Document Title:	National Institute of Justice's Recidivism Forecasting Challenge: Research Paper, Group MNLB
Author(s):	Yujunrong Ma, Kiminori Nakamura, Eung-Joo Lee, Shuvra S. Bhattacharyya
Document Number:	305046
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: Research Report by Team MNLB

Yujunrong Ma[1], Kiminori Nakamura[2], Eung-Joo Lee[1], and Shuvra S. Bhattacharyya[1]

[1] Department of Electrical and Computer Engineering, University of Maryland, College Park
[2]Department of Criminology and Criminal Justice, University of Maryland, College Park

August 2021 Team MNLB

1 Introduction

Actuarial assessment of recidivism risk has been in existence in the U.S. since as early as the 1920s [1] [2]. Using largely time-stable correlates of recidivism, such as age and criminal history, these early risk assessment tools became dominant in the 1970s and 1980s [3]. More recent recidivism risk assessment instruments leverage a wider variety of risk factors, including time-variant dynamic risk factors, such as substance abuse and employment status [4]. Such actuarial risk assessment is now widely used in a variety of decision-making contexts within the criminal justice system, ranging from pretrial release, sentencing, release form confinement, and community supervision conditions and revocation [5]. In recent decades, machine learning (ML) has been increasingly applied to these criminal justice risk prediction tasks by leveraging complex structures and patterns in large amounts of data (e.g., [6] [7]). In addition to predictive accuracy, research has been advancing several aspects of machine learning applications for recidivism prediction, including fairness and interpretability [8] [9] [10] [11].

Responding to the National Institute of Justice's (NIJ) Recidivism Forecasting Challenge, this brief report summarizes the submissions by Team MNLB and presents the results of machine learning models to predict recidivism (rearrest) based on the data from the State of Georgia on individuals who are released from prison and placed under parole supervision. The report is organized as follows: In the next section (Section 2), we first describe our efforts on data processing, including basic feature interpreting and engineering as well as the incorporation of external data associated with the U.S. Census Bureau Public Use Microdata Area (PUMA); In Section 3, we introduce machine learning models that were used for our experiment, as well as strategies such as parameter tuning and weighted sum ensemble; In Section 4, we summarize our experiment results; and in Section 5, we conclude and discuss implications and further thoughts about the recidivism forecasting challenge. The source code can be found at: `https://github.com/May-226/NIJ_challenge`.

2 Data Preprocessing

2.1 Feature engineering

2.1.1 Feature extraction

The dataset provided for the challenge consists of individuals who were released from state prison to parole supervision in Georgia between 2013 and 2015. The dataset includes a wide array of pre-release individual-level characteristics, ranging from demographics, detailed criminal history, supervision history, and current case information to post-release characteristics such as parole conditions and violations. The target prediction outcomes are a new in-state arrest for a felony or misdemeanor offense within 1, 2, and 3 years from release. While the dataset has been generally cleaned and structured for analysis, we have changed data types of the following variables for ease of interpretation:

- $Residence\_PUMA$ was changed from int32 to category(int8);
- $Age\_at\_Release$ was changed from category(int8) to int32;
- $Dependents$ was changed from category(int8) to int32; and
- All the $Prior\_Arrest\_Episodes\_$ and $Prior\_Conviction\_Episodes\_$ were changed from category(int8) to int32.

2.1.2 Addressing missing data

Several variables contain missing values including $Gang\_Affiliated$, $Supervision\_Level\_First$, $Supervision\_Risk\_Score\_First$ and $Prison\_Offense$. For $Supervision\_Risk\_Score\_First$, which only has 330 missing values out of 18,028, we used the listwise deletion by removing all those rows with $Supervision\_Risk\_Score\_First$ missing. For $Gang\_Affiliated$, which has 1,760 missing entries - but its distribution is highly imbalanced (9,414 False vs. 1,477 True) - we set all the missing values to be False. Two remaining variables, $Supervision\_Level\_First$ and $Prison\_Offense$, have 1,212 and 2,321 missing values, respectively. To minimize the sample loss, we imputed the missing entries based on logistic regression predictions.

2.1.3 Standardization

Machine learning algorithms are capable of incorporating different types of features (binary, categorical, or real value) and leverage their relations automatically. Although some models, especially deep learning models, naturally treat those features with larger magnitudes as more important. But usually, they would perform better when the model treats all the features equally at first. To achieve this, we used a standard scaler to transform all the features to the same magnitude by removing the mean and scaling to unit variance.

2.2 Data augmentation

The existing literature on place, crime, and recidivism suggests that social and economic disadvantages as well as the spatial concentration of incarceration and police can be predictive of criminal offenses in general as well as recidivism [12], [13]. Based on the linking code that was provided by NIJ to identify the Public Use Microdata Area (PUMA) for the prisoners' residence at release, the following place-based predictors of recidivism were generated and incorporated. When needed, data at the county level were converted to the PUMA level (`https://www2.census.gov/geo/docs/reference/puma/PUMSEQ10_13.txt`).

- The number of law enforcement officers in Georgia by county from the Uniform Crime Reports 2012: `https://ucr.fbi.gov/crime-in-the-u.s/2012/crime-in-the-u.s.-2012/tables/80tabledatadecpdf/table-80-state-cuts/table_80_full_time_law_enforcement_employees_georgia_by_metropolitan_and_nonmetropolitan_counties_2012.xls`;

- Total prison admission rate in Georgia in 2013 from the Vera Institute of Justice Incarceration Trends Dataset: `https://github.com/vera-institute/incarceration-trends`;

- Population characteristics in 2013, including income-to-poverty ratio, median household income, employment status, educational attainment, percentage of occupied housing units that are renter occupied, percentage of housing that is vacant, population density, and Herfindahl index of racial/ethnic groupings from the U.S. Census Bureau:

https://www2.census.gov/programs-surveys/acs/data/pums/2013/1-Year/.

2.3 Feature selection

Although larger sets of features and predictors of recidivism could potentially improve predictive performance, excessively redundant information can not only result in computational costs, but can also harm the performance of regularization, resulting in overfitting for the training dataset. For ensemble models, such as random forest and gradient boosting decision trees (GBDT), we could calculate the importance by counting the number of times each feature appears in all tree nodes. As an example, the importance output by GBDT is shown in Table 1 and Table 2, for the first year and second year recidivism prediction respectively. Through this process, we dropped those columns with the lowest importance for GBDT.

3 Model Selection

3.1 Tested candidate models

For tabular data, as in the current dataset, several models have been demonstrated to consistently outperform others. We have tested the following:

A. Logistic regression: Logistic regression (LR) is a statistical model that uses a logistic function to model a binary outcome variable. In the logistic model, the output probability of label "1" is a linear combination of all the predictors, which can either be a binary variable or any real value. The training process will allocate each predictor a weight based on optimization techniques such as gradient descent. The basic binary form can be extended into multinomial logistic regression which can accommodate multi-class problems. LR has been a popular benchmark model due to its stable performance as well as the fast

Rank	Feature	Importance
1	Age_at_Release	0.212413
2	Gang_Affiliated	0.139099
3	Prior_Arrest_Episodes_PPViolationCharges	0.134133
4	Prior_Arrest_Episodes_Property	0.105199
5	Prior_Arrest_Episodes_Felony	0.097512
6	Prison_Years	0.036586
7	Residence_PUMA	0.034104
8	Prior_Arrest_Episodes_Misd	0.033652
9	Prison_Offense	0.029626
10	Prior_Conviction_Episodes_Misd	0.019174
11	Supervision_Risk_Score_First	0.017869
12	Prior_Conviction_Episodes_Prop	0.016375
13	Prior_Arrest_Episodes_Drug	0.015474
14	Prior_Arrest_Episodes_Violent	0.014684
15	Condition_MH_SA	0.013581
16	Prior_Revocations_Parole	0.010496
17	Education_Level	0.010381
18	Gender	0.008721
19	Condition_Cog_Ed	0.006403
20	Dependents	0.005976
21	Prior_Conviction_Episodes_Felony	0.005349
22	Prior_Conviction_Episodes_Drug	0.005240
23	Condition_Other	0.004894
24	Prior_Conviction_Episodes_Viol	0.004200
25	Prior_Arrest_Episodes_DVCharges	0.003986
26	Race	0.003959
27	Supervision_Level_First	0.003594
28	Prior_Conviction_Episodes_DomesticViolenceCharges	0.002547
29	Prior_Conviction_Episodes_PPViolationCharges	0.001837
30	Prior_Conviction_Episodes_GunCharges	0.001545
31	Prior_Arrest_Episodes_GunCharges	0.001076
32	Prior_Revocations_Probation	0.000312

Table 1: Feature importance of dataset without supervision features and PUMA data on the 1st year recidivism prediction.

Rank	Feature	Importance
1	Percent_Days_Employed	0.186335
2	Jobs_Per_Year	0.146400
3	Prior_Arrest_Episodes_PPViolationCharges	0.071115
4	Age_at_Release	0.069156
5	Gang_Affiliated	0.062296
6	DrugTests_Meth_Positive	0.043342
7	Prior_Arrest_Episodes_Felony	0.043276
8	Avg_Days_per_DrugTest	0.042210
9	Supervision_Risk_Score_First	0.037508
10	DrugTests_THC_Positive	0.032957
...	...	...

Table 2: Feature importance of dataset with supervision features and PUMA data on the 2nd year recidivism prediction.

run-time.

B. Random forest: The Random forest is an ensemble model consisting of a number of independent single decision trees for classification and regression tasks. Each decision tree is developed to maximize the Gini index or information gain. The ensembling process can significantly improve the regularization performance, as each single tree can cause overfitting if the depth or the minimum samples in each node is not well designed.

C. Gradient boosting decision tree: Gradient boosting decision tree (GBDT) is another ensemble model of decision trees, where the difference from random forest is that GBDT adopts a different optimization strategy for the overall performance instead of a simple voting system. The method aims to minimize the overall loss value on the training set according to the empirical risk minimization principle. It was first developed using a greedy way, which meant it always used the steepest direction of loss gradients. However, the greedy way usually makes the model stuck in a local minimum, which results in a poor regularization performance. Stochastic gradient boosting is currently more widely used to prevent the overfitting problem.

D. Xgboost: Extreme Gradient Boosting (Xgboost) modifies the traditional gradient boosting algorithm. The main differences are as follows: 1) Xgboost calculates second-order gradients of the loss function and its approximation to reach the global minimum; 2) Xgboost introduces both L1 and L2 regularization. Furthermore, Xgboost is much faster than regular gradient boosting for the use of sparse matrices with sparsity aware algorithms as well as other hardware supporting adjustments.

E. Multilayer perceptron: A multilayer perceptron (MLP), also called a feed-forward neural network, is one of the simplest types of deep neural network (DNN). Each perceptron is either an input, intermediate, or output unit, and the activation function can be simply linear combination, or other functions like sigmoid etc. The MLP consists of three or more layers that need to be fully connected.

4 Experiment results

We split the entire dataset into training and test sets of 3:1 ratio, and when training the ML model, we used a cross validation of size 5. This means that we further split training set into 5 groups, and for each iteration, we used 4 groups for training and validation.

We first implemented logistic regression, MLP, GBDT and Xgboost with Random Forest as a baseline. Given the use of the Brier Score as the accuracy metric in this forecasting challenge, we attempted to set the Brier Score as the objective function. Since some of our models do not accept a customized loss function due to software package limitations, we provide model performance based on the ROC curve, which is probability-based similar to the Brier Score. Figures.1-3 present the ROC curves for the 3 years of recidivism prediction, illustrating that, across all 3 years, GBDT, XGboost and logistic regression on average perform slightly better than the other models. Since the performances of

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.

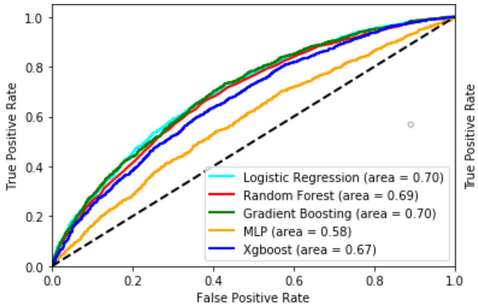

Figure 1: 1st Year ROC curve.

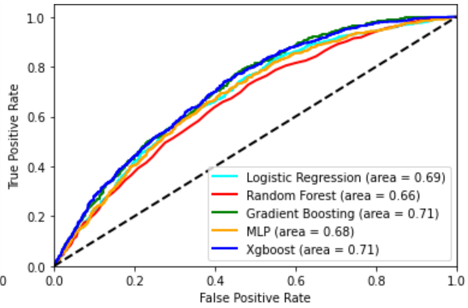

Figure 2: 2nd Year ROC curve.

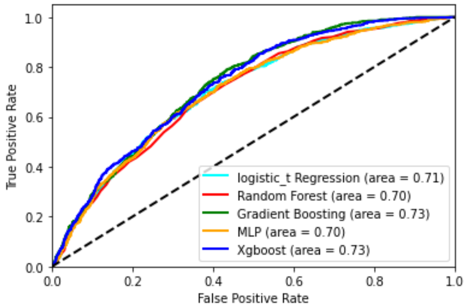

Figure 3: 3rd Year ROC curve.

GBDT, XGboost and logistic regression are very close, we chose to combine the models as a weighted sum. After testing several combinations with a grid search, we settled with the output as $0.5 \times y\_prob(GBDT) + 0.3 \times y\_prob(XGboost) + 0.2 \times y\_prob(LR)$, where $y\_prob(model)$ is the soft prediction of a given model.

5 Conclusions and Discussion

While the interest and use of actuarial risk assessment in criminal justice settings is not new, a recent introduction of machine learning has renewed efforts to improve risk prediction in the field, but it has also raised issues related to equity and other aspects of risk assessment technologies. In this challenge, using prisoner release data, we focused on several avenues to improve recidivism forecasting. First, we used the Brier Score, the designated performance metric for

	Year1		**Year2**		**Year3**	
	AUC	**Brier**	**AUC**	**Brier**	**AUC**	**Brier**
RF	0.69	0.1954	0.66	0.1740	0.70	0.1540
LR	0.70	0.1912	0.69	0.1671	0.71	0.1463
GBDT	0.70	0.1907	0.71	0.1626	0.73	0.1446
MLP	0.58	0.1911	0.68	0.1738	0.70	0.1502
XGboost	0.67	0.1915	0.71	0.1652	0.73	0.1438

Table 3: AUC and Breir score of different models on three datasets. RF: random forest; LR: logistic regression; GBDT: gradient boosting decision tree; MLP: multilayer perceptron; XGboost: extreme gradient boosting

this challenge, as the loss function when we could to directly optimize the metric. Second, we tested multiple machine learning algorithms that are demonstrated to be competitively performing and then developed their weighted ensemble. Third, we expanded the set of features by linking external data sources through Census PUMA.

Several findings and their implications are worth highlighting. First, in line with the existing literature (e.g., [14]), those features that are related to age or criminal history are consistently identified as important predictors. The models tend to allocate a fairly large weight to those features for recidivism prediction. For example, in Table 1, among the top 15 features in importance, 9 features (60%) are directly related to either age or criminal history. The salience of age and criminal history also confirms the advantages of these static risk factors for recidivism risk prediction. They are fairly reliable, easy to be extracted and computed from administrative data, and can contribute to a large portion of predictive power even with the presence of dynamic risk factors [15]. These results also suggest that more specific age-related features, for example, ages at the time of prior arrests/convictions, may contribute to further enhancement of prediction performance.

Second, as seen in Table 3, the prediction performance tends to improve as the length of follow-up period increases. The negative correlation between recidivism risk and the time since the last contact with the criminal justice system (i.e., recidivism-free time) has been well documented [16]. Given the potential role in predictive performance, using recidivism-free time as part of risk assessment can be a promising direction for future research and applications [17].

There are several potential areas where changes can be made to future recidivism forecasting challenges. First, the codebook could be further developed, especially on how features are derived. For example, as part of the supervision features, it was not clear how $Jobs\_Per\_Year$ (Jobs Per Year While on Parole) was created. Features on post-release supervision could be difficult to interpret because the temporal order of supervision and recidivism events is not often clear. If employment status during the parole supervision is a function of recidivism (e.g., a person on parole becomes unemployed directly as a result of rearrest and reincarceration), unknowingly using such features would not result in valid predictions. This type of data leakage is also a concern for other supervision features as well, including drug tests, residence changes, program attendance, etc.

On a related point, the current prediction task was organized sequentially, such that predictions were generated for the first, second, and third year recidivism in a sequential manner, and relevant features for prediction were revealed one year at a time as well. This process reflects the practical settings of predicting recidivism in corrections. For example, recidivism could be predicted by the parole board weighing whether to grant release from incarceration, based on features available during and prior to the incarceration, including static person characteristics, criminal history and in-prison misconduct. Once the person is released to parole supervision, regular assessment (e.g., annually) can be conducted to predict recidivism more dynamically based on features reflecting recent, temporally relevant, changes in the person's life in employment, substance use, treatment and others. The dataset for the current recidivism forecast challenge incorporates many of these features that change over time, which could lead to more dynamically accurate recidivism predictions for timely decision making in community supervision. Moving forward, clearly differentiating the timing of when these post-release features are recorded and when recidivism events occur would improve the value of data as well as the resulting predictions for the field.

References

[1] Ernest W Burgess. Factors determining success or failure on parole. In E. W Burgess A. A. Bruce and A. J. Harno, editors, *The workings of the indeterminate sentence law and the parole system in Illinois.* The Board of Parole, Springfield, IL, 1928.

[2] Daniel Glaser. The efficacy of alternative approaches to parole prediction. *American Sociological Review*, 20(3):283–287, 1955.

[3] James Bonta. Risk-needs assessment and treatment. In A. T. Harland, editor, *Choosing correctional options that work:Defining the demand and evaluating the supply*, pages 18–32. Sage Publications, Inc, 1996.

[4] James Bonta and Donald Arthur Andrews. *The psychology of criminal conduct.* Routledge, 2016.

[5] Sarah L Desmarais and Jay P Singh. Risk assessment instruments validated and implemented in correctional settings in the united states. *Lexington, KY: Council of State Governments*, 2013.

[6] Richard Berk, Drougas Berk, and Drougas. *Machine learning risk assessments in criminal justice settings.* Springer, 2019.

[7] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The quarterly journal of economics*, 133(1):237–293, 2018.

[8] Jiaming Zeng, Berk Ustun, and Cynthia Rudin. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(3):689–722, 2017.

[9] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[10] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.

[11] Berk Ustun and Cynthia Rudin. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, 102(3):349–391, 2016.

[12] National Research Council. *The growth of incarceration in the United States: Exploring causes and consequences.* The National Academies Press, 2014.

[13] National Academies of Sciences, Engineering, and Medicine. *Proactive policing: Effects on crime and communities.* The National Academies Press, 2018.

[14] Paul Gendreau, Tracy Little, and Claire Goggin. A meta-analysis of the predictors of adult offender recidivism: What works! *Criminology*, 34(4):575–608, 1996.

[15] Michael S Caudy, Joseph M Durso, and Faye S Taxman. How well do dynamic needs predict recidivism? implications for risk assessment and risk reduction. *Journal of Criminal Justice*, 41(6):458–466, 2013.

[16] Michael D Maltz. *Recidivism.* Academic Press, 1984.

[17] Nicole E Frisch-Scott and Kiminori Nakamura. Time for a change: Examining the relationships between recidivism-free time, recidivism risk, and risk assessment. *Justice Quarterly*, pages 1–24, 2021.