Machine Learning in Policing and Justice System: Navigating Fairness and Societal Impact

JINGYI ZHANG U26578499
JYZ0328@BU.EDU

Introduction

Machine learning (ML) has made significant strides in transforming various industries, and its application in policing and the justice system is no exception. This article delves into the intricate landscape of machine learning within law enforcement, exploring the factors influencing fairness and the tools available to detect and mitigate bias.

Data Quality and Bias, Algorithmic Transparency and Human Oversight are the factors that can impact fairness of use of machine learning. Fairness Metrics as well as Bias Detection and Mitigation are Fairness Tools to detect and adjust bias. Discriminatory, Inequities and Trust Erosion are the potential problems, which would cause lost in public confidence,lost in trust in the justice system and would contribute to historical injustices.

Factors Influencing Fairness

1.Data Quality and Bias

Machine learning models heavily depend on the data they are trained on. Biased or incomplete datasets can perpetuate existing inequalities in the justice system. It is imperative to scrutinize and address biases present in historical data to prevent discriminatory outcomes. For example, facial recognition algorithms have shown biases, with higher error rates for individuals with darker skin tones. Mis-identification can lead to wrongful arrests and reinforces racial disparities.[1] Another example is that natural language processing models trained on historical texts may inherit and propagate biases present in those texts, which could include gender, racial, or socioeconomic biases. [2]

2.Algorithmic Transparency

The lack of transparency in some machine learning algorithms poses challenges to ensuring fairness. Understanding how algorithms make decisions is crucial for identifying and rectifying any unintended biases. Some social media platforms use algorithms to determine the content users see, but these algorithms are often proprietary and lack transparency. This may lead to information filtering and personalized content delivery, influencing users' opinions and perspectives.[3] The operations of credit scoring models are often non-transparent, making it challenging to understand the specific factors determining individual credit ratings. This may lead to unfair treatment of certain groups.[4] In a word, transparency play an important role in deciding fairness in various areas.

3.Human Oversight

While machine learning models can enhance efficiency, the importance of human oversight cannot be overstated. Human intervention is essential to interpret complex situations, ethical decision-making, and to rectify biases that may emerge. Machine learning models are typically trained on vast amounts of data, but they may face limitations when confronted with complex, ambiguous, or novel situations. Human oversight provides an understanding and interpretation of these contexts, leveraging experience and intuition to navigate uncertainty.In decisions involving ethical considerations, machine learning models may lack a profound understanding of ethical nuances. Human oversight utilizes ethical principles, cultural backgrounds, and societal values to make judgments, ensuring decisions align with ethical standards.[5] For instance, in autonomous vehicles, human supervision is essential to deal with unexpected and complex situations on the road. Humans are better able to adapt to ambiguous scenarios, such as unexpected obstacles or bad weather conditions.

Fairness Tools

1. Fairness Metrics

Implementing fairness metrics allows for the quantitative assessment of how different demographic groups are affected by machine learning models. This helps identify disparities and enables fine-tuning to achieve equitable outcomes.

That is to say, Fairness metrics provide a quantitative method to evaluate the impact of machine learning models on different demographic groups. Through specific numerical indicators, we can more accurately understand whether there are inequalities or differences between model outputs for different groups. This aids in early detection of potential issues, allowing for Fine-Tuning of model to ensure fairness for all groups. Implementing fairness metrics also encourages the consideration of fairness during the algorithm design phase. This helps prevent potential fairness issues and ensures that models are designed with fairness in mind from the outset. For example, consider a loan approval model where fairness metrics can be used to assess whether there are disparities in the loan approval rates for different racial or gender groups. Through these metrics, potential unfairness can be identified and adjustments can be made to ensure equal loan approval.[6]

2. Bias Detection and Mitigation

Specialized tools are available to detect and mitigate bias in machine learning models. These tools analyze the input data, model predictions, and output to identify and rectify potential biases, ensuring that decisions are fair and just.

In handling and maintaining fairness, it is essential to use specially designed tools. These tools typically integrate advanced algorithms and methods, enabling more effective identification and addressing of potential biases. The task of these tools goes beyond just detecting biases; it also includes correcting potential issues. This ensures that not only biases are discovered but proactive measures are taken to repair and enhance the model, ensuring fairness and justice in decision-making. This means that by using these specialized tools, we can rely more confidently on machine learning models to make decisions that align with ethical and fairness standards.

AI Fairness 360 (AIF360) and Google's What-If Tool are all good examples.AIF360 is an open-source toolkit developed by IBM that provides a comprehensive set of metrics and algorithms to help detect and mitigate bias in machine learning models. It includes tools for measuring and addressing bias in data and models, supporting fairness-aware machine learning. The What-If Tool by Google is an interactive visual interface designed to probe and understand machine learning models. It allows users to analyze and detect bias in model predictions, exploring different scenarios to understand the impact of changes on fairness. [7]

Articulating the Problem and Societal Impact

The deployment of machine learning in policing and the justice system raises crucial concerns about the potential amplification of existing biases and the perpetuation of systemic inequalities. The following points highlight the gravity of the issue.

1.Discriminatory Outcomes

Unchecked biases in machine learning models can lead to discriminatory outcomes, disproportionately impacting certain demographic groups. This not only erodes trust in the justice system but also perpetuates historical injustices.

When unchecked biases exist in machine learning models, it refers to the models inadvertently learning and perpetuating biases present in the training data. These biases may involve race, gender, socioeconomic status, or other demographic factors. When these biases are not adequately addressed, they can lead to discriminatory outcomes. For example, a hiring algorithm may unintentionally favor candidates of a specific gender or race, resulting in unequal employment opportunities.Similarly, a housing pricing model might unfairly impact residents of certain communities due to historical unjust pricing.

2.Reinforcement of Inequities

If historical data used to train machine learning models reflects societal biases, these biases may be perpetuated or even exacerbated. The justice system has a responsibility to break away from historical injustices rather than reinforcing them.

When machine learning models are trained on historical data reflecting societal biases, there is a risk of perpetuating or exacerbating these biases. The data used for training often mirrors existing inequalities related to race, gender, socioeconomic status, or other demographic factors. If not addressed, machine learning models may unintentionally reinforce and even amplify these historical inequities. For instance, Predictive policing algorithms, if trained on historical crime data, may perpetuate biases present in law enforcement practices. This can result in over-policing of certain communities and contribute to a cycle of inequitable law enforcement. Another example is that machine learning models used to guide sentencing in the criminal justice system may inadvertently reflect historical biases in sentencing practices. This could lead to disparities in sentencing outcomes for different demographic groups.

The justice system carries a significant responsibility in this context, as it should actively work to break away from historical injustices rather than becoming a vehicle for their perpetuation. Ensuring fairness in the deployment of machine learning models is crucial for promoting justice and equity within societal systems.[8]

3.Trust Erosion

Unfair decisions made by machine learning models can erode public trust in the justice system. Transparency, accountability, and proactive measures to address bias are essential to maintain public confidence.

Public trust is foundational to the effective functioning of the justice system. When machine learning models produce decisions that are perceived as unfair or biased, it can lead to a erosion of trust among the public. Transparency in how these models operate, accountability for their outcomes, and the proactive identification and correction of biases are crucial in mitigating this erosion.

The lack of transparency in the decision-making process of some machine learning algorithms, often referred to as "black box" models, can contribute to mistrust. If individuals cannot understand how a decision was reached, it becomes challenging for them to accept and have confidence in the outcome. Therefore, disclosing the inner workings of these models and making them interpretable is essential for maintaining transparency.

Accountability is another key aspect. When unfair decisions occur, holding the responsible entities accountable helps rebuild trust. This involves investigating and rectifying instances of bias, providing explanations for decisions, and ensuring that there are consequences for unjust outcomes.

Proactive measures to address bias are critical for preventing unfair outcomes in the first place. This includes continuous monitoring and auditing of algorithms, adjusting training data to remove biases, and implementing fairness-aware machine learning techniques. Proactively addressing bias demonstrates a commitment to fairness and helps maintain public confidence. [4]

4 social impact

If people perceive that machine learning models yield unfair, discriminatory, or untrustworthy outcomes, it can undermine confidence in the objectivity and fairness of the decision-making process. In critical fields such as law, the erosion of trust is particularly concerning, as fairness and justice are fundamental principles.

Unchecked biases in machine learning models not only impact current decision-making but also have the potential to perpetuate historical injustices. If the historical data used to train the model contains biases or reflects systemic inequalities, the model may inadvertently learn and perpetuate these patterns. This perpetuation exacerbates existing disparities, hindering progress towards a more equitable and just society.

In summary, unchecked biases in machine learning models, coupled with inequalities and the erosion of trust, can lead to discriminatory outcomes that disproportionately affect specific demographic groups. This not only diminishes trust in the justice system but also perpetuates historical injustices, underscoring the importance of addressing biases in machine learning for the sake of fairness and social justice.

Conclusion

As machine learning becomes more integrated into the policing and justice system, it is paramount to navigate the delicate balance between efficiency and fairness. Addressing biases, utilizing fairness tools, and promoting transparency are essential steps to ensure that the societal impact is positive and that machine learning contributes to a more equitable justice system. The responsibility lies not just in the hands of technologists but in the collaborative efforts of policymakers, ethicists, and society as a whole.

Reference

[1] "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification" by Joy Buolamwini and Timnit Gebru, Conference on Fairness, Accountability, and Transparency (FAT), 2018.
https://www.semanticscholar.org/paper/Gender-Shades%3A-Intersectional-Accuracy-Disparities-Buolamwini-Gebru/18858cc936947fc96b5c06bbe3c6c2faa5614540
[2] "Semantics derived automatically from language corpora necessarily contain human biases" by Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan, Science, Volume 356, Issue 6334, 2017.
https://www.researchgate.net/publication/316973825_Semantics_derived_automatically_from_language_corpora_contain_human-like_biases
[3]Algorithmic Accountability: A Primer
Robyn CaplanJoan DonovanLauren HansonJeanna Matthews
https://datasociety.net/library/algorithmic-accountability-a-primer/
[4]Discrimination in Online Ad Delivery" by Latanya Sweeney, Communications of the ACM, 2013
https://dl.acm.org/doi/10.1145/2447976.2447990
[5]The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation
https://www.eff.org/deeplinks/2018/02/malicious-use-artificial-intelligence-forecasting-prevention-and-mitigation
[6]Fairness and Abstraction in Sociotechnical Systems

Andrew D. Selbst&danah boyd&Sorelle A. Friedler
http://sorelle.friedler.net/papers/sts_fat2019.pdf
[7] "A Survey on Bias and Fairness in Machine Learning" by Mehrabi et al. (2019)
https://arxiv.org/abs/1908.09635
[8]Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy" by Cathy O'Neil (2016)
https://blogs.scientificamerican.com/roots-of-unity/review-weapons-of-math-destruction/