Can Explanations Unveil the Bias in the System?

Ujjwal Dubey, Raksha Rank

Luddy School of Informatics, Computing, and Engineering
Indiana University, Bloomington

Abstract

Ethics and fairness has been a perennial debate in the AI industry for sometime now. Our motivation behind the project is to understand and answer the question, *'Can Explainable AI explain fairness?'* In this paper, we aim to show our analysis on COMPAS dataset and implement explainable AI tools to examine if explanations can unravel the bias present in the system.

Introduction

"There's a simple way to solve the crime problem: obey the law; punish those who do not", Rush Limbaugh.

The biggest challenge is to analyze who obeys the law and who doesn't. Due to advancements in the field of Machine Learning and Artificial Intelligence predicting the outcome from a given Data set has taken a new edge. Nonetheless, researchers and scientists are now trying to add the explainability to the AI and ML models. Judges, probation, and parole officials throughout the country are increasingly utilizing algorithms to determine if a criminal defendant will become a recidivist. Hundreds of risk assessment algorithms are now in use. Many states have developed their examinations, and several academics have developed tools. COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is a popular algorithm used by courts and parole officials to assess the possibility of re-offending of criminal defendants (recidivism). Based on a two-year follow-up investigation, it has been demonstrated that the algorithm is biased in favor of white defendants and against black convicts (i.e., those who actually committed crimes or violent crimes after 2 years).

We selected the COMPAS dataset to study because it is one of the most widely used scores in the United States, and it is increasingly being utilized in pretrial and sentencing proceedings, the so-called "front-end" of the criminal justice system.

As we know, Logistic Regression and Random Forest classifiers are some of the algorithms which are widely used for classification problems. We have used Explainable AI Tools on these two algorithms. Since these algorithms and Machine Learning models are difficult for humans to comprehend, tools to explain models and assess fairness have grown popular. While explainable AI and fair AI technologies have different functions, their aims and expectations are quite similar.

Furthermore, ML practitioners expressed a great demand for tools and metrics that they could use to account for and evaluate fairness pro-actively while creating and testing their algorithms in survey research done by Holstein et al. Our model has been created in the quest to develop a toolkit which gives Fair Clarification that can be used by a wide range of judges, probation and parole officials, and machine learning practitioners to evaluate their models and data for fairness and to fully comprehend the dataset from the interpretations presented by the tools.

Motivation

As we know, AI-based algorithms assist people in making better decisions, but sometimes we humans may not always grasp how AI arrived at that conclusion. Here, we expect the XAI to explain how these algorithms arrive at these findings and what variables/features influenced them. Furthermore, there may arise a question of 'How fair the AI model's prediction is?' In this case, XAI can contribute towards the analysis of the fairness in the results as well as following explanations. XAI and People centered AI together can take the system's ethical and legal standards and accuracy to a higher level of interpretability.

If humans start to make their own interpretation just based on the AI model's prediction, in few cases, they might end up with any unethical and illegal conclusion so, we expect XAI to contribute by explaining the AI model's output based on different features, Ground truth, Cost ratio, and Fairness which will develop the humans trust on the AI/ML models. Thus, XAI can strongly influence our long term goal towards Ethical and Responsible AI.

To gain a deeper understanding of the task in hand, we would start with analyzing the dataset from an ethical perspective. We know that a biased dataset tends to generate biased predictions, here, in our case classification. We would examine the features of COMPAS before importing the data and further try to modify the column feature names that may distort the decisions our model would make. We will rename the columns that contain unethical column names to ethical column names. The new input feature names make the dataset ethical and still retain enough information for us to analyze accuracy and fairness with SHAP and WIT.

We aim to explore various stages at which bias can be inserted or subtly be present without former knowledge in the AI model pipeline. Using various XAI tools to understand the decisions that the model is making, and WIT for checking the accuracy and fairness components along with feature analysis, we aim to learn and demonstrate the effect of explanations on fairness of AI models.

Literature Review

The base paper *Can Explainable AI Explain Unfairness? A Framework for Evaluating Explainable AI* discusses about evaluating fairness post explanations. This paper's main contribution is a rubric designed to assist XAI in explaining the issues of Fairness. The rubric is intended to help developers of XAI tools understand user needs; to help ML developers as they review their models for accuracy and fairness; and to help lay users critique the results of ML models. The authors further mention the apparent conflict between individual and group fairness. The authors work on three XAI tools: LIME, AI Explainability 360, and ad-hoc explainability tools. They train the 3 types of models on COMPAS dataset, Logistic Regression, Random Forest and Neural Network to predict if a person is likely to reoffend.

The paper *Transparency in Fair Machine Learning: the Case of Explainable Recommender Systems* talks about various kinds of bias that can enter the system. Some of those are :

- Confirmation bias: It is a tendency to intentionally search for and include certain data and perform analysis in such a way as to make a predefined conclusion and prove a predetermined assumption.
- Selection bias: This happens when the sample is not collected randomly and because of a subjective selection technique, the data does not represent the whole population under study. Based on the elimination of samples or inclusion of certain samples, the resulting bias can be of the omission or inclusion type, respectively.
- Implicit bias: This type of bias is associated with the unconscious tendency to favor a certain group of people against others based on characteristics such as race, gender, and age.
- Over-generalization bias: This form of bias can come from making a certain conclusion based on information that is too general, especially when the sample size is small or is not specific enough.
- Automation bias: The tendency to favor decisions made from an automated system over the contradictory correct decision made without the automation.
- Reporting bias: This form of bias is the result of an error made in the reporting of the result when a certain positive finding is favored over the negative results.

The book *FAIRNESS AND MACHINE LEARNING, Limitations and Opportunities by Solon Barocas, Moritz Hardt, Arvind Narayanan* very vividly describes how bias in the feedback may reflect cultural prejudices which are quite hard to detect. It quotes a well known study that hints, Google searches for Black-sounding names such as "Latanya Farrell" were much more likely to results in ads for arrest records ("Latanya Far- rell, Arrested?") than searches for white-sounding names ("Kristen Haring"). One of the probable reasons given by the authors was the adherence to stereotypes by the people while clicking on ads. Further the advertising systems are also designed in a way that maximises clicks.

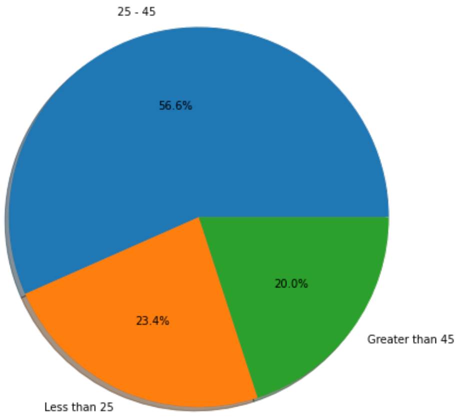


Figure 1: Graphical Representation of Age from cox-violent-parsed-filt

Dataset Analysis

The dataset which we are using is the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) dataset. The dataset includes several features such as Person ID, Case ID, Name, Sex, Ethnicity, Age, Marital Status, Custody Status, Decile score and many more.

If we graphically plot some important features of the Data set, we observe that 56% defendants belong to the age group of 25 to 45, 23% defendants in the data set are below 25 years, and 20% Greater than 45 years (Figure 1). So, the majority of defendants are in the age group of 25 to 45 years. Comparing the Decile Score of all the three age groups (Figure 2) it can be seen that defendants with high Decile Scores belong to the age category of Less than 25.

Similarly, we observe that there are more Male defendants than Female (Figure 3). From (Figure 4) we see that African-Americans make up the majority of defendants, followed by Caucasians and Hispanics. Discussing the score text that is a likelihood of re-offending (recidivism) we can see that there are three categories low, medium, and high and most of the defendants belong to the lower count in score text(Figure 5). Further, dividing the score text data based on race (Figure 6) we see that Caucasians have the lowest probability of re-offending and African-American have the highest probability of re-offending.

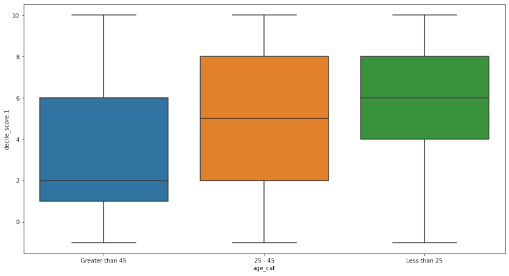


Figure 2: Box Plot of Age vs Decile Score from cox-violent-parsed-filt

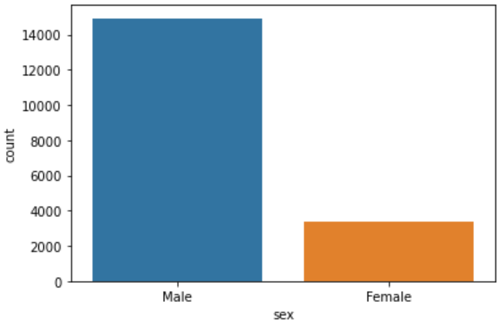


Figure 3: Graphical Representation of Sex from cox-violent-parsed-filt

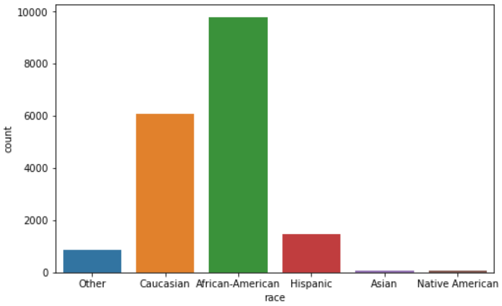


Figure 4: Graphical Representation of Ethnicity from cox-violent-parsed-filt

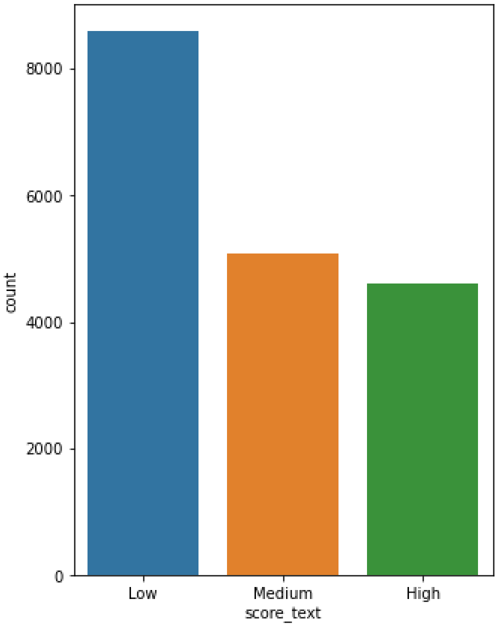


Figure 5: Graphical Representation of score text from cox-violent-parsed-filt

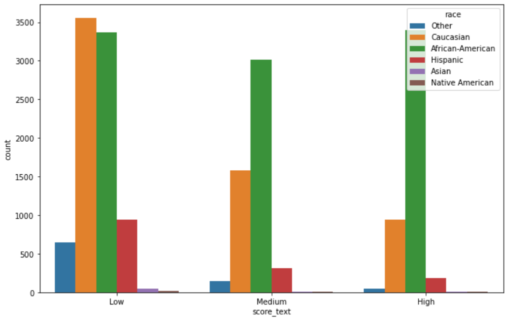


Figure 6: Graphical Representation of score text from cox-violent-parsed-filt

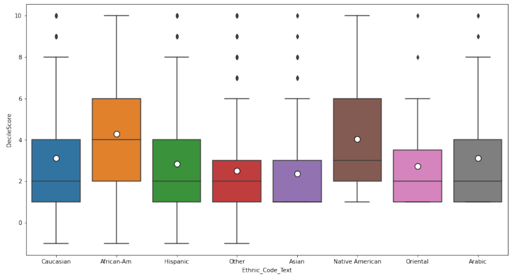


Figure 7: Box Plot of Ethnicity vs Decile Score from cox-violent-parsed-filt

On plotting a graph of Ethnicity with respect to the Decile score (Figure 7) we observe that the range of African American is the highest followed by Native Americans. One reason for the Native Americans to be in the Higher Decile range could be insufficient Data for that particular race as we already saw in the previous plots (Figure 4).

Methodology

To predict if a defendant is likely to re-offend, we create a binary classifier which predicts the label 0/1. The tried several classification methods and added explanations to them using inherent functionality or by using external tools.

Classification Methods

- **Logistic Regression:** The method of modeling the likelihood of a discrete result given an input variable is known as logistic regression. The most frequent logistic regression models have a binary result, which might be true or false, yes or no, and so forth. Multinomial logistic regression can be used to model situations with more than two discrete outcomes. Logistic regression is a handy analytical tool for determining if a fresh sample fits best into a category in classification tasks.
- **Random Forest:** ”A Random Forest is a classifier that combines a several decision trees on different subsets of a dataset and averages them to increase the dataset's predicted accuracy.” Instead than depending on a single decision tree, the random forest collects the forecasts from each tree and estimates the final output based on the majority votes of estimation.The higher the number of trees in the forest, the more accurate it is and the problem of overfitting is avoided.
- **Deep Neural Network:** An artificial neural network (ANN) having numerous layers between the input and output layers is known as a deep neural network (DNN). Neural networks come in a variety of shapes and sizes, but they always include the same basic components: neurons, synapses, weights, biases, and functions. These components work similarly to human brains and can be trained just like any other machine learning algorithm.

XAI Methods and Tools

There are several methods and tools available that helps in adding explanations to the model.

- **Ad hoc explanations:** Logistic Regression and Random Forests have inherent explanation functionality which gives feature importance.
- **SHAP:** is a tool for analyzing the output of any machine learning model and examining how the algorithm arrived at individual datapoint predictions. The SHAP explainer may help us to determine the contribution of each feature.
- **LIME:** generates local explanations for black-box models by generating locally perturbed input data and investigating how model behavior changes toward this data. It further provides explanations of what degree and in what direction any particular attribute influences the model's prediction.
- **WIT:** is a visualization tool for analyzing a Machine Learning model that can be run inside a Python notebook. With the What-If Tool, we will measure accuracy, use partial dependence plots to examine how specific attributes affect your model's prediction(i.e. the performance of the model), and assess human-centered ML models from a fairness standpoint.
  Since, WIT provides the tools to represent what we view as bias so that we can build the most impartial AI systems possible. So, our major focus will be on WIT's analysis.
- **AIF 360:** is a comprehensive open-source toolkit of metrics to check for unwanted bias in datasets and machine learning models, and state-of-the-art algorithms to mitigate such bias.

Results and Discussion

After performing Logistic Regression, to predict score text the weights that is being used by the model. We can see that (Figure 8) Ethnic_Code_Text and Scale_ID have the most negative weights and RawScore, RecSupervisionLevelText have the most positive weights, which symbolizes that these are some of the features that our model has given the most importance to predict score text.

Similarly when we discuss about the implementation of Random Forest, to analyse the weights that is being used by the model. We see that (Figure 9) RawScore and Scale_ID are being given the highest importance by the model.

Our main objective was to understand the biasness in the system, if any. We do see in the Confusion Matrix of the results(Figure 10) using the Logistic Regression model, that the model is little biased. We can infer this by looking at the false positives on the top-right of the matrix, where the model has predicted that the defendant who is of the ethnicity African-American will re-offend whereas the actual data shows contrasting results. To reduce the biasness of our model, we need to minimize the false positives. As we can see from the Confusion matrices of the two models, random forest does seem to be much better than logistic regression as it has reduced the false positives by a significant margin.

We used LIME - Local Interpretable Model-Agnostic Explanations for adding explainability to our model. In LIME

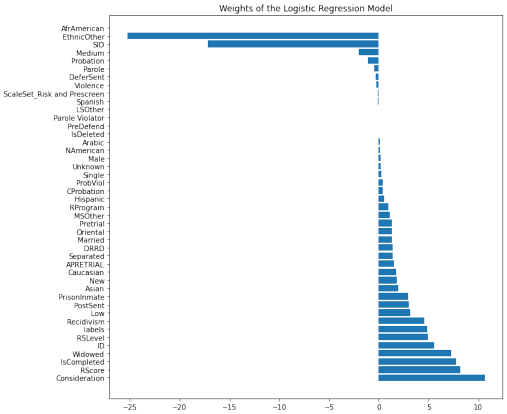


Figure 8: Features weight obtained from the Logistic Regression Model

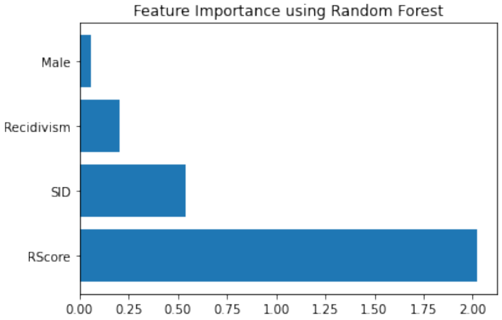
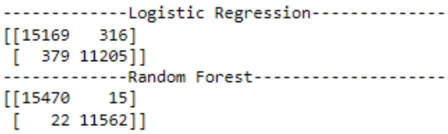
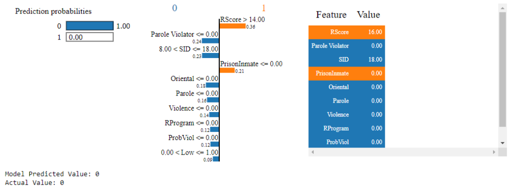


Figure 9: Features weight obtained from the Logistic Regression Model



```
-------------Logistic Regression--------------
[[15169   316]
 [  379 11205]]
-------------Random Forest--------------------
[[15470    15]
 [   22 11562]]
```

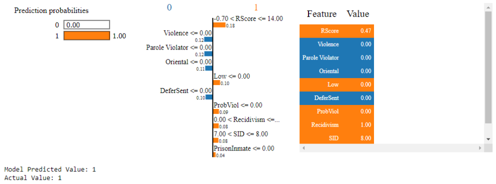
Figure 10: Features weight obtained from the Logistic Regression Model



Feature	Value
RScore	16.00
Parole Violator	0.00
SID	18.00
PrisonInmate	0.00
Oriental	0.00
Parole	0.00
Violence	0.00
RProgram	0.00
ProbViol	0.00

```
Model Predicted Value: 0
Actual Value: 0
```

Figure 11: LIME's explainability for instance 1



Feature	Value
RScore	0.47
Violence	0.00
Parole Violator	0.00
Oriental	0.00
Low	0.00
DeferSent	0.00
ProbViol	0.00
Recidivism	1.00
SID	8.00

```
Model Predicted Value: 1
Actual Value: 1
```

Figure 12: LIME's explainability for instance 2.

we used LimeTabularExplainer which Explains predictions on tabular (i.e. matrix) data. For numerical features, perturb them by sampling from a Normal(0,1) and doing the inverse operation of mean-centering and scaling, according to the means and standard deviations in the training data. For categorical features, perturb by sampling according to the training distribution, and making a binary feature that is 1 when the value is the same as the instance being explained. In LIME we are using explain_instance to explain our model. Basically it generates explanations for a prediction.First, we generate neighborhood data by randomly perturbing features from the instance. We then learn locally weighted linear models on this neighborhood data to explain each of the classes in an interpretable way. And in explain_instance we are using predict_fn – prediction function. For classifiers, this should be a function that takes a numpy array and outputs prediction probabilities. For regressors, this takes a numpy array and returns the predictions. For ScikitClassifiers, this is classifier.predict_proba(). For ScikitRegressors, this is regressor.predict(). The prediction function needs to work on multiple feature vectors (the vectors randomly perturbed from the data_row). 0 and 1 are our targets i.e. Score text(Low=0, medium and High=1) Here we can see(Figure 11) for this particular instance our model has predicted that the probability of this defendant having low chance of Recidivism is 100%. The main attributes contributing to this can be seen on the left side of the horizontal bar graph.

Similarly, for another instance and our model has again predicted (Figure 12) that the probability of this defendant having high chance of Recidivism is 100%. The main attributes contributing to this can be seen on the right side of the horizontal bar graph.

We created SHAP plots for the whole dataset. As per the plot in (Figure 13), we can see that it shows the feature importance for each data point.

We performed various analysis using the What if Tool(WIT). WIT allows us to see feature distribution plots using sampled data points. (Figure 14) depicts the data distribution colored by the feature 'COMPAS Determination'.

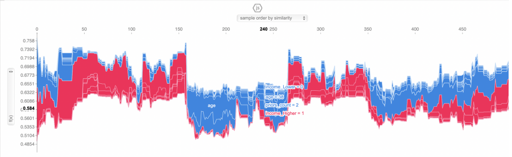
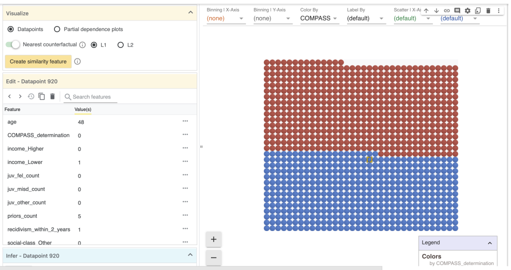


Figure 13: SHAP Values for the complete dataset using SHAP Explainer



Visualize

● Datapoints ○ Partial dependence plots

Nearest counterfactual ● L1 ○ L2

Create similarity feature

Edit - Datapoint 920

Feature	Value(s)
age	48
COMPASS_determination	0
income_Higher	0
income_Lower	1
juv_fel_count	0
juv_misd_count	0
juv_other_count	0
priors_count	5
recidivism_within_2_years	1
social-class_Other	0

Infer - Datapoint 920

Binning I X-Axis: (none)
Binning I Y-Axis: (none)
Color By: COMPASS
Label By: (default)
Scatter I X-A: (default)
(default)

Legend
Colors by COMPASS_determination

Figure 14: What If Tool showing the data distribution colored by COMPAS Determination

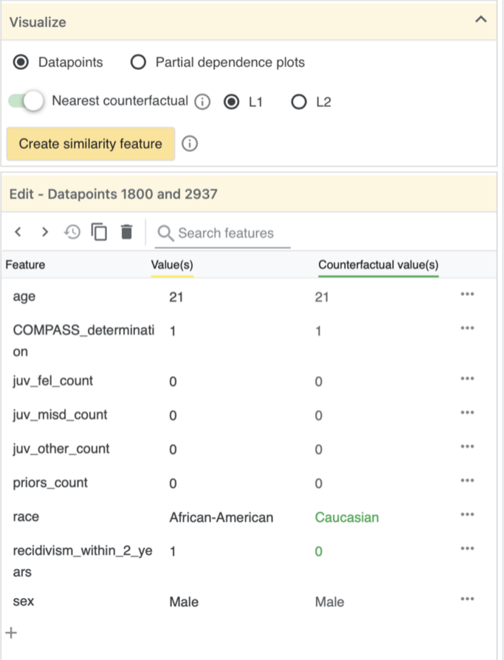
One of the key features of WIT is its capability to generate nearest counterfactual. It allows generating counterfactual based on weights given to a feature. From (Figure 15), we can see that the 2 data points only differ in race, i.e. one is 'African American' and converting it to Caucasian' in its counterfactual changes it's recidivism from 1 to 0 showing the heavy influence of the race in the prediction.
(Figure 16) talks about the fairness and performance metric provided by the WIT. It even permits setting of custom threshold for the analysis. Keeping in focus the race feature, evaluating the fairness, we can see that even though overall accuracy for both 'African American' and 'Caucasian' is the almost similar,'African American' race see a lot of False Positives as compared to the 'Caucasian' race proving the bias in the dataset.

Conclusion & Future Work

So far we have focused on evaluating bias using different tools and techniques. From our interpretation we conclude that black defendants were often predicted to be at a higher risk of recidivism than they actually were. Black defendants who did not recidivate over a two-year period were nearly twice as likely to be misclassified as higher risk compared to their white counterparts.

White defendants were often predicted to be less risky than they were. White defendants who re-offended within the next two years were mistakenly labeled low risk almost twice as often as black re-offenders.

We successfully concluded that explanations aid in detecting bias which brings us to the question *Can Explanations mitigate Bias as well?* Infact, AIF 360 does provide reweighing possibilities for the features that helps in mitigating the bias which would a prospective future work.



Visualize

● Datapoints ○ Partial dependence plots

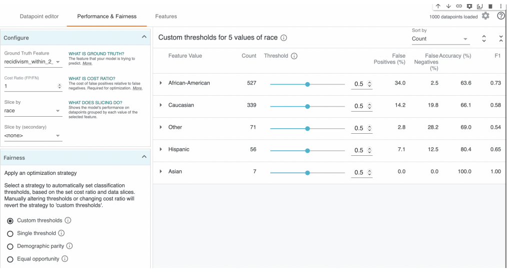
Nearest counterfactual ● L1 ○ L2

Create similarity feature

Edit - Datapoints 1800 and 2937

Feature	Value(s)	Counterfactual value(s)
age	21	21
COMPASS_determination	1	1
juv_fel_count	0	0
juv_misd_count	0	0
juv_other_count	0	0
priors_count	0	0
race	African-American	Caucasian
recidivism_within_2_years	1	0
sex	Male	Male

Figure 15: What If Tool depicting the counterfactual



Datapoint editor | Performance & Fairness | Features

1000 datapoints loaded

Configure

Ground Truth Feature: recidivism_within_2_
WHAT IS GROUND TRUTH? The feature that your model is trying to predict. More.

Cost Ratio (FP/FN): 1
WHAT IS COST RATIO? The cost of false positives relative to false negatives. Required for optimization. More.

Slice by: race
WHAT DOES SLICING DO? Shows the model's performance on datapoints grouped by each value of the selected feature.

Slice by (secondary): <none>

Fairness

Apply an optimization strategy

Select a strategy to automatically set classification thresholds, based on the set cost ratio and data slices. Manually altering thresholds or changing cost ratio will revert the strategy to 'custom thresholds'.

● Custom thresholds
○ Single threshold
○ Demographic parity
○ Equal opportunity

Custom thresholds for 5 values of race

Sort by: Count

Feature Value	Count	Threshold	False Positives (%)	False Negatives (%)	Accuracy (%)	F1
African-American	527	0.5	34.0	2.5	63.6	0.73
Caucasian	339	0.5	14.2	19.8	66.1	0.58
Other	71	0.5	2.8	28.2	69.0	0.54
Hispanic	56	0.5	7.1	12.5	80.4	0.65
Asian	7	0.5	0.0	0.0	100.0	1.00

Figure 16: What If Tool showing the performance and Fairness parameters

Contribution

Both the team members equally contributed and worked towards the project.

Acknowledgement

Authors are highly thankful to Prof. David Leake, Luddy School of Informatics, Computing, and Engineering, Indiana University-Bloomington for his valuable suggestions.

References

1. https://arxiv.org/pdf/2106.07483.pdf
2. Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. Fairness and Machine Learning. fairmlbook.org. http://www.fairmlbook.org.
3. Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta,Aleksandra Mojsilovic, et al. 2018. AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv preprint arXiv:1810.01943 (2018).
4. Reuben Binns. 2019. On the Apparent Conflict Between Individual and Group Fairness. arXiv preprint arXiv:1912.06883 (2019).
5. Andrea Brennen. 2020. What Do People Really Want When They Say They Want" Explainable AI?" We Asked 60 Stakeholders.. In Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems. 1–7.
6. Sam Corbett-Davies and Sharad Goel. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. (jul 2018). arXiv:1808.00023 http://arxiv.org/abs/1808.00023
7. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm
8. https://pair-code.github.io/what-if-tool/ai-fairness.html
9. https://docs.responsibly.ai/notebooks/demo-compas-analysis.html
10. Pratik Gajaneand Mykola Pechenizkiy.2018. On Formalizing Fairnessin Prediction with Machine Learning. Technical Report. arXiv:1710.03184v3
11. Alex Hern. 2018. Google's solution to accidental algorithmic racism: ban gorillas. (2018).
12. Ehsan Toreini, Mhairi Aitken, Kovila Coopamootoo, Karen Elliott, Carlos Gon- zalez Zelaya, and Aad van Moorsel. 2020. The Relationship between Trust in AI and Trustworthy Machine Learning Technologies. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (Barcelona, Spain) (FAT* '20). Association for Computing Machinery, New York, NY, USA, 272–283. https://doi.org/10.1145/3351095.3372834
13. Hands-On Explainable AI (XAI) with Python: Interpret, visualize, explain, and integrate reliable AI for fair, secure, and trustworthy AI apps by Denis Rothman
14. https://www.science direct.com/topics/computer-science/logistic-regression
15. Bianca Zadrozny.2004.Learning and evaluating Classifiers under sample selection bias. In Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004. 903–910. https://doi.org/10.1145/1015330.1015425