Addressing Algorithmic Bias in Recidivism Score Predictions

Shreshth Saini* [1] **Albert Joe*** [1] **Jiachen Wang*** [1] **SayedMorteza Malaekeh*** [1]

Abstract

This paper aims to mitigate the inherent bias in recidivism score predictions generated by the National Institute of Justice. The existing algorithms tend to exhibit biases towards gender and racial/ethnic groups, which can have profound implications on the lives of individuals affected by these predictions. We leverage innovative machine learning techniques to rectify and minimize these biases. In our comprehensive evaluation of various models using various metrics, CatBoost algorithm emerged as the standout performer, exhibiting the highest accuracy rate at 74%. Upon closer examination of the contributing factors, it became evident that employment status, changes in jobs per year, and age were pivotal elements in shaping the predictive outcomes. A noteworthy achievement of our model lies in its significant reduction of bias, particularly concerning predictions related to black individuals.

1 Introduction and Background

1.1 Introduction

In the pursuit of a fair and just criminal justice system, the assessment and prediction of recidivism play a crucial role. However, concerns have arisen over the inherent biases embedded in existing recidivism score prediction algorithms developed by the National Institute of Justice. These algorithms, while aiming to provide objective and data-driven insights, have exhibited pronounced biases that disproportionately impact gender and racial/ethnic groups. The repercussions of such biases are profound, as they not only influence individual lives but also contribute to systemic inequalities within the criminal justice system. This paper undertakes a critical mission to overhaul the existing landscape by developing an algorithmic decision-making model that is not only accurate but, more importantly, transparent and unbiased.

This paper delves into our meticulous process of algorithm selection, validation, and assessment. We evaluated seven algorithms, including logistic regression, Multi-Linear Perceptron Neural Network, KNN, SVM, Bayesian Classifier, XGBoost, and CatBoost. These models centered on forecasting recidivism outcomes for individuals released from prison to parole supervision in the State of Georgia between January 1, 2013, and December 31, 2015. This unique and comprehensive dataset, generously provided by the NIJ, captures a wealth of information encompassing diverse socio-demographic variables. It includes crucial indicators such as racial and ethnic demographics, gender distribution, as well as intricate details like the results of drug tests.

The inclusion of socio-demographic factors allows for a nuanced exploration of how various aspects of an individual's background may contribute to or influence recidivism risk. Notably, our findings reveal that CatBoost not only excelled in terms of traditional accuracy metrics but also demonstrated significant efficacy in mitigating racial bias, establishing it as the optimal model for our purposes. Our study reveals a significant finding: employment and job-related factors exert the most influence in predicting recidivism compared to other features. This underscores the crucial role of addressing economic conditions and employment opportunities for those on parole. Informed by these insights, future policy decisions should prioritize targeted interventions that comprehensively address factors influencing post-release trajectories, particularly focusing on employment and economic conditions.

In the pages that follow, we present a brief overview of the background information and literature review, account of our methodologies and datasets, results, and key observations, contributing to the ongoing discourse on advancing the fairness and effectiveness of recidivism prediction models in the criminal justice domain.

1.2 Literature Review

The existing body of research on recidivism risk has predominantly concentrated on individual-level factors within the context of supervision or reentry programs (e.g., Aos et al. 2007; Bonta and Andrews 2017; Lowenkamp et al. 2010) These works underscore the influence of personal factors, including criminal history, sex, race, and age, on

*Equal contribution [1]Department of Electrical and Computer Engineering, The University of Texas at Austin, TX, USA. Corresponds to: SayedMorteza Malaekeh <malaekeh@utexas.edu>

the likelihood of returning to criminal activities or violating supervision conditions (Piquero et al., 2013; Wang et al., 2010). However, this literature reveals a notable limitation in its reliance on static individual-level characteristics as primary predictors, which are inherently poor targets for intervention and may introduce bias into risk assessments.

Berk and Elzarka (2020) have drawn attention to the pitfalls of risk assessments relying solely on static factors, highlighting the potential for inaccuracies and unfairness. They enumerated several issues of racial bias, such as those arising from disproportionate police contact with residents in disadvantaged neighborhoods, a factor acknowledged by Grogger and Ridgeway (2006). This underscores the critical gap in the literature concerning the comprehensive exploration of contextual factors and the potential impact on algorithmic risk assessments.

Furthermore, limited attention has been given to geographic variations in recidivism, with research primarily relying on census-based social disorganization indicators. This narrow purview, as observed in studies like Clark (2016), Grunwald et al. (2010), and Huebner and Pleggenkuhle (2015), overlooks the broader environmental backdrop. To address this limitation, it becomes imperative to extend the examination of recidivism patterns to a broader scope, such as the state level.

In summary, the existing literature exhibits a predominant emphasis on a narrow set of individual-level factors, a limited exploration of racial bias in algorithmic risk assessments, and a confined scope. This study aims to bridge these gaps by leveraging a comprehensive dataset that encompasses diverse dimensions of individual, community, and environmental factors. By doing so, we aspire to provide a more nuanced and comprehensive understanding of recidivism risk, moving beyond traditional limitations and contributing to a more holistic approach in the field.

2 Data

2.1 Data Description

The dataset is provided by the National Institute of Justice (NIJ) Recidivism Forecasting Challenge, tasking participants with creating data-driven algorithms to effectively estimate the likelihood of parolees who were recently released from prison reoffending three years post-release. The dataset comprises 26,000 individuals who were released from Georgia prisons under discretionary parole for post-incarceration supervision spanning from January 1st, 2013, to December 31st, 2015. Each observation comes with 48 distinct predictor variables, involving supervision case information, prison case information, prior Georgia criminal history, and supervision activities. NIJ split the dataset into a training and test set with a 70/30 proportion.

2.2 Data Preparation

We prepare the dataset for model training. This involves cleaning ‘NaN’ values, converting the data into proper format, and creating new features. We first group the 48 features into 4 categories based on data type: binary, numerical, categorical, and ordinal.

2.2.1 Data Cleaning

For binary features, ’NaN’ values are replaced with ’False’. Within numerical features, we mainly deal with drug test and employment. For ’NaN’ in the feature *Avg_Days_per_DrugTest* (indicates the average days between two drug test), replace it with the average value if any of the drug test is positive; otherwise, replace it with 0. The features *Jobs_Per_Year* and *Percent_Days_Employed* are related: *Jobs_Per_Year* is given 0 if *Percent_Days_Employed* is 0 or ’NaN’, given median if *Percent_Days_Employed* is more than 0; *Percent_Days_Employed* should be 0 if *Jobs_Per_Year* is 0, median if *Jobs_Per_Year* is more than 0. Categorical feature *Prison_Offense* has its ’NaN’ converted into ‘Unknown’. All ordinal features from the dataset are clean. After data cleaning, each observation has features with reasonable values.

2.2.2 Data Format Convertion

We then convert the data into a format that can be used by the model. Convert 'True' and 'False' in binary features into 1 and 0. One hot encoding is applied for categorical and ordinal features. Numerical features are normalized properly given their meaning to make sure all within reasonable range.

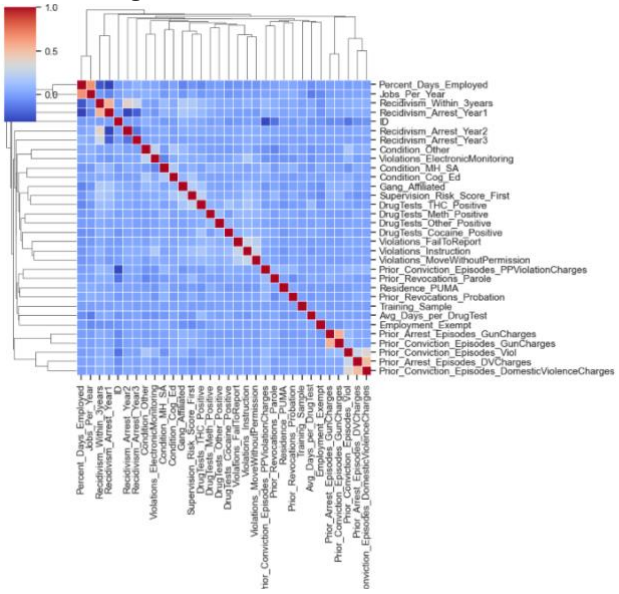

Figure 1. Feature Correlation Matrix.

Figure 1 shows the correlation matrix after data preparation. Weak correlations can be found between almost all feature pairs.

2.2.3 New Features

We also created new features to facilitate the training. Table 1 shows new features added and their definitions, respectively.

Table 1. New Features

Feature	Definition
Required_Drug_Tests	Requirement of drug test from parolee
Prison_Years_Pct_Age	Percentage of parolee's lifetime in
Prior_Arrest_Sum	Sum of all prior arrest episodes
Prior_Conviction_Sum	Sum of all prior conviction episodes
Employed_Indicator	Parolee employment status
Residence_Change_Violati	Parolee move without permission

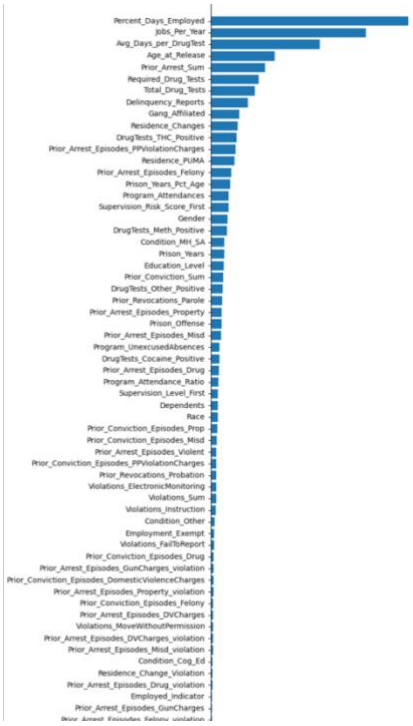

Figure 2. Feature Importance. Top Three Important Features: *Percent_Days_Employed*, *Job_per_Year*, and *Avg_Days_per_Drug_Test*.

2.3 Important Features

Figure 2 shows the feature importance from Logistic Regrssion. Among them the most important 3 features are: *Percent_Days_Employed*, *Job_per_Year*, and *Avg_Days_per_Drug_Test.* This indicates the fact that employment plays a role in recidivism: the more days the parolee stays employed, the lower probability recidivism would occur. Figure 3 demonstrates the distribution of *Percent_Days_Employed*. In current dataset, most parolees are either fully employed or unemployed through the year. Other parolees are evenly distributed between 0 and 100%. Figure 4 gives the distribution of feature *Job_per_Year*. The histogram is in accordance with Figure 3, which further proves our assumption in data cleaning. More than half of the parolees have an average of 0 jobs per year, which could be an indicator of recidivism. Shown in Figure 5, drug test is another high risk factor of recidivism: the fewer days between two drug test, the higher risk of a parolee to commit criminal offenses.

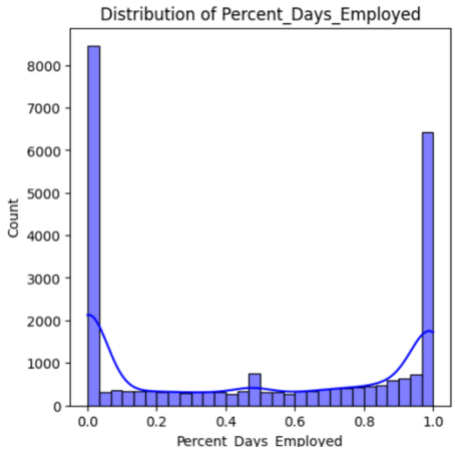

Figure 3. Distribution of the Most Important Feature, *Percent_Days_Employed.*

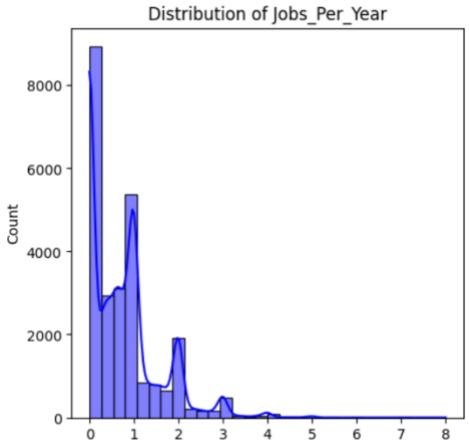

Figure 4. Distribution of *Jobs_per_Year*.

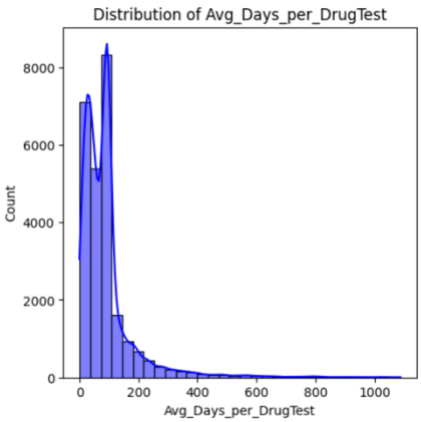

Figure 5. Distribution of *Average_Days_per_Drug_Test.*

3 Model

3.1 CatBoost

CatBoost is a a variant of gradient boosting that can handle both categorical and numerical features (Prokhorenkova et al., 2018). CatBoost introduced two pivotal algorithmic advancements: ordered boosting, which is a permutation-driven alternative to the traditional algorithm, and an innovative approach for handling categorical features. These techniques were developed to mitigate prediction shift resulting from a specific form of target leakage inherent in all existing implementations of gradient boosting algorithms. CatBoost employs an algorithm known as Symmetric Weighted Quantile Sketch (SWQS) to automatically manage missing values within the dataset. This approach serves to mitigate overfitting and enhance the overall performance of the dataset.

CatBoost adopts ordered target statistics (TS) to handle categorical features. One-hot encoding can lead to infeasibly large number of new features in the case of high cardinality features. TS addresses this issue by grouping categories into a limited number of clusters, i.e. to estimate expected target value in each category, and then apply one-hot encoding. Ordered TS is a more effective strategy. The authors introduce an artificial "time", i.e., a random permutation of the training examples. Then, for each example, all the available "history" are used to compute the TS. Note that, if use only one random permutation, then preceding examples have TS with much higher variance than subsequent ones. To this end, CatBoost uses different permutations for different steps of gradient boosting.

3.2 Other Models Adopted

Besides CatBoost, we also trained other machine learning models to tackle this task, including Linear Regression (LR), Extreme Gradient Boosting (XGBoost), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Bayesian classifier, and MultiLayer Perceptron (MLP). Further details of their theories will not be given in this report due to the limited space. Results of these classifiers are given in Section 4.

3.3 Model Performance Metrics

NIJ designed Brier score to assess the model validity, which is defined as:

$$Brier\ Score = \frac{1}{n}\sum_{t=1}^{n}(f_t - A_t)^2$$

where n is the count of individuals in the test dataset, f_t is the forecasted probability of recidivism for individual t, and A_t is the actual outcome (0,1) for individual t. Since the Brier score is a measure of error, models are aimed to minimize this metric.

4 Results

The results section unfolds in three distinctive parts. Initially, we concentrate on the model evaluations, encompassing Logistic Regression, Multi-Linear Perceptron Neural Network, KNN, SVM, Bayesian Classifier, XGBoost, and CatBoost. Assessment metrics include AUC, Accuracy, Precision, Recall, and F1-Score. Subsequently, we present the logistic regression results in a tabular format, allowing for a nuanced interpretation of the impact of each variable on the likelihood of recidivism. Following this, we highlight the best-performing model based on our selection criteria and elucidate the key findings. Lastly, we delve into the Parole Board Recommendation Model, and evaluating its algorithmic bias based on race within the model.

4.1 Evaluation and Model Performance

Table 2 reveals that among the evaluated models, ensemble methods such as CatBoost and XGBoost stand out, outperforming others based on the selection criteria. This superiority can be attributed to their adeptness in handling complex datasets. CatBoost demonstrates a slight edge, potentially attributed to its proficiency in managing categorical variables. With an accuracy of 0.7483 and a Brier score of 0.1686, CatBoost slightly outperforms XGBoost, which exhibits an accuracy of 0.7451 and a Brier score of 0.1701. Moving forward, we focus our attention on CatBoost and present a detailed analysis of its results, aiming to glean insights into its predictive performance and contributions to our model.

Table 2. Machine Learning Algorithms Evaluation

Model	AUC	Brier	ACC	Prec.	Recall	F1
LR	0.78	0.1862	0.7197	0.7303	0.8117	0.7689
XGBoost	0.82	0.1701	0.7451	0.7518	0.8302	0.7891
CatBoost	0.82	0.1686	0.7483	0.7535	0.8347	0.7920
SVM	0.78	0.1864	0.7192	0.7277	0.8166	0.7696
KNN	0.71	0.2119	0.6662	0.6707	0.8227	0.7389
Bayesian	0.71	0.2796	0.6694	0.6953	0.7551	0.7240
MLP	0.77	0.1985	0.7182	0.7087	0.8715	0.7827

3.2 A focus on Logistic Regression

In Figure 6, the logistic regression models consistently reveal significant predictors of recidivism. In both Model 1, without controlling for confounding variables, and Model 2, which includes controls for various factors, similar patterns emerge. Key findings include the significant negative impact of employment on recidivism, the increased likelihood for males compared to females, and the nuanced relationship between race and recidivism. Interestingly, being white compared to black is associated with an increased likelihood of recidivism, even after accounting for other variables.

Furthermore, both model provides additional insights, indicating that factors such as positive drug tests, fewer jobs per year, a higher number of prison years, and older age contribute significantly to the reduction of recidivism likelihood.

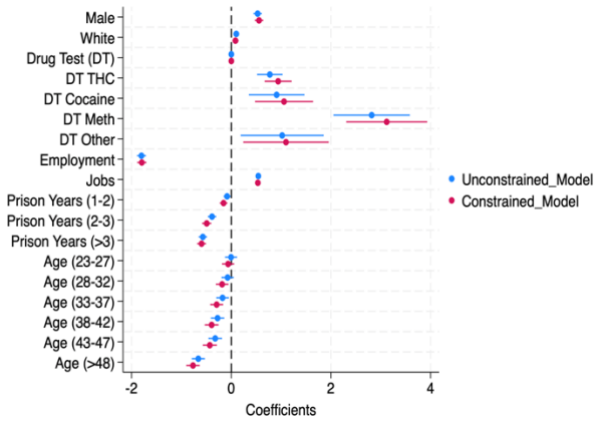

Figure 6. Coefficients from Logistic Regression (Controls in unconstrained variables include types of offense, number of dependent, education, residence changes, program attendances, delinquency, violations, conditions, and revocations. Standard errors in parentheses)

These consistent findings across both models underscore the robustness of the identified predictors and provide a comprehensive view of the factors influencing recidivism. Such insights are crucial for informing targeted interventions and policy decisions within the criminal justice system.

3.3 A focus on CatBoost

Figures 7 and 8 show the Receiver Operating Curve (ROC) and the Confusion Matrix of CatBoost algorithm. CatBoost's superior performance can be attributed to its adept handling of categorical variables, a feature that distinguishes it in predictive modeling. Unlike some other algorithms, CatBoost is designed to naturally manage categorical features without the need for extensive pre-processing. This capability is crucial when dealing with real-world datasets where categorical variables often play a significant role.

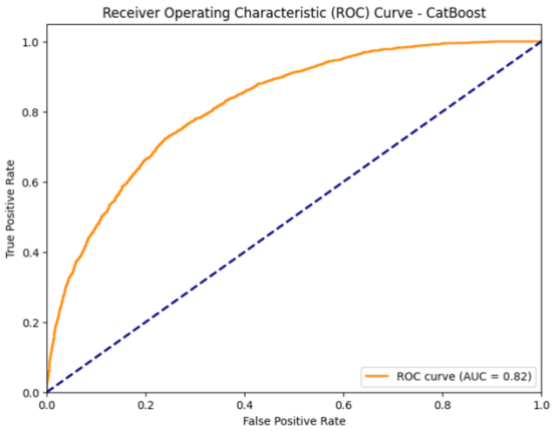

Figure 7. ROC curve of CatBoost Algorithm

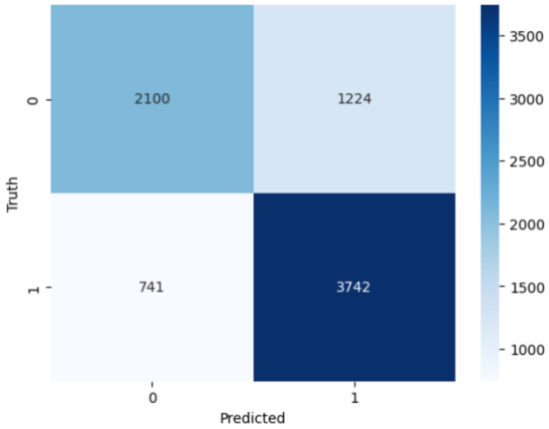

Figure 8. Confusion Matrix of CatBoost Algorithm

In terms of hyperparameter tuning, we employed a grid search approach to systematically explore various combinations of hyperparameters and identify the optimal configuration for CatBoost. This involved testing different

values for parameters such as learning rate, depth, and the number of trees. The grid search process systematically evaluates these hyperparameter combinations and selects the set that maximizes the chosen evaluation metric, in this case, likely a combination of accuracy, precision, recall, or F1-score.

By leveraging CatBoost's inherent strength in managing categorical variables and fine-tuning its hyperparameters through a grid search, we enhanced the model's predictive capabilities, resulting in its superior performance in the context of recidivism prediction.

In Figure 9, the SHapley Additive exPlanations (SHAP) values shed light on the influential predictors, with employment emerging as the most impactful feature. According to Figure 6 from the logistic regression results, the negative impact of employment days suggests that an increase in employment days correlates with a lower likelihood of recidivism. This insight carries significant policy implications, highlighting the potential effectiveness of creating job opportunities for individuals on parole as a preventive measure against recidivism.

Moreover, the number of jobs per year emerges as another robust predictor, with its negative impact indicating that individuals who frequently change jobs might be less satisfied or face challenges, potentially leading to a higher likelihood of engaging in criminal activities. This observation underscores the importance of stability and satisfaction in employment as potential contributors to reducing recidivism rates. Importantly, the analysis reveals

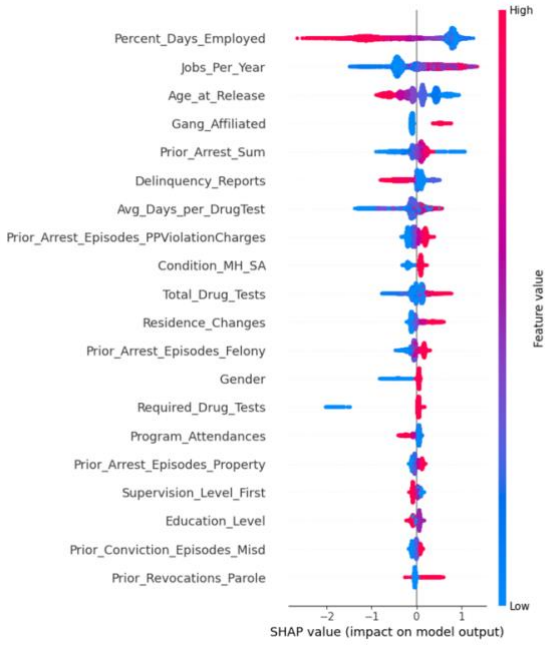

Figure 9. Feature Impacts on Recidivism Prediction in CatBoost Algorithm

that race does not rank among the top predictive features, suggesting that, within the scope of this model, it does not exert a substantial impact on recidivism prediction.

3.4 Parole Board Recommendation Model

In the reduced features Catboost model, designed to address data unavailability at the time of the parole board, the analysis focused on variables that were accessible and relevant at that specific point in the decision-making process. This approach aimed to enhance the model's practical applicability and interpretability, given the constraints of real-world scenarios where certain information may not be readily accessible at a particular juncture. The Brier Score of 0.1996, an accuracy of 0.693, and an AUC of 0.74 showcase the model's robust performance.

Race does not emerge as a prominent predictor, and the analysis of False Positive and False Negative rates reveals comparable patterns for black and white individuals. Specifically, for black individuals, the False Positive rate is 21.5%, and the False Negative rate is 9.81%. In contrast, for white individuals, the False Positive rate is 19.1%, and the False Negative rate is 10.5%.

Despite there being more black prisoners than white prisoners, the model exhibits slightly higher False Positive rates and lower False Negative rates for black individuals. While the AUC is slightly lower for blacks, the overall differences are not substantial.

Previous analyses suggested a 45 percent higher likelihood of black defendants being assigned higher risk scores, which contrasts with our model's results. Figures 10 and 11 illustrate the normalized distribution of predicted probabilities and the Kernel Density Estimate (KDE) plot

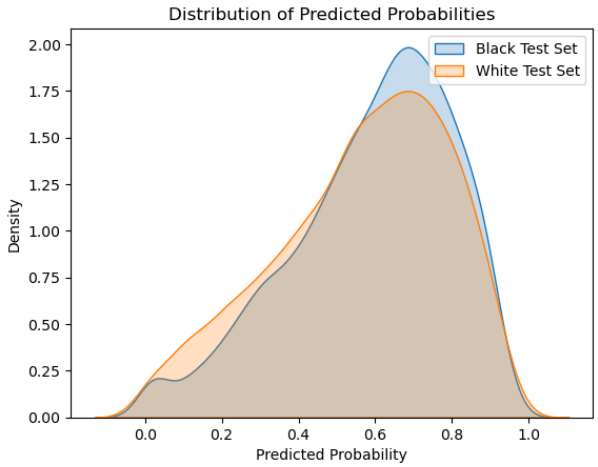

Figure 10. Distribution of Predicted Probabilities in Reduced Features CatBoost Algorithm across Race

of the False Positive rate by predicted probability. The plots show a very similar distribution, with a slight increase for black individuals, indicating successful mitigation of algorithmic bias against race.

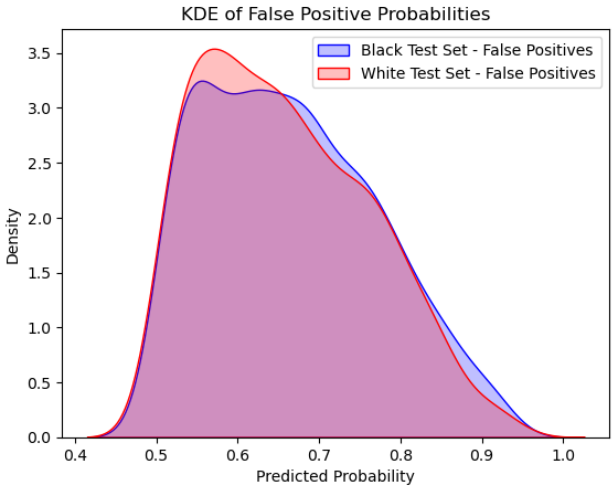

Figure 11. KDE plot of False Positive Probabilities in Reduced Features CatBoost Algorithm across Race

Moreover, the evaluation of Type I and Type II errors, along with the distribution of predicted probabilities, indicates a model that is well-calibrated, achieving similar performance across different racial groups. This fosters fairness and mitigates bias in predicting recidivism.

4 Conclusion and Takeaways

In conclusion, this paper tries to address and rectify biases in recidivism risk scores by the development of an advanced model employing state-of-the-art classification methods. CatBoost emerged as the superior choice between all the models evaluated in this study, showcasing impressive performance with an accuracy of 0.7483 and a Brier Score of 0.1686. Beyond the technical achievements, the practical impact of our improved model is significant. By mitigating biases, we contribute to a fairer and more equitable criminal justice system. The enhanced accuracy and reliability of predictions empower decision-makers to make more informed choices, ultimately fostering positive outcomes in parole board decision-making. This paper underscores the importance of leveraging innovative machine learning techniques to address societal challenges, emphasizing the potential for advanced models to drive positive change in critical domains like criminal justice.

Data Availability Statement

To ensure the integrity of our study, all data, methods, and results utilized are readily accessible through our GitHub repository: https://github.com/shreshthsaini/UBR-UnBiased-and-Robust-Recidivism-Prediction.

Acknowledgments

This project was undertaken as a requirement for the Applied Machine Learning course offered by the Department of Electrical and Computer Engineering at the University of Texas at Austin, under the guidance of Professor Ghosh. The knowledge and skills gained in this course have been instrumental in the successful execution of this study.

References

Berk, R. A., & Elzarka, A. A. (2020). Almost politically acceptable criminal justice risk assessment. *Criminology & Public Policy*, *19*(4), 1231–1257. https://doi.org/10.1111/1745-9133.12500

Boutwell, B. B., Nelson, E. J., Emo, B., Vaughn, M. G., Schootman, M., Rosenfeld, R., & Lewis, R. D. (2016). The intersection of aggregate-level lead exposure and crime. *Environmental Research*, *148*, 79–85. https://doi.org/10.1016/j.envres.2016.03.023

Clark, V. (2016). Predicting two types of recidivism among newly released prisoners. *Crime & Delinquency*, *62*(10), 1364–1400. https://doi.org/10.1177/0011128714555760

Evidence-Based public Policy options to reduce future prison construction, criminal justice costs, and crime rates (Oct. 2006). (2007). *Federal Sentencing Reporter*, *19*(4), 275–290. https://doi.org/10.1525/fsr.2007.19.4.275

Grogger, J., & Ridgeway, G. (2006). Testing for racial profiling in traffic stops from behind a veil of darkness. *Journal of the American Statistical Association*, *101*(475), 878–887. https://doi.org/10.1198/016214506000000168

Grunwald, H. E., Lockwood, B., Harris, P. W., & Mennis, J. (2010). Influences of neighborhood context, individual history and parenting behavior on recidivism among juvenile offenders. *Journal of Youth and Adolescence*, *39*(9), 1067–1079. https://doi.org/10.1007/s10964-010-9518-5

Huebner, B. M., & Pleggenkuhle, B. (2013). Residential Location, Household Composition, and Recidivism: An Analysis by gender. *Justice Quarterly*, *32*(5), 818–844. https://doi.org/10.1080/07418825.2013.827231

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems*, *31*.

Lowenkamp, C. T., Makarios, M. D., Latessa, E. J., Lemke, R., & Smith, P. (2010). Community corrections facilities for juvenile offenders in Ohio. *Criminal*

Justice and Behavior, *37*(6), 695–708. https://doi.org/10.1177/0093854810363721

Mears, D. P., Wang, X., Hay, C., & Bales, W. D. (2008). SOCIAL ECOLOGY AND RECIDIVISM: IMPLICATIONS FOR PRISONER REENTRY. *Criminology*, *46*(2), 301–340. https://doi.org/10.1111/j.1745-9125.2008.00111.x

Piquero, A. R., Jennings, W. G., Diamond, B., & Reingle, J. M. (2013). A systematic review of age, sex, ethnicity, and race as predictors of Violent Recidivism. *International Journal of Offender Therapy and Comparative Criminology*, *59*(1), 5–26. https://doi.org/10.1177/0306624x13514733