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

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

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. Over View

In our project, we embarked on an in-depth analysis of the COMPAS dataset, a critical examination of the Criminal Offender Management Profiling for Alternative Sanctions. Our project can be divided into several key phases:

A. Data Collection:

We began by acquiring the COMPAS dataset, which contains a wealth of information related to criminal cases, defendant demographics, and recidivism risk assessments.

B. Data Cleaning and Preprocessing:

The initial step in our project involved cleaning the dataset to address missing values and outliers, ensuring the data's reliability. We also standardized and normalized certain features to facilitate more effective analysis.

C. Feature Extraction:

One of the essential steps in data preparation was feature extraction. This included creating new variables that encapsulated meaningful information, such as combining juvenile felony and misdemeanor counts to gauge the seriousness of prior offenses. These new features allowed for a more comprehensive assessment of defendant characteristics.

D. Bias Analysis:

A significant focus of our project was on analyzing potential bias in the dataset, particularly with regard to African Americans and their trials. We sought to investigate if racial disparities existed in the COMPAS scores and their impact on sentencing outcomes. Our analysis aimed to answer important questions about fairness and equity within the criminal justice system.

E. Interpretation and Visualization:

We presented our findings through the lens of data visualization, which provided clear insights into the distribution of COMPAS scores, their relationship to race, and their impact on recidivism predictions. These visualizations allowed for a better understanding of the dataset's dynamics.

F. Future Directions and Impact

We will discuss about the problem and our idea about how to overcome the Bias and make the COMPAS algorithm less vulnerable to Bias

III. 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 52 columns and 18316 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

IV. Data Preprocessing

A. Feature Selection:

A crucial step in the data preparation process that enables the model to be fine-tuned for best performance is feature subset selection. During this phase, the emphasis is on locating and keeping the most pertinent and significant data columns while removing those that add noise or add complexity without adding anything of value. 39 columns, including [compas screening date, "id," "name," "first," "last," "dob," "screening date," and others] that were less important for decision-making have been removed. // This reduces the Dimensionality of our Dataset to a great extent and improve the performance of the models.

B. Data Cleaning

Data preprocessing is the cornerstone of any successful data analysis or machine learning endeavor. It's the process of refining raw data into a structured and optimized form, making it ready for deeper exploration and modeling. In our project, we tackled this essential phase with precision and care. Wwe did the following prepossessing steps

- Handling Null values
- Changing column names
- Changing categorical values to numeric values using label Encoder

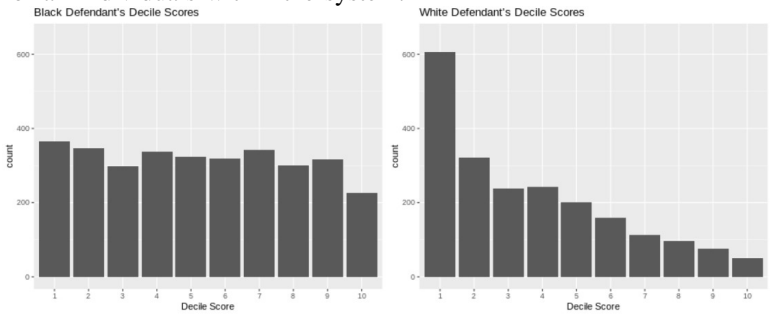
We began by identifying and handling null or missing values, ensuring that our data was complete and reliable. Next, we addressed the clarity and intelligibility of our dataset by renaming columns to more meaningful and informative names. This not only enhanced the interpretability of the data but also laid the groundwork for more intuitive analysis. Additionally, we harnessed the power of label encoding to transform categorical values into numerical representations, a fundamental step in making our data compatible with machine learning algorithms. These preprocessing measures together acted as the foundation upon which we built our analytical framework, empowering us to derive valuable insights from the data and make informed decisions with confidence.

V. Feature Extraction

In pursuit of a more comprehensive understanding of our dataset, we strategically combined two columns to derive a new one that encapsulated meaningful information. By subtracting the values of 'juv misd count' from 'juv fel count,' we created a novel column, 'juv fel seriousness.' This new feature provided valuable insights into the seriousness of prior juvenile offenses, a factor that could play a crucial role in assessing recidivism risk. This strategic column combination approach demonstrated our commitment to extracting richer insights from the data and refining our dataset for more informed analysis.

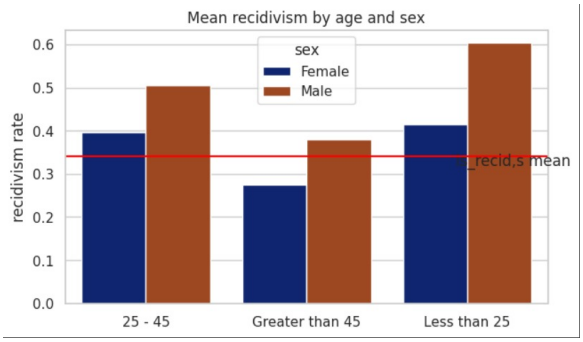
VI. Analysing Bias

The COMPAS dataset has been a focal point of scrutiny due to concerns of racial bias in the criminal justice system. Numerous studies and analyses have indicated potential disparities, particularly affecting African American individuals. One of the key areas of concern is the assignment of risk assessment scores, which can significantly influence sentencing outcomes. Research has shown that African Americans, even when controlling for other factors, may receive higher risk scores compared to their counterparts. These disparities have ignited discussions on equity, fairness, and the need for reform within the criminal justice system. The data-driven insights gleaned from such analyses underscore the importance of addressing racial bias and striving for a more equitable legal framework that ensures equal treatment for all individuals within the system.

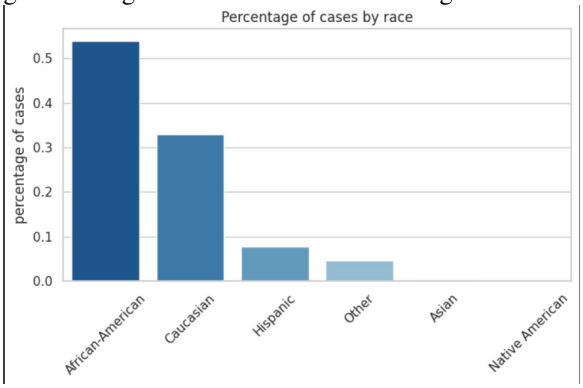


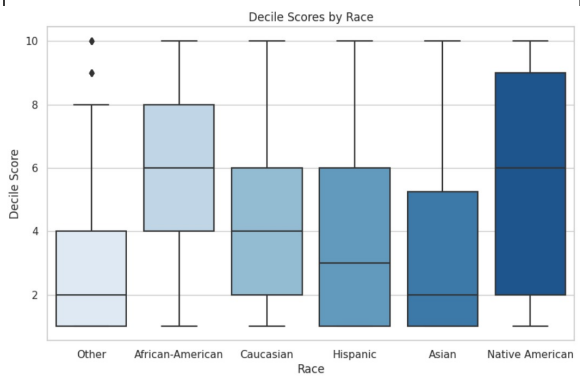
Our Findings:

- Compared to white defendants, black defendants have a 45 percent higher likelihood of receiving a higher score that accounts for the severity of their offense, prior arrests, and potential future criminal behavior.
- 19.4 percent more women than men are likely to receive a higher grade



- people under 25 are 2.5 times as likely to get a higher score as middle aged defendants





VII. Preliminary Trials

In our preliminary trial, we employed two distinct machine learning models, K-Nearest Neighbors (KNN) and Decision Tree, to gain initial insights into the predictive potential of the COMPAS dataset. The KNN model exhibited an accuracy rate of 61%, while the Decision Tree model surpassed it with an accuracy of 82%. These preliminary findings provide a promising glimpse into the dataset's suitability for predictive modeling. The notable contrast in accuracy rates between the models hints at the potential complexity and non-linearity of the relationships within the data. As we delve deeper into the project, these initial results will guide our model selection and refinement, offering a foundation upon which we can build more accurate and robust predictive systems for assessing recidivism risk.

VIII. Progress and Upcoming Endeavours

Addressing Bias: The Road Ahead with Random Forest and Boosting Methods

In our commitment to promoting fairness and addressing bias within the COMPAS dataset, we're embarking on a promising journey. As a future plan, we aim to employ advanced techniques, particularly Random Forest and Boosting methods, to train and fine-tune our predictive models. These methods have the capability to iteratively learn and correct the biases that may exist in the data. By retraining our models, we intend to develop more equitable and reliable risk assessment tools. Our goal is to minimize the disparities that have been identified, especially concerning racial bias, and enhance the accuracy and fairness of our predictions. This exciting path forward will not only refine the quality of our models but also contribute to the ongoing dialogue on fairness and equity within the criminal justice system. Our future endeavors aim to be a testament to our dedication to data-driven solutions that prioritize justice and fairness for all.

IX. Evaluation Methodology - Future Endeavours

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)