Ethics of AI
Prof. Maria De-Arteaga
Zeyu Li, Hayoung Kim, Rajkumar Rajavel, Anukul Kumar Singh
Apr. 26, 2024

Algorithmic Bias in Predictive Policing
- An Assessment on Toronto's Cannabis Arrests

Introduction

The unsettling scene of Yeshimabeit Milner witnessing her former classmates being forcibly detained by police in 2008 at Edison Senior High catalyzed her future in data-based activism. This event in Miami, marked by allegations of excessive force used against predominantly Haitian and African-American students, underscores a systemic issue within law enforcement: the embedding of racial biases within policing practices. These experiences have propelled initiatives like Data for Black Lives, aimed at combating ingrained biases within predictive policing systems, revealing a harsh reality—while these tools are touted for their precision, they often perpetuate discrimination through embedded algorithmic biases.

Characterization of Use

Predictive Policing in the Criminal Justice System

Predictive policing has fundamentally reshaped law enforcement strategies through two main approaches: location-based and person-based predictive tools. Initially, location-based models like PredPol analyze geographical data to forecast crime spots, essentially creating a "crime weather forecast" by updating predictions throughout the day. While insightful, our primary focus lies on the more complex and contentious person-based models.

Person-Based Predictive Policing: A Closer Look

Person-based predictive tools delve deeper, utilizing personal data such as age, gender, marital status, history of substance abuse, and criminal records to predict future criminal activity. Tools like COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) typify this approach. Widely used in jurisdictions across the

United States, COMPAS evaluates the likelihood of recidivism by assigning a risk score from 1 to 10, directly influencing judicial decisions on bail, sentencing, and parole.

These tools are deeply integrated into the sociotechnical systems of the criminal justice system, serving both policing and judicial functions. They are employed to not only guide pretrial and sentencing decisions but also help in probation and parole supervision by prioritizing high-risk individuals. The deployment of these models has been both widespread and impactful, driven by a need for efficiency and cost reduction following the 2008 financial downturn and subsequent budget cuts to law enforcement.

Ethical Consideration

1. Systematic Discrimination - Relevance, Generalizations, and Compounding Injustice

Systematic discrimination within predictive policing arises primarily from the misuse of arrest data, challenging the ethical principle of relevance. Arrest records, often employed as proxy indicators for future criminality, do not necessarily correlate with true criminal behavior, leading to significant misjudgments. For instance, these records might reflect historical biases in law enforcement rather than actual crime rates, especially in minority communities that have been subject to disproportionate policing. This reliance on inappropriate data underlines the flaw in relevance, where characteristics such as location or prior arrests are used to predict criminal propensities inaccurately.

Further exacerbating the issue, predictive algorithms employ generalizations that categorize individuals based on broad and often inaccurate demographic traits like race or zip code. This practice not only overlooks individual behaviors but also perpetuates stereotypes, thus embedding past biases into current predictive models and impacting policing strategies. Such generalizations fail to provide the granularity needed to make fair assessments and instead cement systemic biases.

Moreover, these biases contribute to compounding injustice, where the affected communities continue to bear the brunt of historical prejudices. For example, skewed data may direct more police patrols to specific neighborhoods, increasing the likelihood of encounters and arrests based on biased data rather than actual criminal activity. This feedback loop not only enforces existing prejudices but also deepens social and legal inequalities, trapping these communities in a relentless cycle of discrimination and injustice.

2. Mismeasurement Errors

Mismeasurement errors in predictive policing algorithms stem from their dependence on flawed or incomplete data sources that inadequately capture genuine criminal activity or individual propensities towards crime. These algorithms typically use arrest rates as a fundamental dataset, which disproportionately affects Black communities due to ingrained racial biases in policing practices. Consequently, this leads to higher predictions of crime rates in these areas, based not on objective assessments but on historically biased data.

Moreover, the algorithms' reliance on proxies such as zip codes, educational backgrounds, and socioeconomic statuses, which are legally permissible substitutes for race, perpetuates racial stereotypes and systemic injustices. These proxies indirectly measure criminal behavior through demographic factors associated with historical discrimination, thus exacerbating mismeasurement errors and contributing to a cycle of criminalization and incarceration based primarily on biased data interpretations. This cycle not only misrepresents the true nature of criminal behavior in diverse communities but also reinforces the racial prejudices embedded within the criminal justice system.

3. Sampling Bias and Differential Subgroup Validity

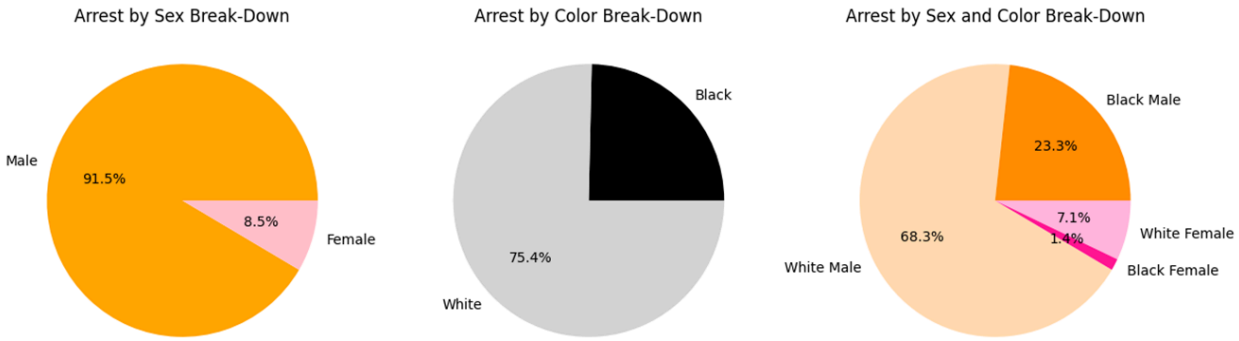
Sampling bias is distinctly evident in the development and application of predictive policing tools, such as Static 99, which was calibrated in a Canadian context where only about 3% of the population is Black—strikingly unrepresentative of the U.S., where Black individuals comprise approximately 12% of the population. This leads to differential subgroup validity, where the algorithm’s accuracy and performance vary significantly across different demographic groups, generally underperforming for minority populations.

Such biases are magnified when predictive models rely on outdated or culturally irrelevant predictors like landline phone ownership to assess court appearance likelihoods. These predictors do not adapt to current societal norms or technological advancements, thus skewing risk assessments and potentially leading to unjust outcomes. This reliance on data from demographically different regions like Canada or Europe exacerbates disparities, resulting in flawed risk assessments, arrest patterns, and sentencing decisions within diverse American contexts.

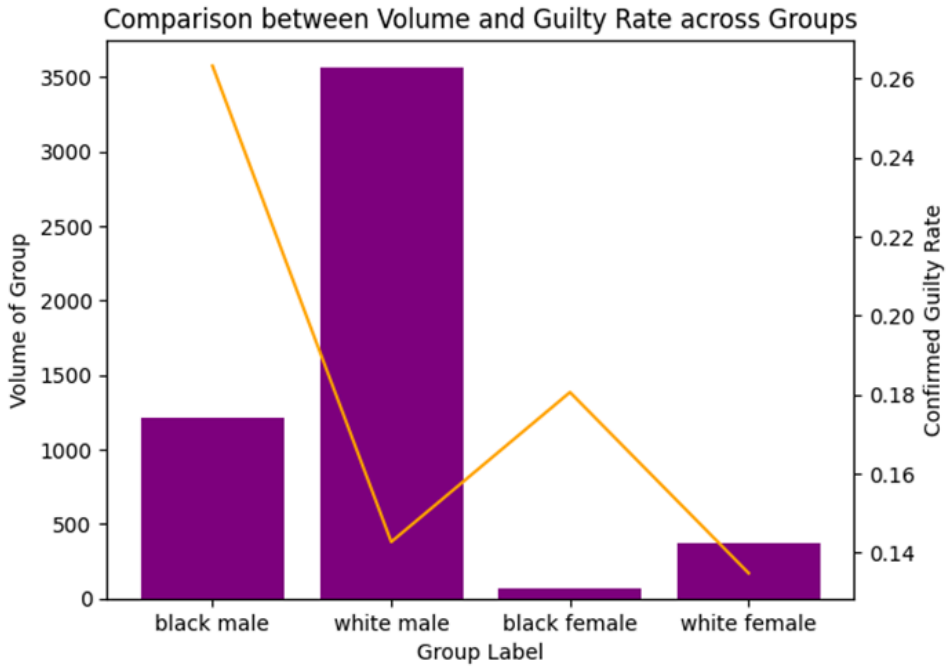
Empirical Bias Assessment

To assess the potential bias in the ML model development phase for predictive policing tools, we conduct a mock system development experiment using Cannabis-related arrest data from Kaggle collected by the police department in Toronto.

First thing to notice before we run any analysis is the great disparity between different ethnic and sexual groups. As shown in the graph below, the demographic distribution of arrested suspects largely lean on white male population. This rests as a potential sampling bias to our model.



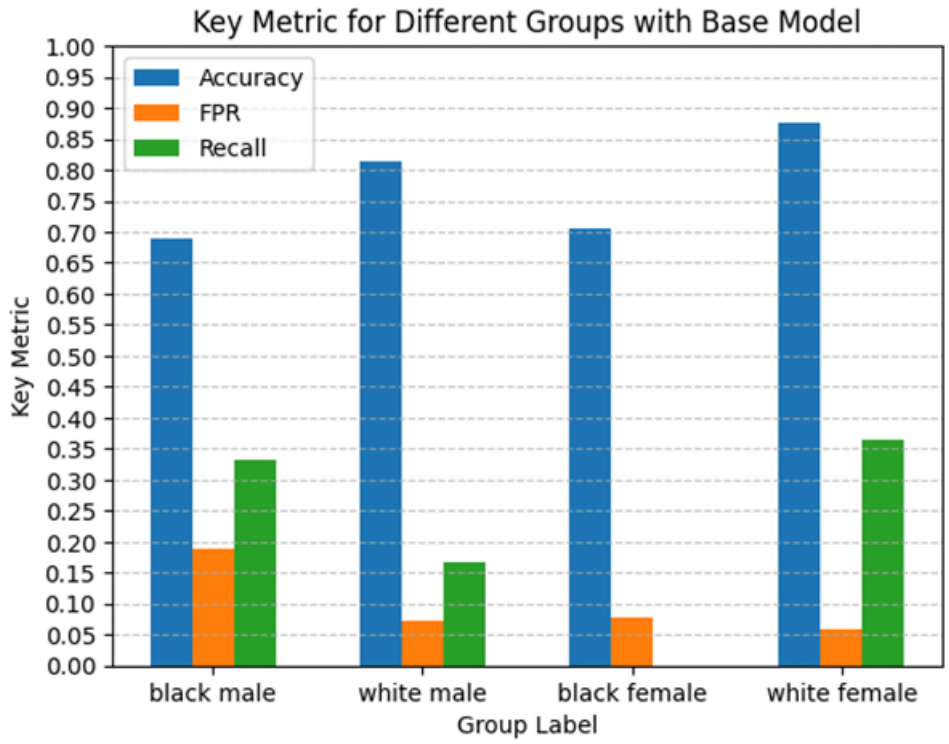
With some exploratory data analysis, it appears that there also subgroup validity embedded in the arrest record. While white male stand as the largest group of arrests, the true guilty rate is higher for black population.



The base Model

Bearing these potential problems in mind, we devised the “base” model that uses random forest classification to predict guiltiness based on a series of variables including age, arrest records, sex, and color.

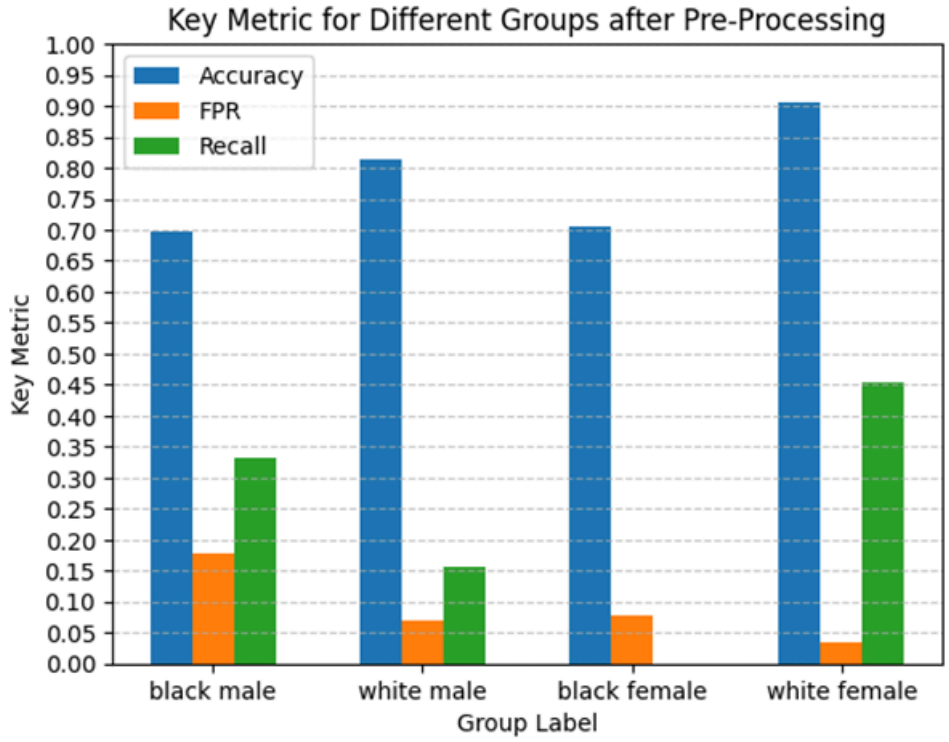
The base model achieved 79% accuracy. However, accuracy is commonly a fallacy. From the graph shown below, we can observe an unsatisfactory prediction performance on true addicts. The model has merely put a great number of people as innocent to achieve accuracy, as there are only about 25% actual guilty instances from the population. Nonetheless, this base model generates high False Positive Rates (innocent people being wrongfully caught) in black males, and cannot identify black female addicts at all.



Reweighting

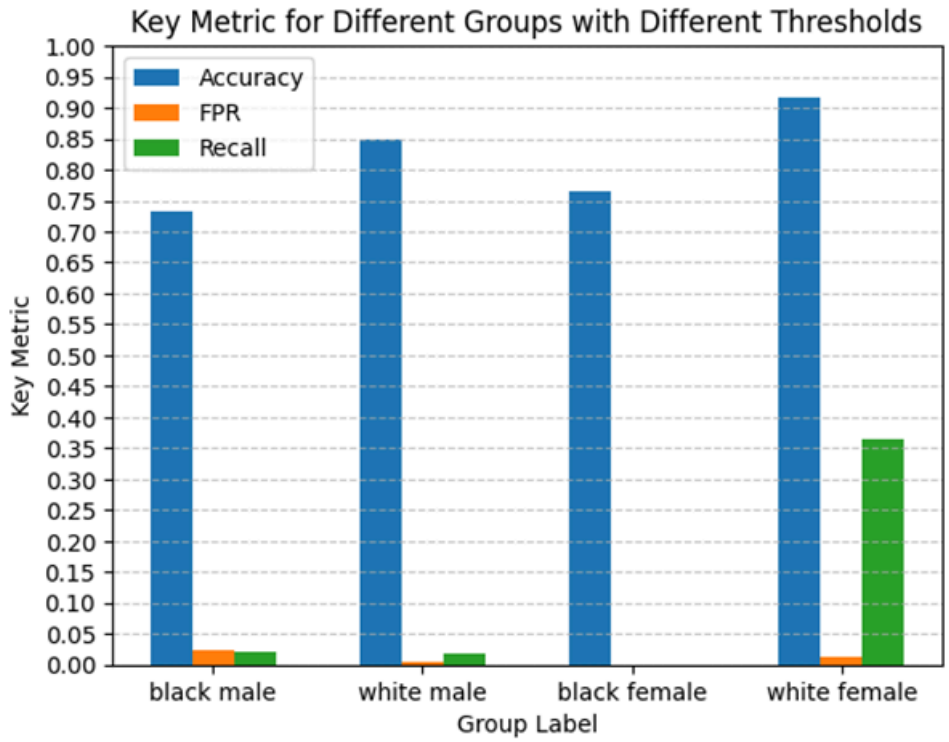
The first data-processing technique we used is re-weighting. As mentioned, the subgroups are largely disparate in volume, so we re-weighted their representation in the model to make sure the model gets equivalently trained by subgroups. However, the result of pre-processing does not seem impactful. Key metrics like recall and FPR are

almost identical to those in the base model, nor did the new model improve the overall accuracy. This would imply that although different groups are not sampled to actual ratio, sampling bias may not be the true cause of the model's poor performance.



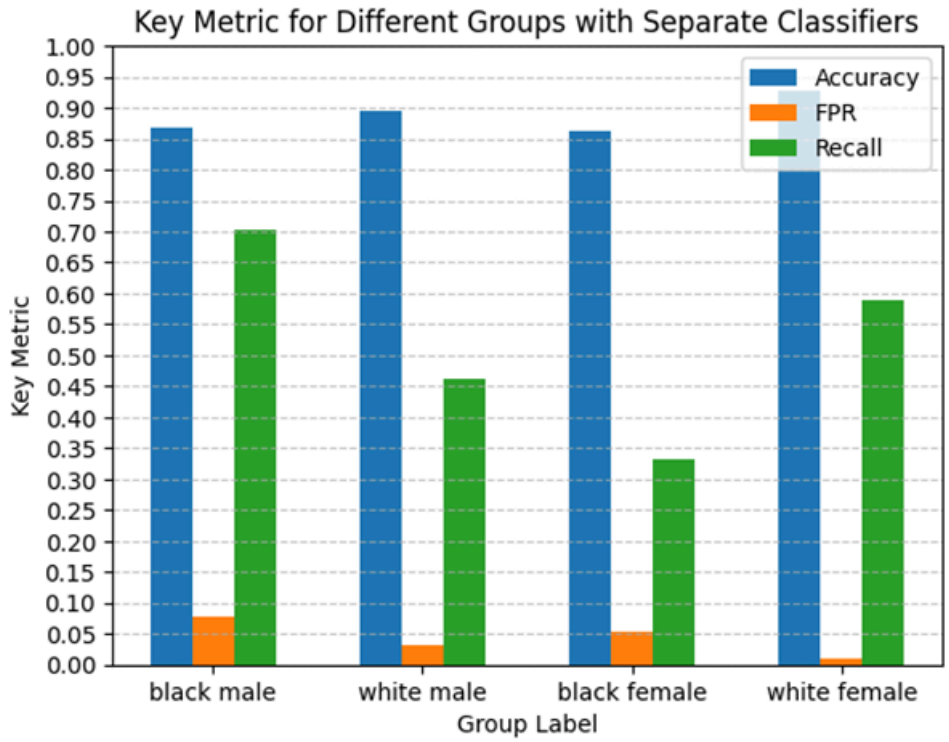
Differential Classification Thresholds

To further mitigate the difference of prediction power among groups, we employed differential classification thresholds to testify potential biases in the final evaluation. Here we pushed differences in FPR for the groups within a negligible range, and selected the set of thresholds to achieve highest possible accuracy. However, this approach even drove the model further done to recognize virtually all instances as innocent. This model could only identify a small portion of white female addicts, and achieved fairness at the cruel cost of model predictability. There we see a failure of the models that takes sex and color as a lump sum of variables in a large model, which calls for the devision of group-based classifiers.



Sub Classifiers

The last attempt was to set different classifiers for each group. This new model has significantly improved recall rate across all groups, showing it can effectively predict and arrest real addicts from suspects. The model has also notably reduced all FPR to below 10%, meaning it on average take more than 10 correct arrests to make a wrong arrest.



Conclusion

The last model certifies that fairness is not a necessary breach of accuracy. Adding variables to the model, or even expanding the model into small sub-models, will not harm, but improve model predictions. This calls for a need to correctly identify demographic features of covariates of a predictive policing model. Sex and color are not the end, there are more long-established bias against people of education background, living locations, nationalities, etc. Some of the biases were construed with a historical record, but making a fair predictive model can make sure only real bad people are caught. Ethical considerations are so deeply needed in such socially important ML/AI uses to plug out the deep-rooted human biases, and to justify subgroup fairness and equality.

Sources

Characterization of use:
https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/

History of predictive policing:
https://medium.com/@Vera_Kerber/a-brief-history-of-predictive-policing-in-the-united-states-ec3568e5c42c

Empirical Bias Assessment - Data source:
https://www.kaggle.com/datasets/utkarshx27/arrests-for-marijuana-possession