the Equality AI GitHub repository

Footnote: The use case is intended for demonstration purposes only to highlight responsible AI methods to address bias and model for fairness. The output model is not intended to be adopted without utilizing additional bias reduction methods (i.e. inclusion of additional data sources), further model development, validation, and adherence to institutional governance processes.

Instructions

1. Define Research Question

Existing cross-sectional study of the emergency department (ED) data from [National Hospital Ambulatory Medical Care Survey \(NHAMCS\)](#) revealed substantial disparity in hospital admission rates between blacks and whites in the United States, with black patients 7% less likely than whites to have their care needs classified as immediate/emergent, and 10% less likely than whites to be admitted to hospital following an ED visit. Algorithms trained on biased data to predict hospital and ICU admissions will learn these biases and reflect them in their predictions.

Our research questions were whether machine learning models trained on these biased data to predict hospital and ICU admissions will learn these biases and reflect them in their predictions, and whether bias mitigation methods improve the fairness of machine learning models.

To answer this question, we started by fitting a risk prediction of hospitalization or ICU admission and assessing the model's fairness before and after mitigation.

2. **Connect to Source Data:** our cohort included 12,258 patients (17,668 whites and 7,624 blacks) who visited the emergency department in the year 2019. The hospital admission was 18.8% for whites and 10.6% for blacks.

Use our sample data and provide information about the protected attribute (e.g. race, gender), privileged group (e.g. white, male), and target attribute in your dataset. This information will be used for computation of fairness metrics. We have provided the `data_fetch()` to easily connect to our sample data from the NHAMCS, a sample of cross-sectional probability for U.S. emergency departments surveyed in 2018 and 2019.

Protected attribute: RACERETH (race)
Privileged group: 1 (white)
Target attribute: HOS (hospitalization)

3. **Select Fairness and Mitigation Strategy:** we need to choose an appropriate fairness metric, `fairness_tree_metric()`, and bias mitigation method, `mitigation_mapping_metric()`, for this research question.

We illustrate the selection of fairness metric using our `fairness_tree_metric()` with the hospital admission use case. Our `fairness_tree_metric()` provides a recommendation about the fairness metric based on a series of questions that relates to the dataset, protected attribute, and analytic goals.

1. Does your algorithm use an individual's sensitive variable information (intentional discrimination) to make a decision?

We answered: No, because it is unlikely racial information is used to predict hospital and ICU admission.

2. Do you want to assess if your population is disadvantaged by multiple sources of discrimination such as race, class, gender, religion, and other inner traits?

We answered: No, existing evidence from literature only suggests disparity in hospitalization between races.

3. Are there any standards or regulations enforced to avoid discrimination in regard to the decision being made?

We answered: No, we assume that it is unlikely that standards or regulations exist.

4. Is there a reliable label or ground truth for the outcome of interest? Is there no historical or measurement bias?

We answered: No, existing literature suggests historical biases between races.

5. Do you have features/explanatory variables in your data that provide information about the outcome variable while at the same time are correlated with the sensitive variable?

We answered: No, existing literature does not suggest correlation between races and other explainable variables.

Selected fairness metric: Statistical Parity

Once the fairness metric is chosen, we then pass that fairness metric to `mitigation_mapping_method()` to determine the appropriate bias mitigation methods. Three bias mitigation methods are available for mitigation, namely, sampling, reweighing, and disparate impact remover.

Selected mitigation methods: Reweighing

4. **Data Preparation:** in our `responsible_mlops`, data preparation involves four key steps, namely feature selection, random splitting of data into training and testing sets, missing value imputation, and data balancing. We have provided the following functions:

`data_prepare_nhamcs()`, `train_test_split()`, `data_balancing()`.

The following features are chosen for the machine learning model: sex, age, type of residence, source of payment, arrival mode, arrival day and time, initial vital signs (body temperature, heart rate, respiratory rate, blood pressure, pulse oximetry), triage level, pain scale, 72-hour revisit rates, and 22 types of comorbidities (such as cancer or diabetes).

Using the `train_test_split()`, the data set was randomly split into training and testing sets with 70% of data used for model training and 30% retained for testing.

Three methods are available for handling missing values in the `data_prepare_nhamcs()`, which include

1. complete case analysis (i.e., remove all instances with one or more missing values),
2. multiple imputation by chained equations (i.e., multiple imputations of each missing value through an iterative series of prediction models)
3. multiple imputation with random forest (a random forest based imputation algorithm for missing values)

Two methods are available for data balancing in our `data_balancing()`, namely over sampling the minority class and under sampling the majority class. Under-sampling is used for this use case.


5. **Fit Prediction Model:** once a fairness metric is chosen, we assess the (un)fairness of a machine learning model trained on this dataset prior to applying bias mitigation. Two machine learning models are available for prediction in our `ml_method()`: random forest model and generalized boosted regression modeling (GBM).

GBM was chosen for this use case.

6. **Compute Model Results and Fairness Score:** `fairness_scores()` and `ml_results()` calculates and stores the fairness metric of choice, True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN), Precision, Recall, F1-score, and Accuracy into a single variable.

The predictive performance of the machine learning model such as classification accuracy, recall, precision, F1 score, and AUC are computed using the testing set. The fairness score of the machine learning model is also computed using the selected fairness metric of Statistical Parity.

7. **Run Mitigation:** using the `bias_mitigation()` will apply the “repair” selected to the dataset.



The reweighing method is selected for bias mitigation

Finally, we assess the (un)fairness of the same machine learning model after bias mitigation is applied. By comparing the predictions before and after mitigation, we will be able to assess whether and to what extent the fairness of hospital admission predictions across different racial groups can be improved. Furthermore, the trade-offs between the accuracy and fairness of the machine learning model will be examined.

8. **Compute Model Results and Fairness Score After Mitigation:** predictive performance and the fairness score of the machine learning model are re-computed after mitigation.
9. **Compare Model Results and Fairness Score Before and After Mitigation:** compare the predictive performance and fairness of GBM before and after applying the reweighing bias mitigation method.

Results and Discussions

Figure 2 shows the proportions of whites and blacks predicted to be hospitalized by GBM before and after bias mitigation. The proportions of predicted hospital admission pre-mitigation were 45.7% for whites and 35.1% for blacks, compared to 44.8% for whites and 36.2% for blacks post-mitigation. Consequently, the prediction gap (proportion of predicted hospitalization for whites minus proportion of predicted hospitalization for blacks) narrowed from 10.6% pre-mitigation to 8.6% post-mitigation.

Figure 3 shows the AUC of the GBM and statistical parity ratio before and after bias mitigation. A moderate improvement in statistical parity ratio (from 66.4% before mitigation to 71.5% after mitigation) is observed whereas the AUC is almost unchanged.

The reweighing bias mitigation method has demonstrated its effectiveness in improving statistical parity ratio while maintaining the performance of the machine learning model.

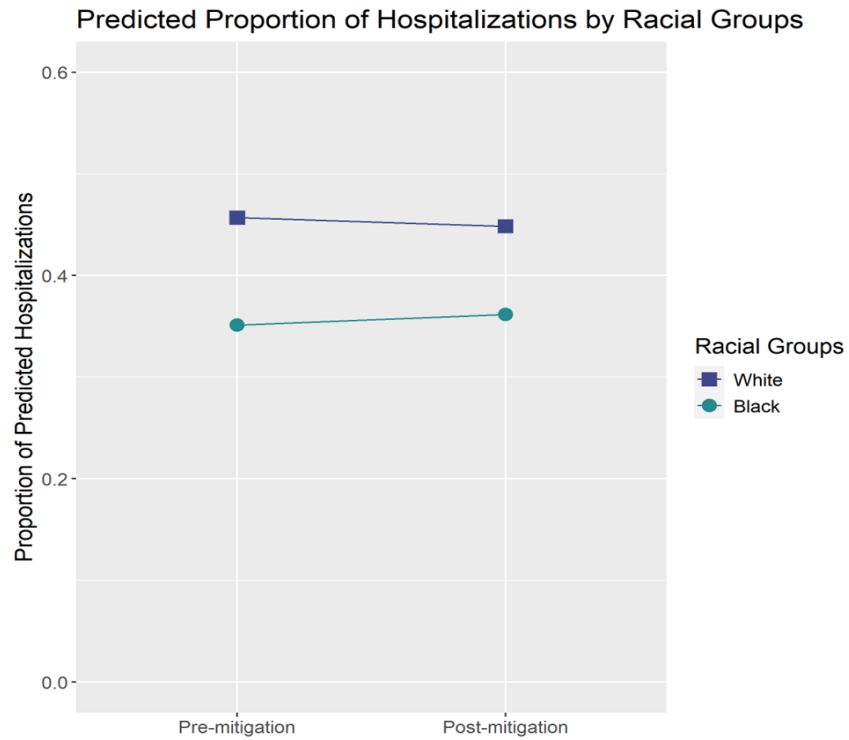


Figure 2: A comparison of proportion of predicted hospitalization by racial/ethnic groups before and after bias mitigation

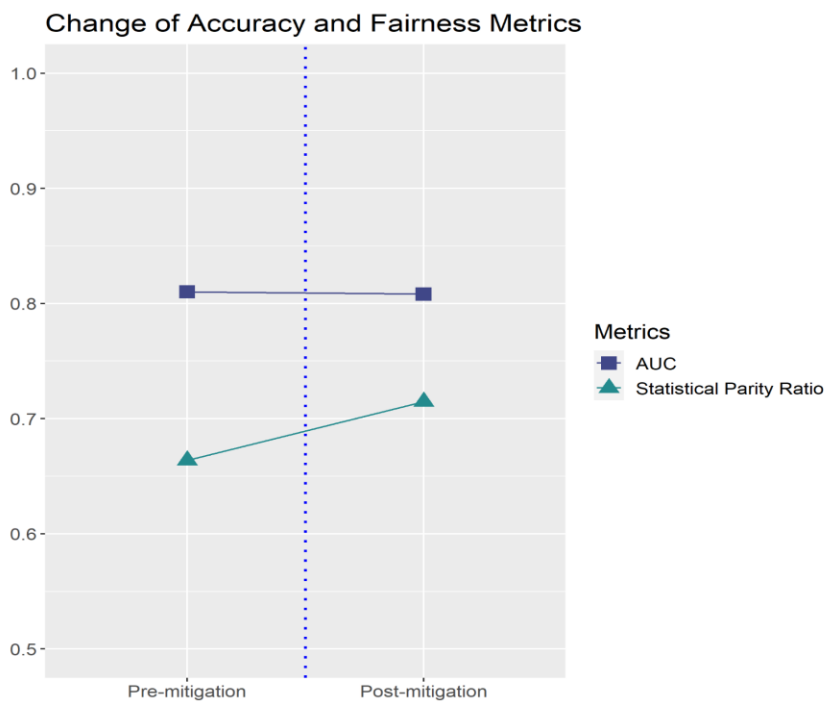


Figure 3: AUC and statistical parity ratio before and after bias mitigation

Limitations

1. Choosing the appropriate fairness metrics given a real-world scenario can be a challenging task and will likely involve some degree of subjectivity. While our “fairness metric tree” is designed to assist users in choosing the metric, it is recommended that data scientists should also discuss their choice of fairness metrics with domain experts.
2. It is possible that multiple fairness metrics are appropriate given a real-world scenario whereas the current implementation of the fairness metric tree can only recommend a single fairness metric. We will extend the functionality of the fairness metric tree in the future.
3. The study only included three pre-processing mitigating techniques. Exploration of in-processing, and post-processing mitigation techniques need to be studied.

References

- Kamiran, F., Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1–33
- Wiśniewski J., Biecek P. (2021). “fairmodels: A Flexible Tool For Bias Detection.” arxiv.
- Makhlouf K., Zhioua S., Palamidessi C. (2021). Machine learning fairness notions: Bridging the gap with real-world applications. *Information Processing & Management* 58, 5, 102642

Authors



[Equality AI \(EAI\)](#) is a public benefit corporation dedicated to helping data scientists close the health disparity gap, and to do this we have conducted extensive literature review and theoretical analysis on dozens fairness metrics and mitigation methods. Theoretical properties of those fairness mitigation methods were analyzed to determine their suitability under various conditions.