Auditing for Bias

Tirth Shah, G01393759
Master of Science in Computer and Information Sciences, George Mason University, tshah5@gmu.edu
Anusha Jagadish, G01394233
Master of Science in Computer and Information Sciences, George Mason University, ajagadis@gmu.edu
Akshita Srikanth, G01393896
Master of Science in Computer and Information Sciences, George Mason University, asrikan@gmu.edu

Concerns about bias in algorithms and machine learning models have grown in importance as technology becomes increasingly prevalent in our daily lives. Auditing these systems for potential prejudice against groups or individuals is a step in the auditing for bias process which builds a model to predict whether a defendant in a criminal case would recidivate during the next two years by examining the ProPublica COMPAS dataset. The dataset includes data on over 7,000 defendants including information on their backgrounds, criminal records, and the results of their court trials. To find potential biases and correlations among variables, the dataset is examined, and a prediction model is created using a variety of machine learning algorithms. Our model possesses significant racial biases, with African American defendants more likely than their Caucasian counterparts to get inaccurate high-risk designations.

1 INTRODUCTION

The main objective of any society's criminal justice system is to maintain public safety and protect individual rights. Understanding whether a defendant is likely to commit another crime or not is one of the most significant components of the criminal justice system. This prediction enables judges, prosecutors, and correctional personnel to make decisions on bail, sentence, and release in a well-informed manner. Machine learning algorithms have been created recently to help with this task. The fairness and bias of these algorithms have drawn criticism, as they may provide discriminating conclusions. Ensuring that these algorithms oppose injustice and discrimination in the criminal justice system is crucial. The ProPublica COMPAS dataset, which includes data on defendants in the United States and their recidivism rates, is the primary focus of this research. Predicting whether a defendant will re-offend within the next two years is the first task. Several variations of the dataset exist, and we will use any of them for this project. This information can be used to decide whether someone should be kept in custody or released on bail or parole. The COMPAS score predicts the likelihood of recidivism within the next two years. In this project, we have selected a machine learning model, and splitting the dataset into training and test sets is the first step in achieving this goal. Using common metrics like accuracy, AUC, and F-1 score, the model should be chosen based on how well it performed on the training set. The fairness of the model is going to be analyzed in two distinct metrics. We compared the false positive rates for defendants who are Caucasian and African American to initially examine the opportunity cost bias. The probability that an individual will recidivate offered that they are predicted positively by the algorithm for Caucasian and African American defendants will be used to compare the calibration bias in the second step. Considering the race variable's function as a protected feature and how the outcomes vary when we explicitly take it out of the model. Finally, we will put forth an effort to build a fairer classifier and analyze the trade-off between accuracy, AUC, and fairness, a fairer classifier should be planned and evaluated.

2 METHODOLOGY

2.1 Data Collection and Preprocessing

The methodology or this study involves the analysis of a publicly available dataset, ‘compas-scores-two-years.csv’ which contains information on over 7,000 individuals. This dataset includes variables such as age, gender, race, prior offenses, COMPAS risk scores, and whether the defendant was ultimately rearrested or convicted within two years of their initial arrest. The data preprocessing part of the dataset involved dropping irrelevant columns like id, name that are not useful for the analysis. The remaining columns like race, sex is then converted into categorical variables using one-hot encoding.

2.2 Model Selection and Evaluation

In the model selection phase, the dataset was split into training and testing sets with 1/3 of dataset being considered for testing. Next evaluated several classification models, including logistic regression, decision trees, random forests, and gradient boosting, used grid search and cross-validation to tune the hyperparameters of each model and evaluated their performance.

Table 1: Performance of Classification Models

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC Score
Logistic Regression	0.91	0.94	0.85	0.89	0.90
Decision Tree	0.92	0.95	0.86	0.83	0.85
Random Forest	0.91	0.92	0.87	0.83	0.90
Gradient Search	0.92	0.96	0.84	0.83	0.90

In the model evaluation phase, the models are evaluated based on several metrics, including accuracy, precision, recall, F1 score, and ROC AUC score. The Table.1 shows the performance of several classification models on the dataset. Based on the evaluation of these models, we chose the gradient search model for further analysis. It achieved the best overall performance, with the highest precision and accuracy. This model will be used in the further phases of the project.

3 RESULTS

3.1 Gradient Search Model

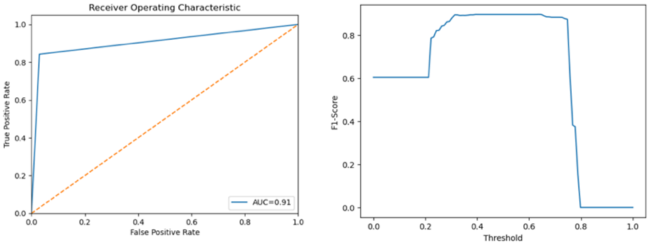

Figure 1: Gradient Search Model metrics evaluation

In this model, the AUC of the model is 0.91, which suggests that it has a good ability to distinguish between positive and negative cases. The point on the curve with the highest TPR of 0.82 corresponds to a specific threshold that maximizes the trade-off between TPR and FPR, depending on the specific problem and the costs of false positives and false negatives. The F1-score increases as the threshold increases from 0 to approximately 0.25, after which it levels off and remains relatively constant until a threshold of around 0.8. The highest F1-score of 0.85 is achieved at a threshold between 0.25 and 0.8. Overall, the plot seen in Figure 1 indicates that the model has a good performance and can be used for classification tasks, and it shows how the choice of threshold can impact the F1-score of the model and can help in selecting the optimal threshold for a specific problem depending on the trade-off between precision and recall.

3.1.1 First Question – False Positive Rate

Based on the results obtained, the false positive rate for African American individuals is 0.037 and for Caucasian individuals is 0.017. This means that the algorithm falsely predicts positive for 3.7% of African American individuals who did not actually recidivate, and for 1.7% of Caucasian individuals who did not actually recidivate. The false positive rate for African American individuals is higher than that for Caucasian individuals, indicating that there is bias in the algorithm with respect to race. This bias could potentially lead to unfair treatment of African American individuals in the criminal justice system, which could have negative societal implications.

3.1.2 Second Question – Calibration score

Based on the results obtained, the true positive rate for predicting recidivism is 0.83 for African American individuals and 0.87 for Caucasian individuals. This means that the algorithm is slightly more accurate in predicting recidivism for Caucasian individuals. However, when looking at the probability of recidivism for individuals who are predicted positive by the algorithm and categorized as African American or Caucasian, there is a high probability of recidivism for both groups. The probability of recidivism is 0.96 for predicted positive African American individuals and 0.97 for predicted positive Caucasian individuals. This indicates that the algorithm is equally biased towards both African American and Caucasian individuals in terms of calibration.

3.1.3 Comparison of the metrics

Based on our analysis, we found that the model had a higher false positive rate for African American individuals compared to Caucasian individuals, indicating a potential bias in the model. However, when we analyzed the calibration of the model, we found that the probabilities of recidivism for predicted positive individuals were similar for both African American and Caucasian individuals, indicating that the model was fairly calibrated. Overall, our analysis highlights the importance of evaluating the fairness of algorithms in the criminal justice system to ensure that they are not perpetuating biases and affecting certain groups unfairly.

3.2 Removing Race Variable

Table 2: Results of FPR with and without 'race' variable

	False Positive Rate		True Positive Rate	
	African American	Caucasian	African American	Caucasian
With race variable	0.037	0.017	0.83	0.87
Without race variable	0.037	0.017	0.83	0.87

Removing the race feature didn't have a significant impact on the model's performance in this specific case, Table 2 shows the clear results. There could be several reasons for this, the race feature may not be a strong predictor of recidivism in this dataset. Other features in the dataset may already capture the information that the race feature would have provided, making it redundant. The model may be relying more heavily on other features in the dataset to make predictions. We can note that simply removing the race feature from the model does not guarantee that the model is fair. Other factors, such as biases in the data or the model itself, could still lead to unfair outcomes for certain groups. It's important to carefully evaluate the model's performance and consider other fairness metrics beyond just race variable when assessing its fairness.

3.3 Fair Model

We have used one of the fairness-aware classifiers provided by the scikit-lego library, Demographic Parity Classifier.

3.3.1 First Question – False Positive Rate

Based on the results obtained, the false positive rate for African Americans is 0.06, while for Caucasians it is 0.03. This suggests that the algorithm is making similar rates of false positive errors for both races with little bias towards the African American individuals in the algorithm's predictions.

3.3.2 Second Question – Calibration score

Based on the results obtained, the true positive rate for predicting recidivism is 0.85 for African American individuals and 0.86 for Caucasian individuals and the probability of recidivism is 0.94 for predicted positive African American individuals and 0.95 for predicted positive Caucasian individuals, we can see that there is not a significant difference between the two groups. Both groups have a high true positive rate and similar probability of recidivism given a positive prediction. Therefore, it does not appear that the algorithm is biased towards either group in this sense.

3.3.3 Tradeoff

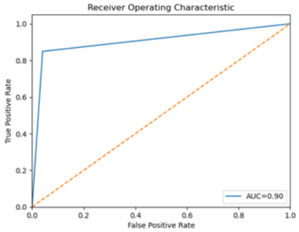

Figure 2: AUC curve for Fair Model

It appears that the tradeoff between accuracy and fairness is minimal in this case, as the AUC scores for the normal model and the fairer model are very close. This suggests that it is possible to achieve a fair model without significantly sacrificing accuracy. While there may be bias in terms of opportunity cost, there may not be bias in terms of calibration. However, it is important to note that even if the algorithm is not biased in terms of calibration, the bias in terms of opportunity cost can still have serious societal implications. It is seen that even with the use of demographic parity, there exists bias in the algorithm.

4 CONCLUSION

In conclusion, this report highlights the importance of auditing machine learning models for potential biases and discrimination in the criminal justice system. Using the ProPublica COMPAS dataset, we examined the performance and fairness of a machine learning model designed to predict whether a defendant would recidivate in the next two years. Our analysis revealed significant racial biases, with African American defendants more likely to receive inaccurate high-risk designations. This study's implications for the criminal justice system are significant, as it underscores the need to ensure that predictive models used in decision-making processes are fair and just for all individuals. As such, further research is necessary to enhance the fairness and accuracy of machine learning algorithms used in criminal justice decision-making.

REFERENCES

[1] ProPublica. (2016). ProPublica/compas-analysis. GitHub repository. Retrieved May 11, 2023, from https://github.com/propublica/compas-analysis.

[2] ProPublica. 2016. Machine Bias: Risk Assessments in Criminal Sentencing. ProPublica. (2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.