# Ethics of Artificial Intelligence in Crime

**Catherine Back** cback@ucsd.edu
**Kavya Sriram** kasriram@ucsd.edu
**Aj Falak** pfalak@ucsd.edu
**Yuancheng Cao** yuc094@ucsd.edu
**Emily Ramond** eramond@deloitte.com
**Greg Thein** gthein@deloitte.com

## Abstract

Predictive models in pretrial risk assessment influence judicial decisions but often inherit racial biases from historical criminal justice data. This work examines racial bias in these models and applies bias mitigation techniques to improve fairness. Using the Pretrial Release dataset (244,271 records, 112 features), we trained a random forest model with 100 estimators, achieving 83.27% accuracy on a 20% test set. To mitigate bias, we applied Reweighing as a pre-processing technique and Calibrated Equalized Odds Postprocessing as a post-processing method. Reweighing reduced the mean outcome difference from 0.009 to -0.00 in the training set, improving fairness at the data level. However, in the testing set, 1 - min(DI, 1/DI) fluctuated between 0.0266 and 0.15, showing instability across classification thresholds. After postprocessing, equal opportunity difference improved, reducing from 0.0136 to 0.0031 in the test set, while balanced accuracy remained at 0.6357. The trade-off between fairness and accuracy was more controlled in postprocessing, making it more effective for a highly imbalanced dataset. Given that fairness adjustments remained stable across validation and testing, Calibrated Equalized Odds Postprocessing is the preferred approach. Future work should explore more diverse datasets, threshold tuning, and hybrid methods to balance fairness and model performance.

Website: https://cao1224.github.io/ucsd_capstone_project/

Code: https://github.com/cao1224/ucsd_capstone_project

# 1 Introduction

## 1.1 Background

With the increasing use of predictive algorithms and artificial intelligence (AI) in different sectors, it is critical that these algorithms are audited for bias. AI applications in the criminal justice system have the potential to significantly impact people's lives, necessitating a more careful approach to guarantee their responsible use. Therefore, it is crucial to apply ethical and trustworthy solutions in these high-stakes decisions. Pretrial risk-assessment algorithms evaluate a defendant's likelihood of failing to appear in court or committing a new crime before trial, directly influencing judicial outcomes such as bail eligibility or detention. While these tools are intended to reduce human subjectivity, they often perpetuate systemic biases embedded in historical criminal justice data. For instance, ProPublica's 2016 analysis of the COMPAS risk-assessment tool revealed that Black defendants were twice as likely as White defendants to be falsely flagged as high-risk for recidivism Angwin et al. (2016). Such disparities arise from biases inherent in policing practices, such as the over-surveillance of minority communities and socioeconomic inequalities reflected in arrest records. When these patterns are encoded into training data, models are at risk of amplifying existing inequalities instead of mitigating them, leading to discriminatory outcomes such as disproportionately denying bail to minority defendants.

Unfair algorithmic predictions can perpetuate cycles of poverty and incarceration: defendants denied pretrial release face higher risks of job loss, housing instability, or loss of child custody, even if later exonerated. Furthermore, biased predictions undermine public trust in judicial fairness. To address these concerns, models must be rigorously audited for disparities in error rates (e.g., higher false positives for Black defendants) and outcomes (e.g., unequal bail denial rates across racial groups). To keep pretrial decisions fair, AI must be constantly tested and improved so that it doesn't reinforce discrimination against legally protected groups.

Applying machine learning to pretrial risk assessment is expected to cause three major harms. First, biased training data (e.g., over-policing in minority neighborhoods) can skew arrest records and thus misrepresent actual crime rates. Second, proxy discrimination occurs when variables such as a zip code or the number of prior arrests act as race proxies, indirectly coding for bias even when protected attributes are excluded. Third, when overprediction of crime in minority areas becomes a justification for increased policing, a feedback loop occurs that generates more skewed data for future models.

This paper focuses on race as a protected attribute due to its disproportionate influence on pretrial outcomes in the U.S. justice system. The target label, 'reoffend,' measures whether a defendant is predicted to commit a new crime pretrial. However, this label is often misleading: rearrest rates (used as proxies for reoffending) overstate the extent of actual criminal behavior due to systemic policing biases. For example, Black individuals are more likely to be arrested for low-level offenses than White individuals. To mitigate these risks, we propose using the AIF360 (AI Fairness 360) toolkit to audit and address bias across three stages of the machine learning pipeline. In pre-processing, techniques such as reweighing samples adjust training data to balance representation across racial groups. In post-processing, decision thresholds are calibrated to equalize error rates between groups.

By integrating these fairness-aware techniques, this work aims to reduce discriminatory outcomes while preserving predictive utility. Current pretrial tools often lack transparency in bias testing, leaving courts unaware of systemic flaws. Our approach prioritizes race, an attribute strongly tied to inequities in the U.S. justice system, to deliver actionable solutions for judicial stakeholders.

## 1.2 Literature Review

Pretrial risk assessment tools aim to predict the likelihood that a defendant will fail to appear in court or commit new crimes before trial. The Public Safety Assessment (PSA), developed by the Laura and John Arnold Foundation (LJAF), is a framework that uses nine objective factors—such as age at the time of arrest, prior convictions, and history of failure to appear—to generate risk scores. The PSA excludes demographic variables such as race, ethnicity, and geography to mitigate bias Laura and John Arnold Foundation (2199). The California courts' pretrial pilot program report evaluates the implementation of the PSA and its accuracy in assessing risk levels across different populations Judicial Council of California (2022).

In the field of machine learning, random forest models are recognized for their effectiveness in risk prediction tasks. Random forest modeling involves constructing a larger number of decision trees during training, where each tree provides a classification, and the forest selects the most voted classification. This integrated approach improves prediction accuracy and controls overfitting. Research funded by the National Institute of Justice (NIJ) has demonstrated the application of random forest models in criminal justice settings. For example, a risk prediction tool developed for the Philadelphia Adult Probation and Parole Department uses a random forest model to predict probationers' behavior over a two-year period. The model gathers available information about probationers and predicts their likely behavior, helping to assign appropriate levels of supervision Ritter (2013).

Beyond risk assessment in probation systems, random forest models have shown strong performance in crime prediction tasks. Hossain et al. (2020) applied a random forest model with 100 trees and random undersampling, achieving 99.16% accuracy in crime prediction using spatiotemporal data. Their model utilized key features such as day, time, type of crime, and address, demonstrating that random forest effectively captures crime patterns across different locations and times. Similarly, Aldossari et al. (2020) compared random forest with decision trees and Naive Bayes models for crime category prediction in Chicago. While the decision tree model achieved 91.68% accuracy using all available features, their study emphasized the importance of feature selection, using backward feature elimination to improve model interpretability. This reinforces the advantage of tree-based models like random forest which can handle complex feature interactions without extensive preprocessing.

Moreover, random forest models are particularly useful when dealing with imbalanced datasets, which is a common issue in crime prediction since certain crime types happen far more often than others. The model's bagging approach and feature randomness help create diverse decision trees, making it more reliable even for less common crimes. However, while these models offer strong predictive power, they also rely on historical crime data, which often reflects existing biases in law enforcement practices. Such behavior has raised concerns about fairness in machine learning-based risk assessments and has led to ongoing discussions about ways to detect and reduce bias in predictive systems.

The fairness debate in machine learning risk assessment has led to discussion of bias mitigation techniques. Three main strategies have been proposed: (1) excluding race-related variables, (2) adjusting algorithms to equalize predictions across racial groups, and (3) rejecting algorithmic approaches altogether Mayson (2019). However, these approaches fail to address the core problem: risk assessment inherently projects past inequalities into future projections. Historical arrest and conviction data are disproportionately biased against minority populations due to systemic factors such as over-policing in certain neighborhoods.

Fairness in crime prediction and pretrial risk assessment is typically measured using three key metrics: demographic parity, equalized odds, and predictive parity. Demographic parity requires that predictions be independent of protected attributes like race, meaning that all groups receive similar risk classifications. However, this can lead to low accuracy if risk scores differ across groups. Equalized odds ensures that false positive and false negative rates are similar across racial groups, aiming to prevent disproportionately harsh treatment of any group Chouldechova (2016). Predictive parity, commonly used in criminal justice settings, ensures that individuals with the same risk score have a similar likelihood of reoffending.

Several bias mitigation techniques have been explored to reduce disparities in pretrial risk assessment. One approach is pre-processing the training data to remove or balance biased historical trends. Kamiran and Calders (2012) proposed reweighing or modifying the training set to reduce dependency on protected attributes. However, this approach risks losing predictive signals and may not fully eliminate hidden biases. Another strategy is in-processing methods, which modify the model itself to adjust decision boundaries for different demographic groups. Hardt, Price and Srebro (2016) introduced an optimization-based approach to enforce equalized odds by adjusting model predictions to ensure similar error rates across groups. Finally, post-processing techniques adjust predictions after training to improve fairness metrics. One common approach is threshold adjustment, where risk scores are calibrated differently for each group to meet fairness constraints. In practice, law enforcement and judicial systems often rely on a combination of these techniques. Still, the choice depends on legal constraints, public policy, and acceptable trade-offs between fairness and accuracy.

In the following section, we describe our dataset, detailing its sources, structure, and key variables. Section 2 outlines our methodology, including data pre-processing steps, exploratory data analysis, feature engineering, and model development using a random forest classifier. We also discuss the integration of bias mitigation techniques to improve fairness in risk assessment. In Section 3, we evaluate the impact of feature selection on model performance, analyze fairness metrics, and examine the effectiveness of debiasing strategies in improving equitable outcomes.

### 1.3 Relevant Data Description

The Pretrial Release dataset, created by the Division of Criminal Justice Services (DCJS), is designed to support research on bail reform policies introduced in the 2019 Criminal Justice Legislation. This dataset is particularly useful for evaluating fairness in pretrial decision-making, as it includes protected attributes such as race. The dataset contains records from 2023, making it a relevant and up-to-date source for studying the impact of pretrial decisions on different demographic groups.

The dataset consists of 244,271 records and 112 features, covering various aspects of pretrial decisions and outcomes. Some of the key features include demographics, pretrial decisions, outcomes, and financial factors. The demographic attributes provide essential context for fairness analysis and include race, ethnicity, gender, age at arrest, and age at crime. These variables allow for assessing disparities in pretrial decisions based on personal characteristics. Pretrial decision-related features capture information on how individuals are processed within the system. These include bail set and posted at arraignment, the release type, and the supervision type. Outcome variables measure the effectiveness and fairness of pretrial decisions. Some of the key outcome features include Failure to Appear (FTA), reoffending, release decision at arraignment, and rearrest. These variables enable an analysis of how different pretrial decisions correlate with public safety and court compliance. Financial factors such as bail amount and bond type are also included, providing insight into how monetary constraints influence pretrial release. These features help in understanding whether financial conditions disproportionately affect certain demographic groups. By leveraging these features, models can help identify potential biases in bail and release decisions, offering insights into disparities across different demographic groups.

## 2 Methods

### 2.1 Pre-processing

Our first step involved a comprehensive analysis of missing values, which revealed 52 columns containing more than 1% missing data. Particularly concerning were columns such as `First_Bail_Set_Cash` (200,656 missing values), `First_Bench_Warrant_Date` (218,556 missing values), and `Days_Arraign_Remand_First_Released` (236,156 missing values).

To address these issues, we began by removing rows with missing values in the critical `rearrest` column, as this column served as our primary outcome variable. We also standardized the handling of missing values by replacing empty strings with NaN values in key categorical columns, including `Top_Severity_at_Arraign`, `Disposition_Type`, `Disposition_Detail`, and `Dismissal_Reason`.

To streamline our dataset, we removed irrelevant administrative columns such as Court_Name and Court_ORI, which contained information that didn't necessarily matter for our model. The Court_Name and Court_ORI (a nine-character identifier containing both alpha and numeric characters assigned by FBI CJIS, which validates legal authorization to access Criminal Justice Information (CJI) and identifies the agency in all transactions)—wouldn't help much in determining whether someone would reoffend.

A significant step in our pre-processing involved the removal of columns with more than 90% missing values. While this reduced the dimensionality of our dataset, we determined that these columns provided minimal analytical value due to their sparseness and their removal would not significantly impact our analysis.

Data type standardization was another form of pre-processing that our group conducted. We converted all date-related columns to the datetime format, including First_Arraign_Date, First_Bench_Warrant_Date, and Disposition_Date. Count-based columns were converted to integer types, ensuring appropriate numerical handling. We also implemented consistent column naming conventions, replacing spaces with underscores and standardizing the format across all variables.

The final preprocessed dataset represents a significant improvement in data quality and usability. We successfully reduced the number of columns from 112 to 79 while maintaining 244,721 records. The resulting dataset features consistent data types, eliminated high-missing-value columns, and includes meaningful derived features. This cleaned dataset provides a solid foundation for our analysis of pretrial outcomes, balancing the need for comprehensive information with data quality and manageability. The pre-processing steps taken ensure that our subsequent analyses will be based on reliable, well-structured data, enabling more accurate insights into the pretrial release system in New York State.

## 2.2 Exploratory Data Analysis

Our EDA of the New York State Pretrial Release Dataset revealed several significant patterns and insights about pretrial outcomes and demographic distributions. Our analysis focused on key variables including rearrest patterns, demographic characteristics, geographic distribution, and the impact of prior criminal history.

**Rearrest** As shown in Figure 1, the analysis of rearrest outcomes indicated that the majority of individuals (83.7%) had no arrests during their pretrial period. Among those who were rearrested, 8.1% were charged with misdemeanors, 6.0% with non-violent felonies, and 2.2% with violent felonies. This distribution highlights that a significant majority of individuals comply with pretrial release conditions, with only a small percentage engaging in serious offenses during this period.

**Age Distribution by Race** As shown in Figure 2, demographic analysis revealed important patterns in different segments of the population. The age distribution analysis showed a

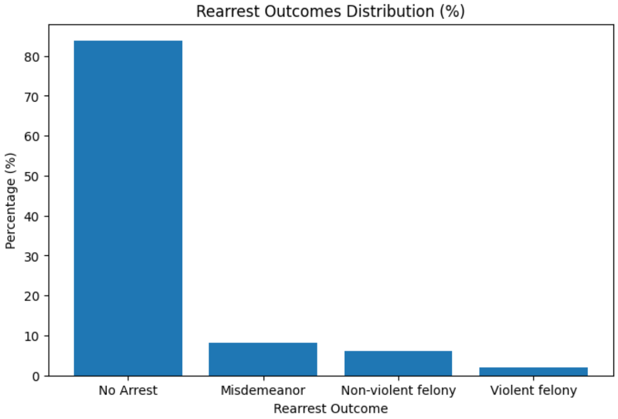


Figure 1: Distribution of rearrest outcomes in percentage.

concentration of cases among individuals aged 20–39, with a median age of 31 years. When examining age distributions across racial groups, we found that Black defendants tended to be younger on average (median age 29) compared to White defendants (median age 33). Asian/Pacific Islander defendants had the highest median age at 35 years. The age-crime relationship was also higher among younger defendants, with the 18–25 age group showing higher rates of pretrial rearrest across all racial categories.

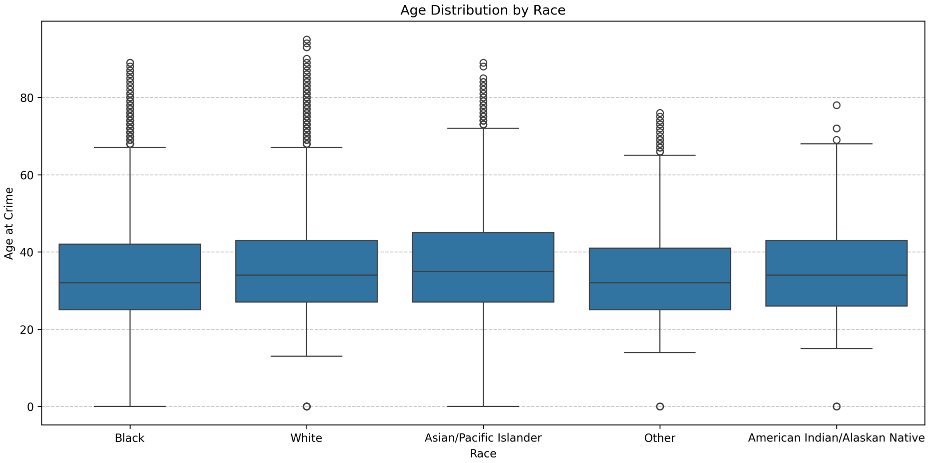
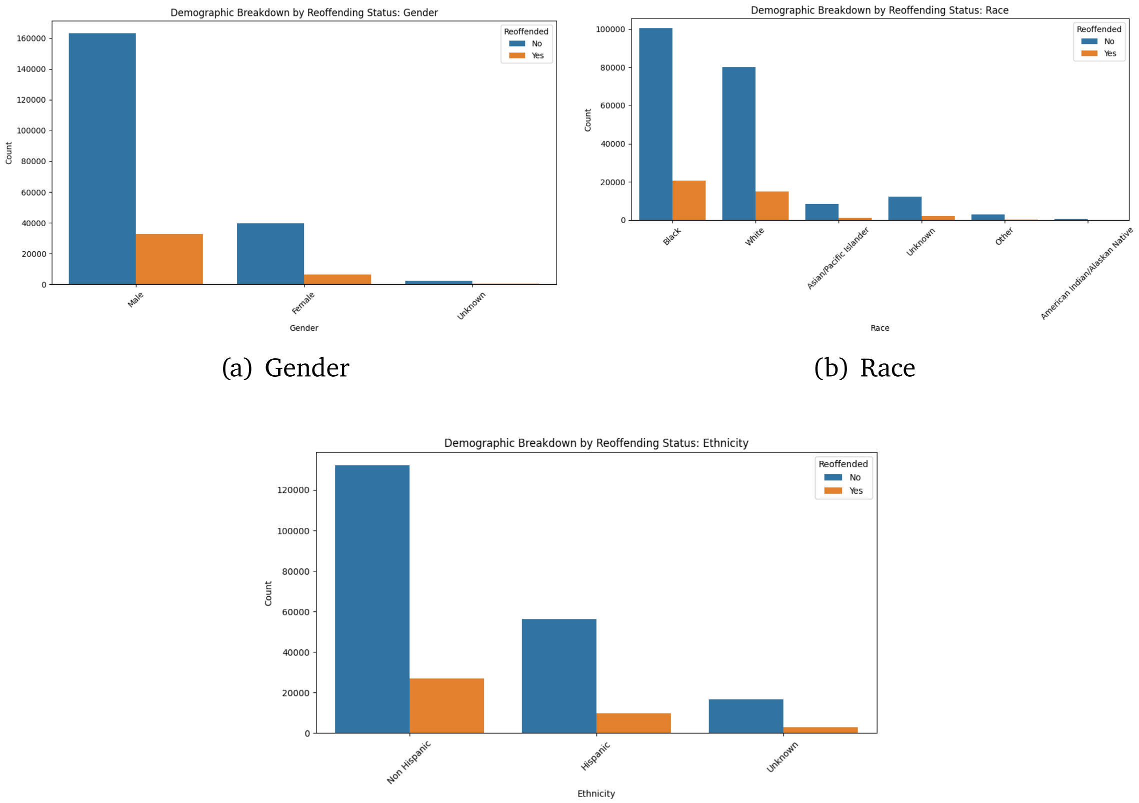


Figure 2: Age distribution at the time of crime across different racial groups, with median, interquartile range, and outliers represented.

**Note:** The races with '0' at the age of crime are attributed to missing data.

**Demographics** Gender distribution (Figure 3(a)) indicated a significant disparity, with males representing approximately 80% of cases and females 19%, with a small percentage (1%) listed as unknown. Further analysis of gender and reoffending patterns revealed that male defendants had a higher rate of pretrial rearrest (18.2%) compared to female defendants (12.4%). This gender gap in reoffending rates remained consistent across different



(a) Gender

(b) Race

(c) Ethnicity

Figure 3: Demographic breakdown of reoffending status by gender, race, and ethnicity.

age groups and charge severities. Notably, the gender disparity in reoffending rates was most pronounced in the 18–25 age group, where male defendants showed a 22.1% rearrest rate compared to 14.8% for females.

Racial and ethnic distributions (Figures 3(b) and 3(c)) showed that Black individuals accounted for 49.5% of cases, followed by White individuals at 38.9%, with other racial categories constituting the remaining percentage. The Hispanic population represented 24.8% of cases, while non-Hispanic individuals accounted for 65.2%. Analysis of reoffending rates across racial groups revealed disparities, with Black defendants showing a higher rate of pretrial rearrest (19.3%) compared to White defendants (14.7%) and Asian/Pacific Islander defendants (11.2%).

**Geographic Distribution** Although the dataset exclusively comprised individuals from New York, analyzing its geographic distribution provided insights into substantial variations across the state's judicial districts. This analysis (Figure 4) demonstrated substantial variations across New York State's judicial districts. The five boroughs of New York City accounted for the majority of cases, with particular concentrations in New York County (25.3%), Kings County (20.1%), and Queens County (15.8%). This distribution reflects both population density and varying levels of law enforcement activity across a variety of regions.

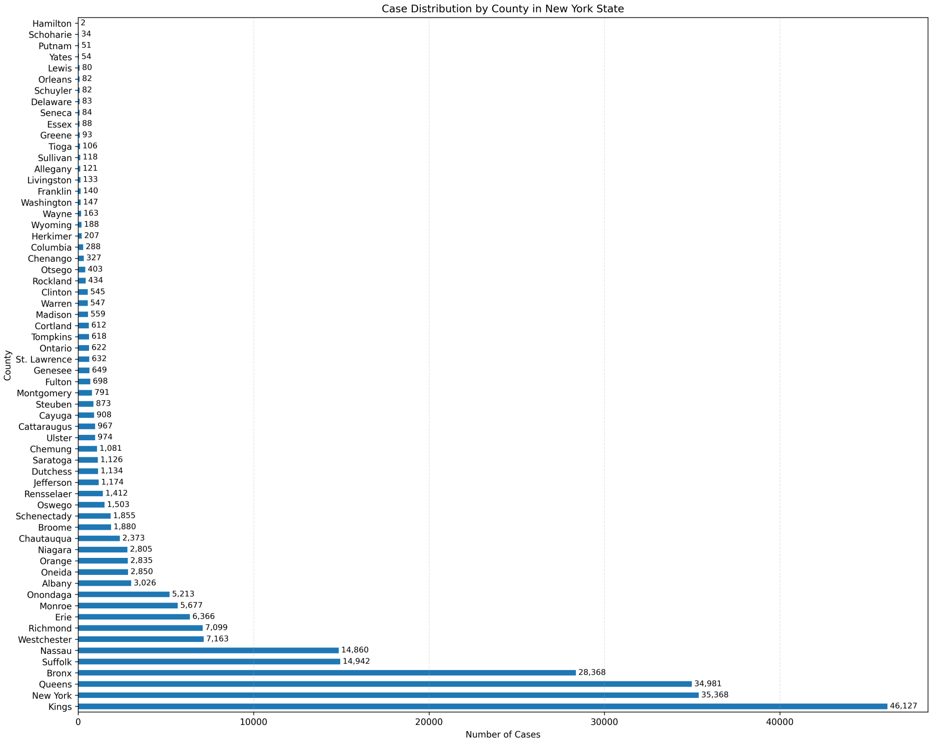


Figure 4: Case distribution by county in New York State, showing the number of cases per county with Kings County having the highest count.

**Criminal History** One of the most significant findings emerged from our analysis of prior criminal history's impact on pretrial outcomes. Individuals with prior offenses showed different patterns of pretrial conduct compared to those without prior records. The data revealed that defendants with prior offenses had a reoffending rate of approximately 21.5%, while those without prior offenses had a significantly lower rate of about 11.6%. This difference suggests that prior criminal history serves as a meaningful indicator of pretrial conduct.

## 2.3 Feature Engineering

We came up with these feature engineering ideas by combining domain expertise, exploratory data analysis (EDA), and insights from the literature review. Some ideas came from patterns we found in the data, while others were inspired by known factors that influenced the outcome of the case. Here's how each feature took shape.

The creation of the `Has_Prior_Offense` feature was driven by the well-established finding that prior criminal history is a strong predictor of recidivism SERC (2022); U.S. Department of Justice (2003). Instead of treating each prior offense type separately (`prior_vfo_cnt`, prior_nonvfo_cnt, and prior_misd_cnt), we consolidated them into a single binary indicator. This new feature provided an indicator of previous criminal activity, where a value of 1 indicated the presence of any prior offense and 0 indicated no prior criminal history. This transformation helped capture the impact of criminal history on reoffending risk in a simplified way.

We also engineered new timing features to capture the timing aspects of criminal behavior. By calculating the time difference between the arrest date and crime date (Time_To_Arrest), we created a feature that might indicate a faster or slower time to arrest. During the data exploration stage, we observed that the gap between the crime date and the arrest date might reflect differences in law enforcement efficiency or the type of crime committed. Similarly, we computed the length of the pretrial period (Pretrial_Duration) by measuring the time between arraignment and disposition dates, providing insight into how the duration of pretrial release might affect outcomes. Longer pretrial periods could influence a defendant's ability to prepare a defense, their likelihood of accepting a plea bargain, or even their future behavior.

Another important feature engineering step involved categorizing case resolution times. We analyzed the Days_Arraign_to_Dispo variable, which tracks the time from arraignment to case disposition. Based on the distribution of these times (with a mean of approximately 59 days and a median of 33 days), we created a categorical feature called Resolution_Category with four distinct categories:

- **Very Fast Resolution**: Cases resolved in less time than the 25th percentile
- **Fast Resolution**: Cases resolved between the 25th percentile and median
- **Moderate Resolution**: Cases resolved between the median and 75th percentile
- **Slow Resolution**: Cases taking longer than the 75th percentile

This categorization helps capture the speed of case resolution in a more interpretable way, potentially revealing patterns in how case processing time relates to pretrial outcomes.

| Resolution Category | Number of Cases |
|---|---|
| Fast Resolution | 69,611 |
| Slow Resolution | 63,646 |
| Moderate Resolution | 59,173 |
| Very Fast Resolution | 52,290 |

Table 1: Distribution of Cases by Resolution Category

**Note:** The categorization of the speed of resolution was determined by our project team based on general perceptions of legal severity. Therefore, it may not fully represent the legal definition of what a 'very fast' resolution looks like.

Moving on, legal classifications often involve complex categorical codes, making direct analysis challenging. Our final step in the feature engineering process involved creating a severity score based on the type of law code associated with each case.

We developed a custom mapping function that assigned severity levels (1–5) to different law codes based on their general severity in the legal system. This custom mapping function assigned severity levels based on the general severity of each law code in the legal system, enabling straightforward comparisons between cases and improving the interpretability of model decisions. The mapping was structured as follows:

- **Severity Level 5 (Most Severe):**
  - PL (Penal Law)
  - CPL (Criminal Procedure Law)
  - LAB (Labor Law)
  - COR (Corrections Law)
  - HHC (Housing Code/Habeas Corpus Law)
- **Severity Level 4:**
  - VTL (Vehicle and Traffic Law)
  - TAX (Tax Law)
  - ECL (Environmental Conservation Law)
  - PHL (Public Health Law)
  - WC (Workers' Compensation Law)
- **Severity Level 3:**
  - GB (General Business Law)
  - RPA (Real Property Actions)
  - LOC (Local Law)
  - AGM (Attorney General's Manual)
  - MTA (Metropolitan Transportation Authority Law)
- **Severity Level 2:**
  - AC (Arbitration Court)
  - AM (Administrative Law)
  - SCL (State Civil Law)
  - TOH (Town or Housing Law)
  - SW (Social Welfare Law)
  - PRR (Public Records Law)
- **Severity Level 1 (Least Severe):**
  - ABC (Alcoholic Beverage Control Law)
  - RSS (Revised Statutes Supplement)
  - CAN (Canon Law)
  - TIS (Transportation and Infrastructure Law)
  - ED (Education Law)

This mapping created a new feature, which we called `crime_severity`, that converted categorical law codes into an ordinal scale, making it more suitable for our machine learning models while preserving the inherent severity hierarchy of different law types.

**Note:** The categorization of severity levels was determined by our project team based on general perceptions of legal severity. Therefore, it may not fully represent the actual severity used in formal legal contexts.

| Severity Level | Number of Cases |
|---|---|
| 5 | 222,344 |
| 4 | 20,224 |
| 3 | 13 |
| 2 | 475 |
| 1 | 164 |

Table 2: Distribution of Cases by Crime Severity Level

## 2.4 Random Forest Model Development

Our model development process focused on creating a **random forest classifier** to predict the risk of pretrial reoffending. We chose the random forest algorithm because it handles both numerical and categorical features effectively, provides rankings of feature importance, and captures complex relationships among the various features included in our model.

We first created binary indicators for our criminal history features. Using `lambda` functions, we transformed the count-based variables (`prior_misd_cnt`, `prior_nonvfo_cnt`, and `prior_vfo_cnt`) into binary indicators where 0 represented no prior offenses and 1 represented one or more prior offenses. These transformations resulted in three new features: `prior_misd_binary`, `prior_nonvfo_binary`, and `prior_vfo_binary`.

For our model input, we selected **seven key features** that captured different aspects of pretrial risk:

- `prior_misd_binary`
- `prior_vfo_binary`
- `prior_nonvfo_binary`
- `pend_vfo`
- `pend_nonvfo`
- `pend_misd`
- `crime_severity`

Our target variable, `reoffend`, was structured as a binary indicator of whether a defendant was rearrested during the pretrial period.

We implemented the random forest classifier using `scikit-learn`'s `RandomForestClassifier` with `n_estimators=100` and a `random_state=42` for reproducibility. The data was split into training (50%), validation (30%), and test (20%) sets using `train_test_split` twice—first to separate the training set and then to divide the remaining data into validation and test sets.

## 2.5 Addressing Imbalanced Labels

To handle the class imbalance in our dataset, we applied two oversampling techniques: Random Oversampling and SMOTE (Synthetic Minority Over-sampling Technique). These methods help balance the dataset by increasing the number of minority class samples, reducing model bias toward the majority class.

First, we implemented SMOTE, which generates synthetic samples for the minority class rather than duplicating existing ones. This technique helps improve model generalization and prevents overfitting. We used `imbalanced_make_pipeline` to integrate SMOTE with the random forest classifier, setting `class_weight='balanced'` to further adjust the model's learning based on class distribution. We employed the entropy criterion to effectively manage the information gain.

After oversampling, we performed hyperparameter tuning to improve model performance. We used Stratified K-Fold Cross-Validation (k=5) to ensure balanced splits across all folds. A grid search was applied to optimize key parameters, including the number of estimators (`n_estimators`), maximum tree depth (`max_depth`), and random state. The best parameters found were `max_depth=6`, `n_estimators=200`, and `random_state=13`, selected based on recall score, as improving recall for the minority class (reoffenders) was our primary goal. The impact of these methods on model performance is discussed in the Results section.

## 2.6 Initial Statistical Analysis

After the initial model development, we completed a statistical assessment of the models' predictions on the test set. The proportion of Black individuals predicted to be reoffenders was higher than the other races at a value of 0.01. White individuals had a rate of 0.008, and Asian/Pacific Islander and Native American/Alaska Natives had rates around 0.006. We conducted an ANOVA test across all groups to determine whether the differences in these proportions were statistically significant. The test yielded a p-value of 0.0002, indicating a significant difference. The t-test comparing Black individuals and White individuals yielded the value of 0.0003, which further confirmed this significance. The FPR rates across races also showed some variance. Black individuals had the highest FPR rate of 0.0093, followed by white individuals with an FPR of 0.0067, and the remaining groups had similar values, such as Asian/Pacific Islander at 0.0052 and Native American/Alaskan Native at 0.0075. Based on these FPR results, we concluded that there was some racial bias. FPR is a useful metric for assessing a model's fairness, as if there is an imbalance in the values, it means a group is potentially being disproportionately targeted. Unfair outcomes, such as the unjust detention of low-risk individuals misclassified as high-risk, can arise in a criminal justice context. Due to these potential serious impacts, we considered this difference across racial groups noteworthy. In the next step of the process, bias mitigation, we used the racial-bias information accordingly. In creating our AIF360 dataset, we selected race as the protected attribute due to this observed difference and aimed to mitigate any bias found across these groups in the model's predictions. These findings identified it as a factor that should not unfairly influence the outcomes of the model, an essential step in debiasing the model.

### 2.7 Bias Mitigation Techniques

To address potential bias in our model, we applied Reweighing as a pre-processing technique and Calibrated Equalized Odds Postprocessing as a post-processing technique using AIF360. We selected these methods to maintain overall model performance while ensuring fairer predictions across different racial groups. We selected Reweighing as a pre-processing method because it adjusts the training dataset before model learning. This technique assigns different weights to instances in the dataset based on group membership, ensuring that underrepresented groups contribute more significantly to the model's learning process. Reweighing helps correct imbalances in the training data so that the model does not develop bias toward the majority group. We specified the unprivileged group as *Race = 0* and the privileged group as *Race = 1*. The algorithm first computes the necessary weights for each sample, transforming the training dataset accordingly. Once reweighing is applied, the classifier is trained on the adjusted dataset, leading to a more balanced decision-making process.

In addition to reweighing, we implemented Calibrated Equalized Odds Postprocessing to ensure fairness at the prediction level. This technique modifies the final predictions of a trained model so that false positive and false negative rates are more balanced across groups. The process begins by training the model on the original dataset without reweighing but using the fine-tuned version with oversampling. After training, a calibration model is created using AIF360 to align the predictions with equalized odds constraints across groups. This calibration adjusts the predicted probabilities, modifying classification outputs to reduce disparities in false positive and false negative rates between privileged and unprivileged groups. The adjusted predictions are then evaluated to ensure fair classification rates while maintaining model accuracy.

## 3 Results and Discussion

### 3.1 Model Performance

The model's performance was evaluated using several key metrics. On the test set, the random forest classifier achieved an overall accuracy of 0.8327, but the balanced accuracy is only 0.5067, indicating poor performance on the minority class. Looking at the classification report, precision, recall, and F1-score vary significantly between classes. For Class 0 (no reoffend), the model performs well with a precision of 0.84, a recall of 0.99, and an F1-score of 0.91. However, for Class 1 (reoffend), the recall is extremely low at 0.02, meaning the model fails to identify most reoffenders correctly. Precision for Class 1 is 0.30, but due to the low recall, the F1-score is only 0.04.

The imbalance in predictions suggests that the model is heavily biased toward Class 0. This happens because the dataset is likely imbalanced, with far fewer reoffenders than non-reoffenders. Since accuracy is dominated by the majority class, it does not fully reflect model performance in this case. The low recall for Class 1 means the model is not useful for predicting reoffenders.

## 3.2 Feature Importance Analysis

Feature importance analysis revealed the following contributions to the model:

- **Age_at_Arrest**: 0.48 (Strongest predictor)
- **pend_misd**: 0.21
- **pend_nonvfo**: 0.11
- **prior_misd_binary**: 0.07
- **prior_nonvfo_binary**: 0.03
- **Top_Arraign_Law (PL)**: 0.02
- **Top_Arraign_Law (VTL)**: 0.01

The `Age_at_Arrest` feature emerged as the most influential predictor, while pending case indicators and prior criminal history also played significant roles. The `Top_Arraign_Law` features contributed slightly less to the model's predictions.

## 3.3 Results After Balancing the Dataset

Applying random oversampling and SMOTE improved the model's ability to detect reoffenders (Class 1). Before fine-tuning, the model achieved a recall of 0.50 for Class 1, which means it correctly identified half of the reoffenders. However, precision remained low at 0.26, leading to an F1-score of 0.35. For Class 0, recall dropped to 0.73, but precision improved to 0.88, resulting in an F1-score of 0.80.

After fine-tuning, recall for Class 1 increased to 0.66, meaning the model captured more actual reoffenders. The precision remained at 0.26, leading to a slight improvement in F1-score (0.38). For Class 0, recall dropped further to 0.63, but precision increased to 0.90, resulting in an F1-score of 0.75. The balanced accuracy improved to 0.6466, and the ROC-AUC score reached 0.6914, indicating a better overall trade-off between sensitivity and specificity. Fine-tuning provided the best balance between precision and recall, ensuring better identification of reoffenders while maintaining reasonable accuracy for non-reoffenders. Since this is the best result we achieved, we will continue using these optimized parameters (*max_depth=6*, *n_estimators=200*, *random_state=13*) for future predictions.

## 3.4 Accuracy and Fairness Metrics

We assessed the impact of bias mitigation by comparing fairness metrics before and after applying Reweighing and Calibrated Equalized Odds Postprocessing. These methods were used to address disparities between unprivileged (Race 0) and privileged (Race 1) groups while preserving overall model performance.

Before debiasing, the mean outcome differences between groups were 0.009 in the training and validation sets and 0.012 in the test set, indicating a small bias. After applying

| Dataset | Before Debiasing | After Reweighing |
|---|---|---|
| Train Set | 0.009 | -0.000000 |
| Validation Set | 0.009 | - |
| Test Set | 0.012 | - |

Table 3: Mean outcome differences between unprivileged and privileged groups before and after reweighing.

Reweighing, the difference in the training set dropped to 0.000000, suggesting that the dataset became balanced at the training stage. The full comparison of mean outcome differences before and after reweighing is presented in Table 3.

The first figure (5) represents the validation set, while the second figure (6) represents the testing set, both using a model trained with the Reweighing technique. In both plots, balanced accuracy (blue) and fairness metrics (red) are plotted against classification thresholds to evaluate trade-offs between performance and fairness.

In Figure 5, the balanced accuracy reaches its peak at a classification threshold of around 0.16 and then gradually declines. The Average Odds Difference fluctuates but stabilizes near the optimal threshold. In Figure 6, the overall trend is similar, but 1 - min(DI, 1/DI) is more volatile across classification thresholds. This result indicates that fairness constraints generalize less consistently in the testing set. The balanced accuracy also peaks near 0.16 but shows a sharper decline compared to validation.

Comparing both figures, the validation set provides a smoother balance between accuracy and fairness, whereas the testing set shows higher variance in fairness metrics. This suggests that while reweighing improves fairness in training and validation, its generalization to unseen test data is more uncertain.

| Metric | Validation | Testing |
|---|---|---|
| Threshold (Best Balanced Accuracy) | 0.1600 | 0.1600 |
| Best Balanced Accuracy | 0.6281 | 0.6357 |
| 1-min(DI, 1/DI) | 0.0581 | 0.0266 |
| Average Odds Difference | 0.0083 | -0.0050 |
| Statistical Parity Difference | 0.0234 | 0.0106 |
| Equal Opportunity Difference | -0.0107 | -0.0234 |
| Theil Index | 0.1239 | 0.1231 |

Table 4: Validation and testing performance metrics after reweighing.

The impact of reweighing on the validation and testing performance is summarized in Table 4. The balanced accuracy improved slightly after reweighing, reaching 0.6357 in testing. The statistical parity difference and equal opportunity difference decreased, reducing disparities in prediction rates across groups. Additionally, the Theil index, which measures entropy-based fairness, remained stable at 0.123, indicating that reweighing did not significantly impact overall model uncertainty.

To further adjust for fairness at the prediction level, we applied Calibrated Equalized Odds

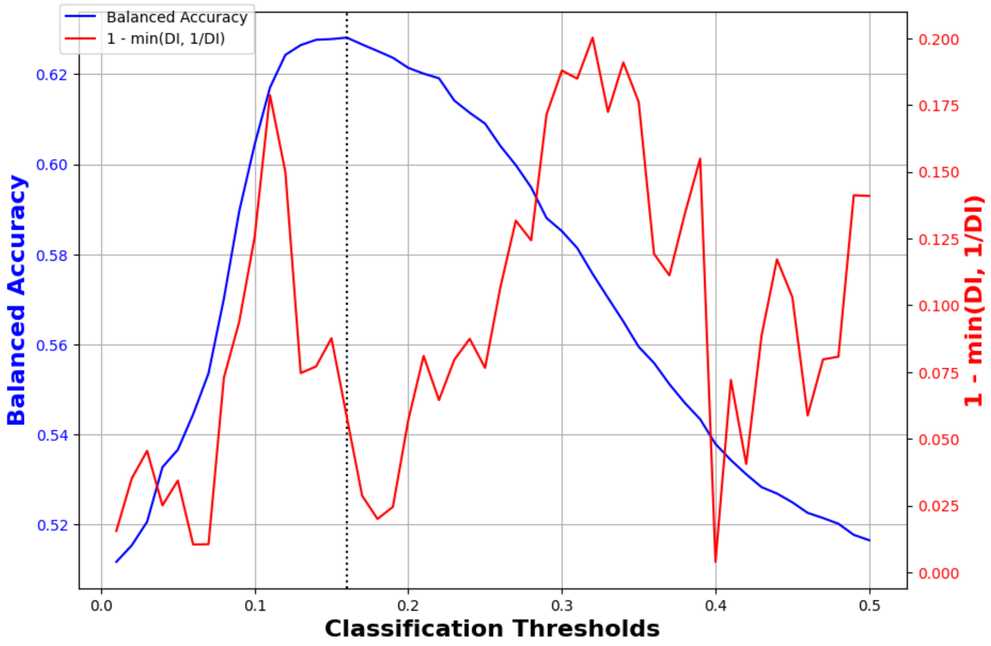


Figure 5: Validation set performance after applying Reweighing

Postprocessing, Table 5 presents the group false positive rate (GFPR) and group false negative rate (GFNR) differences before postprocessing.

| Dataset | GFPR Difference | GFNR Difference |
|---|---|---|
| Train Set | 0.0171 | 0.0001 |
| Validation Set | 0.0167 | -0.0006 |
| Test Set | 0.0136 | 0.0024 |

Table 5: False positive and false negative rate differences before postprocessing.

After postprocessing, differences in group false positive rates (GFPR) and group false negative rates (GFNR) reduced, meaning the model became more consistent in treating privileged and unprivileged groups equally.

| Dataset | GFPR Difference | GFNR Difference |
|---|---|---|
| Validation Set | 0.0162 | -0.0001 |
| Test Set | 0.0132 | 0.0031 |

Table 6: False positive and false negative rate differences after postprocessing.

The transformed results, as shown in Table 6, indicate a drop in GFPR difference to 0.0132 in the test set, with a slight increase in GFNR difference to 0.0031. This adjustment helped balance errors across groups, reducing prediction disparities.

The validation (7) and testing (8) graphs compare balanced accuracy and equal opportunity difference before and after applying Calibrated Equalized Odds Postprocessing. Both graphs

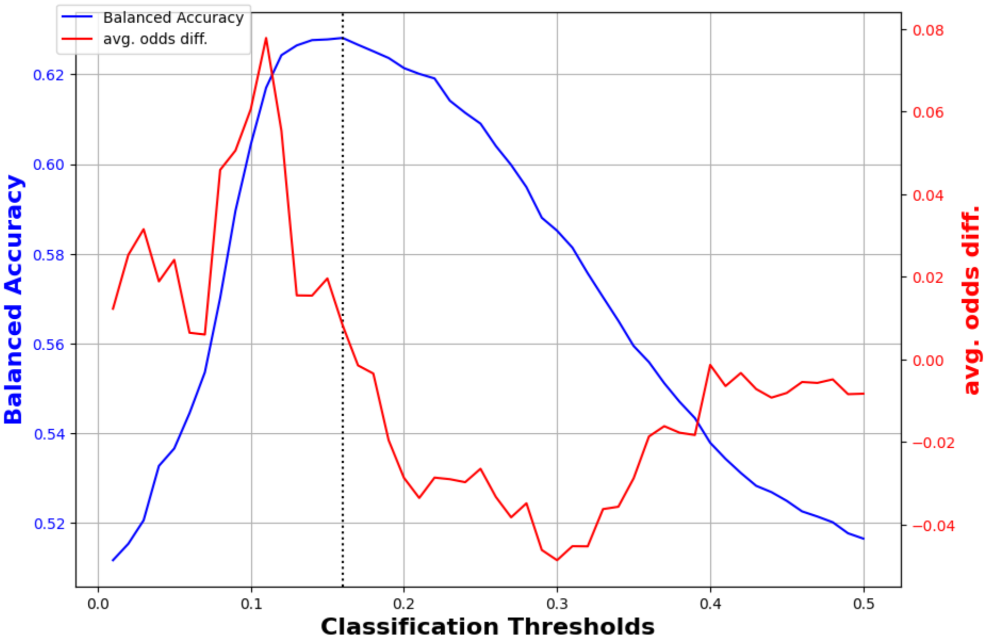


Figure 6: Testing set performance after applying Reweighing

indicate that postprocessing reduces the equal opportunity difference (red dashed line), meaning it improves fairness across groups. However, balanced accuracy (blue dashed line) remains similar or slightly decreases after postprocessing.

Reweighing effectively adjusted the dataset, improving fairness in training, but fairness metrics fluctuated in testing. Calibrated Equalized Odds Postprocessing directly corrected prediction disparities, leading to more stable fairness outcomes in both validation and testing. Given the high class imbalance, postprocessing is preferable because it ensures fairness without modifying data distribution, making it more applicable when fairness constraints must be met.

## 3.5 Limitations and Next Steps

One of the major challenges we faced in this project was difficulty accessing existing models used in classification problems regarding criminal justice. Since the data used to train these models is typically sensitive, there is a lack of public availability of these models. Many of the tools we hoped to assess in our project were inaccessible, which is why we used multiple sources with descriptions of models and features to be used to create a custom model that we moved forward with in the process. We were also limited in finding models that could validate our findings. Our custom model approach introduced new uncertainties and complexity factors without external validation.

Criminal justice-related databases also tend to be imbalanced. In this case, our dataset had significantly more cases of offending (203,730) in comparison to cases of reoffending (39,490). This limitation can cause issues in training and development as the model can

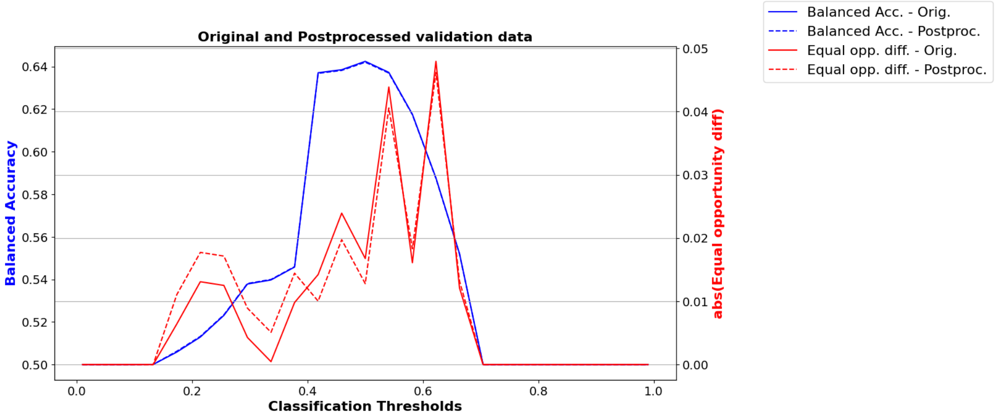


Figure 7: Balanced Accuracy and Equal Opportunity Difference in the validation set before and after postprocessing

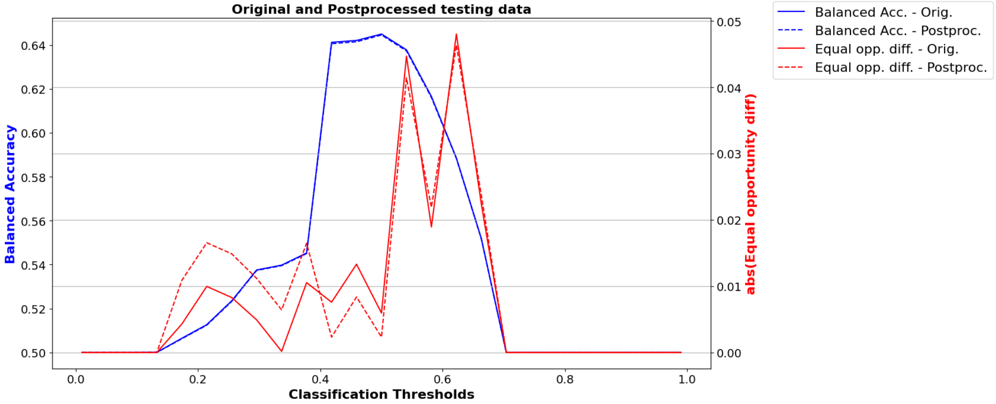


Figure 8: Balanced Accuracy and Equal Opportunity Difference in the testing set before and after postprocessing

over-predict the more common result and not learn the trends in reoffending. This disadvantage also makes it harder to achieve a higher accuracy while not compromising on fairness. Not only that, the result impacts the model's applicability to real-world scenarios, which can be risky and lead to serious consequences in its use case.

As with any predictive task in criminal justice, there are ethical considerations, primarily regarding the model's potential for biased predictions. The model may unintentionally inherit underlying historical biases from the training data, negatively impacting its fairness. Although addressing and reducing bias is a central goal of our project, it remains important to acknowledge this limitation. Certain biases may still exist despite strong efforts to reduce them, so they need to be regularly checked for.

In the future, we would like to pursue gaining access to actual criminal justice predictive models currently in use. As previously mentioned, there is limited availability due to the sensitive nature, but with some proper licensing and allowances, we may be able to gain access to a model and validate our findings from this research. This way, we would know our custom model aligns with the most up-to-date practices and techniques in the field, and our findings about bias can be more directly translated to real-world use cases.

One of the next steps to take would be exploring some more diverse datasets. Currently, the dataset in use is very specific to the NY region. Having a larger database makes our conclusions more generalizable as they capture more complexity in different contexts. With various locations and demographics, the data can help the model be more robust. It can also capture regional differences and provide a more comprehensive representation of behavioral patterns.

It would also be important to continue the model's development to improve its accuracy and try to mitigate any new biases brought on by that process. It is important to avoid any potential overfitting to certain features and introducing new bias. Continuous evaluation of bias throughout the model development process can ensure the model's predictions remain fair.

# 4 Conclusion

This paper explores the application of machine learning fairness techniques to reduce racial bias in pretrial risk assessment algorithms. By identifying bias in these models, we analyze how racial disparities occur and assess the differences in false positive and false negative rates across various racial groups. To address these disparities, we implemented preprocessing, in-processing, and post-processing methods to balance racial representation and ensure the equalization of error rates. Additionally, we measured the effectiveness of bias mitigation techniques and evaluated the trade-offs between fairness and accuracy in the machine learning models.

The initial random forest classification model achieved 83.27% accuracy, but its balanced accuracy was only 50.67%. This value indicates poor performance on the minority class, which in our case are the reoffenders. The model performed well with 0.84 precision, and for non-reoffenders, it achieved a 0.99 recall and an F1-score of 0.91. However, it performed very poorly for reoffenders, with a recall of 0.02 and an F1-score of 0.04, due to the imbalance in the dataset. The feature that was the strongest predictor in the random forest model was `age_at_arrest`, followed by pending misdemeanors and other prior offenses.

The first step in bias mitigation and fairness adjustments was to address the dataset imbalance using random oversampling and SMOTE. This improved the recall for class 1 to 0.50 but led to a decrease in precision, resulting in an F1-score of 0.35. After fine-tuning the model, recall improved to 0.66, and balanced accuracy reached 64.66%. After balancing the data, reweighing was introduced to reduce the bias between the privileged and unprivileged groups. These steps brought the mean outcome difference in the training data to nearly 0, but fairness generalization remained inconsistent in the test set. For post-processing bias mitigation, we used calibrated equalized odds to balance the false positive and false negative rate differences. This technique made the predictions more consistent across the two groups while maintaining overall model performance.

Our findings highlight the challenges of predicting reoffending fairly and accurately. This improved model can help policymakers and criminal justice stakeholders develop more equitable risk assessment tools that reduce bias while maintaining model performance. This research contributes to the field of fairness in machine learning by demonstrating the trade-off between accuracy and fairness in decisions that affect many lives. It emphasizes the importance of balancing model performance with ethical considerations. By applying fairness metrics and debiasing techniques, this study provides insight into addressing bias in predictive modeling within the criminal justice system.

# References

**Aldossari, Bshayer S., Futun M. Alqahtani, Noura S. Alshahrani, Manar M. Alhammam, Razan M. Alzamanan, Nida Aslam, and Irfanullah.** 2020. "A Comparative Study of Decision Tree and Naive Bayes Machine Learning Model for Crime Category Prediction in Chicago." In *Proceedings of 2020 6th International Conference on Computing and Data Engineering*. New York, NY, USA Association for Computing Machinery. [Link]

**Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner.** 2016. "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks." *ProPublica*. [Link]

**Chouldechova, Alexandra.** 2016. "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments." [Link]

**Hardt, Moritz, Eric Price, and Nathan Srebro.** 2016. "Equality of Opportunity in Supervised Learning." [Link]

**Hossain, Sohrab, Ahmed Abtahee, Imran Kashem, Mohammed Moshiul Hoque, and Iqbal H. Sarker.** 2020. "Crime Prediction Using Spatio-Temporal Data." In *Computing Science, Communication and Security*. Singapore Springer Singapore

**Judicial Council of California.** 2022. "Pretrial Pilot Program: Risk Assessment Tool Validation." Technical Report. [Link]

**Kamiran, Faisal, and Toon Calders.** 2012. "Data preprocessing techniques for classification without discrimination." *Knowledge and Information Systems* 33 (1): 1–33. [Link]

**Laura and John Arnold Foundation.**, "Public Safety Assessment: Factors." [Link]

**Mayson, Sandra Gabriel.** 2019. "Bias In, Bias Out." *Yale Law Journal* 128, p. 2218. [Link]

**Ritter, Nancy.** 2013. "Predicting Recidivism Risk: New Tool in Philadelphia Shows Great Promise." *NIJ Journal*(271): 4–11. [Link]

**SERC, MIT.** 2022. "Risk Prediction in Criminal Justice." *MIT SERC PubPub*. [Link]

**U.S. Department of Justice.** 2003. *Recidivism of Prisoners Released in 1994*. Bureau of Justice Statistics. [Link]

# Appendices

This appendix provides definitions and formulas for the key metrics used in our bias mitigation and model evaluation.

## A.1 Synthetic Minority Over-sampling Technique (SMOTE)

SMOTE is a technique used to address class imbalance by generating synthetic examples for the minority class. It selects a random sample from the minority class and interpolates new samples along the line segment joining it with one of its k-nearest neighbors.

$$x_{\text{new}} = x_i + \lambda(x_{nn} - x_i) \tag{1}$$

where:

- $x_i$ is a randomly selected minority class instance,
- $x_{nn}$ is one of its k-nearest neighbors,
- $\lambda$ is a random number in the range $[0, 1]$.

## A.2 Reweighing

Reweighing is a pre-processing technique that assigns different weights to instances in the dataset to reduce bias. Samples from unprivileged groups are assigned higher weights, and those from privileged groups receive lower weights to ensure a balanced dataset.

## A.3 Calibrated Equalized Odds Postprocessing

This post-processing technique adjusts the prediction probabilities to satisfy equalized odds, ensuring similar false positive and false negative rates across groups. It applies probability calibration to correct classification disparities without retraining the model.

## A.4 Balanced Accuracy

Balanced accuracy is the average of recall scores for each class, ensuring performance is fairly evaluated across imbalanced classes.

$$\text{Balanced Accuracy} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \tag{2}$$

## A.5 1-min(DI, 1/DI)

This metric is derived from the disparate impact (DI) measure, which quantifies fairness in classification outcomes. It is calculated as:

$$\text{DI} = \frac{P(\hat{Y}=1|A=0)}{P(\hat{Y}=1|A=1)} \tag{3}$$

where:

- $P(\hat{Y}=1|A=0)$ is the probability of a positive outcome for the unprivileged group,
- $P(\hat{Y}=1|A=1)$ is the probability of a positive outcome for the privileged group.

Calculate the metric 1-min(DI, 1/DI) as follows:(DI, 1/DI)

$$1 - \min(DI, \frac{1}{DI}) \tag{4}$$

1-min(DI, 1/DI) $<$ 0.2 is typically desired for classifier predictions to be fair.

## A.6 Average Odds Difference

Average Odds Difference measures the average difference in true positive rates (TPR) and false positive rates (FPR) between unprivileged and privileged groups. Average Odds Difference must be close to zero for the classifier to be fair.

$$\text{Avg Odds Diff} = \frac{1}{2}((TPR_0 - TPR_1) + (FPR_0 - FPR_1)) \tag{5}$$

where:

- $TPR_0, TPR_1$ are the true positive rates for the unprivileged and privileged groups.
- $FPR_0, FPR_1$ are the false positive rates for the unprivileged and privileged groups.

## A.7 Equal Opportunity Difference

Equal Opportunity Difference focuses only on true positive rates (TPR) and measures the fairness gap in positive outcomes.

$$\text{Equal Opportunity Difference} = TPR_0 - TPR_1 \tag{6}$$

A value of **0** indicates perfect fairness.

## A.8 Theil Index

The Theil Index measures inequality in information distribution. It quantifies prediction disparity and is based on entropy.

$$T = \sum_i p_i \log \frac{p_i}{\bar{p}} \tag{7}$$

where:

- $p_i$ is the proportion of positive outcomes for group $i$,
- $\bar{p}$ is the overall mean proportion.

## A.9 GFPR Difference (Group False Positive Rate Difference)

GFPR Difference measures the disparity in false positive rates between unprivileged and privileged groups.

$$\text{GFPR Difference} = FPR_0 - FPR_1 \tag{8}$$

## A.10 GFNR Difference (Group False Negative Rate Difference)

GFNR Difference measures the difference in false negative rates across groups.

$$\text{GFNR Difference} = FNR_0 - FNR_1 \tag{9}$$

where:

- $FNR_0$ is the false negative rate for the unprivileged group,
- $FNR_1$ is the false negative rate for the privileged group.