Auditing for Bias

PRATISH MASHANKAR, PRANAY SHIVENDRA SHARMA, and NISHAD VISHWANATH MAIN, George Mason University, USA

This report analyzes a dataset from ProPublica to predict recidivism risk. The dataset includes 7,214 individuals and their characteristics like age, sex, race, and recidivism status. We performed data preprocessing, feature engineering, selection, cleaning, imbalance handling, and scaling. Various models, including linear and non-linear classifiers, ensembles, and boosting algorithms, were tested. The highest accuracy of 69.37% was achieved with a gradient boosting machine. False positive rates were computed for African-American and Caucasian individuals, revealing bias towards the former. We also explored bias mitigation using a Diversity Parity fair classifier. These biases have societal implications, perpetuating systemic racism and discrimination.

CCS Concepts: • **Data Mining** → Classification, Clustering, Boosting; • **Machine Learning**; • **Bias analysis**;

Additional Key Words and Phrases: ensemble classifiers, ProPublica, COMPASS, diversity parity classifier

1 INTRODUCTION

The COMPAS algorithm is a widely used tool in the criminal justice system to predict the likelihood of a defendant reoffending within two years. However, concerns have been raised about its impartiality and accuracy, particularly in terms of racial bias. ProPublica published a report in 2016 that revealed significant racial bias in the COMPAS tool, with African American defendants being twice as likely as white defendants to be classified as high risk, despite having similar recidivism rates. This report aims to replicate ProPublica's analysis and examine the impartiality and accuracy of the COMPAS tool, as well as explore the potential implications of these biases in the criminal justice system. The report's findings suggest that while a more equitable algorithm could reduce bias, there is a trade-off between fairness and accuracy. The report highlights the need for more equitable algorithms in the criminal justice system to ensure that justice is served to all individuals equally.

2 METHODOLOGY

2.1 Data Visualization

To analyze the ProPublica dataset, we plotted bar graphs for sex ratio, age distribution, and recidivism rates. The dataset consisted of 5,819 males and 1,395 females. Age distribution showed 4,109 individuals in the 25-45 age group, 1,576 over 45 years old, and 1,529 under 25 years old. Recidivism rates revealed 3,743 non-recidivists and 3,471 recidivists, with a higher proportion of African Americans among the recidivists. A pie chart demonstrated the race versus recidivism rate: 58.7% African Americans recidivated as against 29.5% Caucasians, 7.1% Hispanics, 0.3% Asians and Native Americans, and 4.1% from other races. Figure 1(a)(b) for reference

Authors' address: Pratish Mashankar, pmashank@gmu.edu; Pranay Shivendra Sharma, psharm24@gmu.edu; Nishad Vishwanath Main, nmain2@gmu.edu, George Mason University, 4400 University Dr, Fairfax, Virginia, USA, 22030.

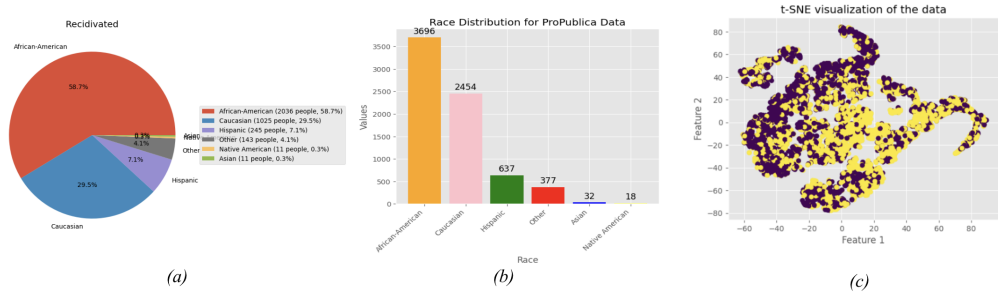

Fig. 1. Visualizations for ProPublica data: (a) Percentage breakdown of race among people who have recidivated., (b) Distribution of race, (c) t-SNE visualization of the data reduced to 2 components.

2.2 Data Preprocessing

To prepare the data for training, we created a new feature representing the number of days a defendant was jailed, by calculating the difference between the 'c_jail_out' and 'c_jail_in' columns and performed feature selection, choosing 10 relevant features correlated with the label. We dropped NaN values and rows with days before screening below 30 days, handled class imbalance using RandomUnderSampler, and scaled all features. Finally, we split the data with 1/3 for testing. We further split the training data reserving 20% for Validation. We utilized t-SNE for post-processing visualizations, reducing the data to two dimensions and creating a scatter plot. Our objective was to identify linearly separable data for classification purposes. Figure 1(c) for reference

2.3 Model Training

We aimed to predict whether a convict will recidivate using our preprocessed ProPublica dataset. We used accuracy as a metric to determine the best model for our classification task. The validation results are reflected in Table 1.

2.3.1 Linear Models. We tested logistic regression and support vector machine (SVM) as linear classifiers. Even though Logistic regression achieved satisfactory accuracy, both of their performance was poor due to the non-linear relationship between the features and the target.

2.3.2 Non Linear Models. Next, we experimented with non-linear classifiers such as K-Nearest Neighbors whose best value of 'K=24' was determined using 10-fold cross-validation. Its accuracy was better than Decision Tree probably indicating overfitting and poor generalization of the latter.

2.3.3 Ensemble Models and Boosting. We employed ensemble methods like Random Forest where we used 10-fold cross-validation to determine 180 estimators as optimal. AdaBoost, employing decision stumps as weak learners, optimized to 30 estimators via 5-fold cross-validation. Gradient Boosting Machine (GBM), using GridSearchCV for hyperparameter tuning, attained the highest accuracy of 69.54%. XGBoost, optimized using GridSearchCV and 10-fold cross-validation, yielded 68.44% accuracy. GBM outperformed XGBoost, potentially due to XGBoost's greater susceptibility to overfitting stemming from its numerous hyperparameters. The test accuracy for GBM was 69.37%

Table 1. Accuracy of Machine Learning Models

Type	Model	Accuracy
Linear	Logistic Regression	68.32%
Linear	Support Vector Machine	67.91%
Non-Linear	Decision Tree	67.27%
Non-Linear	k Nearest Neighbors	68.17%
Ensemble	Random Forest	67.02%
Ensemble	Adaboost	68.51%
Ensemble	Gradient Boosting Machine	69.54%
Ensemble	XGBoost	68.44%

Table 2. Opportunity cost vs Calibration vs Accuracy for Gradient Boosting Machine (GBM) Classifier

Race	**GBM Model**			**GBM Model without race**		
	OC bias	**Bias in Cal**	**Accuracy**	**OC bias**	**Bias in Cal**	**Accuracy**
AA	206/470=0.44	479/685=0.7	69.37%	199/470=0.42	413/550=0.75	69.52%
C	78/364=0.21	154/232=0.66		79/364=0.22	184/295=0.62	

Table 3. Opportunity cost vs Calibration vs Accuracy for the Diversity Parity and GBM Model

Race	**Fairclassifier Model**		
	OC bias	**Bias in Cal**	**Accuracy**
AA	137/470=0.29	413/550=0.75	68.62%
C	111/364=0.3	184/295=0.62	

3 IDENTIFYING BIAS

3.1 Bias in terms of opportunity cost

We calculated the algorithm's false positive probabilities for individuals who did not actually recidivate, based on their race (African-American or Caucasian). Table 2 presents the calculations along with the numerators and denominators. The results reveal a significant disparity between the false positive rates for African-American and Caucasian individuals, with the former being twice as high as the latter. This indicates bias in the algorithm towards African-American individuals. This bias carries severe societal implications, reinforcing systemic racism and discrimination. Unfairly labeling African-American individuals as high-risk can result in denial of opportunities such as employment or parole, perpetuating cycles of poverty and incarceration.

3.2 Bias in terms of calibration

We computed recidivism probabilities for individuals based on the algorithm's positive prediction and race (African-American or Caucasian). Table 2 shows the results used for these calculations. The small difference in probabilities between African-American and Caucasian individuals indicates no bias in the algorithm's calibration towards either group. However, these findings hold significant societal implications within the criminal justice system. In the case of COMPAS, using these probabilities to argue against bias towards African Americans has faced criticism from researchers and activists. They argue that these calculations present partial truths, potentially perpetuating existing biases and inequalities such as systemic racism and economic disparity.

3.3 Bias analysis

Both measures are essential for evaluating algorithm performance in the criminal justice system. The false positive rate identifies unjustified harsh sentencing, while the probability of recidivism given a positive prediction measures accuracy in identifying high-risk individuals. In this domain, the false positive rate is particularly relevant as it uncovers potential biases. A significantly higher false positive rate for African-Americans suggests bias in the algorithm's sentencing decisions. Conversely, in finance domain, false positives deny credit to qualified borrowers. Here, focusing on the probability of recidivism (loan default in context of finance) is crucial. Maximizing accuracy in this measure improves credit decisions and enhances financial stability.

4 AFFECT OF RACE VARIABLE

Removing the race variable did not significantly impact the algorithm's accuracy or probabilities as shown in Table 2. However, this raised concerns about potential indirect correlations with race and other dataset features, perpetuating inherent biases. Avoiding direct use of protected features doesn't always ensure unbiased results due to historical discrimination and systemic biases. These biases can manifest in data collection, labeling, model development, and result interpretation. For instance, historical discrimination against African Americans may have influenced over-representation in certain categories, like higher recidivism rates, which could indirectly correlate with race in the algorithm's data.

5 USING A FAIR CLASSIFIER

We used demographic parity as a fairness metric to evaluate a fair classifier's performance. The results in Table 3 reflect the numerators and denominators used to calculate the probabilities. This ensures that the classifier's decisions are independent of protected attributes, like race. After running the experiment, we found that the fair classifier achieved a more equitable distribution of false positives between African-American and Caucasian individuals, but resulted in a tradeoff between fairness and accuracy. The fair classifier reduced the difference between false positive rates for African-American and Caucasian individuals, but more improvement is needed to address biases in the original model.

6 CONCLUSION

In conclusion, our analysis of the ProPublica dataset revealed that while machine learning models can effectively predict the risk of recidivism, they can also perpetuate systemic racism and discrimination. We experimented with various models, achieved the highest accuracy with Gradient Boosting Machine, and identified bias towards African-American individuals. We also conducted additional experiments to understand the impact of removing the race variable and using a fair classifier, which resulted in improved fairness but slightly lower accuracy. Our findings emphasize the importance of ethical and responsible machine learning practices to mitigate bias and ensure fairness in algorithmic decision-making. Moving forward, it is crucial to continue research and development in this area to create more inclusive and equitable AI systems that serve all members of society.

7 REFERENCES

- https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal
- https://www.washingtonpost.com/news/can-an-algorithm-be-racist
- https://www.researchgate.net/publication/335716566_Comparative_Analysis_of_Linear_Non_Linear_and_Ensemble_Machine_Learning