Recidivism Forecasting Challenge: Team IdleSpeculation Report

Jeremy Walthers

August 31, 2021

1 Introduction

This report is intended as a summary of the author's submissions as team IdleSpeculation to the National Institute of Justice Recidivism Forecasting Challenge[1]. The basic structure of the challenge was to forecast arrest likelihoods of former inmates in each of the three years subsequent to their release.

The challenge structure naturally introduces some differences in forecasting methodology primarily due to the availability of attributes in each year. Where possible the submissions will be presented as a unified approach with yearly discrepancies highlighted when necessary.

In what follows we will find that minimal variable transformation or selection was applied to the competition data and the bulk of the effort was aimed at building a diverse set of predictions which were blended into yearly estimates via a model stacking technique.

While these yearly estimates did manage to score well in the racial fairness component of the challenge, there were no special steps taken to produce this outcome. A suggestion is made later regarding a potential improvement to this part of the evaluation.

Although the author feels the competition went smoothly and was well executed on the part of the host, a few ideas for future competitions are provided as well as a comment regarding suitability of the competition data for practical use in predicting recidivism.

[1] https://nij.ojp.gov/funding/recidivism-forecasting-challenge

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.

2 Variables

With the exception of a few minor transformations the data used in generating the author's submissions were simply the data provided for each year. Moreover, no manual variable selection was attempted so all models involved were exposed to every attribute.

While the author did review several sources of external data, none actually made it into the models and the reason for this is described in the next section.

2.1 External Data

The rationale for excluding alternate data was along the lines of the following: If there were some valuable external data then it would be necessary to merge it into the existing dataset. Since the data are anonymized at the individual level, the only plausible ways to merge in new sources are through geography or time. The number of distinct geographies and times is small enough that a reasonable model should be able to deduce their impact without additional characteristics. Hence external data was not expected to provide much benefit.

The author does not mean to imply that external data would be useless for predicting recidivism in general. In fact, if the format of the challenge had been slightly different, say if there were hundreds of geographies from all over the country then external data would likely prove very informative.

2.2 Feature Engineering

A few modifications were made to the source data to make it appropriate for the algorithms applied. For algorithms based on decision trees, any categorical variables were replaced with numeric identifier. For example `Education_Level` was altered as follows:

`Education_Level`	`Transformed Value`
High School Diploma	0
Less than HS diploma	1
At least some college	2

For the linear or neural network models, categorical variables were one-hot encoded with each distinct level giving a different column of the data.

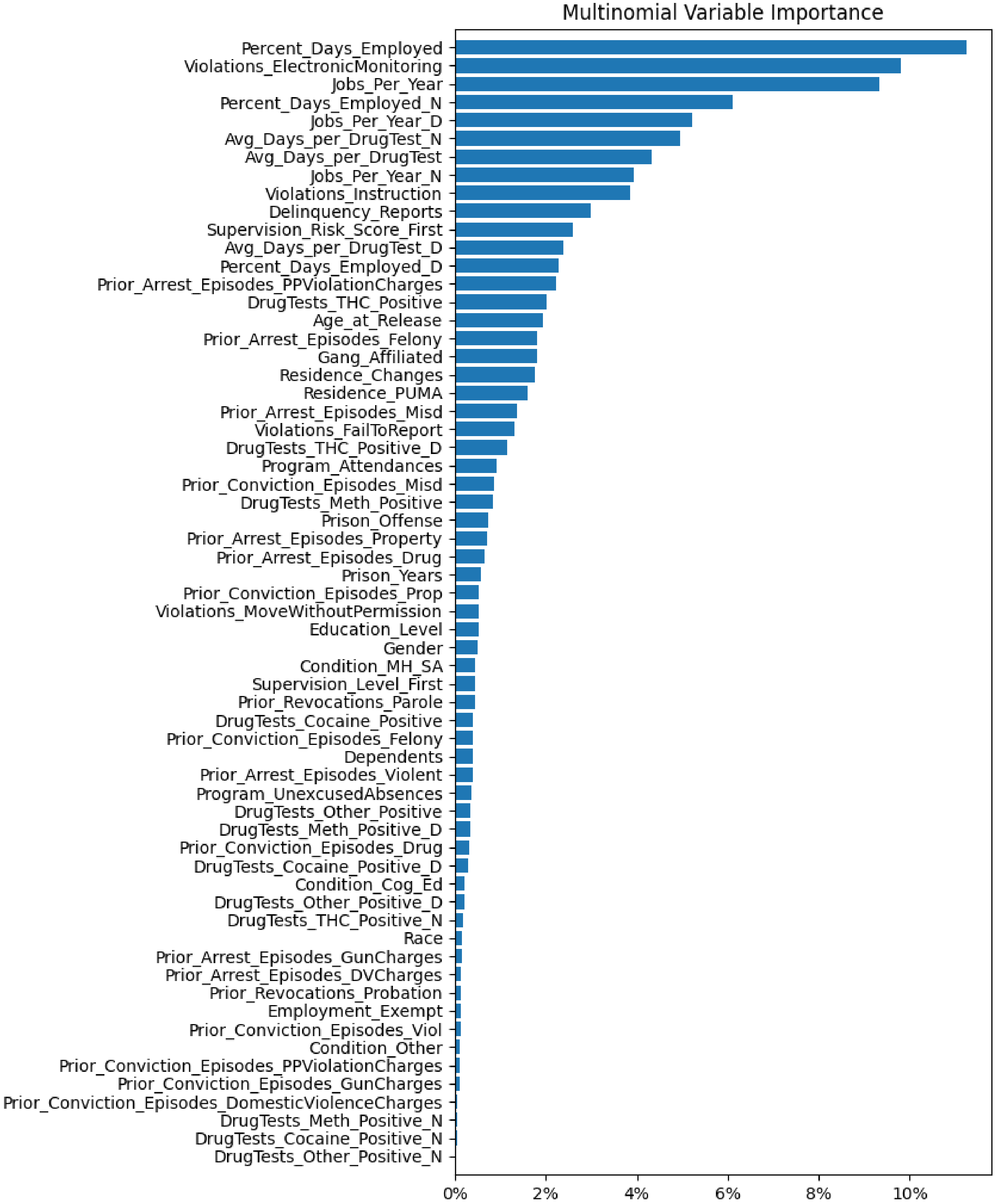
Education_Level	Education_0	Education_1	Education_2
High School Diploma	1	0	0
Less than HS diploma	0	1	0
At least some college	0	0	1

Aside from the previous two types of transformations, the only other variables constructed were derived from the monitoring columns. In the source data some of these columns were expressed as a percentage. The variables constructed from the percentages were the numerator and denominator of the closest rational number. An example for one of the drug test attributes is shown below where _N denotes the numerator and _D the denominator.

DrugTests_THC_Positive	THC_Positive_N	THC_Positive_D
0.333333	1	3
0.266667	4	15
0.048780	2	41

2.3 Feature Importance

As discussed in the next section, the author's submission primarily relied on gradient boosted decision trees. The author is not aware of a method to evaluate the statistical significance of a variable in an algorithm of this type. Instead, on the next page is depicted the reduction in prediction error generated by each split and aggregated per attribute. The graph is derived from a multinomial model which attempts to simultaneously predict recidivism for all three years.



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.

From the previous graph it is evident that some attributes are not particularly important to this multinomial model. No effort was made to eliminate attributes from this model or any others. Every available attribute was used in every model. However, note that some of the algorithms employed possess built-in variable selection mechanisms and consequently may ignore or marginalize variables selectively at their own discretion.

3 Models

The submissions for the competition were generated via a technique know as stacking. In stacking, one first fits several base models on the training data. The outputs of the base models are then used as predictors in another set of meta models. For this competition, the output of this second layer of meta models was combined by a weighted average to yield the submission.

3.1 Base Models

At this level a variety of algorithms were evaluated. These include:

- gradient boosted decision trees
- artificial neural networks
- linear models
- random forests
- extremely randomized trees

Only the first two in the list were incorporated into the solution. The gradient boosted trees were clearly better overall but the neural networks provided some incremental value at the meta model level. The other algorithms showed inferior performance and added no incremental value in the presence of stronger models.

With the goal of increasing diversity of the base models, additional models were built employing the same algorithm but with different targets or error metrics. The alternate target values used were:

- binary outcome for recidivism in a single year
- binary outcome for recidivism within N years
- multinomial outcome with recidivism prediction in all years

The first two were fit with both a logistic and mean square error loss.

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.

3.2 Meta Models

At the meta model level a simple linear regression was found to provide the best results. Cross-validated error decreased somewhat when adding a gradient boosted tree model with conservative parameters. The final prediction was a weighted average of the two with a majority of weight on the linear component.

4 Evaluation

The author admits to not truly understanding the evaluation up until the final scores were revealed. In hindsight some different choices would have been made in the modeling process.

Although the author's submissions scored quite well in the racial fairness metrics for women, there was nothing deliberate in this outcome. Gender and race were given equal footing with all other predictors in the models. The fairness section below contains a suggestion which may improve the evaluation of this component.

4.1 Metrics

An interesting metric to consider for future competitions would be one that measures the quality of predictions for high-risk individuals while de-emphasizing the penalty for low-risk folks. The idea here would be to assist in focusing limited resources within the high-risk category.

On a related note, understanding what types of intervention or assistance are most effective in preventing recidivism is clearly highly valuable although the challenge might need significant restructuring to address this aspect.

4.2 Fairness

The evaluation of the fairness component of the competition involved a 0.5 prediction threshold for identifying false positives. The author feels that the competition sponsors may have been overly optimistic here since predictions seldom passed this threshold. A change that might make the fairness evaluation more relevant to the dataset would be to set this threshold at the average outcome of each race/gender group in the training set. This would ensure an even mix of predictions both above and below the threshold.

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.

5 Future Considerations

5.1 Field Applications

Making suggestions based on a dataset like this is challenging because of the difficulty establishing causality. The models here clearly indicate that employment is the most important determinant of future recidivism. Is employment, or lack thereof, the root cause for recidivism or just the symptom of other issues such as drug abuse?

From a techincal point of view, one aspect about this dataset should be addressed if there is an intention to use it in practice. The issue is that the monitoring attributes are fixed. That is they were the same for each of the three prediction years. Ideally, there would be multiple sets of these attributes each including all information up to the point they would be used to make a prediction.

5.2 Future Competitions

In the author's opinion, the competition went smoothly and was well executed on the part of the host. Some things which might have enhanced the experience for the participants are:

- some type of forum to communicate with the host and other teams
- some notion of the number of people working on the problem
- some sort of leaderboard to compare progress against other teams

6 Conclusion

In closing, the author would like to thank the National Institute of Justice and particularly the individuals whose work made the competition possible. This report has summarized team IdleSpeculation's submission to the recidivism challenge. As previously detailed, the bulk of the effort for this solution was building a wide variety of models and blending them together. A few suggestions were made for improvements to the current and potential future challenges of this type.