IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 6, June 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 6, June 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1206086 |

Ethics and AI: Mitigating Bias in Machine Learning Models

Yogesh Santosh Umalkar, Prof. Swati Chopade

MCA Student, Department of MCA, Veermata Jijabai Technological Institute, Matunga, Mumbai, Maharashtra, India

Guide, Department of MCA, Veermata Jijabai Technological Institute, Matunga, Mumbai, Maharashtra, India

ABSTRACT: AI has quickly integrated itself into many fields but the unnoticed biases in algorithms present considerable ethical issues. This research paper aims at understanding the degree of bias in use of machine learning (ML) in making decisions. In this paper, the prejudice in AI systems is illustrated by actual scenarios in the fields of healthcare, criminal justice system, and employment, to demonstrate fair discriminations that occur due to the use of bias models.

We continue exploring various aspects of bias and how it affects ML models using the AIF360 toolkit with concrete examples. Similar to fairness impact ratio, disparate impact and equal opportunity measures calculate bias and evaluate its effect on a specific group of people. Furthermore, AIF360 bias mitigation algorithms, for instance reweighing and adversarial debiasing, concretely detail how biases are dealt with and fairness achieved.

Hence, this research seeks to analyze the interconnection between AI, bias, and fairness in order to provide the necessary knowledge and tools that will allow stakeholders develop efficient, polite and bias free AI systems. Its aim is to advance the creation of AI technologies that are FAIR and therefore, positive socio-economic impact that will not marginalize some structures in society.

KEYWORDS: AI, Machine learning, bias, bias reduction, ethics

I. INTRODUCTION

Through advanced technology, AI continues to grow in popularity within different industries; that today's decision-making process especially concerning justice and employment cannot go without the assistance of the powerful tool. Although, AI has the potential of boosting efficiency and effectiveness the world over, the fact that it is increasingly being oriented towards data-based decision making presents the society with some of the most basic ethical questions. Underpinning these worries is the problem of bias which is inherent to the algorithmic approach that AI employs in specific activities resulting in AI reinforcement of the existing social injustices.

More specifically, Machine Learning or ML as part of AI is a process of allowing the algorithms to emulate patterns and make the necessary suggestions based on big data. However, given the fact that ML algorithm learning is dependent on the occurrence and patterns of past events, the algorithms are conditioned to exhibit the biases present in the data used for training. It goes without saying that bias in training data will lead to similar bias in the models produced 'these models are a product of the environment that they are trained in, and if the training environment is unjust and prejudiced, then the models will be the same'. The nature of this phenomenon can be rather problematic and result in discrimination with possible drastic negative implications for people and vulnerable populations.

For instance, in the criminal justice system, AI algorithms in risk assessments might rule some individuals from certain race or pocketless backgrounds as high risks thus denying them bail or elongating their term. Likewise in employment, when AI is applied for resume filtering, possibly prejudiced algorithms could consistently filter out worthy applicants on grounds of gender or color.

Bias in AI is a problem that goes beyond individual cases, as it can perpetuate the existing inequalities and slow down the development of a society based on fairness and justice. Thus, it is crucial to confront the problem of bias in the creation of ML models directly. This research paper aims at undertaking an extensive analysis of how bias enters into the ML models with a focus on the causes of the phenomenon and its consequences on the systems and users.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 6, June 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1206086 |

In this paper, to identify the specific sources of bias in the machine learning models, we are planning to examine the numerous failures and analyze the existing scientific literature and case studies. We will also discuss the various forms of bias including disparate impact, disparate treatment and algorithmic bias.

However, besides providing a description of the problem, this paper will also discuss possible recommendations. In this case, we will discover some techniques for bias reduction including data pre-processing, algorithm design with a focus on fairness; and post-processing. Thus, we will also touch upon the need for the ethical standards and regulatory measures to be set in the sphere of AI advancement and application.

It is with this spirit that this research paper hopes to contribute toward the current discourse on AI ethics and fairness particular to bias in ML models by unveiling the subtle issues with the models and providing potential solutions toward the problem. Thus, our long-term aim is to help all the aforementioned groups of stakeholders, including researchers, practitioners, policymakers, and the public, to create and use AI systems that are not only effective but also fair and ready to address the existing challenges.

II. LITERATURE SURVEY/EXISTING SYSTEM

AI systems and their use have been on the rise in recent years and thus a lot of studies have been conducted on the problem of bias. Many experts have noted that there are several main kinds of bias that hold AI algorithms hostage.

These include:
Bias in Training Data: This is maybe the most familiar source of bias and takes place when the data we used to train the model is not similar to the data the model will encounter. This is because past prejudices including social and historical may be inherent in the data feed into the algorithm and the result will reflect the same.

Inherent Bias in Algorithms: While calling for equally adjusted datasets, the algorithms that the big players implement can inadvertently have biases that come from their configuration or assumptions. For instance, optimization working for certain objectives can lead to certain population domination over the others.

Bias in Developers/Creators: People's prejudice in one or another sense can become a part of the development process depending on the decisions made during data gathering, defining components and functions, or designing algorithms. Such biases could inevitably define the behavior of the obtained AI system.

The studies for bias in ML models have suggested different methods to measure and evaluate bias. Some of the most widely used metrics include: Some of the most widely used metrics include:

Statistical Parity Difference: Calculates the absolute amount of difference in the probability of the occurrence of a positive event in favor of one group against another.

Equal Opportunity Difference: Measures how algorithms' true positive rates have changed across groups, and in particular relative to their balanced accuracy for classification problems.

Disparate Impact: Ascertains the proportionate of positive decision for people in a specific category often employed in the assessment of discrimination in lending/employment credit models.

Scholars have also come up with various methods aimed at reducing bias in ML models which are grouped into pre-processing, in-processing, and post-processing techniques.

Pre-processing: These techniques work with the idea of trying to remove any bias that could be in the data used to train the model before the data is fed into the model, through the various methods like reweighing of the data or sampling.

In-processing: These techniques integrate the fairness constraints into the learning algorithm so as to force the model to learn with fairness as part of the objectives.

Post-processing: Implementation methods should regulate the model's behavior after training for fairness purposes, such as altering the decision boundaries.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 6, June 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1206086 |

Although, there are some advancements in these fields currently available solutions are not sufficient enough to eliminate bias in AI entirely. Some of the issues that have been identified include, the absence of robust tools that can be used to measure the effectiveness of bias reduction strategies in relation to other strategies within various fields. Also, key objectives such as fairness and accuracy are often inversely proportional; in other words, it is hard to achieve both at the same time.

The last crucial problem is complexity, non-understandability of many AI systems, and the marked absence of critical self-describing and self-justifying. AI decisions are worst when they are not transparent and this make it hard to remove bias and this in the end lowers the confidence that users have in the system.

Overall, AI research has made significant progress in constructing fair AI systems; however, further studies are needed in the future as a way to produce even more reliable methods of bias suppression, taking into account the context, as well as dependably increase transparency and explainability. Besides, the problem of AI is that it was almost exclusively developed by computer scientists, while the society and other scientists were left out of the equation; thus, the cooperation between the specialists in various fields, such as computer science, social science, ethicists, and policymakers is imperative.

III. PROPOSED METHODOLOGY AND DISCUSSION

In this research, we proceed with a data mining task with the primary goal of detecting bias in ML models using only the AIF360 toolkit and a dataset of US homicide reports. Therefore, our analysis is aimed at identifying and dealing with sources of bias in the prediction of the attributes of the perpetrator on the basis of the victim.

Descriptive statistics of the datasets and exploratory data analysis or pre-analysis.
The data set we use obtained from Kaggle includes data from 1980 to 2014 and it has more than 630,000 records. Each record mobilises paramount evidential data concerning a homicide occurrence such as the aspects touching on the victim (e. g., age, sex, race) and the offender (e. g., age, sex, race). This dataset is useful for the purpose of seeing if there is any bias that might creep into the decision-making process of estimating perpetrator attributes, especially aspects such as race and sex.

Our process of analysis starts with the exploratory data analysis (EDA). This involves assessing the various descriptive statistics of the dataset and distribution of important variables and some analyses of victim and perpetrator characteristics. This is because EDA allows one to detect sources of bias that may be existed in data including imbalanced class distributions or relations between sensitive attributes and outcomes. For example, we may find minority race groups being represented differently in the victims or the perpetrators' column which may affect the model's outputs.

The Baseline Model and the Fairness Assessment
After EDA, we set a 'naïve' ML model that has no component dealing with bias removal mechanisms. This first model, which is often logistic regression or random forest, is compared with other methods to determine the effectiveness of the bias reducing techniques.

Using the AIF360 toolkit, we analyze the fairness of the baseline model on which we have worked. We utilize the Fairness metrics from AIF360: disparate impact, statistical parity difference, and Equal Opportunity Difference. These metrics allow assessing the magnitude of the deviation of the model's predictions using victim attributes as a means of differentiation. For instance, we can find out that the model has disparate impact such that prediction of a particular race for perpetrator varies with the victim's race.

Bias Mitigation Strategies
Leveraging from the fairness assessment information, we proceed to begin the process of bias reduction programs formulated in AIF360. These algorithms encompass:
Pre-processing: Some of the algorithms called "Reweighing" and "Disparate Impact Remover" are used to adjust the patterns of bias in the training data prior to training the model. Reweighing influences the weights of the instances based on their values of protected attributes, while Disparate Impact Remover aims at modifying features to meet the state of ratios.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 6, June 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1206086 |

In-processing: There are approaches such as Adversarial Debiasing and Prejudice Remover which incorporate the fairness constraints by design during the training of algorithms. Post-processing methods such as Adversarial Debiasing incorporate an adversary, which aims to lessen the adversary's ability to guess the protected attribute from the model's output.

Post-processing: There are methods such as "Equalized Odds Postprocessing", "Reject Option Classification", to eliminate unfairness after the training of the model. Equalized Odds Postprocessing adjusts the probability to achieve the equality of respective false positive and true positive rates, and Reject Option Classification will postpone decision making when a model has uncertainty to produce skewed results.

We adopt these algorithms with great care to the homicide dataset and assess their effects on the measures of unfairness and predictive accuracy. Thus, this iterative process enables one to identify the best action plans for addressing certain types of biases in the data.

Comparative Analysis and Conclusion

The main focus of our elaborate work involves comparing the effectiveness of the different strategies that aim to reduce the effects of bias. We evaluate their performance in eradicating bias on a range of fairness measures while at the same time keeping the model's accuracy intact. This is possible due to the comparative analysis, which permits to state that the recovery from bias always comes at the cost of diminished precision, while keeping the page as balanced as possible.

In addition, we establish how well each technique works in various conditions that include the degree of bias in a given dataset or the sensitivity towards certain fairness measures. This more refined depiction thus provides detailed descriptions that highlight the sociology of the two methods with respect to the nature of their performance, and where and when each is best suited to be used.

Thus, after analyzing concrete data and using such new instrument as AIF360, the study advances the emerging field of AI fairness. The results obtained in this work aim to increase understanding of how bias exists and works in ML models and offer practical recommendations for constructing AI systems free from prejudice. In this spirit, we seek to carry out research, both theoretical and empirical, that will enable others—especially researchers, professionals, and public policymakers—to find their way through the many and varied conceptual and practical problems that attend AI systems development and use in the world as it is today and as we imagine it could or should be.

IV. RESULTS

Our analysis of the US homicide dataset through the lens of the AIF360 toolkit reveals a stark reality: the baseline machine learning (ML) model shows biases in the prediction of the perpetrator's properties depending on the victim's profile. These biases are evident in several key findings: These biases are evident in several key findings:

Racial Bias: Self's model shows a clear pattern of predicting a higher probability that a white offender was involved if the victim too is white. Quantifying this bias is what the statistical parity difference metric does, and when using this metric, the disparity between the model's predictions of white and non-white victims is significant.

Disparate Impact: The deciding of disparate impact also provides concrete evidence of racism, which tells that the model predictions are more often burdensome to a specific race. By this, it means that the model will often predict that people of certain racial background are perpetrators even when they are not so that it increases the chances of discriminating against individuals who have such backgrounds.

Gender Bias: The gender test also exposes a similar issue but with less severity as observed in the case of race. The model is likely to attend more importance to some attributes of the perpetrator when dealing with victims of a specific gender thus discriminating against the other groups.

This argument proves that the issue of bias requires better strategies and solutions to eliminate from the AI systems. When applying different from the AIF360 toolkit, we successfully achieved positive changes in fairness of the models.
Pre-processing: Training where the weights of the training data are adjusted based on protected attributes such as race and gender shows a minimal disparity and a considerable difference in statistical parity, thus being able to reduce discriminative impacts and improve equality when giving out predictions.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 6, June 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1206086 |

In-processing: It is worth mentioning that adversarial debiasing is an approach where a model is trained to compete with a detection adversarial trained to identify bias; the experiment shows that the technique reduces both racial- & gender-bias in the course of training. This method helps in providing representation learning that will not focus a lot on protected attributes so that both fair predictions are made.

Post-processing: That is why, for example, after using the equalized odds postprocessing method, a further refinement of the decision-making model's conclusions is achieved, and false positives and true positives are comparable for all the subjects distinguished. This assists in reducing discrimination and achieving equality while handling the case.

However, our analysis also sheds light on a more fundamental conflict, namely, that between fairness and the accuracy of models. At times, it is possible to balance the bias by applying ways and means that help in bias reduction and this in a way can lower the standard accuracy margin of the model. This trade-off emphasizes the need to choose and fine-tune the measures for reducing model bias that correspond to a specific setting and use case of the AI system at hand to achieve the best balance between the two.

Therefore, utilising the bias analysis of the employed dataset of homicide rates in the USA, it is possible to determine the presence of elevated levels of bias in the ML models and outcomes of applying different bias mitigation strategies. In this paper, we have used AIF360 toolkit and data analytical methods to measure bias levels, address bias and analyze the effect of bias mitigation on models' fairness and performance. Thus, the results highlight the importance of presenting further papers on efforts in the creation of fair AI systems to spread fair use of technologies in artificial intelligence and utilization of its potential for the peoples' overall benefit without causing new prejudices.

V. CONCLUSIONS

This research paper proves that bias removal in ML models is possible utilizing the AI Fairness 360 toolkit. The accuracy of the different fairness metrics and bias mitigation algorithms focuses on the mitigation and improvement of fairness of the AI systems, such as reweighing and adversarial debiasing. Thus, it can be concluded that eliminating bias means constant research and interdisciplinary approaches. Further research should be made to produce better bias identification and reduction methods, especially for intersectional kinds of bias. It is necessary to reveal the metrics for fairness and create ethical frameworks that allow tracking of AI bias and utilization of AI that will positively affect the lives of all people.

REFERENCES

1. Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... & Zhang, Z. (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv preprint arXiv:1810.01943.
2. Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. In Advances in Neural Information Processing Systems (pp. 3995-4004).
3. Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 259-268).
4. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In Advances in Neural Information Processing Systems (pp. 3315-3323).
5. Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems, 33(1), 1-33.
6. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. In International Conference on Machine Learning (pp. 325-333).
7. Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 335-340).

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details