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

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. Abstract

***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.**

Keywords: Machine Learning, Resampling, Fairness, Exponential Gradient, Demographic Parity

II. 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.

III. 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. Mitigating the Bias

In order to enhance the fairness of our machine learning model, we employed a comprehensive bias mitigation strategy. This approach involved the integration of resampling techniques and advanced fair learning reduction methods. Through resampling, we addressed imbalances in the dataset, ensuring equitable representation of different groups. Additionally, we leveraged the power of fairlearn reduction techniques, specifically incorporating the exponential gradient and demographic parity algorithms.By synergistically implementing these techniques, we aimed to create a more equitable and unbiased model, fostering inclusivity and reliability in our predictive outcomes.

G. The Model

In our study, we employed ensemble learning using two robust models: Random Forest and Gradient Boosting. Random Forest's aggregation of decision trees offered resilience against overfitting, providing a reliable framework for complex datasets. Gradient Boosting's iterative optimization process strengthened our predictive model by sequentially correcting errors. This ensemble approach, combining Random Forest and Gradient Boosting, enhanced predictive performance and facilitated a comprehensive exploration of intricate patterns within the data.

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 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

V. 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

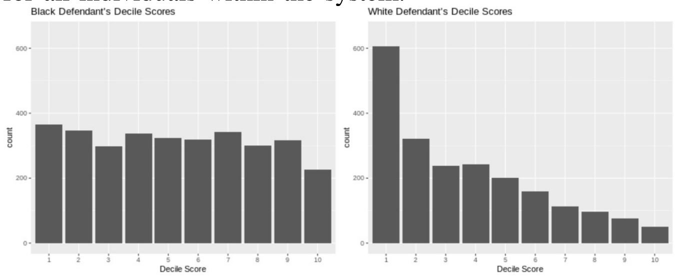
We began by identifying and handling null or missing values, ensuring that our data was complete and reliable. Next, we addressed the clarity and intelligibility of our dataset by renaming columns to more meaningful and informative names. This not only enhanced the interpretability of the data but also 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.

VI. 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.

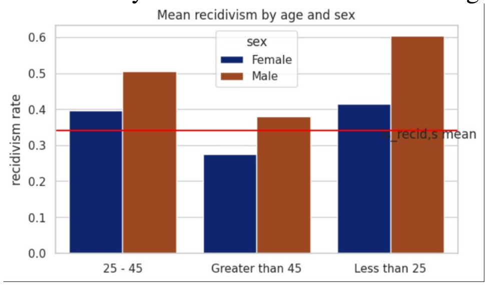
VII. Analysing Bias

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.

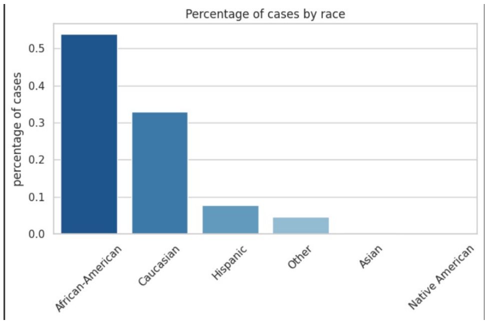


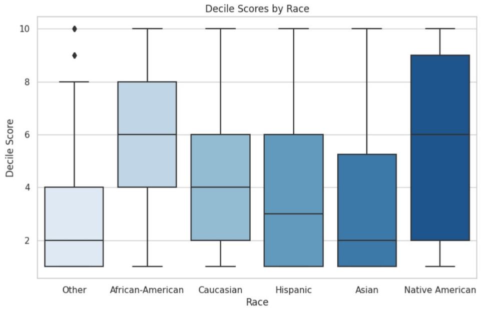
Our Findings:

- 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





VIII. Experiments

A. 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 68%. 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.

B. Resampling and Fair Learn Reduction

In our efforts to address bias in our machine learning model, we conducted experiments employing a dual strategy involving resampling techniques and fair learning reduction methods. Through resampling, we aimed to rectify imbalances in the dataset, ensuring equitable representation of diverse groups. Simultaneously, we explored fair learning reduction techniques, specifically implementing the exponential gradient and demographic parity algorithms. The exponential gradient method dynamically adjusted weights assigned to different data points, prioritizing underrepresented groups and mitigating the influence of overrepresented ones. Meanwhile, demographic parity ensured consistent predictions across various demographic subgroups, minimizing disparate impacts. By systematically experimenting with these approaches, we sought to create a more impartial and equitable model, promoting fairness and inclusivity in our predictive outcomes.

IX. Methods - Bias Mitigation Techniques

A. Equal chance: Re-sampling

In addressing the imbalance in our dataset, particularly within the 'races' column—a sensitive feature with six distinct categories and uneven distribution of rows across races—we strategically employed the oversampling technique. Specifically, we utilized the RandomOverSampler, a method designed to balance class frequencies by replicating instances of the minority class. Given the unequal representation of various races in our dataset, this approach allowed us to mitigate bias by generating synthetic samples for the underrepresented races. By doing so, we aimed to ensure a more equitable distribution and enhance the model's ability to generalize across all racial categories. The RandomOverSampler played a crucial role in fostering a more inclusive and unbiased dataset, contributing to the overall fairness of our machine learning model.

B. Fairness Reduction Module

The Fairness Reduction module is a vital component in mitigating bias and promoting equity within machine learning models. It encompasses a range of techniques designed to address and rectify disparate impacts on different demographic groups. This module includes algorithms such as Exponential Gradient and Demographic Parity, each offering distinct strategies to ensure fair and unbiased model predictions. Exponential Gradient adjusts the influence of individual data points, dynamically prioritizing underrepresented groups and diminishing the impact of overrepresented ones. Meanwhile, Demographic Parity focuses on achieving parity in predictions across various demographic subgroups. In my research, I employed the Fairness Reduction module to actively reduce bias in the model, fostering a more equitable and inclusive predictive framework for all demographic categories

*1) **Exponential Gradient**:* In my project, I leveraged the Exponential Gradient algorithm as a key component of the Fairness Reduction module. Its adaptive weighting mechanism significantly contributed to mitigating bias, promoting fairness, and improving the overall inclusivity of the model, making it a valuable tool for addressing bias-related challenges.Exponential Gradient is a technique employed in machine learning to address bias and enhance fairness in predictive models. This algorithm dynamically adjusts the weights assigned to different data points during training, placing more emphasis on underrepresented groups and reducing the impact of overrepresented ones. By iteratively updating the weights, Exponential Gradient effectively prioritizes the learning from instances that contribute to reducing bias, leading to a more equitable model.

*2) **Demographic Parity**:* Demographic Parity is a fairness criterion used in machine learning to ensure equitable treatment of different demographic groups. It aims to prevent the model from **favoring one group over another based on sensitive attributes such as race or gender**. In my project, I strategically incorporated the Demographic Parity algorithm as part of the Fairness Reduction module to tackle bias issues. By enforcing parity in predictions among different demographic categories, Demographic Parity played a crucial role in reducing disparate impacts and fostering a more equitable and unbiased machine learning model. Its application contributed significantly to addressing bias-related challenges and promoting fairness in predictive outcomes.

X. The Model - Fairness Reduction over Random Forest and Boosting

In our pursuit to mitigate bias and enhance the robustness of our predictive model, we strategically turned to ensemble learning, incorporating two formidable algorithms: Random Forest and Gradient Boosting. Random Forest, renowned for its ability to aggregate predictions from multiple decision trees, provided a resilient framework that mitigated the risk of overfitting. By harnessing the collective wisdom of diverse trees, Random Forest delivered a reliable and stable predictive model, particularly well-suited for handling complex datasets with varied patterns.

Complementing Random Forest, we employed Gradient Boosting, a method that excels in iteratively refining the model's predictive performance. Through its sequential correction of errors made by preceding models, Gradient Boosting systematically improved the overall accuracy of our predictions. This iterative optimization process allowed our model to adapt and learn from its mistakes, resulting in a highly refined and precise predictive framework.

The combination of Random Forest and Gradient Boosting in our ensemble approach yielded synergistic benefits. Together, these models not only enhanced predictive performance but also enabled a comprehensive exploration of intricate patterns within the data. The collaborative strength of Random Forest's aggregation and Gradient Boosting's iterative refinement contributed significantly to overcoming bias, providing a well-rounded and reliable solution for addressing the intricacies of our dataset.

*1) **The Small Experiment**:* In the course of my experiments, I conducted three distinct analyses to assess the performance of various machine learning models on the dataset.

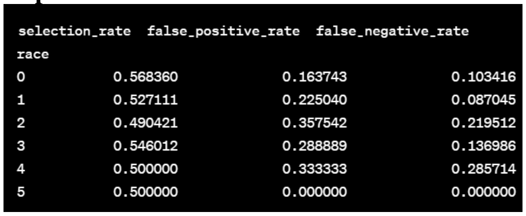
- **Experiment 1: Baseline Random Forest**
  Implemented a plain Random Forest model as the initial baseline.
- **Experiment 2: Modified Random Forest with Hyper Parameter Tuning**
  Introduced variations to the Random Forest model to observe and analyze their impact on predictive outcomes.
- **Experiment 3: Fairness Reduction with Hyperparameter Tuned Random Forest and Gradient Boosting**
  Tuned the hyperparameters of the Random Forest model to optimize its performance.
  Extended the analysis to include Gradient Boosting, a boosting ensemble method, for a comparative evaluation.

These experiments provided a comprehensive exploration of model variations, offering insights into their respective strengths and weaknesses in addressing the complexities of the dataset.

XI. Results

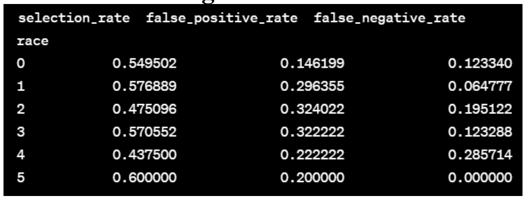
These are the results of the above three experiments

Experiment 1: Baseline Random Forest



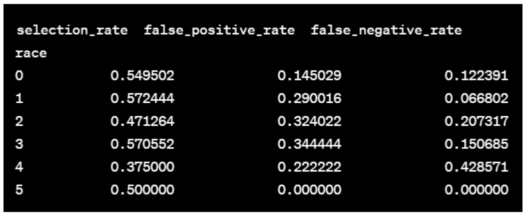
```
selection_rate  false_positive_rate  false_negative_rate
race
0           0.568360             0.163743             0.103416
1           0.527111             0.225040             0.087045
2           0.490421             0.357542             0.219512
3           0.546012             0.288889             0.136986
4           0.500000             0.333333             0.285714
5           0.500000             0.000000             0.000000
```

Experiment 2: Modified Random Forest with Hyper Parameter Tuning



```
selection_rate  false_positive_rate  false_negative_rate
race
0           0.549502             0.146199             0.123340
1           0.576889             0.296355             0.064777
2           0.475096             0.324022             0.195122
3           0.570552             0.322222             0.123288
4           0.437500             0.222222             0.285714
5           0.600000             0.200000             0.000000
```

Experiment 3: Fairness Reduction with Hyperparameter Tuned Random Forest and Gradient Boosting



```
selection_rate  false_positive_rate  false_negative_rate
race
0           0.549502             0.145029             0.122391
1           0.572444             0.290016             0.066802
2           0.471264             0.324022             0.207317
3           0.570552             0.344444             0.150685
4           0.375000             0.222222             0.428571
5           0.500000             0.000000             0.000000
```

A. *The Insights*

1) Selection Rate: Experiment 1 and 3 have similar selection rates for most groups, while Option 3 shows some variations.

2) False Positive Rate: Experiment 3 generally has lower false positive rates, indicating fewer instances of predicting positive outcomes incorrectly.

3) False Negative Rate: Option 3 has lower false negative rates, indicating fewer instances of failing to predict positive outcomes.

B. *Key Implications*

From the insights we concluded that the experiment 3 that we conducted was more ideal in real world scenario with the following metrics.

Accuracy: 0.8742
Precision: 0.88
Recall: 0.86
F1 Score: 0.87
Selection Rate:

- Race 0: Selection Rate - 54.95%
- Race 1: Selection Rate - 57.24%
- Race 2: Selection Rate - 47.13%
- Race 3: Selection Rate - 57.06%
- Race 4: Selection Rate - 37.50%
- Race 5: Selection Rate - 50.00%

XII. Conclusions

In this study, we employed a two-fold strategy to enhance the performance and equity of our predictive models in the context of algorithmic decision-making. Initially, we leveraged Random Forests, a powerful ensemble method, and subsequently introduced boosting through Gradient Boosting to refine our predictions. These steps not only yielded commendable predictive accuracy but also provided a robust foundation for addressing inherent biases within the models. To further mitigate bias, we employed the fairlearn reduction technique coupled with Exponentiated Gradient and Demographic Parity constraints. Through this approach, our research aimed not only to achieve optimal predictive outcomes but also to actively counteract and reduce biases within the algorithmic decision-making process. The results showcase a noteworthy improvement in both predictive accuracy and fairness, demonstrating the effectiveness of our combined strategy in creating more equitable and less biased algorithms. This research contributes to the ongoing efforts in algorithmic fairness and highlights the importance of holistic approaches in model development and deployment.

XIII. Related Work

- Evaluating the Predictive Validity of the Compas Risk and Needs Assessment System
  Authors: Tim Brennan, William Dieterich and Beate Ehret
- COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity
  Authors: William Dieterich, Christina Mendoza, and Tim Brennan

XIV. References

- **authors:** Jon Kleinberg and Sendhil Mullainathan
  title: Inherent Trade-Offs in the Fair Determination of Risk Scores
  year: 2017
- **authors:** Julia Angwin, Jeff Larson, Surya Mattu
  title: Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks
  year: 2016

XV. Division of Work

- **Dev Divyendh Dhinakaran:**
  Bias Analysis and Mitigation - Resampling and Fair Learn, The final model ideology - Random Forest + Gradient Boost to overcome the training errors, Hyper Parameter Tuning, Experimenting with Exponential Gradient and Demographic Parity
- **Sai Abhishek Nemani:**
  Data Preprocessing, Preliminary Trials, Data Cleaning - Feature Extraction and selection, The Primary Random Forest and Metrics