Machine Judges Reduce Sentencing Bias? A Computational Social Science Evaluation

Mingyang Chen, Zhipeng Wu

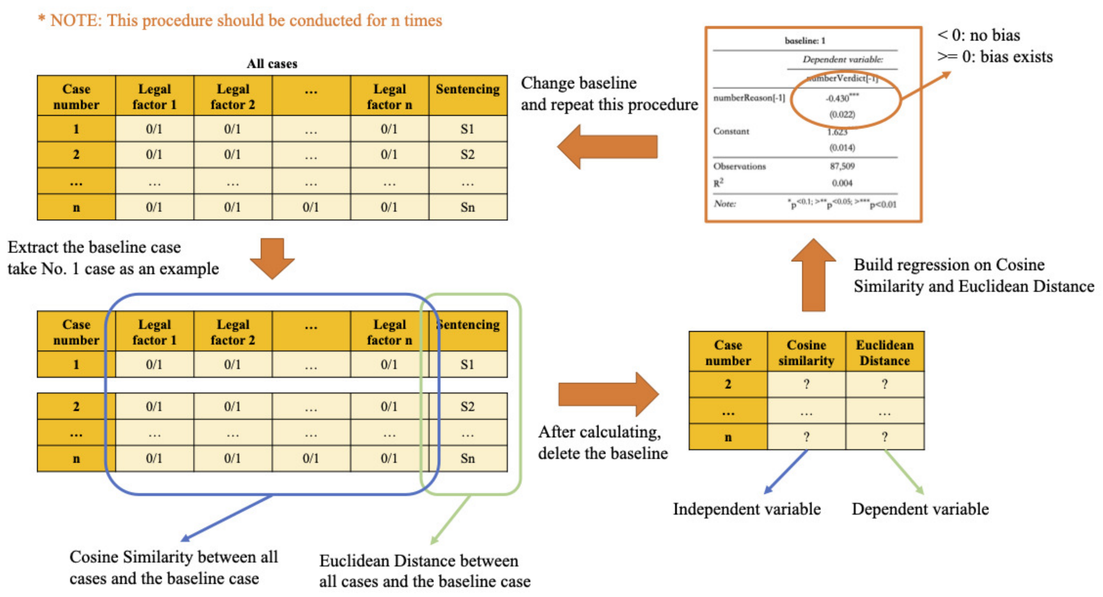
University of Macau, Chongqing University

Objective

Machine learning models have been applied in many criminal justice decisions, and prior research has proved that machine learning models can reduce biases if they are blind. However, prior research focuses on classification tasks in criminal justice. Regression tasks' disparity are much more difficult to be evaluated. Prior research on sentencing bias evaluation only focus on systematic biases and ignore the individualized biases in cases. In this study, we focus on the sentencing task.

Methods

We propose a new method to evaluate whether an individual case is biased based on comparing it with all other cases based on the theory of "Treating Like Cases Alike". We collect all 238,419 theft cases and extract the legal factors and sentencing results. 159,699 cases are used for building a machine learning model, and we test our model' ability of reducing biases on the rest 78,720 cases. We use XGBoost to train our model. We use RR and OR to compare the results of machine judges and human judges. Besides, we let human judges cooperate with machine, which is, only when humans make a wrong decision, then we use machine to adjust the results. Below figure shows the method we developed.



Results

By employing the method, We find if all judges are replaced by machine learning models, the probability of being sentenced an unfair result is 35% lower; if cooperating with judges, 55% biased cases can be sentenced in a more fair way. Machine learning models can reduce individualized biases. Machine can produce new biased cases even when they are blind.

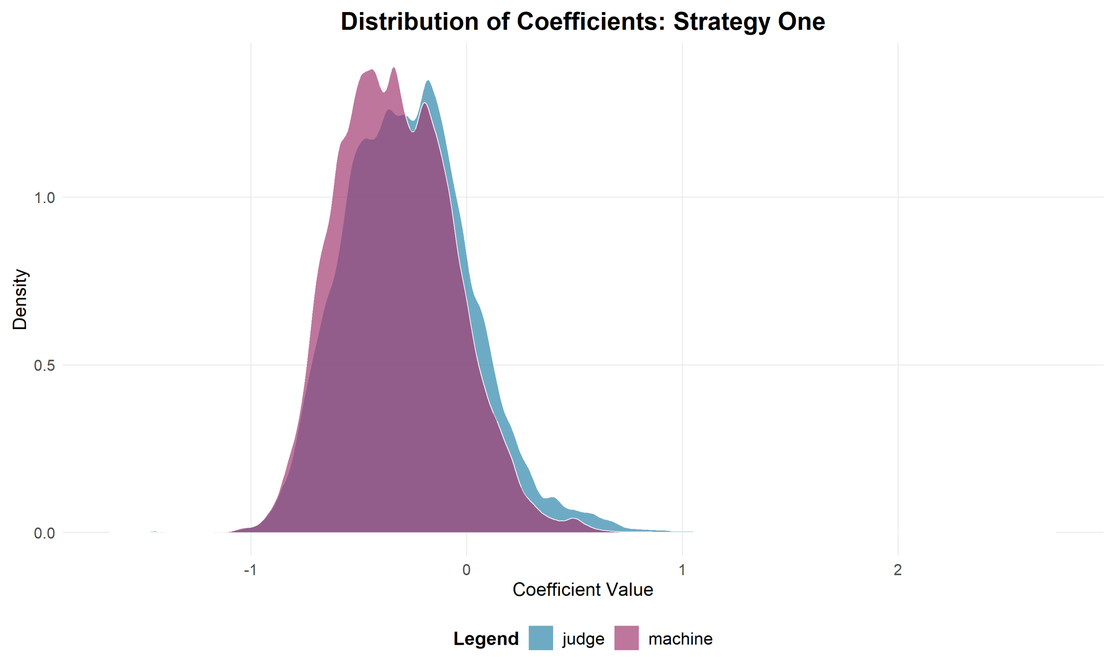
Table 1: Machine v.s. Judges: Strategy One

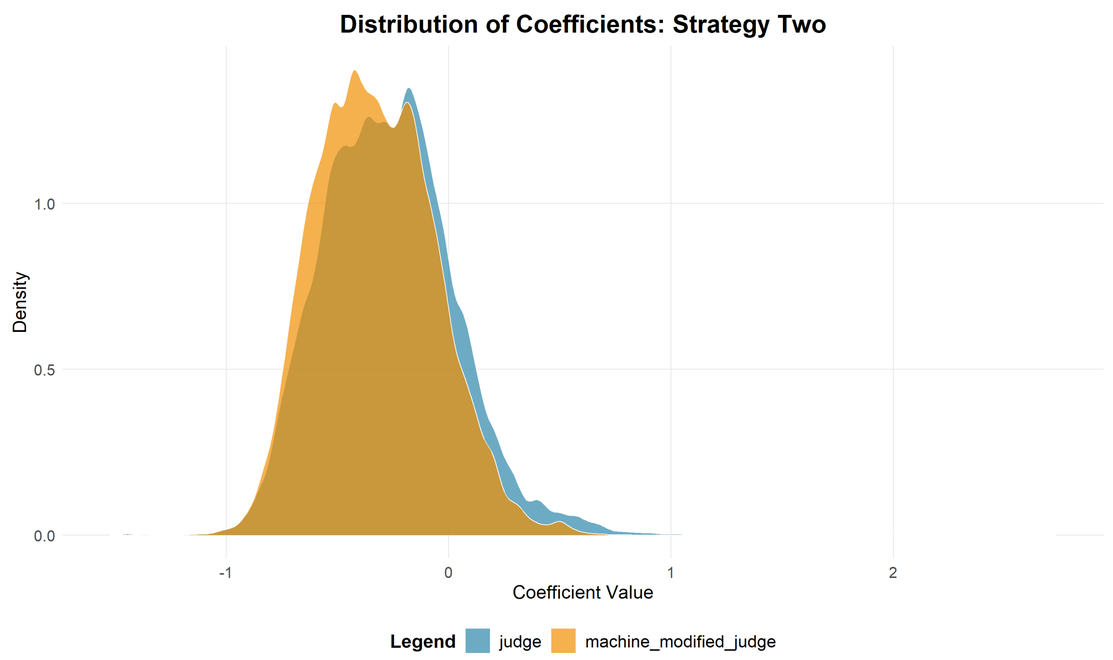
Strategy	RR (95% CI)	OR	Decision	Biased	Unbiased	Total
Strategy One	0.65 (0.64 to 0.67)	0.61 (0.59 to 0.63)	Machine	8644	70076	78720
			Human	13217	65503	78720
Strategy Two	0.45 (0.44 to 0.47)	0.41 (0.40 to 0.42)	Machine	5992	72728	78720
			Human	13217	65503	78720

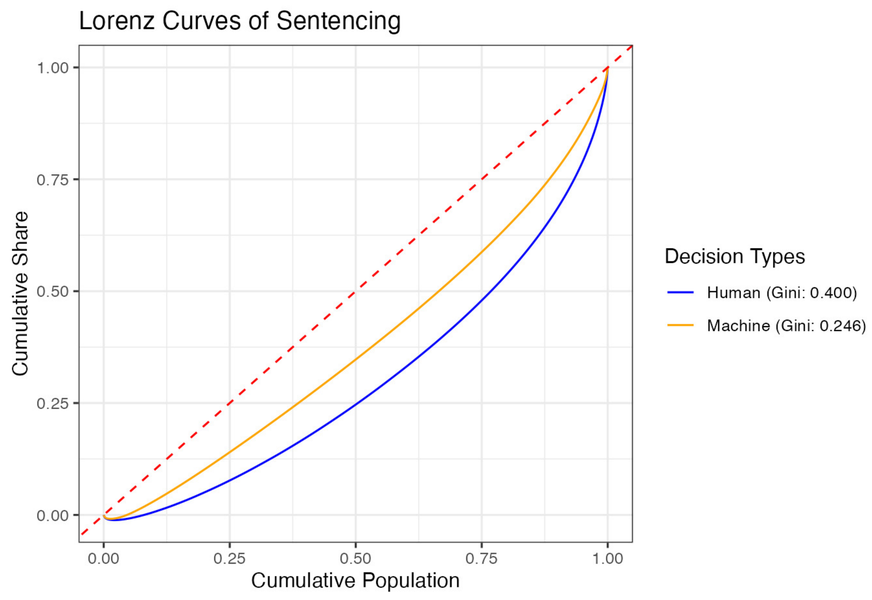
Contact Information

- mc55649@um.edu.mo
- 202530131014T@stu.cqu.edu.cn

Below figures show the visualized results. Strategy One means no cooperation; Strategy Two means machine-human cooperation. The last figure is the result of Gini Coefficient, which is a robustness check.







Conclusion

This paper proposes a new method for evaluating sentencing disparity for machines. Machine judges can reduce sentencing disparity, but can also produce new biases. Overall, machine judges perform better than humans. However, the new disparity is worth considered that whether a CJ system should employ machine learning models and legal researchers need to provide a cost-effectiveness analysis on using machines.