Optimizing Bandwidth with Gradient Boosting and Cross-Validation in RDD: A Study on Drunk Driving Deterrence

By Paul Dahm, David Jones, Jonas Ater, and Espen Hovland

Intro

The paper titled "Punishment and Deterrence: Evidence from Drunk Driving" by Benjamin Hansen examines the effect of increased legal penalties on drunk driving behaviors, using regression discontinuity to estimate the impact of blood alcohol content, or BAC, thresholds on DUI recidivism, which is the likelihood to reoffend. Hansen's analysis leverages a substantial dataset from Washington state involving over 500,000 DUI stops, assessing whether higher sanctions at specific BAC levels effectively deter repeat offenses.

This study situates itself within the broader criminological discourse, as informed by Gary Becker's economic theory of crime, which posits that criminals engage in offenses based on a rational cost-benefit analysis. By increasing the cost of crime via stiffer penalties, law enforcement could theoretically reduce criminal activity. Hansen finds that drivers just above the BAC thresholds, which legally categorize them as drunk, tend to have lower rates of recidivism. Specifically, the paper presents evidence suggesting that having a BAC above the standard legal limit reduces repeat offenses by up to 2 percentage points, and an even higher BAC—above the aggravated DUI threshold—further reduces recidivism by an additional percentage point.

These results are significant in shaping policies regarding the legal limits of blood alcohol concentration for drivers and the structuring of penalties for violations. By demonstrating a clear link between increased sanctions and decreased likelihood of reoffense, Hansen's research supports the efficacy of punitive measures in deterring drunk driving, contributing valuable insights to the ongoing legislative and public debates on optimal strategies to combat this persistent social issue.

The aspect of this paper we want to improve upon is how the bandwidths are chosen. In Hansen’s paper, he used arbitrary RDD bandwidths of 0.05 and 0.025 around the BAC cutoffs, giving no explanation for why these bandwidths were chosen for his work. Using gradient boosting and cross-validation to determine optimal bandwidth we can enhance the study, hopefully leading to more meaningful results.

Literature review

Our decision to use cross validation was influenced by the 2023 cheminformatics paper “Large-scale evaluation of k-fold cross-validation ensembles for uncertainty

estimation” by Thomas-Martin Dutschmann, Lennart Kinzel, Antonius Ter Laak, and Knut Baumann, to learn more about cross-validation and how it can enhance studies or research. The goal of this study was to use machine learning models to help evaluate the uncertainty of compound properties predictions that were found using the ensemble method. The focus on determining the uncertainty level in each prediction is extremely important in determining how to use the predictions, especially in the field of cheminformatics, which often leads to new drug design.The paper uses k-fold cross-validation to analyze these levels of uncertainty for predictions, and was able to gather an accurate understanding for how reliable each prediction was showing that cross validation was a great way to optimize predictive models.

In the 2022 cancer research paper, "SKCV: Stratified K-fold Cross-Validation on ML Classifiers for Predicting Cervical Cancer" by Sashikanta Prusty, Srikanta Patnaik, and Sujit Kumar Dash, we found that using gradient boosting with cross validation provides more accurate results than just gradient boosting. This paper uses data on patients including demographic, behavioral, and clinical features known to be associated with cervical cancer risk, and uses several different machine learning methods to test their predictive capabilities. They then use cross validation to optimize and score the different methods. In their results, they found that using extreme gradient boosting or XGB with cross validation to be among the most successful methods in providing an accurate prediction that a patient may have/be at risk of cervical cancer.

Explain Machine Learning

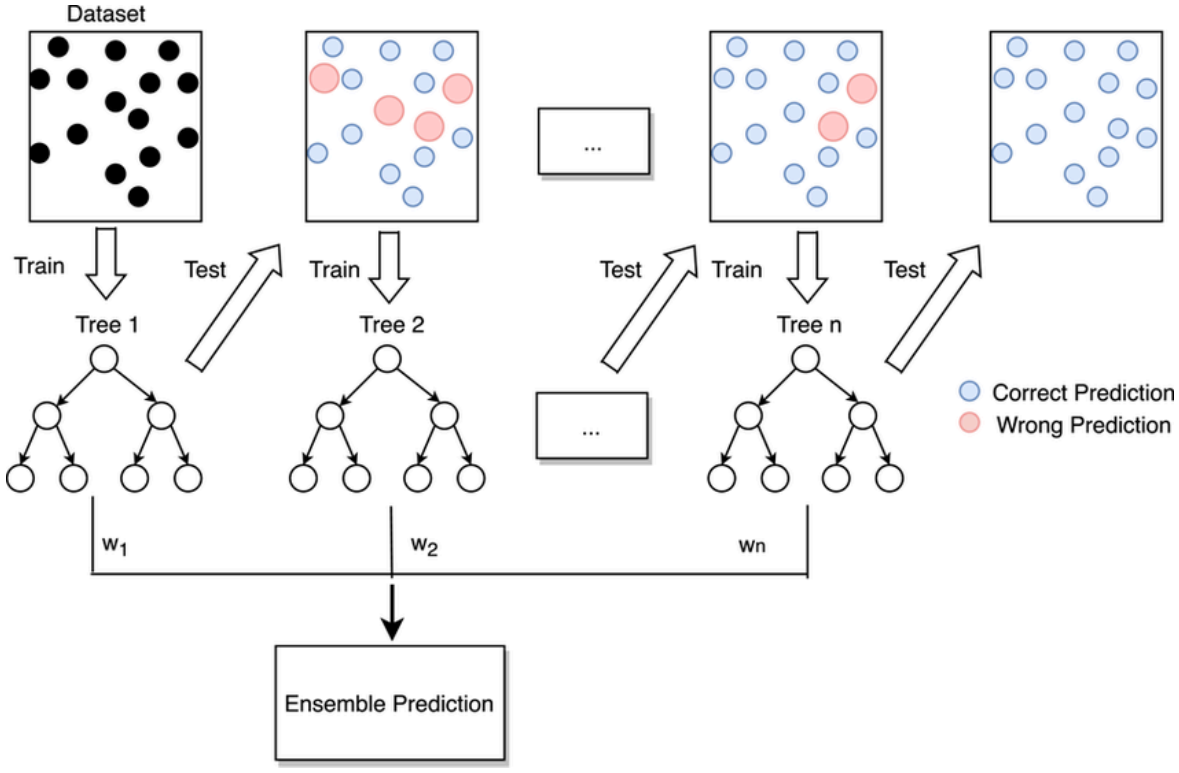
In an RDD study, bandwidths are essential to providing relevant results. In our enhancement of the original study, we used machine learning to determine the optimal bandwidth. The process starts by defining a range of potential bandwidths. These bandwidths determine the subset of data considered around each threshold (e.g., 0.08 and 0.15 BAC levels). Each bandwidth essentially creates a different "window" of data around the threshold.

To begin finding the optimal bandwidth, we began with gradient boosting. The gradient boosting regressor is applied to model the relationship between BAC levels and recidivism within varying bandwidths around both the DUI and Aggravated DUI thresholds. Gradient Boosting begins with a simple prediction and then uses negative mean squared error to calculate the residual error. It then accounts for that error by making a new model that is the original prediction except it is now fitted to these residuals (this is done by adding a decision tree). The results of the new model are

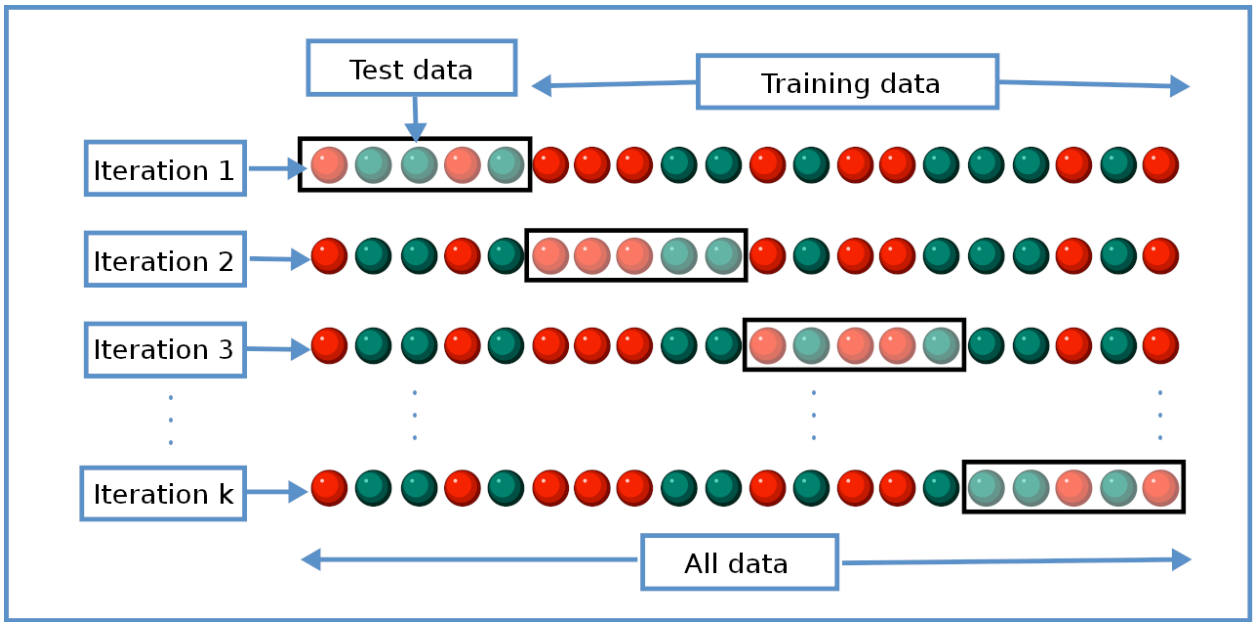
scaled by the learning rate and then added to the predictions of the existing model. This process is repeated until the improvement in the model prediction decreases across a certain threshold.

We then used k-fold cross validation combined with grid search to optimize the gradient boosting model. These tune the model parameters (like the number of trees, depth of each tree, learning rate, etc.) to optimize performance for each bandwidth. The various bandwidths found with gradient boosting are separated into folds. For this study we used 10 folds, the model then trains on 9 folds and validates against the 10th fold. This process is repeated for each fold, by training on 9 folds, then testing against the 9th fold, 8th fold, and so on. The performance of each model is scored with negative mean squared error, and the bandwidth that yields the best average score (i.e. the lowest error) is chosen as the optimal bandwidth. Using cross validation on top of gradient boosting is a good way to avoid overfitting, and ensuring the model's performance is not just the result of the peculiarities of one particular bandwidth.

Gradient Boosting Visualization
By Zhang, Tao, et al.- https://doi.org/10.1029/2020MS002365



Cross Validation Visualization
By Gufosowa - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=82298768



Results

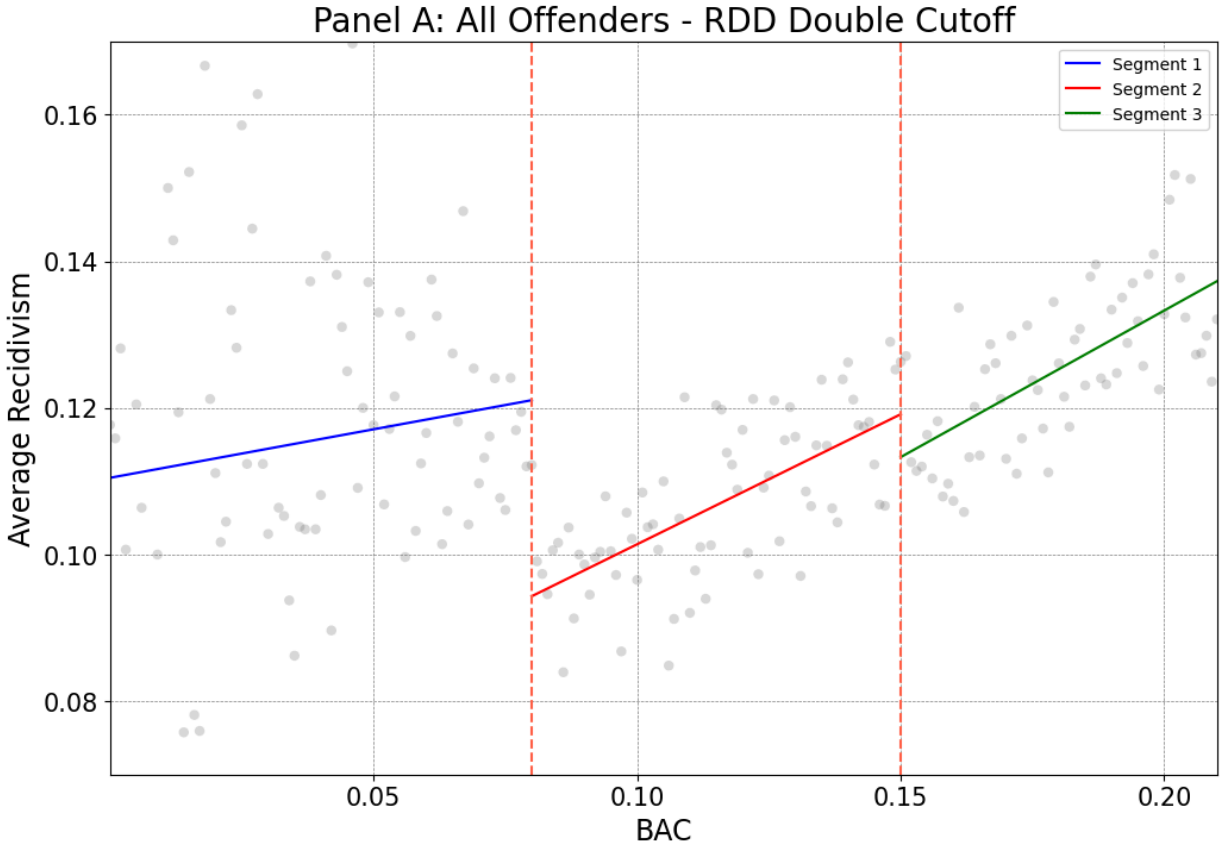
The findings of our replication showed that the estimated effect of getting a DUI on recidivism is -5.82 percentage points when the bandwidth is 0.05. This can be interpreted to mean that if someone is convicted of a DUI, the likelihood of them reoffending decreases by 5.82 percentage points. However, when using a bandwidth of 0.025, the estimated effect was found to be -6.83 percentage points. Finally, when using the optimal bandwidth, the estimated effect was found to be -4.2 percentage points. It is important to note that the estimated effect using the optimal bandwidth yielded a non-significant result. This could be due to unresolved errors in the code, as this was the only estimate that was non-significant. This poses a problem with our machine learning enhancement for the effect of DUI on Recidivism.

The findings of our replication also showed that the estimated effect of getting an Aggravated DUI on recidivism is -0.42 percentage points when the bandwidth is 0.05. This can be interpreted to mean that if someone is convicted of an Aggravated DUI, the likelihood of them reoffending decreases by 0.42 percentage points. However, when the bandwidth is 0.025, the estimated effect was found to be -0.55 percentage points. Finally, when using the optimal bandwidth, the estimated effect was found to be -0.57 percentage points. Note that all of the estimates for the effect of Aggravated DUI on recidivism were all statistically significant.

These estimates were calculated using a Weighted Least Squares (WLS) model. In a WLS model, the closer an observation is to the cutoff, the more “weight” it carries for the estimated effect. While the estimates we calculated did not exactly match the

estimates presented in the paper, our estimates are consistent with those in the paper. In the paper, the estimated effect of DUI on recidivism is larger than the estimated effect of Aggravated DUI on recidivism, and our results are consistent with this. Reasons for the difference in our estimates and the paper's estimates could be that our data source was a slightly altered version of the original data, with correct data that was missing a few columns containing explanatory variables that the original authors used.

Replication Results for Recidivism (bandwidth = 0.05)



Conclusion

We found that the threshold for DUI (0.08) has a greater effect on the reduction in recidivism than the threshold for Aggravated DUI (0.15). Every estimate I calculated yielded a statistically significant result, with the exception of the threshold for DUI at the optimal bandwidth we calculated. I employed a Weighted Least Squares model, which gives higher "weight" to values closer to the cutoff. This is a common method used in RDD analysis.

Our results show the effect of the 0.08 threshold is stronger than the effect of the 0.15 threshold. This can most likely be attributed to how offenders who fall below the limit of an aggravated DUI, are more likely to be non-malicious, being unaware they are above the legal limit or have their decision making impaired by their BAC. After using the optimal bandwidth, only the estimate for the effect of Aggravated DUI was statistically significant.

Citations:

Hansen, Benjamin. “Punishment and Deterrence: Evidence from Drunk Driving.” American Economic Review, vol. 105, no. 4, Apr. 2015, pp. 1581–617, doi:10.1257/aer.20130189.

Scott Cohn, (2021). hansen2015_RDD_replication [R]. GitHub. https://github.com/scottcohn97/hansen2015_RDD_replication/tree/main

Dutschmann, Thomas-Martin, et al. “Large-Scale Evaluation of k-Fold Cross-Validation Ensembles for Uncertainty Estimation.” Journal of Cheminformatics, vol. 15, no. 1, Apr. 2023, doi:10.1186/s13321-023-00709-9.

Prusty, Sashikanta, et al. “SKCV: Stratified K-Fold Cross-Validation on ML Classifiers for Predicting Cervical Cancer.” Frontiers in Nanotechnology, vol. 4, Aug. 2022, doi:10.3389/fnano.2022.972421.

Zhang, Tao, et al. "Improving Convection Trigger Functions in Deep Convective Parameterization Schemes Using Machine Learning." *Journal of Advances in Modeling Earth Systems*, vol. 13, 2021, e2020MS002365. https://doi.org/10.1029/2020MS002365

Contributor Gufosowa."K-fold Cross Validation EN." *Wikipedia*, Wikimedia Foundation, https://en.wikipedia.org/wiki/Cross-validation_%28statistics%29#/media/File:K-fold_cross_validation_EN.svg.