Document Title: National Institute of Justice Recidivism Forecasting Challenge: Team "MCHawks" Performance Analysis

Author(s): Giovanni Circo, Andrew Wheeler

Document Number: 305050

Date Received: July 2022

Award Number: NIJ Recidivism Forecasting Challenge Winning Paper

Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

National Institute of Justice Recidivism Forecasting Challenge

Team “MCHawks” Performance Analysis

Giovanni Circo

University of New Haven

Andrew Wheeler

HMS

Introduction

In April 2021 the National Institute of Justice (NIJ) announced a recidivism forecasting challenge. The goal of this challenge was for participating teams to develop a model to predict future recidivism among a sample of persons released from prison to parole. The primary outcome was prediction of re-arrest for any crime at 1, 2, and 3 years post-release. A secondary outcome was the development of a “fair and accurate” model that balanced false-positive predictions between Black and White individuals. We entered this competition in the small-team category under the name “MCHawks”. Our team’s prediction model placed in the following categories:

- First Place: Year 2, Male Parolees
- First Place: Year 2, Average Accuracy
- Second Place: Year 2, Female Parolees

We also placed second, third, and fifth among the “Accounting for Racial Bias Category” in the Year 1 and Year 2 categories. Below, we detail the development of our models, discuss the features used in building the predictions, and the interpretation of the results. Under our future considerations section, we discuss how predictive modelling can best be applied and how subsequent competitions can better help guide the development of fair and accurate algorithms.

Models

Predicting recidivism presents a difficult challenge under any circumstance, and the development of a reasonable and actionable model presents further challenges. The first major question is the selection of an appropriate modeling strategy. A wide variety of models exist for prediction purposes – including ordinary least-squares regression, logistic regression, generalized additive models, support vector machines, random forests, and gradient boosted models – among many others (James, et al., 2013). While logistic regression remains the most popular and tends to work well in many circumstances (see: Christodoulou et al., 2019) one of its limitations is that it assumes linearity in all model predictors. In cases where there are many potential complex interactions, logistic regression risks underfitting the data. One alternative are tree-based methods such as regression or classification trees. These models are often easy to explain and have more flexibility than classical approaches. However, individual trees can be very non-robust – meaning small changes in the data can cause large changes in the final prediction (James, et al., 2013). Ensemble approaches, like bagging and boosting, represent an approach where many small "weak learners" combine to make a single more effective model. Of these, boosting is a general-purpose approach for improving predictive performance and which can easily be applied to classification trees. A boosted classification tree is built iteratively by fitting feature splits on residuals from prior models. This approach results in many "shallow" trees that, combined, often perform better than a single large tree.

Our winning model for round 2 was fit using a gradient boosted decision tree employing XGBoost (Chen & Guestrin, 2016). XGBoost is a widely used algorithm for fitting gradient boosted models (GBM) and has a strong track record in various prediction competitions. Prior to fitting the model, we first expanded the full training dataset from NIJ into a $18{,}028 \times 135$ data matrix. For the year 2 predictions we filtered out individuals who were arrested at time 1, leaving us with 12,651 observations. No additional variables were added or constructed from the existing data. The first step in fitting a GBM is the tuning of the model parameters. This involves adjusting key parameters in a way to minimize overfitting (learning too much from the training data) and underfitting (learning too little from the training data). In more general terms, this step reflects a strategy of balancing bias and variance (James, et al., 2013). We employed a grid search strategy, where we iteratively fit models using different combinations of the major XGBoost parameters and calculated an estimate of out-of-sample performance via 5-fold cross validation. The most important parameters tuned were:

- max_depth: 6
- eta: 0.005
- min_child_weight: 5
- max_delta_step: 5
- subsample: .75

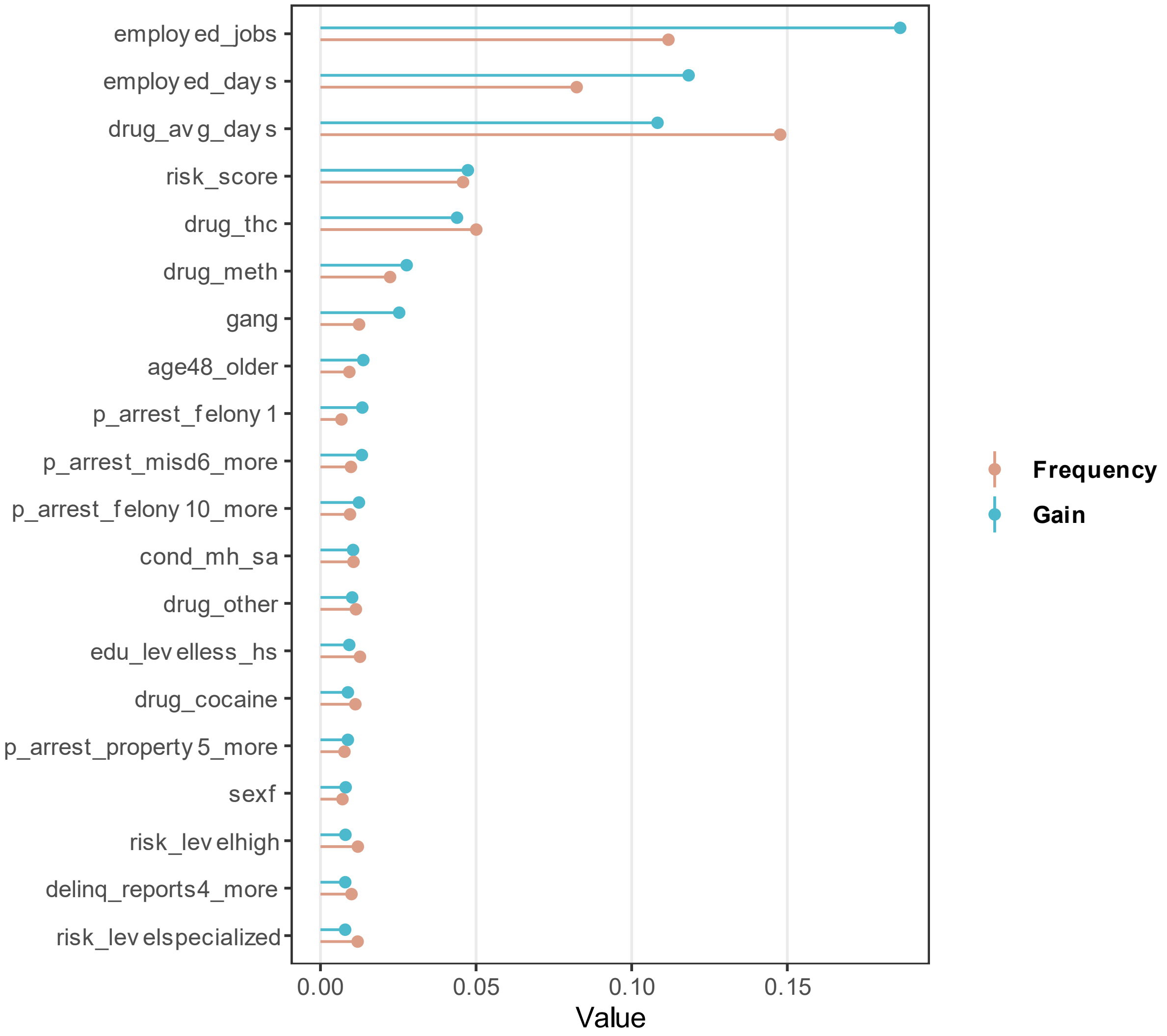
max_depth and min_child_weight control the overall model complexity, eta controls the stepsize per iteration, and max_delta_step and subsample help control overfitting. Rather than attempting to optimize the Brier score directly, we used area under the curve (AUC) to determine the optimally fitting model. The AUC value for this model was approximately .74. We then used this dataset to make predictions on the full training dataset.

Variables

Feature Importance

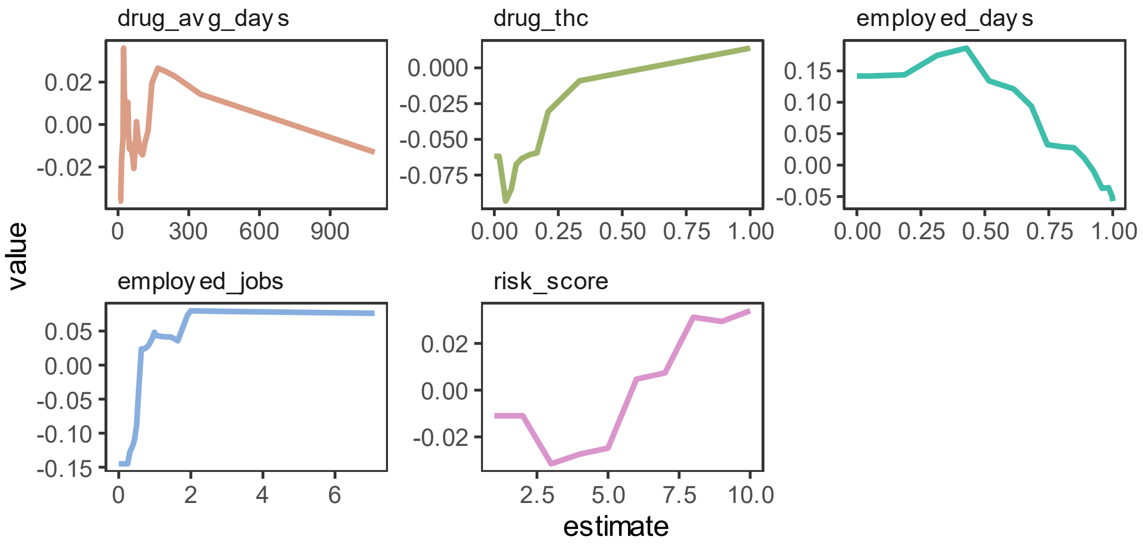
While the final model was fit on all 135 features, only a small subset of these features were especially useful for the final prediction. Because the model used does not rely on the conventional frequentist framework, we cannot evaluate whether any individual predictor is statistically "significant." However, we can evaluate the individual contribution to the final prediction for each feature. To accomplish this, we evaluated the gain and frequency for each feature. Gain represents the proportional contribution of a feature to the overall model, based on the reduction in overall error for each of the feature's splits. Frequency is the percentage of times a feature is present in a tree (Molnar, et. al., 2018). Together, these indicate which features are most important to the final prediction. Below, Figure 1 displays the gain (blue) and frequency (orange) of the top 20 predictors in our model arranged by gain.

Figure 1. Top 20 Features, by Gain and Frequency



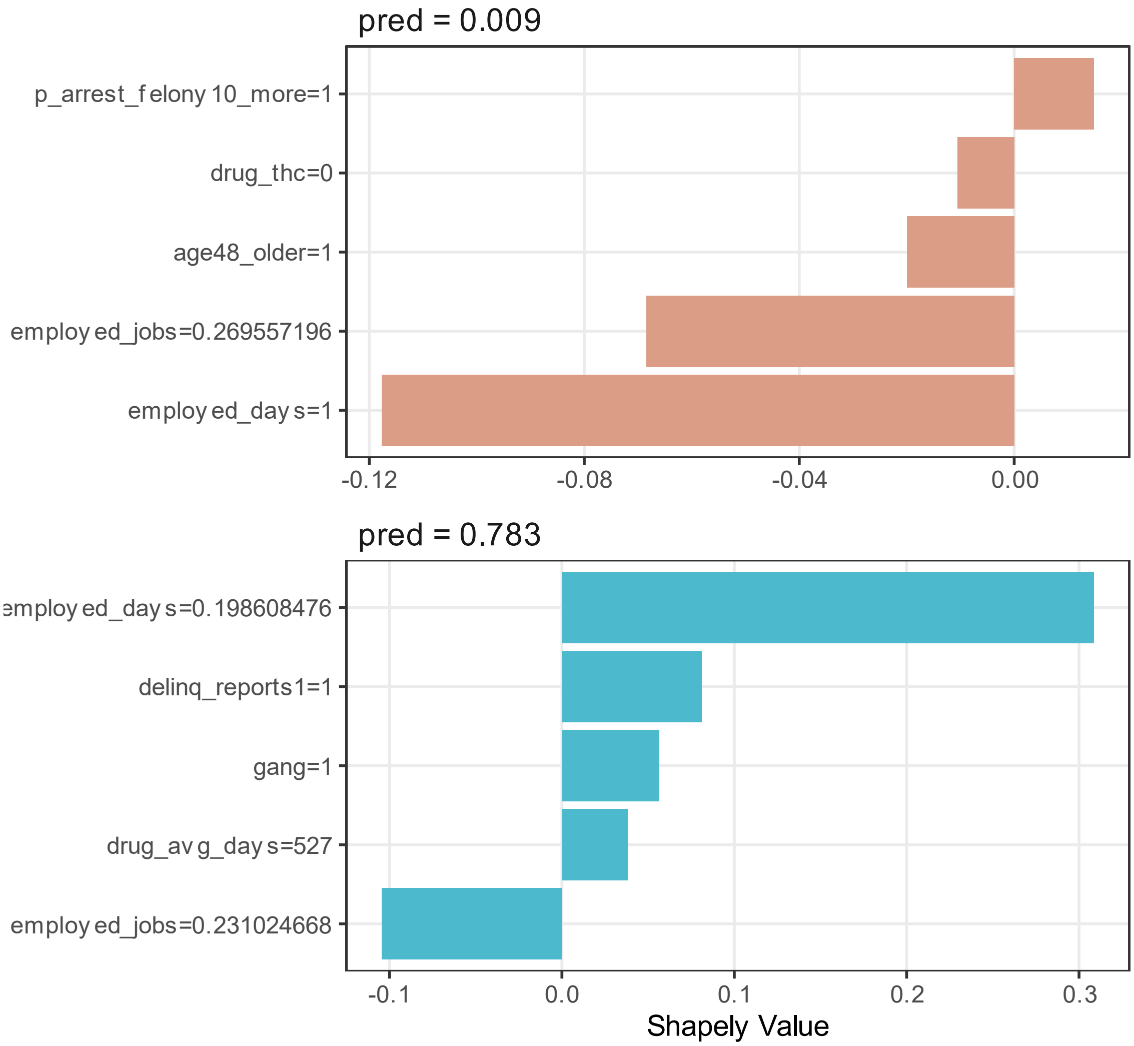
Among these, the five most important features were: (1) the number of jobs, per-year, while on parole, (2) the percentage of days employed while on parole, (3) the average number of days on parole between drug tests, (4), their supervision risk score, and (5) the percentage of drug tests positive for THC/marijuana. Interestingly, race (white vs. black) is not among even the top 20 predictors in the model. Based on this, the first feature might represent an irregular work history - for example, an individual who cannot hold a steady job - while the second feature reflects largely a more consistent work history. Below, Figure 2 displays the accumulated local effects (ALE) for these top 5 predictors. An ALE plot presents the change in predictions over a grid of values using data instances within plausible ranges present in the data. For example: In Figure 2 it is evident that as the proportion of days employed increases from 50% to 100%, the probability of re-arrest decreases substantially.

Figure 2. Accumulated Local Effects, Top 5 Features



Finally, we can also examine the effect of all variables on an *individual's* prediction using Shapely values. A Shapely value computes the average contribution for a feature across all possible combinations. For example, for an individual observation, it determines the contribution of each feature between the actual prediction and the mean prediction. This can be useful to determine why the model predicts high or low values for individual observations (Molnar, et. al., 2018). Consider the following problem: we have predictions for two individuals. One individual is given a very high prediction of recidivism at .783, while another individual is given a very low prediction of recidivism at 0.009. What about these individuals is different that makes their score differ by so much? Using Shapely values, we can determine the individual factors that determine their risk scores. For simplicity, we examine just the top 5 most important features however nearly *all* variables play some part in an individual's prediction score. On the other hand, the low-risk individual's prediction was largely predicated on a stable work history (employed_days = 1), low job turnover (employed_jobs = 0.26), and older age (age_48_older = 1). Using Shapely values can help analysts understand what factors play the largest role in their individual prediction.

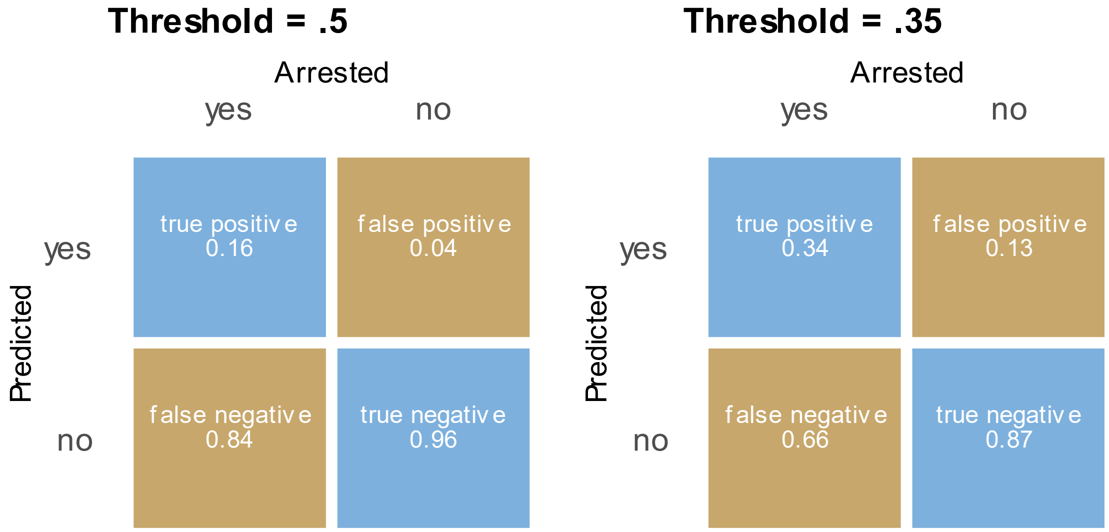
Figure 3. Shapely Values for Two Individuals, Top 5 Factors



Fair and Accurate Threshold

For the "fair and accurate" metric, the contest required the use of a .5 threshold for a positive prediction. In this case, all individuals who received a prediction of .5 or higher would be forecast in the recidivism category. While seemingly logical, this threshold is likely overly simplistic. In real-world practice, choosing thresholds for classification problems require balancing the cost of wrongfully classifying an observation. For example, setting the metric to a .5 threshold misclassifies nearly 84% of individuals who were arrested at time 2 as "false negatives" (not predicted to be arrested, but were arrested). Lowering the threshold to .35 reduces the false negative rate to 66%, but also increases the percentage of false positives to 13%. Below, a series of confusion matrixes illustrate that selection of the threshold has significant impacts on all four outcomes. It should be clear that there is no 'best' option here – rather the given thresholds should consider real-world costs, as well as factoring in domain knowledge and expertise.

Figure 4. Confusion Matrix, by Threshold Value



For scoring purposes, the fair and accurate score was calculated as $(1 - BS) * FP$ where BS is the model Brier score, and FP is the difference in false-positive rates between black and white individuals. These were separately scored based on a prediction threshold of .5. In order to meet the fair and accurate requirement outlined in the NIJ contest, we biased our predictions to never exceed .499. This means all individuals were, in the fair and accurate category, never forecast to recidivate and, therefore, were never misclassified as a false-positive. Indeed, if no one is forecast to be arrested, then the false-positive rate is exactly 0. In addition, because predictions from our model rarely exceeded .5, biasing these predictions to .4999 and under did not substantially impact our overall Brier score.

Very similar to the findings in Mohler and Porter (2021), we evaluated different models attempting to optimize this fairness metric directly in the loss function in each round. We found that while such a loss could be estimated for the in-sample data, out of sample data did not produce better results on the fairness metric compared to simply truncating the scores below the 0.5 threshold. This is partially because even if a model is trained to be fair on a particular sample, there is always some variance in the predictions applying that model to new data. In particular among the female sample, only a very small number of individuals in the unbiased predictions were above the 0.5 threshold.

For a simplified example, the out of sample prediction may only have 10 white and 10 black females above the 0.5 threshold. The actual results may subsequently then be 6/10 white recidivate and 8/10 black recidivate. This is a difference of 20% in the false positive rate between the two groups and would penalize the overall Brier score by a very large margin. (Although is very weak evidence of differences in the false positive rates between the groups due to the small sample size.) Winning solutions typically only bested each other in the 3rd or 4th decimal place, whereas such a false positive penalty could easily shift the final metric by a tenth (or much more). The penalty to the Brier score to simply bias the predicted probabilities to be under 0.5 is much smaller than the potential variance due to randomness in the fairness metric – which no matter what modelling strategy one employs will never be able to be perfectly balance in future predictions. This is further exacerbated in samples with very few individuals classified in the high risk category.

Future Considerations

One important factor to note is the performance of other, similar models. A competing model using a logistic regression with L1 regularization performed nearly as well as the considerably more complex XGBoost model. While this logistic regression model did not place in the top 5 for any of the accuracy categories, its performance on test data was remarkably similar. Given that tuning and running complex GBMs can consume significant time, more simple alternatives may work nearly as well. That being said, XGBoost remains relatively fast, and even faster GBMs are now available - for example, Microsoft's LightGBM (Ke, et al., 2017). Another major consideration is the development of interpretable machine learning models. Relying on diagnostic plots like ALE or Shapely values can provide a more intuitive view into the inner workings of the model. In addition, it helps users to understanding why specific individuals are assigned a risk value.

In terms of future considerations for the fairness metric, there are two obvious issues with the current metric as proposed by NIJ. First, the threshold of 0.5 to categorize an individual as high risk is likely misinformed. One would need to conduct a cost-benefit analysis to how such predictions will be used by criminal justice agents to determine what an appropriate threshold would be. For instance, if these are used to assign more intensive supervision for parolees, there is a limit on how many individuals can be assigned high risk. One may then take the threshold that approximately fills up this queue based on historical values.

Another approach would be to conduct a cost-benefit analysis on a case by case basis. The cost of assigning more intensive supervision (both in terms of labor costs for the agency, as well as for costs of the additional government oversight to the parolee) should be balanced with the estimated benefits of that assignment (e.g. reduced recidivism). This may produce a threshold either much higher or much lower than simply a probability of 0.5. Given the difficulty of conducting such a cost benefit analysis, future competitions may estimate such metrics at different thresholds and average overall results together.

A second consideration is the fairness metric considered itself. The multiplicative term results in large variances, whereas historical fairness penalties to model terms tend to be additive (Mohler & Porter, 2021). So here a slight change to the metric we believe would be more appropriate in practice:

$$BS + \lambda \cdot FP$$

Where lambda is a term to adjust how much one penalizes the false penalty term relative to the Brier score. Given the typical differences among solutions on the leaderboard for the competition, a lambda value of 0.1 could still result in meaningful shifts among winning solutions but result in much less variance in the overall score due to small sample differences in the FP rate between groups.

This however does not solve the problem of determining overall if a model succeeds or fails in producing racially equitable outcomes (for whatever metric one wishes to use). One may consider such questions entirely separately from accuracy of the model, instead of blending accuracy and fairness results together. We believe an important future research goal is determining metrics to not only build such models, but to *monitor* those models as used in practice. While minor deviations from false positive rates in small samples may not signify problems, where to draw the line in practice and how to determine if equitable outcomes are being achieved in real life, noisy data are more difficult.

References

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of clinical epidemiology*, *110*, 12-22.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, *30*, 3146-3154.

Mohler, G., & Porter, M.D. (2021). A note on the multiplicative fairness score in the NIJ recidivism forecasting challenge. *Crime Science* 10, 17.

Molnar, C., Casalicchio, G., & Bischl, B. (2018). iml: An R package for interpretable machine learning. *Journal of Open Source Software*, *3*(26), 786.