Mitigating Racial Bias in Recidivism Prediction:

Machine Learning & Threshold Selection Approach

Yijia Lyu

June 2020

Abstract

The algorithm COMPAS, which is used by many courts in USA to assess the recidivism likelihood of defendants, has been accused of racial bias. This paper assesses COMPAS's classification setting in defining 'low', 'medium', 'high' risks and tries to select proper threshold for each racial group so that the tool can improve its algorithm fairness whilst having good predictive performance. Then it also uses machine learning methods separately by splitting race feature as a protected attribute to examine if such method will reduce bias. By examining the accuracy, FN, FP, the result shows that COMPAS is relatively biased towards the black defendants. Reselecting thresholds can improve fairness in equal accuracy by setting different thresholds for the white and the black, but the false prediction disparities are still high. Machine learning separation performs excellent in reducing the FN differences between the white and the black group but the FP for the black defendants is still twice the white defendants.

Introduction

Crime has been a significant disruptive factor in social security and our daily life. Keeping track of crime occurrences and criminals is therefore essential to build safer communities. Washington courts reported that 63.3% of the sentences in 2007 involved cases related to recidivism. This means that if our systems are able to capture and supervise criminals who have high-risks to re-commit crimes, we are able to reduce safety concerns and better allocate policing and other resources. Multiple statistical and machine learning algorithm have concluded a number of related factors to predict who is likely to reoffend after release have been developed. Many jurisdictions in the United States have adopted one predictive tool called COMPAS to estimate defendants' likelihood of recidivism and use it as a reference to determine final sentencing on defendants. But as the algorithm of COMPAS is business confidential, it raises concern on prediction accuracy and algorithm fairness. Verification studies (Larson et.al, 2016) criticize this tool for being racially unfair and argue that the black community suffers great discrimination. Therefore, it is important to identify how it is unfair towards different race groups and reduce these bias for such real-world application.

This paper verifies the algorithm fairness on this widely-adopted recidivism evaluation software, COMPAS. Then it uses two bias-correction methods: 1) threshold selection method, 2) attribute-protected machine learning method, to mitigate the racial bias in this tool.

Literature Review

1.1 Statistical Risk Assessment on Recidivism Prediction

Criminologists have long proposed and attempted to estimate and predict the recidivism risk of criminals, especially for cases that may cause extraordinary harm to the society. According to the definition put by Urahn (2011), recidivism is the 'act of re-engaging in criminal offending'. Prior research has incorporated key factors of recidivism as demographic factors including race, age, gender, etc. (Langan & Levin, 2002), criminal history like the term of imprisonment (Blumstein, Cohen, & Farrington, 1988; Piquero, Farrington &Blumstein, 2003), other individual-level factors such as antisocial attitudes, associates, personality (Serin, Lloyd, Helmus, Derkzen, & Luong, 2013), social bonds (Yang, Liu, & Coid, 2010) or socioeconomic status (Hanson &

Harris, 2000). Since 1980s, estimations on sexual and violent offence recidivism have been a frequent topic in criminology. Many estimation studies start from the psychopathy point of view. For example, the U.S. Sentencing Commission (2005) employed a prediction tool called CHC for federal judges to 'measure offender culpability, deter criminal conduct, and protect the public from further crime of the defendant'. Similarly, in 1995, the Canadian forensic researcher, Quinsey, combined the Psychopathy Checklist (PCL-R; Hare, 1991) with a number of relevant variables (Rice et al., 1990) to perform multivariate statistics and calculation of actuarial estimates of risk. Then in 2006, collaborating with his colleagues, Quinsey developed a prediction instrument, Sex Offender Risk Appraisal Guide (SORAG), which was later adopted by many scholars, to assess criminals' risk score of violent and sexual recidivism based on fourteen items. Each item is scored individually and aggregated together following an assigned weight of each item. Multiple datasets gathered from German, USA, Canada, Belgium are used to test the validity of SORAG and many replication studies are built upon this guide (Rettenberger & Eher, 2007; Ducro & Pham, 2006). However, the magnitude of such sample size is often restricted to only hundreds, the risk factors are static, ignoring the dynamic predictors, and the scope for such prediction only limits to a few extreme crime categories. Most importantly, the coefficients of each item should be validated with external data, otherwise it may lead to inappropriate or even erroneous causal relationship explanation. With these concerns on accuracy and reliability, although these predictive studies were frequently discussed in academia, the variables to evaluate the recidivism were regarded only as a guideline while the risk scores themselves were rarely applied to jurisdiction at this stage.

1.2 Extending the Scope: Machine Learning Predictions on Recidivism

With the wide application of machine learning, the predictive tools in recidivism have also transited from clinical judgement to algorithm decision-making. A surge of novel data mining techniques including logistic regression, random forests, support vector machines, neural networks and the search algorithm are found to outperform the traditional methods (Attewell &Monaghan, 2015). These novel methods not only enlarge the scale of data being fed in, but also extend the scope to a wider range of crime types by decreasing unexplained variables in the dependent variables (ibid). However, a statistically expected outcome may not be a perfect match in an actual policy, predictive power is accompanied with errors and the cost of these errors

needs to be evaluated (Berk, 2012). When assessing the performance of a predictive algorithm, apart from forecasting accuracy, the ratio of false positives (false alarms) to false negatives (missing cases) is another key metric. Bradley (1997) put that the cost of misclassification is more important than the rate of misclassification. Overestimating the false positives can lead to great amount of resource waste, leaving the low-risk defendants take unfair consequences including loss of freedom, decreased life quality or loss in future employment. But on the contrary, underestimating the false negatives may also put many lives in danger. The trade-off between the false positives and false negatives is at practitioners' discretion and vary between jurisdictions (Barnes & Hyatt, 2012). Researchers suggest practitioners to have the rate pre-determined on an agreeable level, such as 5:1 (Berk et al., 2005).

1.3 Real-world Application, Algorithmic Bias and Unfairness

Despite such discussion on prediction errors, many county-level jurisdictions in the United States have adopted machine learning or deep learning algorithms as a sentencing reference. By identifying which criminals are at high risks of re-committing crimes and predicting what types of crimes they may commit, the judges are referring to the result of this risk assessment to determine the final sentence imposed on the defendants. In 2012, the Wisconsin Department of Corrections launched COMPAS, an algorithmic software developed by Northpointe, and used it in each step in the prison system from sentencing to parole. This software was also employed in the jurisdiction systems in New York State, California, Florida and some others, but in neither of the states or country was the tool evaluated statistically-carefully. Brennan and his two colleagues (2009) published a validation study of COMPAS and found that the accuracy rate of the tool was 68%, however, COMPAS was 67% accurate in black men while it has a 69% accuracy in white men – although this algorithm did not include race as a variable. In a later analysis on COMPAS produced by Larson et, al (2016), the result showed that black defendants who did not recommit crimes over a two-year period were nearly twice as likely to be mistakenly labeled as higher risks compared to white counterparts (45% vs 23%). White defendants who were misclassified as low risk re-offenders almost twice as the black (48% vs 28%). In violent recidivism, compared to the black defendants, the white violent recidivists were 63% more likely to be misclassified as low risk.

Although compared to the early stage predictions, the accuracy and reliability of recidivism prediction seem to be dramatically improving with machine learning models, various validation studies on the widely-adopted prediction software COMPAS have proved us that the black communities suffer from significant algorithm unfairness. Even if we have removed the explicit race factor as an input variable, the systematic inequality still profoundly affects the individuals, groups and society. Accuracy is no longer the only concern in predictive models as existing bias towards certain groups might be further perpetuated through advanced machine learning algorithms. Algorithm fairness, defined as anti-classification (protected attributes like gender, race should not be used to make decisions), parity (the ratio of false positives and negatives should be equal across protected attribute groups) and calibration (conditional estimates are independent from protected attributes), is therefore crucial to measure model performance (Davies & Goel, 2018).

Therefore, examining and reducing the algorithmic bias is an urgent task if we decide to apply such machine prediction result in pretrial, parole, and sentencing decisions. Addressing these issues, this study will explore the reasons that lead to such algorithm unfairness in recidivism, build harm-reduction framework in machine learning models that mitigate such racial disparities to improve algorithm fairness while maintaining accuracy.

Methods and Data

2.1 Key Definitions

• Recidivism

Following the result of a recent recidivism study by the U.S. Sentencing Commission (2019) that most recidivists commit a new crime within the first two years after their release, we define recidivism in this study as a new arrest in two-year period. With this definition, we validate whether the COMPAS scores of criminal suspects in our dataset correspond with their criminal behaviors in the next two years. In this paper we use (p) to stand for prevalence and refer to proportion of individuals who recidivate in a given population.

• High, Medium, Low Decile Score

According to COMPAS, all scores range from 1 to 10. Decile scores between 1 to 4 are classified as low scores, scores between 5 to 7 are labelled as medium scores, and scores between 8 to 10 are regarded as high scores. We use this classification to verify the algorithm bias when we look at the raw data. But when in the threshold selection section, we define high risk as score greater than the threshold value.

• Algorithm Fairness

Predictive parity and calibration are two key definitions when assessing algorithm fairness, but they are not often compatible.

Predictive parity (PPV) means that statistical features such as classification error, false positive rate, false negative rate, precision, etc., should be equal across groups when classified with protected attributes.

$$\mathbb{P}(Y = 1 \mid S > s_{\mathrm{HR}}, R = b) = \mathbb{P}(Y = 1 \mid S > s_{\mathrm{HR}}, R = w).$$

Calibration means that predictive outcomes should be independent of protected attributes like demographic features, for example, race or gender. In the context of pretrial, we refer calibration to be same probability of true recidivism across different racial groups among defendants who are given with a reoffend risk score after release. A score s = s(x) is calibrated if it satisfies the following equation:

$$\Pr(Y = 1 \mid s(X),\, X_p) = \Pr(Y = 1 \mid s(X)).$$

2.2 Research Questions

This research focus on the three key research questions:

a) Is COMPAS racially biased when scoring recidivism on defendants?

b) Following algorithm fairness definition, is current risk classifier setting for low, medium, high risks (1-4, 5-7, 8-10) fair? If not, how should the thresholds change?

c) How might protected attributes (race/gender) separation improve the machine learning prediction performances in ways that reduce false predictions?

2.3 Data Sources

As the goal of this study is to assess recidivism algorithm fairness of COMPAS and improve model fairness towards different racial groups, we use data from Broward County in Florida whose jurisdiction is known to use COMPAS risk evaluation tool to decide pretrial release. Florida is also a state that has strong open-records laws. The data records consist of 11757 criminal suspects who were assessed at the pretrial stage and scored for recidivism risk using the COMPAS in Broward County during 2013 and 2014. The dataset is publicly available, including information of person id, assessment id, case id, agency text, last name, first name, middle name, sex code, ethnic code, date of birth and other nineteen variables. Then jail data and public incarceration data from 2013 to 2016 is used to match with previous records based on identifiable variables like name, case id, etc.

2.4 Data Processing

With datasets above, we are able to build a criminal history dataset. However, some records are not correctly matched for a number of reasons: (1) The basic information of defendants were entered with spelling or numeric mistakes; (2) There were no corresponding crime charge cases to COMPAS scores. (3) Recidivism information is less than two years. The error rate is close to 3.8% with confidence interval to positive and negative 1.75%, we remove mismatching records and merge the matched ones into a final dataset with population of 7214 defendants.

2.5 Methods

To answer the three questions, we first examine COMPAS score distribution for each race group and compare the accuracy disparity, FP and FT for the two race groups. This helps us identify if later bias-mitigation measures are effective. Then we build on the current COMPAS algorithm and use threshold selection method to assess if current score classifier is fair to the two groups. Finally, we change the predictive algorithm by testing separate modelling on both race groups and individual race group.

• Threshold Selection

Based on the initial COMPAS fairness result, we use threshold bias correction model to examine if COMPAS's uniform score classification on low, medium, high risks is reasonable. Many risk

assessment tools compare true risks $\hat{U}(x)$ with the risk score s(x) and set the corresponding d(x) = 1 when it satisfies the condition that $\hat{U}(x) >= t$, 't' here represents an appropriate threshold which we adjust in different context. In this study, 't' is a threshold that differentiate high-risk recidivists. We use iterative threshold selection to check if COMPAS's score classifier for high, medium, low risks is the optimal and fair towards both race groups.

By iterating threshold computation on each decile score, we are able to compare the corresponding accuracy, FP, FT for each race group. Then we follow the statistical parity and calibration as algorithm fairness criteria to decide the optimal threshold as risk classifier and discuss the threshold classifier feasibility in real life.

• Machine Learning: Decision Tree

Decision Tree is used in supervised machine learning and the method is easy to interpret. We first separate the data into three datasets on both races, black only, white only. Then we use Decision Tree model to train and predict the recidivism result on the dataset with both races. Next, we use cross validation to offset the dataset size limitation after data splitting for the black and the white and conduct Decision Tree modelling to train and predict the recidivism result on the black group and the white group individually. Finally, we compare the three training results and evaluate if separating protected attributes (race) mitigates the bias.

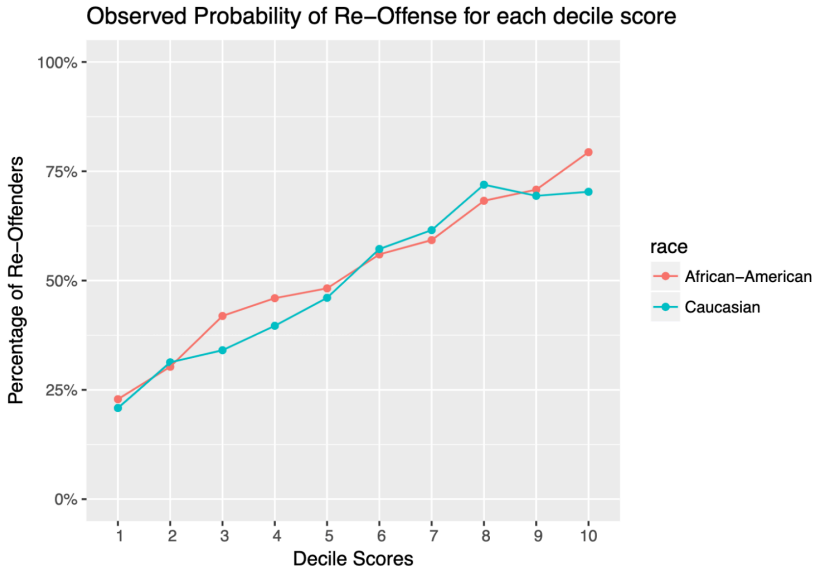
Results

3.1.1 Raw data overview

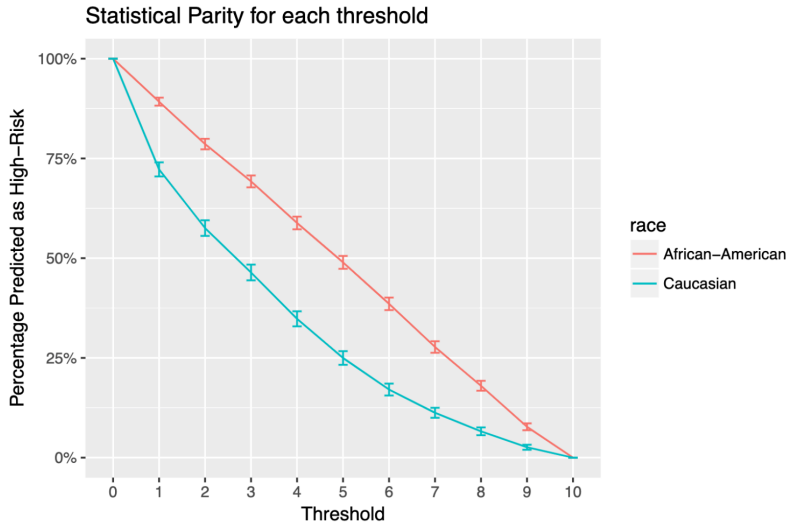
decile_score	African-American	Asian	Caucasian	Hispanic	Native American	Other
1	398	15	681	196	NA	150
2	393	4	361	113	4	66
3	346	5	273	86	1	36
4	385	NA	285	52	1	46
5	365	1	241	52	NA	22
6	384	3	194	37	2	21
7	400	1	143	34	4	10
8	359	2	114	26	1	10
9	380	NA	98	20	2	8
10	286	1	64	21	3	8

(Figure1 COMPAS score distribution: predicted results)

The graph above shows that in general, COMPAS gave decile scores from 1 to 10 to the black defendants fairly averagely while it gave scores to the white defendants with high discrepancy. Among the 2454 white defendants, 35.4% of them were classified as 'high/medium risk'. In comparison, 58.8% of the 3696 black defendants were labelled as high/medium, suggesting a higher likelihood of recidivism than other races.



(Figure 2 Calibration check - actual recidivists proportion on each decile scores)



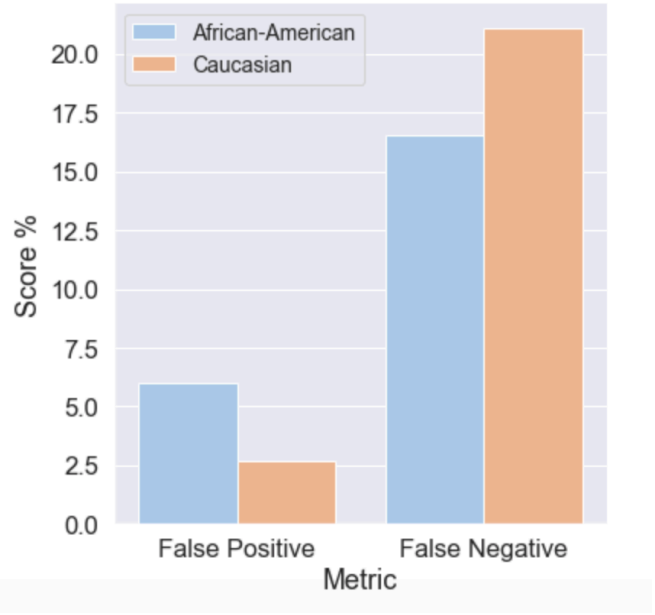
(Figure 3 Statistical parity check)

Then we plot the calibration graph and statistical parity prediction graph for each racial group. Figure 2 examines the calibration and demonstrates that current COMPAS system does not well

calibrate. On each decile score level, prediction accuracy on black and white defendants differs from each other. The differences at score=3, 4, 10 is huge. Figure 3 checks statistical parity by comparing the probability Pr(S) that the algorithm classifies a defendant with a score (S) above certain decile scores. It must be the same with the probability for African-Americans (X = a), as it is for Caucasians (X = c). But the result is as follows:

$$Pr(S > 4 \mid X = a) = 0.59 > Pr(S > 4 \mid X = c) = 0.35$$
$$Pr(S > 7 \mid X = a) = 0.28 > Pr(S > 7 \mid X = c) = 0.11$$

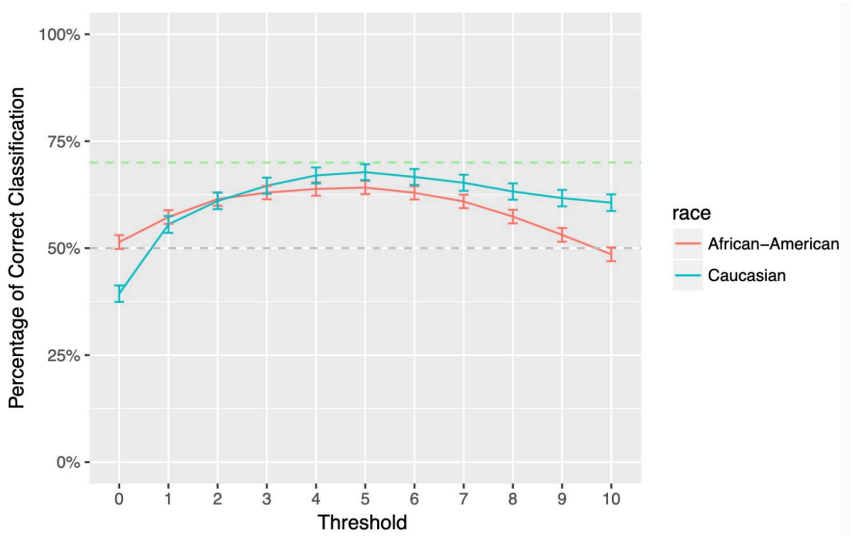
This means that COMPAS does not satisfy the algorithm fairness.



(Figure 4 FP and FN on different racial groups)

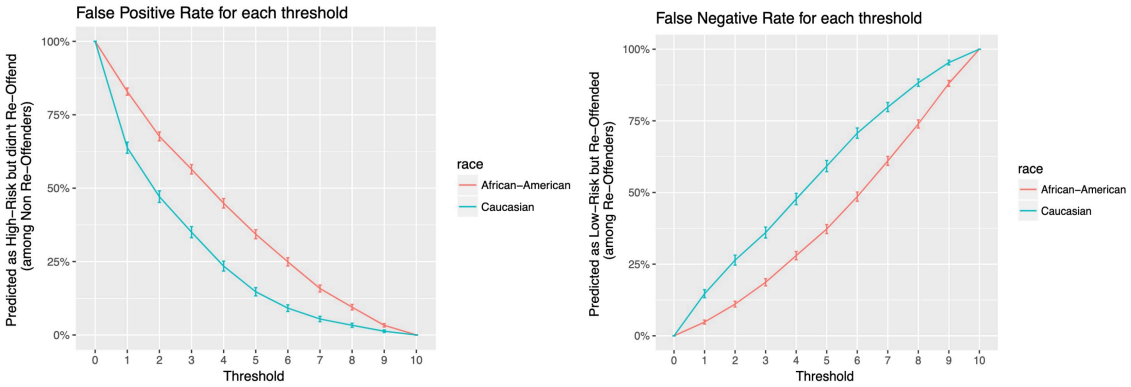
Figure 4 further shows that the black population experiences both a higher false positive rate and lower false negative rate. Its FP is almost twice the rate of the white and its FN is 25% less than Caucasians. This implies that COMPAS underestimates the recidivism likelihood of white re-offenders but overestimate the recidivism probability of the black population. This metric is a clear indicator that the scoring system is biased to different race groups.

3.1.2 Threshold Selection



(Figure 5 Accuracy and threshold for two race groups)

The x-axis of the plot represents different threshold from 1 to 10, and the y-axis represents the accuracy. The graph shows that the accuracy for African-American is better than Caucasian when threshold is under 3, whereas from threshold 4 onwards, the accuracy for Caucasian increases and performs better than African American. There is also substantiate difference in accuracy at threshold above 8. The result shows that for both the two groups, threshold=5 has the highest accuracy. The predictive accuracy when threshold = 5 for the black population is closest to the accuracy when threshold = 3 for the white.

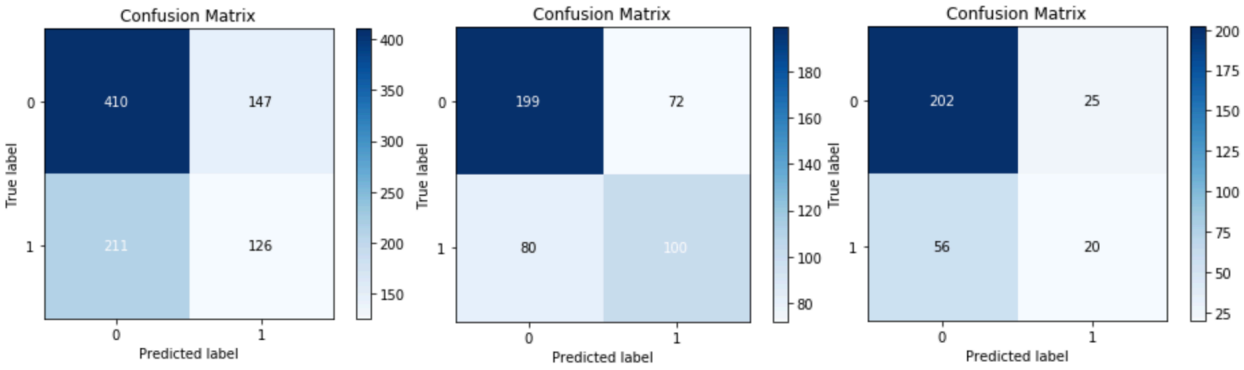


(Figure 6 FP and FN for each threshold for the two race groups)

However, the results of false positive rate and false negative rate for each threshold shows that even if we re-evaluate the threshold to classify risk categories for different race groups and find new threshold values that maximize the predictive accuracy to the two groups on the same level,

the COMPAS tool systematically imposes high FN and low FP on the black population whilst giving high tolerance on the white defendants.

3.1.3 Machine Learning on Classified Race Groups



Both Races Black Only White Only

From the confusion matrix color we can see that compared to mixing the races together, removing the race feature as a protected attribute gives better prediction results.

	Accuracy	False postive	False negative
black&white	59.96%	16.44%	23.60%
black	65.74%	15.96%	17.74%
white	71.31%	8.25%	18.48%

Separating races to conduct modelling increases predictive accuracy whist reducing FP, FN for both racial groups. Although the overall prediction performances are improving after splitting, the black still suffer from nearly two times false positive rate compared to the white.

Conclusion and Discussion

This study first verifies that the COMPAS tool is racial biased in assessing the likelihood of recidivism. The two methods to mitigate bias are effective in different ways. Threshold selection method gives results that threshold = 3 for the white and threshold = 5 for the black might be almost equal in the predictive accuracy to be highest and close, but the goal to satisfy calibration and statistical parity and achieve similar FP and FN cannot be realized at the same time. Also, if we reset the threshold at a low score level to be 'high risks of recidivism', the resources spent on these potential recidivists may be huge. Modification on the classifiers cannot meet all algorithm fairness expectation. Therefore, instead of make adjustments on current score standards inside the COMPAS system, we split the race as a protected attribute in machine learning prediction. Since the algorithm used in COMPAS is confidential, we use decision tree as an explainable machine learning method to test if this separation gives better results. We find that separating race for individual machine learning improve accuracy, reduce FP and FN for both groups. It also reduces the black to white FP/FN ratio compares to the original COMPAS results. But strict algorithm fairness of equal statistical parity and calibration has not been satisfied as the disparity in FP for the black and white group is still large.

The improvement from threshold reselection and machine learning separation gives indication that mixing the races together for prediction may not provide a fair algorithm for recidivism prediction. On the one hand, the imbalanced data for each group in training set can itself lead to bias. On the other hand, setting thresholds or modelling individually may better reflect correlations of other race-related variables in each computation, which can reduce the implicit effect from race. Future research on this area can consider a combination of these two methods.

The limitation in this study lies in several aspects. One is that although we used cross validation to enhance the prediction result, the size of the subset data for the two race groups after splitting is much smaller, which might influence the predictive accuracy. Another problem is on the algorithm confidentiality of COMPAS. We have limited knowledge on what machine learning method the company is actually using, therefore,

we cannot directly make modifications on the COMPAS algorithm but only select a highly-interpretable model to compare and evaluate the effectiveness of the protected-attribute separation modelling method. Thirdly, we might need to identify race-correlated variables to resolve the false positive rate disparity.

Reference

Attewell, P., & Monaghan, D. (2015). Data mining for the social sciences: An introduction. Oakland: University of California Press.

Berk, R. A. (2012). Criminal justice forecasts of risk: A machine learning approach. New York, NY: Springer.

Blumstein, A., Cohen, J., & Farrington, D. P. (1988). Criminal career research: Its value for criminology. Criminology, 26(1), 1–35.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition, 30(7), 1145–1159.

Brennan, T., Dieterich W., Ehret, B. (2009) Evaluating the Predictive Validity of the Compas Risk and Needs Assessment System. Available at: https://doi.org/10.1177/0093854808326545

Davies, S.C. & Goel, S. (2018) The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning.

Ducro, C., & Pham, T. (2006). Evaluation of the SORAG and the Static-99 on Belgian sex offenders committed to a forensic facility. *Sexual Abuse: A Journal of Research and Treatment*, *18*(1), 15.

Hardt, M., Price, E., and Srebro, N (2016). *Equality of opportunity in supervised learning.* CoRR, abs/1610.02413.

Hare, R.D. (1991). Manual for the revised Psychopathy Checklist. Toronto: Multi-Health Systems.

Langan, P., & Levin, D. (2002). Recidivism of prisoners released in 1994. Washington, DC. Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). How We Analyzed the COMPAS Recidivism Algorithm**.** Available at: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Piquero, A., Farrington, D., & Blumstein, A. (2003). The criminal career paradigm. Crime and

Justice, 30, 359–506.

Quinsey, V.L., Harris, G.T., Rice, M.E. & Cormier, C.A. (2006) 2nd Ed. Violent Offenders: Appraising and Managing Risk. Washington D.C: American Psychological Association.

Rettenberger, M., & Eher, R. (2007). Predicting reoffense in sexual offender subtypes: A prospective validation study of the German version of the Sexual Offender Risk Appraisal Guide (SORAG). *Sexual Offender Treatment*, *2*(2), 1-12

Rice, M.E., Harris, G.T. & Quinsey, V.L. (1990). A foloowup of rapists assessed in a maximum security psychiatric facility. Journal of Interpersonal Violence, 5, 435-448

Sentencing Guidelines Commission State of Washington (2008). Recidivism of Adult Felons

Serin, R. C., Lloyd, C. D., Helmus, L., Derkzen, D. M., & Luong, D. (2013). Does intraindividual change predict offender recidivism? Searching for the Holy Grail in assessing offender change. Aggression and Violent Behavior, 18(1), 32–53.

Urahn, S. (2011). State of recidivism: the revolving door of America's prisons. The PEW Center on the States.

U.S. Sentencing Commission. (2005). A comparison of the federal sentencing guidelines criminal history category and the U.S. Parole Commission Salient Factor Score. Washington, DC.

United States Sentencing Commission (2019). *Recidivism Among Federal Offenders: A Comprehensive Overview*. Available at: https://www.ussc.gov/research/research-reports/recidivism-among-federal-offenders-comprehensive-overview

Yang, M., Liu, Y., & Coid, J. (2010). Applying neural networks and other statistical models to the classification of serious offenders and the prediction of recidivism. UK Ministry of

Justice Research Series 6/10.