# FAIRNESS IN RECIDIVISM PREDICTION MODELS

Nimitkumar Jignesh Jogani (G01379296),
Dipak Falgun Meher (G01382367),
Prommy Sultana Hossain (G01373197),
*Computer Science, George Mason University, Fairfax, Virginia, USA*

## 1 PROBLEM STATEMENT

The objective of this project is to investigate the fairness of a machine learning algorithm used to predict recidivism in the criminal justice system. We utilize the ProPublica COMPAS dataset to build a model that predicts the likelihood of a defendant recidivating within the next two years. The performance of the model is evaluated using standard measures such as AUC, accuracy, precision, recall, and F-1 score. [*Page number 1*]

The project examines the fairness of the algorithm in two ways: opportunity cost and calibration. To determine opportunity cost, we calculate the false positive rates for African-American and Caucasian defendants who did not recidivate. In contrast, calibration is measured by comparing the probability of actual recidivism given a positive prediction for both racial groups.[*Page number 2*]

Additionally, we analyze the role of the "race" variable as a protected feature in the model. We examine how the results change when the feature is explicitly removed or included in the model.[*Page number 3*]

Finally, we compare the performance of our model with a fair classifier designed to achieve demographic parity or equal opportunity, depending on the desired fairness objective.[*Page number 3 & 4*]

## 2 METHODOLOGY

The ProPublica COMPAS dataset is utilized for predicting recidivism. We analyzed the dataset and found that it has 52 columns where many of which are irrelevant to predicting if a person would recidivate in the following two years. Categorical data is labeled and encoded to numeric values. The correlation matrix is used to identify the relationship between the different features in the dataset. There are multiple target variables in the dataset like is_recid, events and two_year_recid, and we have chosen two_year_recid as the target variable for our analysis, eliminating others because they provide information that wouldn't be available at the time of the COMPAS assessment. Then we further analyzed the features, normalized them using Min Max Scalar, and used SelectKBest function and chi-squared test as the score function for feature selection. We dropped the irrelevant columns with low scores of the k-best function that includes description columns, date columns, and columns describing individual information. The 14 correlated features are chosen that include 'age', 'c_charge_degree', 'race', 'age_cat', 'score_text', 'sex', 'priors_count', 'days_b_screening_arrest', 'decile_score', and 'length_of_stay'. We have chosen these features because they are strongly correlated with the target variable and are likely to be important predictors of recidivism. The 'length_of_stay' in jail is calculated using 'c_jail_in' and 'c_jail_out' and a new column is created for this. Missing values are replaced with the median of the respective feature column. Columns related to the violent assessment of prisoners are removed as only general recidivism is analyzed. Rows with charge dates outside of 30 days from arrest dates are removed to enhance data quality [2, 4]. The values are normalized using Standard Scalar, and the dataset is split into training and testing sets using the split function from scikit-learn. Several classification models are trained on the training data, including Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), Support Vector Machine (SVM), K-Nearest Neighbour (k-NN), and Multilayer Perceptron (MLP). The accuracy of these models is evaluated using the testing data as seen in table 1.

Table 1. Model selection, 0: not recidivist and 1: recidivist

| Classifier | Accuracy | precision(0,1) | recall(0,1) | f1-score(0,1) | AUC | False Positive |
|---|---|---|---|---|---|---|
| LR | 0.6765 | 0.69, 0.66 | 0.74, 0.60 | 0.72, 0.62 | 0.72 | 287 |
| DT | 0.5984 | 0.64, 0.55 | 0.62, 0.57 | 0.63, 0.56 | 0.58 | 427 |
| RF | 0.6618 | 0.69, 0.63 | 0.70, 0.61 | 0.69, 0.62 | 0.71 | 335 |
| **GB** | 0.6858 | 0.70, 0.67 | 0.75, 0.61 | 0.72, 0.64 | 0.73 | 281 |
| SVM | 0.6765 | 0.68, 0.67 | 0.78, 0.55 | 0.73, 0.61 | 0.71 | 247 |
| K-NN | 0.6367 | 0.66, 0.60 | 0.69, 0.57 | 0.68, 0.59 | 0.66 | 345 |
| MLP | 0.6691 | 0.68, 0.65 | 0.75, 0.57 | 0.71, 0.61 | 0.72 | 276 |

We have used the Gradient Boosting algorithm as our optimum classifier, based on its superior performance in terms of accuracy, precision, AUC, and false positives, as compared to other classifiers on our dataset.

## 3 RESULTS AND EXPERIMENTATION

After preprocessing, the COMPAS dataset had 6172 cases, of which 33% were reserved for testing and the rest were used for training the algorithm. The Gradient boosting algorithm was chosen as the optimum classifier and generated a confusion matrix with True Positives of 560, True Negatives of 837, False Positives of 281, and False Negatives of 359.

To answer the first question, we evaluated the performance of the algorithm in predicting recidivism, false positive rates were calculated, which reflect the algorithm's tendency to predict that individuals will recidivate when they do not. The false positive rate for African Americans was determined to be $\frac{180}{513} = 0.35$, meaning that 35% of African Americans who did not actually recidivate were incorrectly identified as positive cases by the algorithm. On the other hand, the false positive rate for Caucasians was found to be $\frac{68}{423} = 0.16$, indicating that 16% of Caucasians who did not actually recidivate were falsely predicted to do so by the algorithm.

Based on these results, it appears that the algorithm may be biased against African Americans. This is because the false positive rate is higher for African Americans compared to Caucasians which says the algorithm is more likely to incorrectly predict recidivism for African Americans compared to Caucasians. If this algorithm were used in the criminal justice system to predict recidivism rates, it could have significant societal implications. If the algorithm disproportionately misclassifies African Americans as being at a higher risk of recidivism, it could result in unjust sentencing and reinforce racial biases.

In addition to the false positive rates, another measure of bias in the algorithm is its calibration, as asked in question 2. The calibration is calculated by determining the probability that an individual recidivates given that they are predicted positively by the algorithm and are categorized as African-American by the race variable. The true positive rate (TPR) for African-Americans, in this case, is $\frac{381}{561} = 0.68$, while the TPR for Caucasians is $\frac{126}{194} = 0.65$. The difference is quite close which suggests that the algorithm is similarly calibrated for both groups in terms of correctly identifying those who will recidivate.

However, it's important to note that even a small amount of bias can have significant societal implications. In this case, the higher false positive rate for African-Americans could result in a disproportionate number of individuals being wrongly classified as high-risk and subjected to harsher punishments. It is therefore crucial to continue evaluating and improving the fairness of such algorithms to ensure they do not perpetuate societal injustices.

To answer question three which is based on the comparison of the opportunity cost and calibration metrics, it appears that the opportunity cost metric is more appropriate in the domain of prisoner classification, as it considers false positive rates. The model generated 640 misclassified labels, which is 31% of the total classification, with 281 of those being false positives. Further analysis revealed that 35% of these false positive classifications were African American, while 16% were Caucasian. Although no system is perfect, recent research work has shown to provide a new approach to designing fair and accurate risk assessment tools that take into account the underlying social and historical context of the criminal justice system. The author in [3] has provided a system that produces a reduced false positive rate for African American defendants from 44.9% to 28.7%. Hence such a system should be used in the decision-making process of bail and parole where the FPR is considerably low. The model's 18% false negative rate also poses a risk, as it means that people who are likely to recidivate within two years of the COMPAS assessment could be approved for bail and parole.

On the other hand, for decision-makers in the bail and parole system, the calibration metric would be more appropriate. The model classified 69% of the cases correctly, which translates to 1397 out of the 2037 cases that were tested. While no system is perfect, this rate is acceptable for selecting a model to assist decision-makers in approving bail and parole requests. However, it is important for decision-makers to not heavily rely on the model's output and to apply the human factor in their considerations. Ultimately, the decision-makers are responsible for ensuring that the model's output is used fairly and justly, without causing harm or bias to any group.

Answer for question four, removing the "race" feature from the initial model does not have a significant impact on the model's overall accuracy. Specifically, the accuracy slightly changes from 68.58% to 68.88% with and without the "race" feature respectively. Furthermore, the false positive rates for both African Americans and Caucasians only slightly changed from 180 to 177 for African-Americans, and 68 to 69 for Caucasians. These minor changes suggest that the initial model is not heavily relying on the "race" feature to make biased predictions.

It's important to note, that simply removing the "race" feature does not guarantee that the model is completely unbiased. It's crucial to carefully analyze and evaluate the dataset and model to ensure that it's not perpetuating existing societal biases and prejudices.

To solve question five, which was to make our model fair by ensuring that there is no disparity in false positive rates (FPR) between African American and Caucasian groups. To achieve this objective, we used fairlearn's equilized_odds_difference metric [1]. We found that the FPR for both groups were identical; $\frac{133}{513} = 0.26$ African Americans and $\frac{109}{423} = 0.26$ Caucasians being falsely identified as positive. This means that both groups have the same chance of being wrongly classified as positive, provided that they are actually negative.

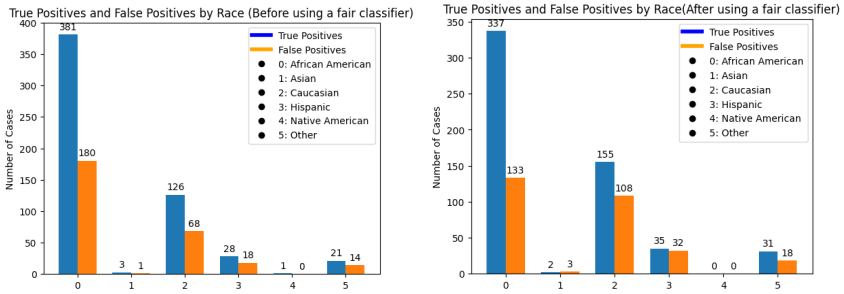


Fig. 1. TPR and FPR distribution over race

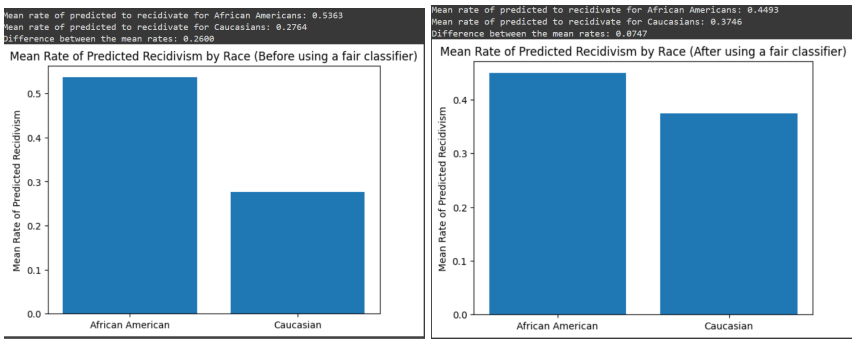


Fig. 2. Mean Rate of Predictive Recidivism

Looking at Figure 1, we can see a comparison of two different models used to predict recidivism among various racial groups. The left graph shows that the original model had a higher number of false positives for African Americans, with 180 individuals being wrongly identified as positive compared to 68 for Caucasians. This indicates that the model had a bias in favor of Caucasians. In contrast, the right graph shows that the fair model had a more balanced distribution of false positives and true positives across both racial groups. Specifically, the number of false positives for African Americans was reduced to 133, while the number for Caucasians was reduced to 108. This suggests that the fair model succeeded in reducing the bias in the original model by decreasing the opportunity cost for African Americans.

Figure 2 demonstrates the average rate of "predicted to recidivate" for African Americans and Caucasians, and how the implementation of a fair classifier resulted in a significant decrease in the disparity between the two groups. Prior to adopting the fair classifier, African Americans had a predicted recidivism rate of almost double that of Caucasians (0.5363 versus 0.2764), with a difference of 0.2599. However, after using a fairer classification method, the mean predicted recidivism rates for African Americans and Caucasians were much closer at 0.4493 and 0.3746, respectively. This indicates a reduced difference of 0.0747 between the two groups in terms of predicted recidivism rates.

The fair model achieved an accuracy of 68.18%, which is slightly lower than the accuracy of the earlier biased model in terms of opportunity cost with 68.58%. However, to achieve fairness, a trade-off had to be made, resulting in a decrease in the f1-score from 0.69 to 0.67. Also, the calibration has slightly changed from $\frac{381}{561} = 0.68$ to $\frac{340}{574} = 0.72$ for African American, and for Caucasian, it has changed from $\frac{126}{194} = 0.65$ to $\frac{156}{266} = 0.59$. The difference is quite close which suggests that the algorithm is similarly calibrated for both groups in terms of correctly identifying those who will recidivate.

The fair model was successful in addressing the issue of significant disparity in FPR between African American and Caucasian groups that were present in the biased model. By ensuring parity in FPR, the fair model treats both groups equally in terms of false positive errors and eliminates the possibility of unintended discrimination or biased outcomes.

---

*CODE:. for this project can be found by clicking on "CODE" word of this sentence.*

## REFERENCES

[1] UCSD Course DSC 167. [n. d.]. *2020. Fairness and Algorithmic Decision Making*.

[2] Allen Downey. 2023. *Case study on evaluating statistical tools that predict recidivism.* Retrieved May 10, 2023 from https://github.com/AllenDowney/RecidivismCaseStudy

[3] Julia Dressel and Journal Science Advances Hany Farid. [n. d.]. Assessing and Addressing Algorithmic Bias in Criminal Justice Risk Assessments, 2020. ([n. d.]).

[4] Marjorie Roswell Jeff Larson and Vaggelis Atlidakis. 2018. *Data and analysis for 'Machine Bias'.* Retrieved May 10, 2023 from https://github.com/propublica/compas-analysis/