Chuer Yang

DATASCI 154: Data Science for Social Impact

Professor Allcott

Spring 2025

DSSI Rearrest Prediction

1. Introduction

Pretrial detention decisions carry serious consequences for both individuals and society. Traditionally, judges have relied on their discretion and experience to determine whether a defendant should be detained or released before trial. In recent years, there has been growing interest in using data-driven algorithms to support these decisions—particularly in predicting the likelihood of rearrest. These tools offer the potential for greater consistency and a reduction in individual biases, but they also raise complex and important questions about fairness. This project explores how accurately we can predict rearrest using data available at arraignment, and how fairness can be defined and achieved across different demographic groups.

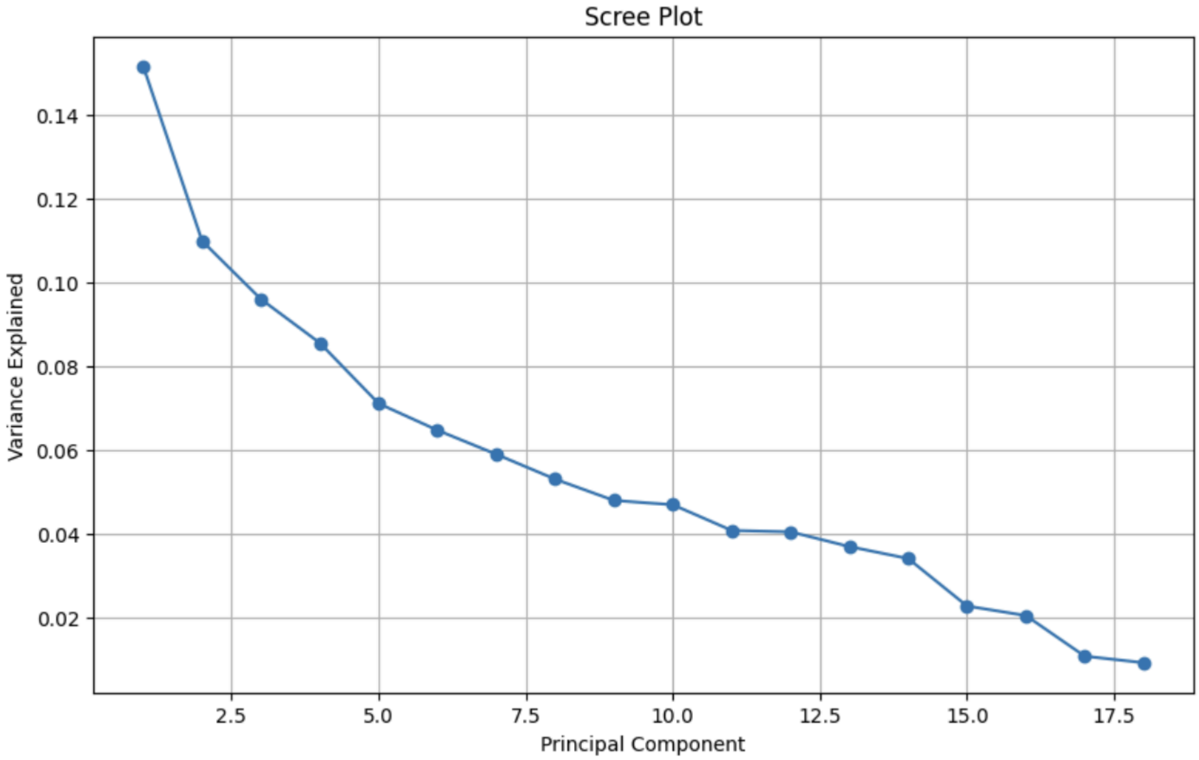
Fairness in this context is a nuanced issue, with multiple competing definitions.Although algorithms can mitigate some forms of human bias, they are not inherently fair. Ensuring fairness requires thoughtful consideration of how fairness is defined, the quality and limitations of available data, and how model outputs are interpreted. Two major challenges arise in this work: defining what fairness means in practice, and choosing an appropriate threshold for classifying individuals as high risk based on their predicted probability of rearrest.

2. Data Processing

To prepare the dataset for modeling, I began by addressing missing values. For both numeric and categorical variables, I used the most frequent value imputation strategy to preserve the structure of the data without aggressive methods like mean or median imputation. I also created a new binary feature, *felony_charge*, to indicate whether the current charge was classified as a felony, capturing a key aspect of the severity of the alleged offense. I dropped the *judge_name* column and one-hot encoded categorical variables such as *charge, charge_severity, and charge_weigh*t.

I conducted a Principal Component Analysis (PCA) on the preprocessed dataset to explore the underlying structure of the features and assess dimensionality.

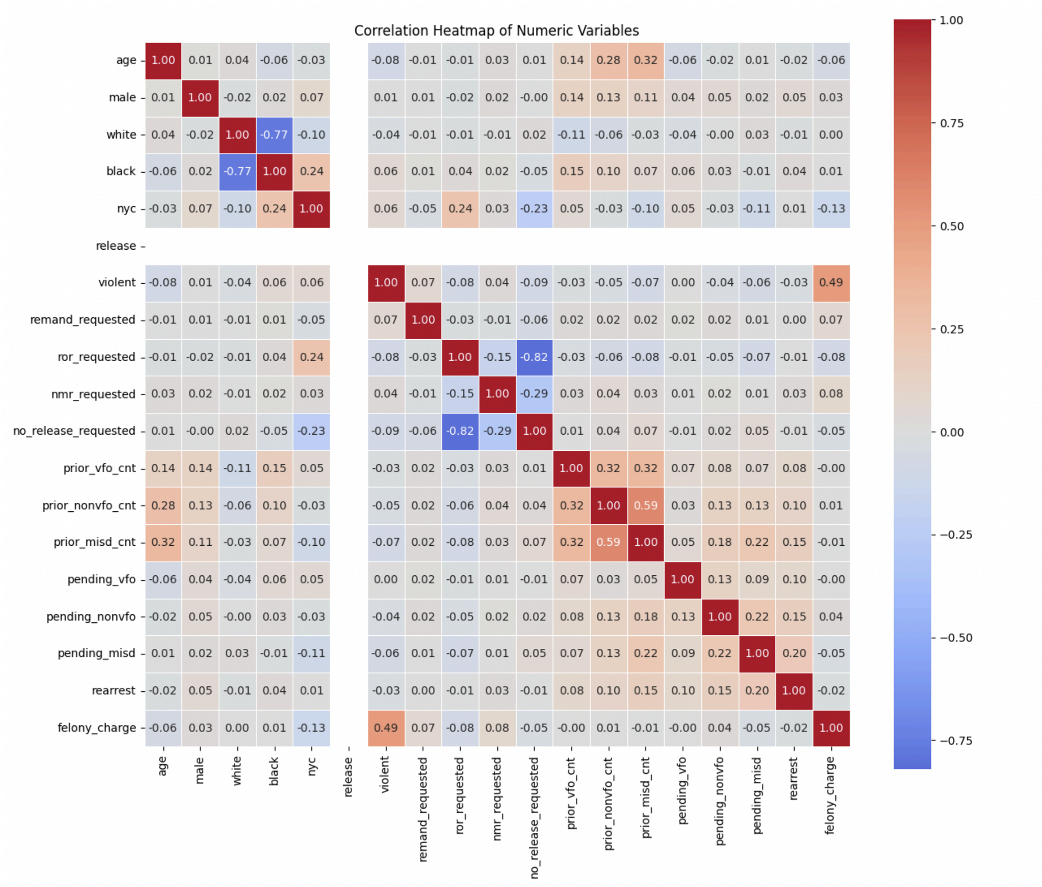
Figure 1: Principal Component Analysis



This analysis helped ensure that the input space was not dominated by redundant or highly correlated variables, and gave insight into how much variance in the data could be captured with fewer dimensions—useful for both model interpretation and fairness audits. Here, the first five components explain 50% of the variance in the data. This suggests that multiple directions contribute to the overall variance, indicating that there are numerous interactions between the features that help explain the variability among individuals.

Finally, I created a correlation heatmap to visualize the relationships between the different features.

Figure 2: Correlation heatmap



The heatmap provided insight into which variables might be collinear and guided decisions about feature selection and dimensionality reduction. The heatmap revealed that black and white are negatively correlated, which is consistent with expectations based on the dataset's composition. The variables *ror_requested* (Prosecutor requested release on recognizance) and *no_release_requested* were also negatively correlated, warranting further investigation to understand the underlying dynamics between these features. Additionally, the positive correlation between prior misdemeanor count and prior nonviolence offense count with age aligns with prior assumptions about criminal history patterns. Finally, it makes sense that violent and felony charge are positively correlated, as more serious offenses tend to align with felony charges. Overall, the data appeared to make sense, reflecting the expected relationships between key features.

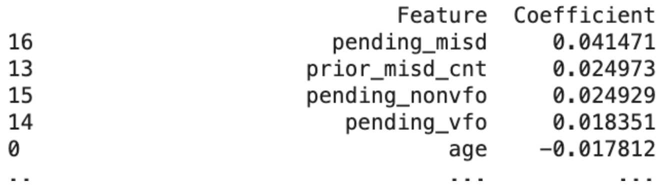
3. Pre-fair Model

I trained the initial model (pre-fair) using all available columns and applied a threshold of 50% recidivism to classify individuals as high-risk or not, following the approach used in class. The target variable is the rearrest indicator, with a value of 1 indicating rearrest and 0 indicating no rearrest. For the model, I utilized Ridge regression with an l_2 norm penalty, which applies shrinkage to the coefficients and helps to prevent overfitting:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2=\mathrm{RSS}+\lambda\sum_{j=1}^{p}\beta_j^2,$$

The data was split into training and test sets, with 70% allocated for training and 30% for testing. To assess the model's generalization, I implemented k-fold cross-validation with k values of 5, 10, and 20. This process ensures that the model is robust and not overly dependent on a particular subset of the data. I standardized the features by centering them to a mean of 0 and scaling to a standard deviation of 1.

Figure 3: Feature weights from Ridge regression



```
                         Feature  Coefficient
16                  pending_misd     0.041471
13                prior_misd_cnt     0.024973
15                pending_nonvfo     0.024929
14                   pending_vfo     0.018351
0                            age    -0.017812
..                           ...          ...
```

This ridge regression model identifies pending misdemeanors, prior misdemeanors, and pending felony charges—both non-violent and violent—as the strongest positive predictors of recidivism, with pending misdemeanors having the highest coefficient. Age is negatively associated with recidivism, indicating that older individuals are less likely to reoffend. The use of ridge regression regularizes the coefficients, but the model still places the most weight on recent and prior criminal activity, reflecting typical patterns in recidivism risk assessment.

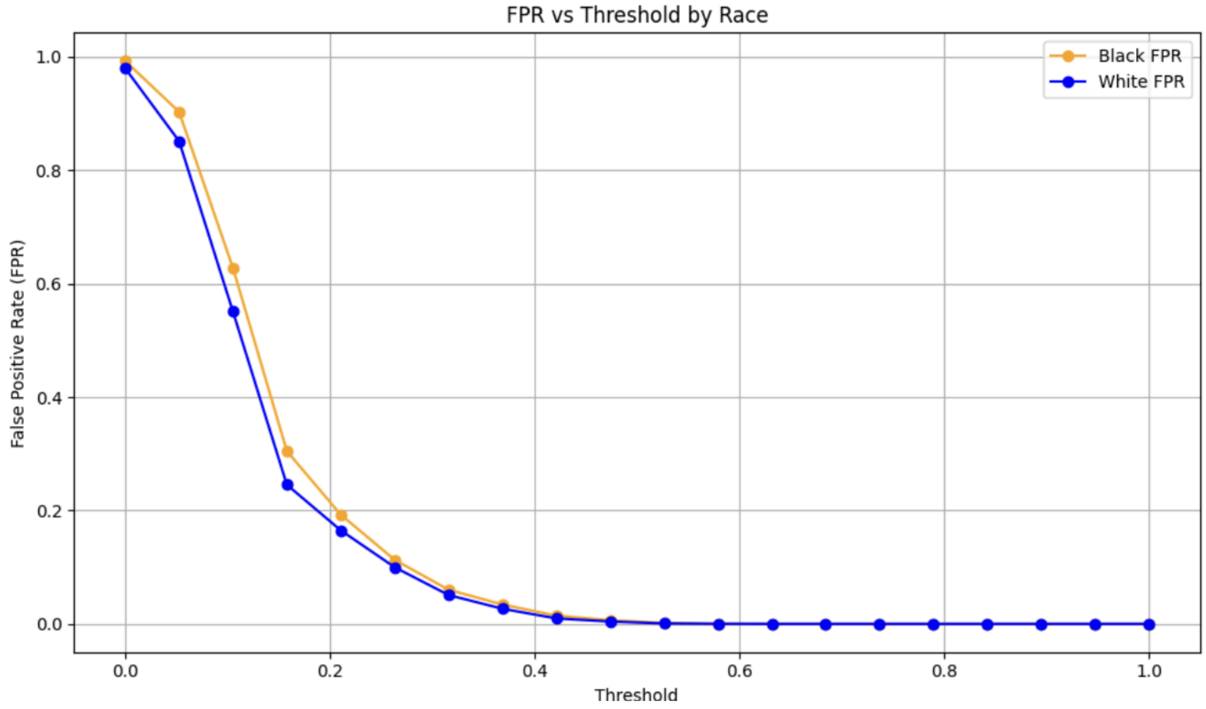
After training the pre-fair model, I evaluated its performance using the test set. The test Mean Squared Error (MSE) was 0.1148, which indicates that the model's average squared error on unseen data is around 11.5% when predicting the probability of rearrest. Given that the target variable is scaled between 0 and 1, this suggests a reasonable level of prediction accuracy. One key observation from the results is the low overfitting. The fact that the cross-validation (CV) scores and the test MSE are very similar—CV: ~0.1145 and Test MSE: 0.1151—indicates that the model is generalizing well to new, unseen data. This similarity suggests that the model has not overfitted to the training data and is capable of making reliable predictions on real-world data.

4. Fair Model

To ensure fairness in the model, I chose the error rate balance definition, which seeks to ensure that false positive and false negative error rates are equal across groups . This approach aims to minimize bias between groups and ensure that the model's performance is consistent for all individuals, regardless of their demographic characteristics. In consultation with Professor Sklansky from the Stanford Law School, he emphasized that a practical way to think about fairness is by considering the unfairness that we are trying to avoid. Based on this advice, I decided to adopt error rate balance as the definition of fairness, which aims to balance the false positive and false negative rates across groups. This balance reduces the likelihood of disproportionately impacting any particular group with incorrect classifications, thus mitigating potential harm. Additionally, there are significant economic implications to the fairness of the model. Falsely identifying someone as high-risk could result in pretrial detention, which can have long-lasting consequences on the individual's life trajectory. This highlights the need to carefully balance the risks of false positives and false negatives, ensuring that the model does not unfairly impact vulnerable populations.

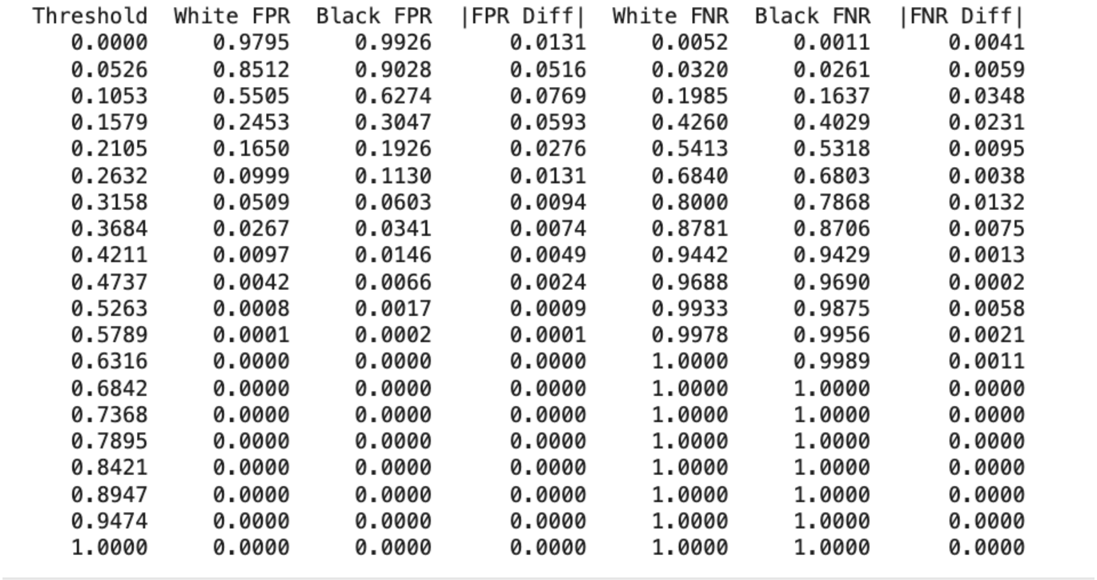
To achieve fairness, I continued using Ridge regression and focused on evaluating the tradeoffs between fairness and accuracy in the model. A crucial part of this process involves selecting the appropriate threshold for classifying an individual as high-risk or low-risk. I iterated through 20 threshold values ranging from 0 to 1, assessing their impact on false positive rates and false negative rates. I plotted the results of this threshold sensitivity analysis to visualize the tradeoff between maximizing accuracy and ensuring fairness, providing insight into how different thresholds influence the model's behavior across different subpopulations.

Figure 4: False positive rate vs. threshold by race



A highly conservative model leads to a larger disparity in false positive rates between Black and White groups. As the threshold is raised and the model becomes more lenient, the gap in false positive rates between the two groups narrows. However, this reduction in the false positive disparity comes at the expense of an increased false negative rate.

Figure 5: False positive and negative rates by race and threshold



```
 Threshold  White FPR  Black FPR  |FPR Diff|  White FNR  Black FNR  |FNR Diff|
    0.0000     0.9795     0.9926      0.0131     0.0052     0.0011      0.0041
    0.0526     0.8512     0.9028      0.0516     0.0320     0.0261      0.0059
    0.1053     0.5505     0.6274      0.0769     0.1985     0.1637      0.0348
    0.1579     0.2453     0.3047      0.0593     0.4260     0.4029      0.0231
    0.2105     0.1650     0.1926      0.0276     0.5413     0.5318      0.0095
    0.2632     0.0999     0.1130      0.0131     0.6840     0.6803      0.0038
    0.3158     0.0509     0.0603      0.0094     0.8000     0.7868      0.0132
    0.3684     0.0267     0.0341      0.0074     0.8781     0.8706      0.0075
    0.4211     0.0097     0.0146      0.0049     0.9442     0.9429      0.0013
    0.4737     0.0042     0.0066      0.0024     0.9688     0.9690      0.0002
    0.5263     0.0008     0.0017      0.0009     0.9933     0.9875      0.0058
    0.5789     0.0001     0.0002      0.0001     0.9978     0.9956      0.0021
    0.6316     0.0000     0.0000      0.0000     1.0000     0.9989      0.0011
    0.6842     0.0000     0.0000      0.0000     1.0000     1.0000      0.0000
    0.7368     0.0000     0.0000      0.0000     1.0000     1.0000      0.0000
    0.7895     0.0000     0.0000      0.0000     1.0000     1.0000      0.0000
    0.8421     0.0000     0.0000      0.0000     1.0000     1.0000      0.0000
    0.8947     0.0000     0.0000      0.0000     1.0000     1.0000      0.0000
    0.9474     0.0000     0.0000      0.0000     1.0000     1.0000      0.0000
    1.0000     0.0000     0.0000      0.0000     1.0000     1.0000      0.0000
```

The optimal threshold I would select is 0.2632, at which the false positive rate (FPR) for White individuals is 0.0999, and for Black individuals, it is 0.1130. The FPR difference between races is small, at just 0.0094. Meanwhile, the false negative rate (FNR) difference is 0.0132, with the White FNR at 0.684 and the Black FNR at 0.6803. This threshold represents a tradeoff between false positive and false negative rates, balancing fairness across both metrics. While the model generally exhibits a relatively lower FPR, there is an inherent tradeoff. The chosen threshold balances the differences in FPR, FNR, and FNR difference between races. At this threshold, the model predicts that 13.33% of the population is high-risk and should be detained, compared to the pre-fair model that predicts 0.5% at a threshold of 50% recidivism.

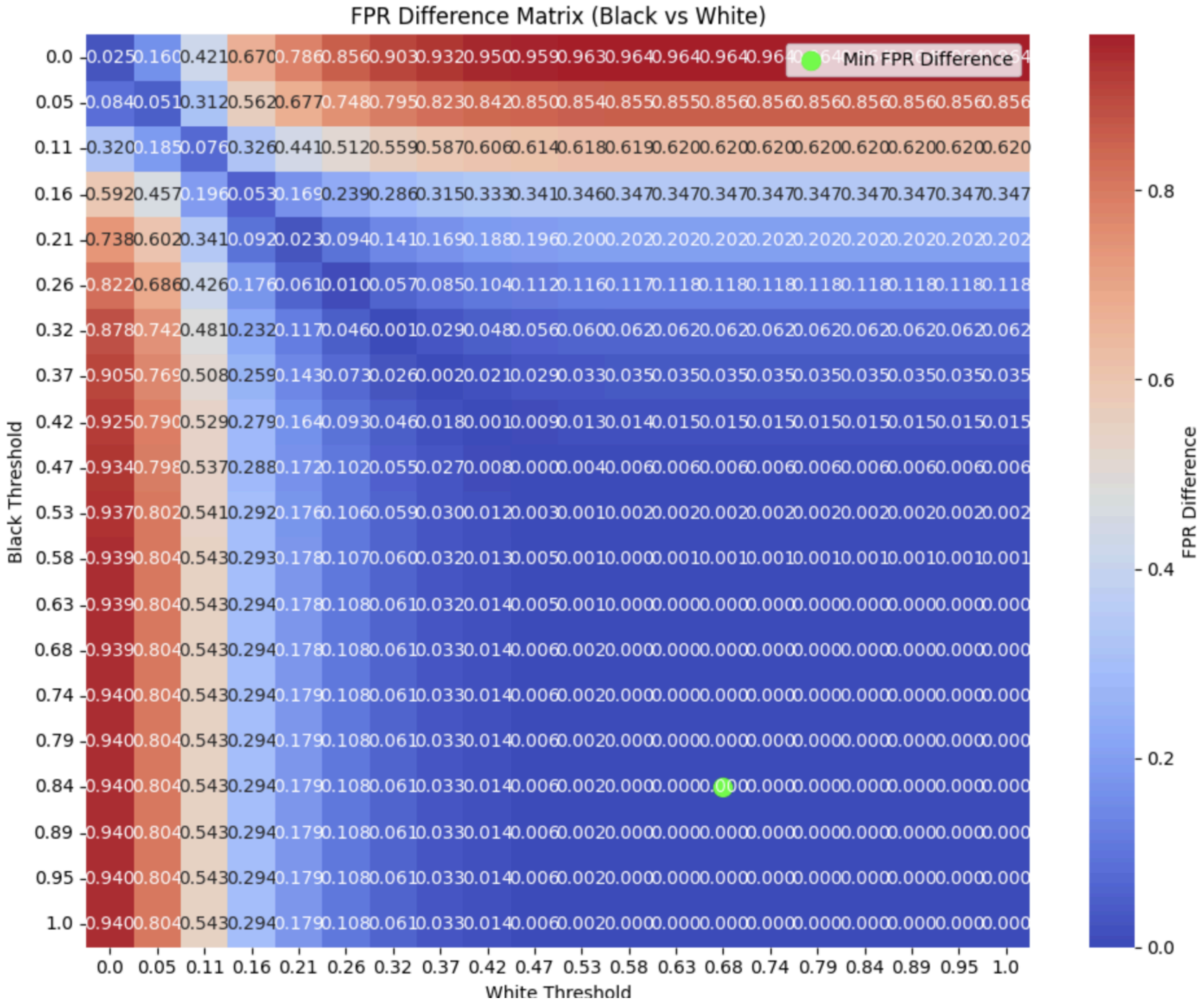
Below is the classification algorithm D(X) which determines whether a defendant with characteristics X will be detained or released before trial.

$$D(X) = \begin{cases} 1 & \text{if } \hat{p}(X) > 0.2632 \quad \text{(Detain)} \\ 0 & \text{otherwise} \quad \text{(Release)} \end{cases}$$

After adjusting the threshold, I found that the model achieves relatively balanced error rates across demographic groups in terms of false positives and false negatives, but there is an inherent tradeoff between the two. The threshold that minimizes this disparity reduces overall accuracy, as a more liberal threshold increases false negatives. When examining calibration fairness, discrepancies in predicted probabilities across groups suggest that the model is not fair by the calibration definition, as similar individuals from different groups are assigned different risk scores despite having similar true probabilities of rearrest. In conclusion, while the model is fair according to the error rate balance definition, it sacrifices accuracy to achieve this fairness and is not fair according to the calibration definition, highlighting the complexities of balancing fairness with model performance.

Another approach to address fairness is to choose different thresholds for different racial or demographic groups in order to minimize disparities in false positive rates (FPR). By adjusting the threshold individually for each group, the model can be tailored to reduce the likelihood of falsely classifying members of that group as high-risk. This method focuses on minimizing the false positive rate for each group, potentially leading to more balanced outcomes across races. However, this approach may still involve tradeoffs, as it could affect other metrics like false negative rates (FNR) or overall accuracy. In practice, setting group-specific thresholds allows for more equitable treatment of different groups, but it requires careful calibration to avoid unintended consequences, such as creating new biases or reducing the model's predictive power in some groups.

Figure 6: False positive rate difference matrix (Black vs. White)



Black Threshold \ White Threshold	0.0	0.05	0.11	0.16	0.21	0.26	0.32	0.37	0.42	0.47	0.53	0.58	0.63	0.68	0.74	0.79	0.84	0.89	0.95	1.0
0.0	0.025	0.160	0.421	0.670	0.786	0.856	0.903	0.932	0.950	0.959	0.963	0.964	0.964	0.964	0.964	0.964	0.964	0.964	0.964	0.964
0.05	0.084	0.051	0.312	0.562	0.677	0.748	0.795	0.823	0.842	0.850	0.854	0.855	0.855	0.856	0.856	0.856	0.856	0.856	0.856	0.856
0.11	0.320	0.185	0.076	0.326	0.441	0.512	0.559	0.587	0.606	0.614	0.618	0.619	0.620	0.620	0.620	0.620	0.620	0.620	0.620	0.620
0.16	0.592	0.457	0.196	0.053	0.169	0.239	0.286	0.315	0.333	0.341	0.346	0.347	0.347	0.347	0.347	0.347	0.347	0.347	0.347	0.347
0.21	0.738	0.602	0.341	0.092	0.023	0.094	0.141	0.169	0.188	0.196	0.200	0.202	0.202	0.202	0.202	0.202	0.202	0.202	0.202	0.202
0.26	0.822	0.686	0.426	0.176	0.061	0.010	0.057	0.085	0.104	0.112	0.116	0.117	0.118	0.118	0.118	0.118	0.118	0.118	0.118	0.118
0.32	0.878	0.742	0.481	0.232	0.117	0.046	0.001	0.029	0.048	0.056	0.060	0.062	0.062	0.062	0.062	0.062	0.062	0.062	0.062	0.062
0.37	0.905	0.769	0.508	0.259	0.143	0.073	0.026	0.002	0.021	0.029	0.033	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035	0.035
0.42	0.925	0.790	0.529	0.279	0.164	0.093	0.046	0.018	0.001	0.009	0.013	0.014	0.015	0.015	0.015	0.015	0.015	0.015	0.015	0.015
0.47	0.934	0.798	0.537	0.288	0.172	0.102	0.055	0.027	0.008	0.000	0.004	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006	0.006
0.53	0.937	0.802	0.541	0.292	0.176	0.106	0.059	0.030	0.012	0.003	0.001	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002
0.58	0.939	0.804	0.543	0.293	0.178	0.107	0.060	0.032	0.013	0.005	0.001	0.000	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
0.63	0.939	0.804	0.543	0.294	0.178	0.108	0.061	0.032	0.014	0.005	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.68	0.939	0.804	0.543	0.294	0.178	0.108	0.061	0.033	0.014	0.006	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.74	0.940	0.804	0.543	0.294	0.179	0.108	0.061	0.033	0.014	0.006	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.79	0.940	0.804	0.543	0.294	0.179	0.108	0.061	0.033	0.014	0.006	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.84	0.940	0.804	0.543	0.294	0.179	0.108	0.061	0.033	0.014	0.006	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.89	0.940	0.804	0.543	0.294	0.179	0.108	0.061	0.033	0.014	0.006	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
0.95	0.940	0.804	0.543	0.294	0.179	0.108	0.061	0.033	0.014	0.006	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1.0	0.940	0.804	0.543	0.294	0.179	0.108	0.061	0.033	0.014	0.006	0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Using different thresholds for different racial or demographic groups raises significant ethical concerns, particularly regarding fairness and discrimination. In the scenario where the threshold for Black individuals is set to 0.84 and for White individuals to 0.68 in order to equalize false positive rates, this adjustment may achieve a statistical balance in error rates. However, it introduces the risk of discrimination because it treats individuals differently based on their race, which could be perceived as unjust or unfair. This practice might be seen as reinforcing systemic biases, especially if it leads to a situation where one group is disproportionately labeled as high-risk or denied release due to a lower threshold. Therefore, such threshold adjustments should be carefully considered and accounted for, with a clear explanation of the potential for discriminatory outcomes and a thorough analysis of their social and legal implications. Balancing fairness with ethical considerations in predictive modeling is crucial, especially in high-stakes areas like criminal justice.

5. Conclusion

In conclusion, the quantitative threshold derived in this analysis serves as a proxy for the statutory standard of evidence used by judges and prosecutors when making pretrial detention decisions without the aid of algorithms. By focusing on error rate balance as the definition of fairness, this study finds that while it is possible to minimize the false positive and false negative disparities between groups, the tradeoffs are significant. Achieving balance in error rates comes at the expense of overall model accuracy, and the selection of a threshold that minimizes disparity results in a model that is more liberal in predicting high-risk individuals, thus increasing false negatives. This reflects the inherent complexity of achieving fairness in predictive algorithms.

However, using arrest as a proxy for recidivism introduces serious ethical and practical challenges, particularly due to the over-policing of Black neighborhoods, which can lead to biased outcomes in predictive modeling. The data used to validate these algorithms is not free from bias, perpetuating existing disparities in the criminal justice system. Moreover, algorithms reduce individuals to a set of features, overlooking the complexities of their unique life stories and personal circumstances. While the algorithmic approach offers a more data-driven solution, it is essential to recognize that certain definitions of fairness, such as statistical parity and calibration, are often mutually exclusive, and selecting one fairness criterion often undermines another.

The tradeoffs involved in designing these algorithms are not insignificant. They require careful consideration of fairness, transparency, and their potential consequences on the justice system. Algorithms should be developed democratically, with input from experts across various sectors of the justice system to ensure a comprehensive understanding of their implications. Even then, algorithms must be used with caution, recognizing the limitations of the training data and the computational challenges in addressing multiple definitions of fairness. Ultimately, the ethical implications of using these tools in high-stakes environments like pretrial detention decisions must be balanced with the goal of reducing bias and enhancing consistency in judicial decision-making.