Recidivism Forecasting with Multi-Target Ensembles

National Institute of Justice (NIJ) Recidivism Forecasting Challenge

Winning Solution for Male, Female, and Overall Categories in Year One, Team CrimeFree

- David Lander, Northquay Capital, TrueFit.AI
- Russell D. Wolfinger, SAS Institute, Cary, NC

Abstract

Classification based on quantitative data is primarily about feature engineering and model ensembling. The former enriches the existing patterns in the data, while the latter produces robust predictions even from a limited amount of data. Our winning solution for the Recidivism Forecasting Challenge included heavy amounts of each; we provide a roadmap to this solution, along with source code and guidance on how to produce similar results on future datasets.

Overview

The key to this competition was using all available information about each parolee. Whether or not one recidivates in their first year is a binary low-signal outcome, while knowing future employment, drug, and other outcomes, including recidivism in future years, allows models to extract far more information about shared factors that would predict these.

Our initial models included a mix of linear, tree, and neural network models, with the best final models being primarily gradient-boosted trees and neural networks. While many features were tested, including second-order feature combinations, high-powered neural networks with a simple grouping of geographies generally performed best.

The breakthrough improvements in performance came when building models to predict all future outcomes. The features that predict recidivism in later years and other outcomes overlap with those that lead to first-year recidivism; building models to predict all known outcomes and then stacking these forward to predict this competition's target was the primary driver of our winning solution.

Variables

Our models included all provided variables. The primary form of feature engineering was converting all textual features to ordered numerical categories, and grouping all PUMA regions by their regional composition. We also mapped future outcomes to a range of extra targets, such as a binary variable for each drug test category equaling 0%, 100%, or being unlisted. Full detail is available in our shared Github repository.

Models

We tested the full range of modeling methodologies—linear models, support vector machines, tree-based models, neural networks, nearest neighbors and t-SNE clustering, etc. Our final ensemble consisted purely of gradient-boosted trees and neural networks; while linear models with combinatorial features showed gains early, ultimately neural networks were better suited to modeling these feature interactions.

Our most important, highest weighted, and highest scoring models were all multi-target networks. The initial approach was a multi-target neural network, predicting not just year-one recidivism but also every other future outcome. Later, we added a gradient-boosted models, trained individually for each fufure outcome, and then stacked back in to a model predicting just the first-year-recidivism target. The gain from these two additions to our ensemble was roughly 0.00030, or half the gap to the second-place solution.

All models were fed into a linear model for averaging, in particular, an elastic net model where all weights must be greater than or equal to zero. Our full solution included a massive 400-model ensemble, trained over more than 10,000 CPU hours. This ensemble process produced another 0.00030 gain in performance, or the other half of the gap between ours and the next best solution.

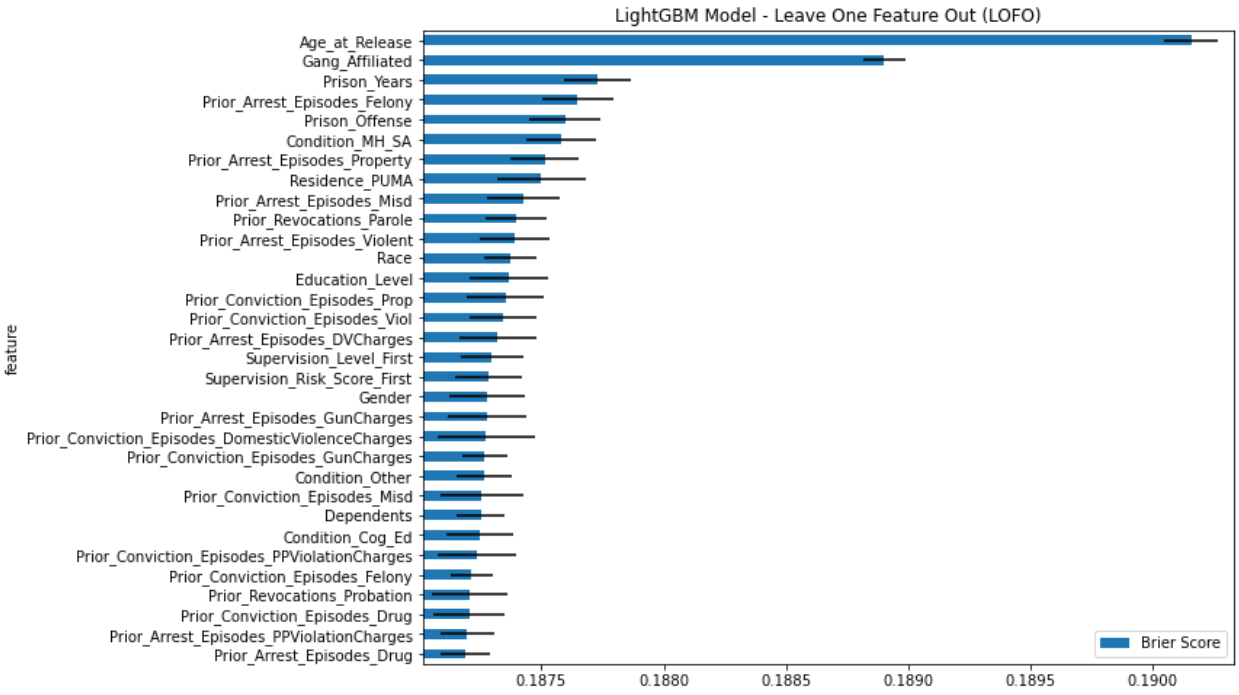
We have provided a more realistic solution at https://github.com/david-1013/RFC that trains within a day and produces a winning solution.

Cross-Validation

Our models were generally trained on 80% of the data, while using the remaining data to estimate model performance. Given how easy it is to overfit any one model to any portion of the data, our process typically included training around 50-100 'folds' of any given model, so that any data point was estimated as the average of at least 10-20 different model predictions.

Feature Importance

Estimating the importance of every feature is possible by producing models that "leave one feature out." This analysis is shown below, along with confidence intervals to determine the statistical significance of each feature; features listed in the table below produce a statistically significance change in performance when left out. The two most important features are age at release and gang affiliation, with dramatic decreases in forecast quality without including these variables.



Feature Leave-Out	Brier Score
Age_at_Release	0.1897
Gang_Affiliated	0.1883
Prison_Years	0.1873
Condition_MH_SA	0.1872
Prior_Arrest_Episodes_Felony	0.1872
Residence_PUMA	0.1871
Education_Level	0.1870
Prison_Offense	0.1870
Prior_Arrest_Episodes_Property	0.1870
Prior_Revocations_Parole	0.1870
Gender	0.1869
Prior_Arrest_Episodes_Misd	0.1869
Prior_Arrest_Episodes_PPViolationCharges	0.1869
Prior_Arrest_Episodes_Violent	0.1869

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. 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.

Source Code

A full solution to the challenge, with performance halfway between our full ensemble and the score of the second-place team, is open-sourced at https://github.com/david-1013/RFC. This source code can be trained on a single machine within a day, and includes the full stack from data intake to model ensembling. Our solution is fully MIT-licensed with no restrictions on its use.

Future Considerations

We found the competition to be well-structured with more than sufficient data to yield highly informative predictions. Based on the above feature analysis, and our approach of predicting all future outcomes, including as much data as possible on age and gang affiliation will produce the best results.

Conclusion

We thank the organizers for an interesting series of challenges. Our main go-forward suggestion would be providing models with as much information as possible. High-powered machine learning techniques are always a plus, and our codebase should provide a self-tuning stack to accomplish this, but knowing an additional employment statistic or number of gang ties would have likely exceeded the gaps between top teams.

Machine learning models are clearly up to the task of predicting recidivism for the purposes of parole judgments or monitoring—the predicted probabilities spanned the full range from near-zero to very likely—and we look forward to seeing more evidence-based methods like this put into practice.