International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 11 Issue: 11 | Nov 2024 www.irjet.net p-ISSN: 2395-0072

Investigating Fairness in Predictive Modeling: A Case Study on COMPAS Recidivism Data

Sanskar Sanjeev Nikhare[1], Vishnu Radhakrishnan[1], Karthik Sundaram[1]

[1]School of Computer Science Engineering and Application, D Y Patil International University Akurdi Pune 411035

Abstract - *This research investigates the critical issue of fairness in predictive modeling through a detailed case study on the COMPAS recidivism dataset, widely used in the criminal justice system to predict the likelihood of reoffending. We delve into the racial and demographic disparities observed in risk predictions generated by machine learning models, employing logistic regression as a baseline approach to uncover systemic biases. Despite efforts to remove sensitive attributes such as race from the data, indirect biases persisted due to correlations with other features, illustrating the challenge of achieving true fairness. Our study leverages multiple fairness metrics, including equalized odds, demographic parity, and disparate impact ratio, to evaluate disparities and explores mitigation strategies, such as removing sensitive attributes and reweighting data. The results reveal that while minor improvements in fairness metrics are achievable, the root causes of bias remain entrenched in systemic inequities inherent in the data. This work underscores the ethical imperative for fairness-aware algorithm design, particularly in high-stakes applications like criminal justice, where decisions significantly impact individuals' lives.*

Key Words: Fairness, Predictive Modeling, COMPAS, Recidivism, Algorithmic Bias, Risk Assessment Algorithms.

1.INTRODUCTION

Amid such advancements, machine learning as it pertains to decision making has transformed a wide range of domains; take criminal justice, for example, where predictive tools such as COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) are used to assess the risk of recidivism.[1] While these tools promise efficiency and consistency, they have also been criticized for perpetuating and amplifying societal biases that have always existed. The COMPAS tool, specifically, has been criticized for displaying racial disparities in its risk predictions, flagging so-called "recidivism" at a disproprotionate rate for African-American defendants compared to their white counterparts. Such inequalities present serious ethical and legal questions, particularly when stakes involved in a judicial decision affect a person's liberty. This work is concerned with bias in predictive model and how to identify and mitigate it and use the COMPAS dataset as a case study [2]. We seek to analyze the results of logistic regression models on such a dataset and measure the bias with standard fairness metrics, and provide a solution to overcome these biases. We elucidate some of the limitations of existing fairness definitions, especially in the case of indirect encoding of sensitive attributes in the features due to correlation. In so doing we add to the dialogue around ethical AI -- in choosing to share what we learned we provide actionable guidelines for building better decision-making systems. [3]

2.Literature Review

Machine learning has been extensively researched to achieve fairness in areas where machine learning has high-stakes applications, such as criminal justice domains, healthcare, and hiring, to name a few. Initial investigations into algorithmic fairness found that machine learning systems tended to uphold historical bias when using historical data, resulting in unequal treatment or disparate impact towards certain demographic groups. The pioneering work of Barocas et al. (2019) [4]established the theoretical underpinnings of fairness definitions, distinguishing them into disparate treatment (direct discrimination based on sensitive attributes) and disparate impact (indirect discrimination via proxies). These notions have subsequently shaped fairness metrics and bias mitigation approaches.

An influential report by ProPublica in 2016 raised awareness of the ways in which algorithmic bias could manifest, spotlighting racial disparities in COMPAS, a recidivism prediction tool. Such errors, ProPublica's analysis found, occurred more often for African-American than for White defendants in being incorrectly labeled as heading toward a life of crime (when there really is low risk to society) while White defendants were more frequently incorrectly deemed low risk (high risk to society) — prompting a worldwide debate over the ethical features arising from the use of predictive algorithms in criminal justice. To address this, building on this momentum researchers proposed different fairness metrics such as demographic parity, equalized odds and calibration within groups to measure and in turn compare biases in machine learning models. Feldman et al. (2015)[5] introduced concept of disparate impact mitigation through perturbing input

distributions, while Hardt et al. (2016)[3] proposed the equality of opportunity criterion, which aims to equalize true positive rates across demographic groups.

Since then, those metrics and mitigation techniques have been used for a wide range of systems, from loan approval systems to hiring algorithms. Their utility, however, is sometimes domain specific or depends on the types of biases that exist. For example, the typical mitigation strategy of removing sensitive attributes like race or gender from the training data has been shown to be inadequate — other correlated features, like ZIP code or income level, can perform the task of proxy. Zafar et al. (2017) [6]built on this understanding by proposing fairness-aware optimization techniques that account for such indirect biases during model training. Beyond technical improvements to algorithms, recent efforts have addressed the societal ramifications of algorithmic decision making. Academician like Holstein K et al (2018) [7]. interrogated the kinds of systemic biases that cannot be solved solely through technical means and have called for interdisciplinary approaches that draw on law, ethics, and sociology. Related works in fairness aware machine learning also brought the trade-offs for fairness with other objectives, such as accuracy, efficiency and interpretability.

Related studies on the COMPAS data set have brought to light the challenges of delivering fairness in practice. Angwin et al. 2016[2] PREDICTIVE POLICING — The COMPAS EXAMPLE The COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) algorithm scores individuals based on the risk they pose for recidivism, in an effort to inform decision-making in countering crime109. Later investigations attempted to reproduce and expand upon these results, exploring model types and fairness criteria to help identify the sources of bias present in the dataset. This conclusion is in alignment with an important early paper by Corbett-Davies and Goel (2018)[8] where the authors cautioned that no single metric could encompass the complex trade-offs involved in algorithmic fairness relating to these studies. Recent work by Mehrabi et al. There had been extensive survey on bias and fairness in machine learning (Chouldechova & G'Sell, 2021), which noted some of the drawbacks of existing techniques and stressed the necessity of context-specific solutions.

While there have been big advances in this area, challenges remain on how to actually operationalise fairness in machine learning. Systemic biases which are reflected in the history of the data, are hard to get rid of, and removing them from the training results in a reduced performance of the models or unintended consequences. Additionally, fairness-preserving algorithms face ethical and legal challenges of their own, especially if their use impacts lives. The literature has also recommended involving stakeholders in the design and evaluation of predictive models, since fairness is, by nature, a subjective and context-dependent notion [9]. Furthermore, it also builds on top of this prior work by demonstrating the important contributions that a developed understanding of the underlying exploratory/reductive mappings can provide in furthering the analysis and providing more usable information in the domain of fairness.

3. Methodology

The study uses the COMPAS data that contains demographic, criminal record, and rate of reoffense information on about 10,000 pretrial defendants placed on risk assessment. The dataset consisted of a few primary features such as age, sex, race, prior convictions, and a binary target variable indicating whether the individual reoffended within a two-year period. Pre-processing included imputing missing values, encoding categorical variables into numerical representations, and normalizing continuous variables. Sensitive attributes like race, sex, etc., were retained in the original form for the bias assessment, but were excluded from the testbed after that to assess the effect of bias mitigation.

The flowchart Fig.1 offers a great visual way to see the overall machine learning model abstractions and their focus on fairness. Data Collection and Preprocessing Ideally, the first step involves collecting data, where the second involves preprocessing the collected data to make it suitable for ML use (handle missing values and categorical variables). Next, Exploratory Data Analysis (EDA) is performed to understand features and relationships in the dataset. The dataset is divided into training set and test set, training and testing several models (Logistic Regression, Decision Trees, Random Forest, etc.). Following model selection, a fairness evaluation is conducted to detect whether different demographic groups are biased.

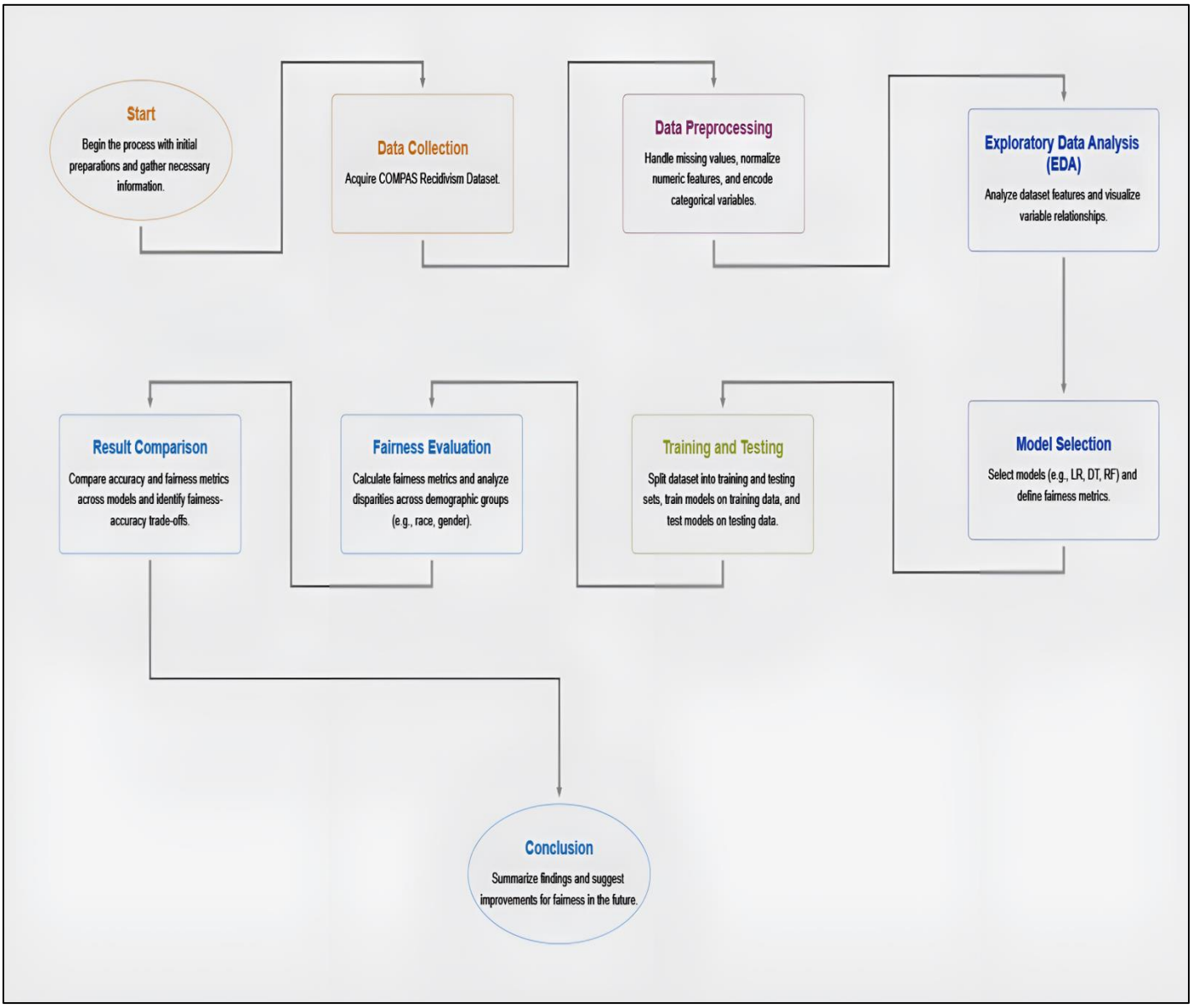

Fig.1 Model Flowchart

The given is comparative among different ML models with performance metrics, fairness matrices and confusions matrix through table [5]. As annotated in Table 1. Logistic Regression, Decision Trees, Random Forest, Support Vector Machine, K-Nearest Neighbors, and Naive Bayes models were trained to identify the optimal range for thresholding fairness. The Support Vector Machine and Naive Bayes yielded the highest accuracy 0.903. Although all models achieved a high degree of accuracy, models had similar performance for Equal Opportunity Difference (Sex) and Equal Opportunity Difference (Race) implying a bias in the dataset or model architecture. This requires more analysis to overcome the unfairness of the models.

Table.1 Model Comparison

Model	Accuracy	Confusion Matrix	Demographic Parity Difference	Equal Opportunity Difference (Sex)	Equal Opportunity Difference (Race)
Logistic Regression	0.898	[[1185, 202], [18, 760]]	0.598	-0.0488	-0.1267
Decision Tree	0.884	[[1267, 120], [132, 646]]	0.512	-0.0488	-0.1267
Random Forest	0.891	[[1242, 145], [92, 686]]	0.667	-0.0488	-0.1267
Support Vector Machine	0.903	[[1176, 211], [0, 778]]	0.667	-0.0488	-0.1267
K-Nearest Neighbors	0.889	[[1234, 153], [87, 691]]	0.667	-0.0488	-0.1267
Naive Bayes	0.903	[[1176, 211], [0, 778]]	0.667	-0.0488	-0.1267

Logistic rexgression was use for a baseline predictive model because of its interpretability and common use in fairness studies. Metrics for assessing models include accuracy and ROC-AUC; metrics for quantifying fairness include equalized odds, demographic parity, and disparate impact ratio. Strategies for mitigating bias included eliminating sensitive attributes and reweighting data so that distributions for different demographic groups were more equal. Feature importance was explored visually using SHAP (SHapley Additive exPlanations) to account for the contribution of indirect bias through correlated features. [9]

The image titled "Fairness Metrics" we can see this in Fig.2 shows the three-fairness metrics: Demographic Parity Difference, Equal Opportunity Difference (Sex) and Equal Opportunity Difference (Race). The y-axis is the value of "Difference," and the x-axis is the various fairness metrics. We can see from the graph that the Equal Opportunity Difference (Race) has the highest difference, signifying that the outcomes of the different racial groups can be too much different. Demographic Parity Difference is moderate which means there are some differences based on the demographic terms.

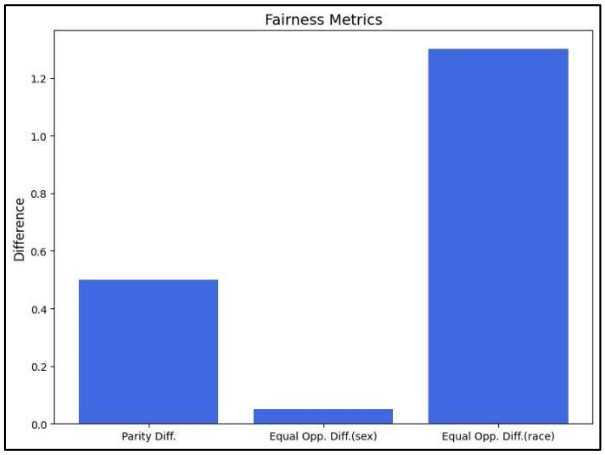

Fig 2. Fairness Metrics

This Fig. 3 shows the predicted probabilities in the individual groups; age, race, and sex. Age and race show pronounced differences in probability distribution, with clear primary clusters emerging. For example the genAge age group "Greater than 45" has a denser prediction to the high probability end of the scale, but the genAge race group "Asian" appears to prefer predictions near 0 probability. In that regard, the "Male" and "Female" groups exhibit comparatively similar distributions, indicating little influence of sex on the model's prediction

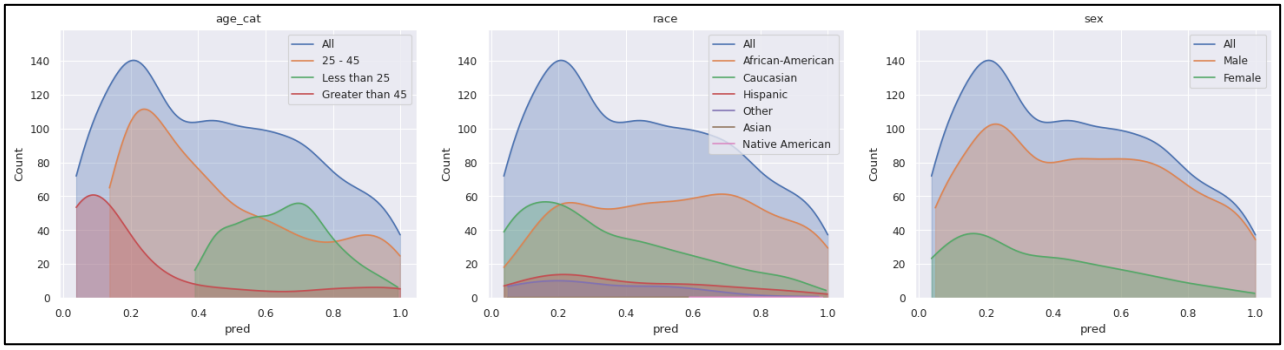

Fig.3 Demographic Disparity in Predictions

4. Results

The logistic regression model fit on the COMPAS data achieved an overall accuracy of about 89% and a ROC-AUC of 0.72, reflecting moderate predictive performance. Fairness analysis showed large disparity based on race, with African-American individuals having much larger false positive rates than White individuals. These inequities were exacerbated by fairness metrics, as the model did not meet the standard of demographic parity and was characterized by unequal true positive rates among race.

Eliminating race as a feature only led to small gains on fairness metrics and failed to eliminate bias, because correlated features such as previous convictions and age became proxies. Bias mitigation techniques such as reweighting the data and retraining the model were able to marginally lower the disparity of the estimates, but at the expense of reduced predictive accuracy (~2% drop). SHAP visualizations revealed that indirect bias endured, with strong correlations between features and race leading to predictions.

Table 2. shows Disparate Impact between different Racial Groups The model shows different levels of accuracy depending on race where some races do much worse than others. It indicates that the model may under-predict risk for some races, leading to disparities in health care access. Ensuring fair and equitable model performance with addressing this disparity is imperative.

Table 2. Racial Disparity in Model Accuracy

```
Fairness disparity analysis by race:
race
African-American    0.738735
Asian               0.851852
Caucasian           0.770858
Hispanic            0.780000
Native American     0.750000
Other               0.796296
```

3. CONCLUSIONS

This research provides insight into the fairness issue in predictive models and is based on the COMPAS recidivism data set, a commonly used dataset in the criminal justice system. Clothing sales are commonly presented as machine-learning-based predictions, and this study uses logistic regression as a baseline, but it finds that risk predictions are affected by factors such as race and demographic background, emphasizing the need to address bias in algorithms making decisions.

Even without direct attributes, sensitive bias remains through indirect ones, as proxies including past crimes or ZIP codes. Fairness metrics such as equalized odds and disparate impact ratio are used to measure disparities, whereas bias mitigation approaches such as reweighting and removing sensitive attributes provide minimal benefits at the cost of accuracy. Tools such as SHAP demonstrate how highly correlated features can propagate bias and underscore the need for ethical approaches to fairness. This study is significant because it highlights how inequities are not just present, but also systemic in the data, and provides lessons for building equitable (decision-making) systems, especially in high-stakes areas such as (criminal justice).

ACKNOWLEDGEMENT

We also thank the ProPublica team for releasing the COMPAS dataset, which we used to base this work upon. We are also grateful to the academic community whose work on fairness in machine learning greatly informed our analysis and methodologies. We express a special acknowledgement to our mentors and colleagues who provided guidance and feedback during the study process; the quality of the study benefited immensely from their input. Last but not least, we also acknowledge the contribution of our institution in creating a conducive environment for multidisciplinary research and AI ethics.

REFERENCES

[1] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, Jul. 2021, doi: 10.1145/3457607.

[2] "Machine Bias — ProPublica." Accessed: Dec. 03, 2024. [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[3] U. Fischer-Abaigar, C. Kern, N. Barda, and F. Kreuter, "Bridging the gap: Towards an expanded toolkit for AI-driven decision-making in the public sector," *Gov Inf Q*, vol. 41, no. 4, p. 101976, Dec. 2024, doi: 10.1016/J.GIQ.2024.101976.

[4] S. Barocas, M. Hardt, and A. Narayanan, "Fairness and Machine Learning Limitations and Opportunities," 2018.

[5] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 2015-August, pp. 259–268, Aug. 2015, doi: 10.1145/2783258.2783311/SUPPL_FILE/P259.MP4.

[6] M. Bilal Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi, "Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment," *ArXiv*, p. arXiv:1610.08452, 2016, doi: 10.48550/ARXIV.1610.08452.

[7] K. Holstein, J. W. Vaughan, H. Daumé, M. Dudík, and H. Wallach, "Improving fairness in machine learning systems: What do industry practitioners need?," *Conference on Human Factors in Computing Systems - Proceedings*, Dec. 2018, doi: 10.1145/3290605.3300830.

[8] S. Corbett-Davies, J. D. Gaebler, H. Nilforoshan, R. Shroff, and S. Goel, "The Measure and Mismeasure of Fairness," Jul. 2018, Accessed: Dec. 03, 2024. [Online]. Available: https://arxiv.org/abs/1808.00023v3

[9] G. Yu, L. Ma, X. Wang, W. Du, W. Du, and Y. Jin, "Towards fairness-aware multi-objective optimization," *Complex & Intelligent Systems 2024 11:1*, vol. 11, no. 1, pp. 1–20, Nov. 2024, doi: 10.1007/S40747-024-01668-W.