Uncovering Bias and Ensuring Fairness: A Comprehensive Analysis of the COMPAS Algorithm

Dev Divyendh Dhinakaran
G01450299
College of Engineering and Computing
George Mason University
ddhinaka@gmu.edu

Sai Abhishek Nemani
G01462099
College of Engineering and Computing
George Mason University
snemani4@gmu.edu

***Abstract*—In this project, we delve into the heart of recidivism and eligibility for pretrial release concerns by conducting a rigorous fairness and bias analysis within the realm of the COMPAS dataset. Our research seeks to unveil latent biases and disparities that may inadvertently perpetuate discrimination in algorithmic decision-making. Through the meticulous examination of predictive algorithm outputs, we aim not only to ensure fairness and transparency at the individual level but also to uphold the broader public trust in the criminal justice system's integrity and equity. This investigation serves as a pivotal exploration of the intricate interplay between data-driven algorithms and justice, ultimately contributing to a more equitable and transparent legal landscape.**

I. Introduction

In recent years, the criminal justice system has witnessed a profound transformation with the increasing adoption of predictive algorithms, which play a pivotal role in aiding decision-making processes. These algorithms, including the widely known COMPAS system, utilize extensive datasets to predict various aspects of a defendant's case, such as the likelihood of reoffending or the potential for pretrial release. While such technological advancements hold promise in enhancing the efficiency of the criminal justice system, they also raise important ethical and fairness concerns. The significance of studying fairness in algorithmic decision-making, particularly in the context of the COMPAS dataset, lies in its potential to uncover biases and disparities that may inadvertently perpetuate discrimination within the system. As these algorithms have gained prominence, the need to scrutinize their outputs for fairness, transparency, and equity has become paramount, not only to ensure justice for individuals but also to maintain public trust in the fairness and integrity of the criminal justice system as a whole.

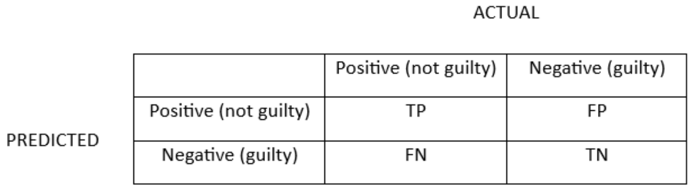
II. Problem Statement

The utilization of algorithms based on the COMPAS dataset in the criminal justice system has raised concerns due to the observed biases in their decision-making processes. These biases can result in unfair treatment and disproportionate impacts on certain demographic groups, undermining the core principles of justice and equity. The current challenge is to create an algorithm that eliminates prejudice and increases fairness in its predictions, resulting in more equitable outcomes in the criminal justice system.

III. Objective

In our project we have Five main objectives:

- Algorithm Development:
  Design and develop a novel predictive algorithm for assessing defendant-related outcomes, such as recidivism risk or pretrial release eligibility, using the COMPAS dataset as a foundational resource.
- Bias Mitigation:
  Implement strategies and techniques to mitigate bias within the newly developed algorithm. This includes identifying and addressing biases related to race, gender, or any other sensitive attributes that may exist in the data.
- Fairness Assessment:
  Utilize fairness metrics and statistical analysis to rigorously evaluate the fairness of the developed algorithm. Assess its performance with respect to various fairness criteria, such as disparate impact and equal opportunity, to ensure equitable predictions.
- Reduce **Type 2 Error**:
  Reduce the False Negative Rate to ensure that there are fewer opportunities for an innocent person to end up in prison.
- Reduce **Type 1 Error**:
  Reduce the False Positive Rate to ensure that the guilty people are punished



		ACTUAL	
		Positive (not guilty)	Negative (guilty)
PREDICTED	Positive (not guilty)	TP	FP
	Negative (guilty)	FN	TN

IV. Dataset Description

COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) Overview:
The COMPAS dataset is a comprehensive collection of data on individuals involved in the criminal justice system. It is largely used to estimate the likelihood of recidivism and influence choices about pretrial release, punishment, and parole. The dataset has received a lot of attention in recent years because of questions about fairness, transparency, and potential bias in the COMPAS system's algorithmic judgments.

Attributes:
The COMPAS dataset has a total of 47 columns and 11757 entries. It includes a variety of factors that provide useful insights into the criminal histories and demographic features of the people it profiles.
The COMPAS dataset comprises a diverse set of attributes essential for assessing individuals' involvement in the criminal justice system. These attributes encompass demographic information such as age, gender, race, and ethnicity. Additionally, the dataset includes in-depth records of criminal history, detailing prior convictions, offense types, and sentencing history. Central to the dataset are risk scores, which estimate the likelihood of recidivism and inform significant legal decisions. Moreover, case-specific details, including the charges filed, jurisdiction, and legal outcomes, are documented. These attributes collectively form a rich resource for analyzing fairness and bias in algorithmic decision-making within the legal domain.

Dataset Link: https://www.kaggle.com/datasets/danofer/compass

V. Anticipated Hurdles:

The challenges often revolve around ethical, legal, and technical considerations.

- Bias and Fairness Challenges:
  The dataset has been criticized for potential racial and gender biases in its predictions. Addressing and mitigating these biases while developing fairer algorithms can be a complex and sensitive task.
- Algorithm Building:
  Developing a predictive algorithm that outperforms existing models while reducing bias can be complex.
- Interpreting Results:
  Interpreting and communicating the results regarding fairness assessments, can be challenging.

VI. Limitations on Previous Studies

Limitation 1: Inherent Trade-off Between Fairness Criteria: The issue is that there is an inherent conflict between two fairness requirements for risk assessments. [1]

- Northpointe's Fairness Criterion: He guarantees that reoffending rates are comparable across risk categories, regardless of race.
- ProPublica's Fairness Criterion: It focuses on total recidivism rates, which differ between black and white convicts.

These two criteria are challenging to satisfy simultaneously, so we need to perform a trade-off.

Limitation 2: Risk Classification Linked to Race-Related Attributes: [2]
Although Northpointe's methodology does not use race as a direct input, certain race-related characteristics, such as prior arrests, have an impact on classification. Because of these characteristics, the computer labels more black defendants as high risk.
As a result, even if they do not reoffend, a disproportionate percentage of black defendants are labeled as high risk.

Limitation 3: Predicted Risk Imbalance for Non-Reoffending Defendants: [1]
Black defendants who do not reoffend are predicted to be riskier than white defendants with a similar non-reoffending status. This discrepancy is a criticism of the algorithm raised by ProPublica.

VII. Evaluation Methodology

As we intend to focus on lowering the Bias factor, we would like to use the following evaluation metrics to assess the **Fairness** of our algorithm:

- **False Positive Rate (FPR):**
  The FPR quantifies the proportion of false positive predictions among all actual negatives.
  $$FPR = FP/(TN + FP) \tag{1}$$
- **False Negative Rate (FNR):**
  The FNR measures the proportion of false negative predictions among all actual positives.
  $$FNR = FN/(TP + FN) \tag{2}$$
- **Area Under the Receiver Operating Characteristic Curve (AUC-ROC)**
  ROC curves plot the true positive rate (TPR) against the false positive rate (FPR) at various decision thresholds.
- **Confusion Matrix**
  A confusion matrix is a table used in machine learning and statistics to evaluate the performance of a classification model, particularly for binary classification problems. They have 4 key features:
  - True Positive (TP)
  - True Negative (TN)
  - False Positive (FP)
  - False Negative (FN)

References

[1] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," in *Proceedings of Innovations in Theoretical Computer Science (ITCS)*, 2017.

[2] S. M. Julia Angwin, Jeff Larson and L. Kirchner, "Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks." 2016.