Machine Learning Enhanced RDD Study

David Jones, Paul Dahm, Jonas Ater, Espen Hovland
Texas A&M University
April 24, 2024

Abstract

Using gradient boosting and cross validation to determine optimal bandwidth, we enhanced the study, "Punishment and Detterrence: Evidence From Drunk Driving" by Benjamin Hansen. The study uses the different thresholds for a DUI (0.08 BAC) and aggravated DUI (0.15 BAC) as cutoffs for a Regression Discontinuity Design analysis in order to examine recidivism (the tendency of a criminal to reoffend).

Machine Learning Method

- Hansen used an RDD bandwidth of 0.05 around the BAC cutoffs.
- We used gradient boosting to model the relationship between BAC levels and recidivism within varying bandwidths around both DUI and Aggravated DUI thresholds
- We then used cross validation (through a grid search) to optimize the gradient boosting model, this systematically tested different combinations of the G-Boost parameters across multiple folds to find the combination that minimizes error.

Findings

Table: Bandwidth Comparison

Independent Variable	Coefficient
DUI threshold w/ original bandwidth (0.05)	-0.026699
Aggrevated DUI w/ original bandwidth (0.05)	-0.00293
DUI threshold w/ optimal bandwidth (0.027222)	-0.021628
Aggravated DUI w/ optimal bandwidth (0.027222)	0.002516

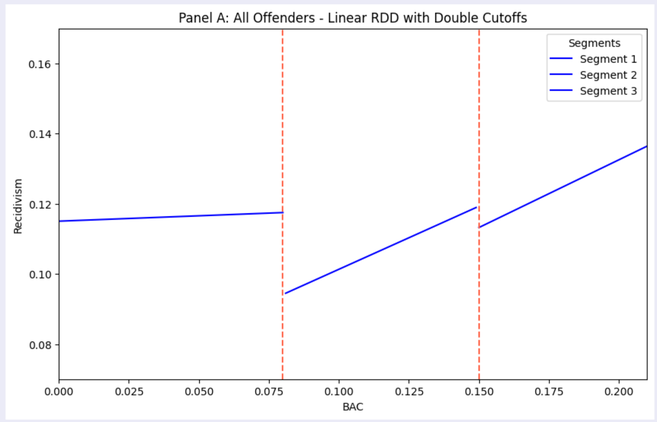

Figure: Original Cutoff

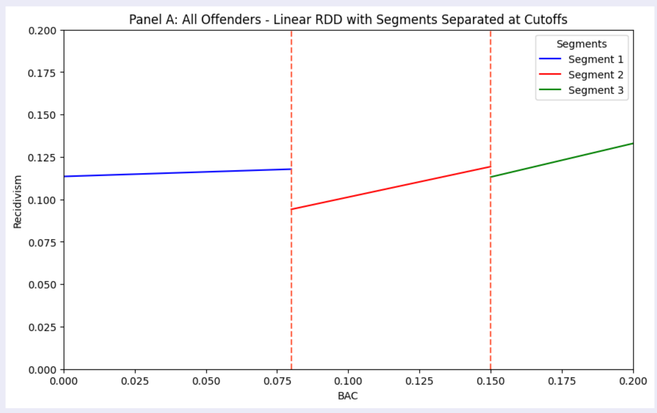

Figure: Optimal Cutoff

Literature Review

- Large-scale evaluation of k-fold cross-validation ensembles for uncertainty estimation (Dutschmann et al. 2023)
- SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer (Prusty et al. 2022)

Results

- After applying the optimal bandwidth, the coefficient for chance of recidivism after the DUI threshold is lower than under the initials study's 0.05 bandwidth.
- After using the optimal bandwidth, the coefficient around aggravated DUI actually became positive

References

- Github link
- Colab link