Unpacking Machine Learning Bias: Sources, Mitigation Techniques and Ethical Implications

Garvit jain, India, jaingarvit352@gamil.com

Abstract - This research paper aims to analyze and mitigate the bias in machine learning models by exploring the different aspects of sources of bias and ethical implications. The paper reviews existing research on the fairness definitions, and techniques to mitigate bias. The authors propose a basic approach that addresses bias from the source and implications of bias. The paper hypothesizes that proper preprocessing techniques, hyperparameters tuning, and optimal model selection can reduce bias in machine learning models. Additionally, the paper proposes that the sources of bias in machine learning models can include biased training data, flawed algorithms, and the lack of diversity in the development teams. The methodology used to achieve the objectives of the study includes analyzing existing literature, conducting experiments, and developing interactive plots to visualize the bias. The research questions addressed in this paper include the impacts of machine learning bias on society, the effectiveness of pre-processing, hyperparameters, and model selection in mitigating bias, and the main sources of bias in machine learning models.

Keywords — Machine learning bias, Mitigation techniques, Ethical implications, Pre-processing approach, Model selection, Sources of bias

I. INTRODUCTION

Machine learning and Artificial intelligence have advanced their presence to the point where they can produce poetry, art, and music, and even hold conversations like humans. However, one of the major challenges in this is the issue of bias in the data available. For machine learning models to learn and make decisions, they need to be fed on data, which is based on patterns of previous human decisions. That explains why a machine learning model can be biased. Several papers exploring different aspects of this problem. While previous research has examined types of bias, sources, and ways to mitigate it, none has addressed all these aspects together. Paper aims to provide a comprehensive analysis of the Sources and implications of bias in machine learning. Paper also proposes to focus on mitigating bias from the source by examining the main sources of bias in machine learning models, such as biased training data, flawed algorithms, and lack of diversity in development teams. Additionally, In this paper I hypothesize that proper preprocessing techniques, hyperparameters tuning, and optimal model selection can significantly reduce bias in machine learning models. The main research questions addressed in this paper are the impacts of machine learning bias on society, the effectiveness of pre-processing, hyperparameters, and model selection in mitigating bias, and the main sources of bias in machine learning models.

II. RELATED WORK

There are several papers that have explored the different aspects of Machine learning bias, its types, sources and how to mitigate bias and ethical implications, but not any talks about all of them together.

Hellström et al [1] explores the different meanings and contexts of bias in machine learning, both in public media and scientific publications. It proposes a taxonomy of bias based on four dimensions: source, type, effect, and value. The authors argue that bias is not always negative or undesirable, but can have positive or useful implications depending on the situation and the goal of machine learning. They suggest that bias should be evaluated and communicated in relation to its source, type, effect, and value, rather than being treated as a generic term. The paper aims to provide a common framework and terminology for discussing and addressing bias in machine learning.

The Yapo et al [3] paper examines the issue of bias in machine learning models that can impact a range of fields, including social media, healthcare, education, and criminal justice. The study investigates the sources and types of bias in machine learning, such as data, algorithm, human, and societal bias, and how they can result in ethical concerns and dilemmas. The paper proposes two frameworks, the issue management process (IMP) and the ethical decision making process (EDMP), for managing bias in machine learning. The research applies the IMP and EDMP frameworks to three case studies, namely Facebook's news feed algorithm, the COMPAS risk assessment tool for criminal sentencing,

 DOI : 10.35291/2454-9150.2023.0102

and Google's facial recognition software. The paper concludes that bias in machine learning can be significantly reduced through deliberate design, development, testing, and monitoring of algorithms, as well as by engaging diverse stakeholders and ensuring transparency and accountability.

Saxena et al [4] in their paper explores the alignment between various definitions of fairness in algorithmic decision-making and public perceptions and preferences of fairness. A large-scale online survey with over 3,000 participants from 50 US states and 10 European countries is conducted to rate the fairness of different scenarios involving algorithmic decisions on loan approval, hiring, and criminal justice. The participants' ratings are compared with four commonly used algorithmic definitions of fairness: anti-classification, classification parity, calibration, and individual fairness. The paper reveals that none of the existing definitions of fairness can fully capture the public's contextual and nuanced views of fairness, and that there are significant variations in fairness perceptions among different demographic groups and regions. The paper suggests that further research is necessary to comprehend and incorporate public values and expectations into the design and evaluation of fair algorithms.

The paper by Feldman, T. (2021)[5] focuses on gender bias in deep learning models across various domains, including healthcare, education, social media, and criminal justice. It surveys existing methods to address gender bias in machine learning, including pre-processing, in-processing, and post-processing techniques, outlining their advantages and limitations. The paper introduces a novel approach called end-to-end bias mitigation, which combines these methods to exploit their strengths and mitigate their weaknesses. The paper applies the end-to-end bias mitigation method to a deep neural network trained on a gender-biased dataset of movie reviews, comparing its performance with the baseline methods using various fairness metrics. The paper finds that the end-to-end bias mitigation approach improves fairness and reduces gender bias more effectively than baseline methods, while maintaining high accuracy and performance.

The Gat at el[6] paper investigates bias in multi-modal classifiers and its potential to favor certain data modalities, leading to unfair outcomes. A novel regularization term is introduced, based on functional entropy that measures the uncertainty of a classifier's output given a modality. The paper proposes a method to maximize the functional entropy of each modality to balance their contributions and reduce bias, using the log-Sobolev inequality and the functional Fisher information. The proposed method is evaluated on three multi-modal datasets: VQA-CPv2, SocialIQ, and Colored MNIST, and compared to several baselines using different fairness metrics. Results show that the proposed method achieves state-of-the-art results in mitigating bias and improving fairness while maintaining high accuracy and performance.

This paper by Chakraborty et al[7] addresses the issue of bias in machine learning systems that can negatively impact certain social groups, including those defined by factors such as race, sex, age, and marital status, in various areas such as criminal justice, credit card approvals, and hiring decisions. To tackle this problem, the authors propose a new algorithm called Fair-SMOTE, which eliminates biased labels and rebalances data distributions based on sensitive attributes to improve fairness. The effectiveness of Fair-SMOTE is evaluated on 15 datasets and 6 learners and compared to other state-of-the-art fairness improvement algorithms. The study finds that Fair-SMOTE is as effective as previous approaches in reducing bias while achieving higher performance, measured in terms of recall and F1 score. The authors claim that this is one of the most comprehensive studies on bias mitigation in machine learning to date.

The paper by Hu et al[8] examines gender bias in machine learning models that use data from human evaluators on an online micro-lending platform. A structural econometric model is developed to estimate the evaluators' preference-based bias and belief-based bias towards female applicants and to understand the decision dynamics. Counterfactual simulations are conducted to assess the impact of gender bias on loan granting outcomes, company profits, and borrower welfare. The paper also trains machine learning algorithms on both real-world data and simulated data to compare their decisions and investigate how evaluators' biases are inherited by the algorithms. The study shows that machine learning algorithms can mitigate both preference-based and belief-based biases and highlights the need for greater transparency and accountability in machine learning applications.

The Taniguchi et al[9] papers discuss a machine learning model that incorporates human cognitive biases to improve its ability to learn from small and biased datasets. The authors implemented a human cognitive model into machine learning algorithms and compared their performance with other popular methods such as naïve Bayes, support vector machine, neural networks, logistic regression and random forests. They focused on the task of spam classification and found that their models achieved superior performance with small and biased samples in comparison with other representative machine learning methods

The authors, Jindong Gu and Daniela Oelke [10], discuss the issue of bias in machine learning from a technical perspective and illustrate the impact that biased data can have on a machine learning model. They explain that bias can get into a machine learning model through the data that is used for building it. If the data or decisions taken on it are biased, then the machine learning model can incorporate this bias into its predictions. The authors also develop interactive plots to visualize the bias learned from synthetic data

The authors, verma et al [11], propose a black-box approach to identify and remove biased training data. They found that machine learning models trained on such debiased data (a subset of the original training data) have low individual

 DOI : 10.35291/2454-9150.2023.0102

discrimination, often 0%, and greater accuracy and lower statistical disparity than models trained on the full historical data. They evaluated their methodology using 6 real-world datasets and found that their approach outperformed seven previous approaches in terms of individual discrimination and accuracy.

Our work is complementary to the work cited above, by focusing mitigation of bias from the source and also providing a good analysis of the definition, origin and implications of bias.

III. SOURCES OF BIAS IN MACHINE LEARNING MODELS

Biased Training Data

One of the most significant challenges in building machine learning models is to ensure that the training data is not biased. When training data is biased, the model is likely to be biased as well, leading to inaccurate or unfair predictions. This is especially true in the case of home loan approval datasets, where biases can result in discrimination against certain groups, such as minorities or low-applicant-income individuals. For example, suppose a home loan approval dataset contains information only about individuals who have a high credit score and a steady income. In that case, the model may learn to associate these characteristics with a higher likelihood of loan approval, leading to biases against individuals who do not meet these criteria. Similarly, if the training data contains more data about homeowners in one area than in another, the model may learn to favor individuals from that area, leading to geographic bias.

Limited or Inappropriate Features

Inadequate or inappropriate features can also lead to bias in machine learning models, including those designed for home loan approval. For instance, suppose a home loan approval dataset only contains information about the borrower's income, credit score, and employment history. In that case, the model may not consider other important factors such as the borrower's debt-to-income ratio or their history of paying rent on time. This can lead to bias against individuals who have a high debt-to-income ratio or those who have rented rather than owned a home in the past.

Unintentional Algorithmic Bias

Another way in which bias can creep into machine learning models is through the algorithm's design. For example, an algorithm designed to optimize loan approval rates may inadvertently favor individuals with certain characteristics, such as a high credit score, while discriminating against others, such as individuals with a low credit score or those who have a non-traditional employment history. This can lead to bias in the model's predictions and create a feedback loop that reinforces the initial bias.

Human Bias

Human bias can also play a significant role in the development of machine learning models, including those used for home loan approval. Biases can be introduced at several stages of the model development process, such as during the selection of training data, the choice of features, or the algorithm's design. For example, if the individuals who selected the training data have an implicit bias towards certain groups, this can lead to a biased dataset and, consequently, a biased model.

Concept Drift

Concept drift can also lead to bias in machine learning models designed for home loan approval. Concept drift occurs when the distribution of the data changes over time, leading to a model that becomes increasingly inaccurate as new data becomes available. For example, if a home loan approval dataset contains information about borrowers' credit scores over time, and the distribution of credit scores changes significantly, the model may become biased towards certain credit score ranges, leading to inaccurate predictions.

IV. BIAS MITIGATION

Data and Analyses

The secondary dataset I utilized pertains to home loan applications and includes customer information provided on the online application form, such as Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History, and other details. The dataset contained a total of 614 customer records, out of which 422 loan applications were approved, resulting in an approval rate of approximately 68.73%. The remaining 192 applications were rejected by the home loan company. The dataset showed that the average income of applicants was $5403.45, while the average loan amount was $1621.24. I also observed negative correlation between loan approval and gender and marital status for female and unmarried applicants, respectively, although the correlation was very low. Conversely, there was a positive but very low correlation between loan approval and male and married applicants, indicating potential preference or bias towards certain applicant attributes by the home loan company. Additionally, there were instances of missing data in the dataset.

Model Selection

I opted to use the most common supervised machine learning models, namely KNN, SVM, Logistic Regression, and Random Forest, to address the problem. For each model, I utilized the optimal parameter values to determine which model and parameters performed the best. I employed GridSearchCV, a library function in the model_selection package of sklearn, to identify the optimal parameter from a list of hyperparameters, with a cross-validation of 5. According to the results, Logistic Regression and Random Forest were the best performing models, with mean cross-validated scores of the best_score at 0.8083 and 0.7875, respectively. The results are presented in Table 1.

 DOI : 10.35291/2454-9150.2023.0102

	mode	best_score	best_params
1	knn	0.60625	{{'algorithm': 'auto', 'weights': 'distance'}}
2	svm	0.7729166667	{'C': 20, 'kernel': 'linear'}
3	random_forest	0.7875	{'class_weight': 'balanced', 'criterion': 'log_loss', 'max_features': 'sqrt', 'n_estimators': 5}
4	logistic_regression	0.8083333333	{'C': 1, 'multi_class': 'auto', 'solver': 'lbfgs'}

Table 1

The table-1 presents information on four algorithms utilized for selecting the best model. It consists of three key columns: "Mode" which denotes the algorithm name, "Best Score" indicating the highest score achieved through cross-validation, and "Best Params" which provides details about the optimal parameters utilized in each algorithm.

FairML

FairML, an end-to-end toolbox for auditing predictive models by quantifying the relative significance of the model's inputs. Created by MIT. FairML leverages model compression and four input ranking algorithms to quantify a model's relative predictive dependence on its inputs. The relative significance of the inputs to a predictive model can then be used to assess the fairness (or discriminatory extent) of such a model. With FairML, analysts can more easily audit cumbersome predictive models that are difficult to interpret.s of black-box algorithms and corresponding input data.

Effectiveness check

I executed the top two performing models twice - once with minimal preprocessing and hyperparameter tuning, and another with thorough preprocessing and hyperparameter tuning - and evaluated the significance of input features in each using FairML. The model with minimal preprocessing and hyperparameter tuning is referred to as M1, while the model with thorough preprocessing and hyperparameter tuning is M2. In M1, I carried out basic/no preprocessing, which entailed removing missing values, and did not provide any hyperparameters to the model in both cases. In contrast, M2 involved proper preprocessing, such as removing missing values, converting strings to numbers, and eliminating outliers, and utilized the best hyperparameters for the corresponding model obtained from gridsearchcv (refer to Table 1). Subsequently, I assessed the results in FairML to determine the relative importance of attributes to each model.

Modifications

The outcomes of FairML reveal that M2 outperforms M1 in logistic regression, although the results were somewhat less impressive for random forest. Therefore, I made some modifications to the models, such as eliminating irrelevant attributes like gender, which has limited significance in the home loan application context. As a result, the models exhibited better performance than before and generated consistent results in all cases.

V. FINAL RESULTS

The logistic regression M1 model demonstrated better performance than either M1 or M2 of random forest, with M2 of logistic regression being left behind due to inconsistent attribute results. In Fig 1, loan amount was found to be more significant than applicant income in the M1 of logistic regression, while self-employment status had little importance for the model. In contrast, the M2 of logistic regression performed better than any other model and delivered consistent results, although it required more preprocessing and hyperparameter tuning.

In comparison to any other model (both M1 and M2), the M1 of random forest assigned more importance to each attribute. However, the M2 of random forest improved and reduced its significance in some attributes, increasing the importance of ApplicantIncome even more than credit history, which is given the highest priority in any loan application. This indicates that the random forest model, whether M1 or M2, does not perform well, even though it is the second-best model for this dataset.

pltalgo	Dependents	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Married_Yes	Education_Not Graduate	Self_Employed_Yes	Property_Area_Semiurban	Property_Area_Urban
lg_withoutprocessing	0.05208333333	0.05625	0.02291666667	0.1041666667	0.1854166667	0.3708333333	0.1166666667	0.025	0.02083333333	0.05208333333	0.01875
lg_withprocessing	0.08541666667	0.17291666667	0.09166666667	0.175	0.26875	0.425	0.1708333333	0.02916666667	0.02916666667	0.07291666667	0.03958333333
rf_withoutprocessing	0.23125	0.42083333333	0.18333333333	0.5520833333	0.65	0.6770833333	0.4729166667	0.12708333333	0.08958333333	0.30833333333	0.20625
rf_withprocessing	0.17708333333	0.45	0.20416666667	0.4916666667	0.475	0.41875	0.2979166667	0.07916666667	0.06666666667	0.2625	0.18125

Table 2

Table 2 showcases the importance assigned by FairML, with "pltaglo" indicating the presence or absence of data preprocessing in the algorithms. The first column displays the attributes of the dataset.

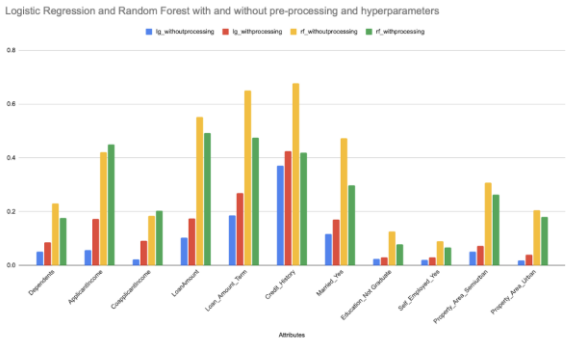

Fig 1

Verdict

Based on the results obtained, I can conclude that selecting the best model, performing proper preprocessing, and tuning the hyperparameters can significantly impact the performance of the model and mitigate the bias to some extent. This suggests that careful attention must be given to these steps in the machine learning pipeline to ensure the model is unbiased and delivers accurate results.

 DOI : 10.35291/2454-9150.2023.0102

VI. IMPLICATIONS OF SOCIETY

The implications of a biased machine learning model on society can be quite significant, as such models have the potential to perpetuate and exacerbate existing social inequalities. Here are some potential implications:

Discrimination

Biased models can unfairly discriminate against certain groups of people, such as minorities or women, by inaccurately predicting their behavior or characteristics. This can lead to situations where people are denied opportunities or access to resources, solely based on their race, gender, or other personal characteristics.

Reinforcing existing biases

Biased models can reinforce existing biases in society, such as stereotypes about certain groups being less intelligent, less reliable, or less likely to succeed. This can perpetuate systemic inequalities and make it harder for disadvantaged groups to overcome social and economic barriers.

Loss of trust

If people believe that the decision-making process is biased or unfair, they may lose trust in the system as a whole. This can lead to a decrease in participation and investment, which can have broader economic and social implications.

Limited access to resources

Biased models can limit access to resources such as education, healthcare, and employment opportunities for certain groups of people. This can create a cycle of poverty and disadvantage that is difficult to break.

Lack of accountability

Biased models can make it difficult to hold decision-makers accountable for their actions. This can create a situation where decisions are made without transparency or oversight, which can lead to abuses of power.

VII. CONCLUSION

In conclusion, this research paper has provided a comprehensive analysis of bias in machine learning models, addressing its sources, ethical implications, and potential mitigation strategies. The paper's unique contributions lie in its holistic examination of bias, its focus on mitigating bias from the source, and its hypothesis on the effectiveness of preprocessing techniques, hyperparameters tuning, and model selection. The theoretical and managerial implications of this research highlight the significance of addressing bias in machine learning systems. However, the limitations of the research underscore the need for further investigations in alternative techniques, standardized metrics, and ethical considerations. Future research in these directions would contribute to advancing the understanding and mitigation of bias in machine learning, ultimately fostering more equitable and fair AI systems.

VIII. REFERENCES

[1] Hellström, T., Dignum, V., & Bensch, S. (2020). Bias in machine learning - what is it good for? European Conference on Artificial Intelligence, 3–10. Retrieved from http://ceur-ws.org/Vol-2659/hellstrom.pdf

[2] Wang, R., Chaudhari, P., & Davatzikos, C. (2023). Bias in machine learning models can be significantly mitigated by careful training: Evidence from neuroimaging studies. Proceedings of the National Academy of Sciences of the United States of America, 120(6). https://doi.org/10.1073/pnas.2211613120

[3] Yapo, A., & Weiss, J. (2018). Ethical Implications of Bias in Machine Learning. Proceedings of the . . . Annual Hawaii International Conference on System Sciences. https://doi.org/10.24251/hicss.2018.668

[4] Saxena, N., Huang, K. E., DeFilippis, E., Radanovic, G., Parkes, D. C., & Liu, Y. (2019). How Do Fairness Definitions Fare? National Conference on Artificial Intelligence. https://doi.org/10.1145/3306618.3314248

[5] Feldman, T. (2021). End-To-End Bias Mitigation: Removing Gender Bias in Deep Learning. arXiv.org. https://doi.org/10.48550/arXiv.2104.02532

[6] Gat, I., Schwartz, I., Schwing, A. G., & Hazan, T. (2020). Removing Bias in Multi-modal Classifiers: Regularization by Maximizing Functional Entropies. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2010.10802

[7] Chakraborty, J., Majumder, S., & Menzies, T. (2021). Bias in machine learning software: why? how? what to do? arXiv (Cornell University). https://doi.org/10.1145/3468264.3468537

[8] Hu, X., Huang, Y., Li, B., & Lu, T. (2022). Uncovering the Source of Machine Bias. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2201.03092

[9] Taniguchi, H., Sato, H., & Shirakawa, T. (2018). A machine learning model with human cognitive biases capable of learning from small and biased datasets. Scientific Reports, 8(1). https://doi.org/10.1038/s41598-018-25679-z

[10] Gu, J., & Oelke, D. (2019). Understanding Bias in Machine Learning. arXiv: Learning. Retrieved from https://arxiv.org/pdf/1909.01866

[11] Verma, S., Ernst, M., & Just, R. (2021). Removing biased data to improve fairness and accuracy. ArXiv, abs/2102.03054.

[12] Saxena, N., Huang, K. E., DeFilippis, E., Radanovic, G., Parkes, D. C., & Liu, Y. (2019). How Do Fairness Definitions Fare? National Conference on Artificial Intelligence. https://doi.org/10.1145/3306618.3314248

[13] O'Reilly-Shah, V. N., Gentry, K. R., Walters, A., Zivot, J. B., Anderson, C., & Tighe, P. J. (2020). Bias and ethical considerations in machine learning and the automation of perioperative risk assessment. BJA: British Journal of Anaesthesia, 125(6), 843–846. https://doi.org/10.1016/j.bja.2020.07.040

[14] Pfohl, S., Foryciarz, A., & Shah, N. H. (2021). An empirical characterization of fair machine learning for clinical risk prediction. Journal of Biomedical Informatics, 113, 103621. https://doi.org/10.1016/j.jbi.2020.103621

 DOI : 10.35291/2454-9150.2023.0102