Mitigating Racial Bias in Recidivism Prediction: Machine Learning & Threshold Selection Approach

Yijia Lyu

Division of Social Science, University of Chicago, lugalyu@uchicago.edu

Introduction

- Crime has been a significant disruptive factor in social security and our daily life. Keeping track of crime occurrences is therefore essential to build safer neighborhoods.

- Many jurisdictions in United States to have adopted predictive tools to estimate defendants' likelihood of recidivism and use it as a reference to determine final sentencing on criminals. But whether such tools provide fair assessment on different racial groups is debatable.

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

Research Questions

a) Is COMPAS racially biased when scoring defendants?
b) Is current risk category setting fair? If not, how should the thresholds change?
c) How might protected attributes (race/gender) separation improve the prediction result in ways that reduce false predictions?

Key Definitions

• Recidivism

Recidivism in this study as a new arrest within two years after release (U.S. Sentencing Commission, 2019).

• High, Medium, Low Decile Score

COMPAS categories its decile score as follows:

Score 1-4 low risks ; 5-7 medium risks ; 8-10 high risks

• Goal of Algorithm Fairness

- Statistical Parity

Statistical features(classification error, false positive rate, false negative rate, precision, etc.), should be equal across groups when classified with protected attributes

- Calibration

Predictive outcomes should be independent of protected attributes like demographic features, for example, race or gender. A score s = s(x) is calibrated if it satisfies the following equation:

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

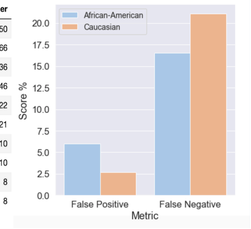
Data & Methods

• Data

Broward County COMPAS records:

11757 criminal suspects records who were assessed at the pretrial stage and scored for recidivism risk using the COMPAS in Broward County during 2013 and 2014.

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



• Methods

-COMPAS data checking

1) Examine COMPAS score distribution for each race group as displayed above

2) Compare the FP and FT for the black defendants and the white defendants as displayed above

-Threshold Selection

Use iterative threshold selection to check if COMPAS's score classifier for high, medium, low risks is the optimal and fair towards different race groups.

1) Iterate threshold computation on each decile score

2) Compare the corresponding accuracy, FP, FT for each race group

3) Following the statistical parity and calibration as algorithm fairness criteria, decide the optimal threshold as risk classifier

4) 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

1) Separate the data into three datasets on both races, black only, white only

2) Use Decision Tree model to train and predict the recidivism result on the dataset with both races

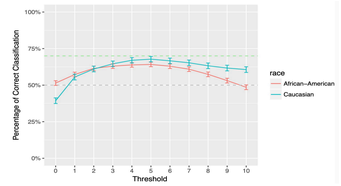
3) Use cross validation to offset the dataset size limitation after data splitting for the black and the white

4) Use Decision Tree model to train and predict the recidivism result on the black group and the white group individually

5) Compare the three training results and evaluate if separating protected attributes (race) mitigates the bias

Results

• Threshold Selection



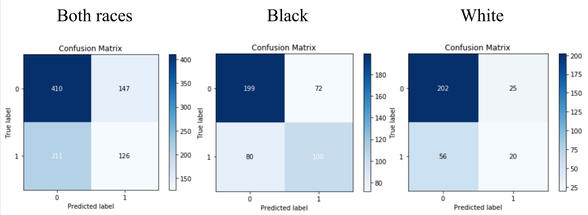
(COMPAS Predictive Accuracy on each threshold)

Accuracy of threshold = 5 for the black population is closest to accuracy of threshold = 3 for the white. But the highest accuracy for both groups lies when threshold=5.

(COMPAS Predictive Accuracy on each threshold)

The black community always suffer higher FP and lower FN compared to the white group in all thresholds.

• Decision Tree



(Confusion Matrix of Modelling Results)

From the confusion matrix color, removing the race feature as a protected attribute for prediction produces better prediction results compared to mix the races together.

	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%

(Statistics of Modelling Results)

Separating races to conduct individual decision tree modelling increases predictive accuracy and reduces 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.

Conclusions

1)The COMPAS tool is significantly racial biased in assessing the likelihood of recidivism.

2) Threshold Selection:

Threshold = 3 for the white and threshold = 5 for black in the current COMPAS system satisfies equal predictive accuracy, but cannot produce equal or close predictions for the two race groups.

Setting the threshold at 3 or 5 would burden the prison resources. Modification on the classifiers cannot meet all algorithm fairness expectation.

3) Separate Machine Learning:

Separating race for individual machine learning improve accuracy, reduce FT and FN for both groups.

It also reduces the black to white FT/FN ratio compares to the COMPAS results.

But strict algorithm fairness of equal statistical parity and calibration has not been satisfied.

Future Improvements

1) Combining threshold selection and protected attributes separation algorithm training.
2) Race-related feature filtering
3) More data sources

References

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

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

Kleinberg, J. M., Mullainathan, S., and Raghavan, M (2016). Inherent trade-offs in the fair determination of risk scores. CoRR, abs/1609.05807, 2016.

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.

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

Acknowledgements

I sincerely thank Dr. Richard Evans for inspiring me to work on this meaningful topic. His kind instructions and helpful suggestions throughout this course help me to maintain enthusiasm and curiosity in conducting this research.