Document Title:	CATBOOST Models for the Recidivism Forecasting Challenge
Author(s):	Kristen Guerrera, Brian Rieksts
Document Number:	305058
Date Received:	July 2022
Award Number:	NIJ Recidivism Forecasting Challenge Winning Paper

CATBOOST Models for the Recidivism Forecasting Challenge

Kristen Guerrera and Brian Rieksts

Abstract

The Department of Justice (DOJ) National Institute of Justice (NIJ) hosted a competition to predict recidivism according to different metrics, gender categories, and time periods. We developed models to predict recidivism using the machine learning algorithm called CATBOOST – an algorithm that has robust performance for models with categorical variables. Among large teams and businesses, we finish in the top 5 positions for 7 of the 9 categories for raw accuracy (5 fourth place finishes and 2 fifth place finishes). We also finished 2nd and 5th in categories accounting for racial bias. Our algorithms found variables relating to employment to be highly influential in the model. This suggests that policies and other interventions related to employment should be evaluated to determine their effectiveness in reducing recidivism.

Introduction

In 2021, the National Institute of Justice (NIJ) hosted a competition to forecast recidivism using person and place-based variables with the goal of improving outcomes for those serving a community supervision sentence. The contest used data from the State of Georgia about persons released from prison to parole supervision for the period January 1, 2013 through December 31, 2015. Contestants submitted forecasts (percent likelihoods) of whether individuals in the dataset recidivated within one year, two years, or three years after release. Each of these 3 categories was judged separately by gender and the average across gender for 9 categories focusing on raw accuracy that varied according to team size. Another 6 categories judged the forecasts with a metric penalizing racial bias.

This paper describes our models, results, and observations. The literature section provides a high-level overview of some relevant topics from the literature. The section on variables and models describes our implementation of the machine learning algorithm CATBOOST to forecast recidivism. Following that, we discuss modeling issues associated with racial bias. Next, we discuss our results and conclusions. Ethical issues are critical when considering a forecasting model of this nature as we describe in the section on future implications. Finally, we discuss considerations for future contests.

Literature Review

The literature relevant to this challenge and our approach covers a broad range of topics. Our approach provides a forecasting methodology, but the literature provides depth on relevant topics. First, many researchers have used data mining techniques to forecast recidivism. Hashim and Nohuddin (2018) provide a survey of research in this area.[1]

This challenge is also framed by providing a data set of tabular data for a classification problem. Gradient boosting has generally been a successful approach for these forecasts.

[1] Hashim, Emmy N; Nohuddin, Puteri N. E, "Data Mining Techniques for Recidivism Prediction: A Survey Paper," *Advanced Science Letters*, Volume 24, Number 3, March 2018, pp. 1616-1618(3).

Friedman (1999) developed the gradient boosted machine (GBM) model.[2] The XGBOOST[3] and CATBOOST[4] algorithms provide an improved implementation. LightGBM[5] is another high performing implementation of gradient boosting, but our approach only considered XGBOOST and CATBOOST.

Another aspect of the competition is addressing racial fairness. Several researchers have also studied issues of fairness in machine learning.[6] In addition to research papers, Google has hosted a competition to promote research in fairness for image recognition.[7]

Variables and Models

We used the class of machine learning algorithms called gradient boosting to predict recidivism. In particular, we used the CATBOOST implementation of gradient boosting. Details of the results of our parameter tuning are available in our code posted to Github.[8] We also explored results with the XGBOOST implementation of gradient boosting. We did not do a rigorous comparison, but XGBOOST model forecasts were as close as within 1% of the CATBOOST model forecasts and occasionally slightly outperformed CATBOOST. In most instances that we used to validate our models, however, CATBOOST outperformed XGBOOST. Our intuition was also that CATBOOST would be the better model given the number of categorical variables in the model.[9]

Our approach focused on using the data provided for the competition without matching to outside data. We considered matching external data to the variable called Residence_PUMA, but instead we modeled the effects across different regions as a categorical variable. We focused on feature engineering for the data provided, but external data would be a potential improvement to our models.

Gradient boosting algorithms can overfit models with categorical variables, particularly those with a high number of levels, which then results in a large number of possible interactions. We modeled variables as numerical data even if all values above a certain threshold are grouped together in the original dataset. For example, we modeled Prior_Arrest_Episodes_Felony with values 1 through 10 even though the last value is actually "a 10 or more" felony prior arrest episodes. The trees from gradient boosting have the flexibility to model non-linear relationships if using the value 10 for "10 or more" creates such a relationship. The CATBOOST algorithm also has a design that is robust to modeling categorical variables. A potential enhancement that was not included would be using numerical external

---

[2] Friedman, Jerome, "Greedy Function Approximation: A Gradient Boosting Machine," IMS 1999 Reitz Lecture, February 24, 1999. https://statweb.stanford.edu/~jhf/ftp/stobst.pdf, Accessed on September 16, 2021.
[3] https://xgboost.readthedocs.io/en/latest/, Accessed on September 16, 2021.
[4] https://catboost.ai/, Accessed on September 16, 2021.
[5] https://lightgbm.readthedocs.io/en/latest/, Accessed on September 16, 2021.
[6] Canton, Simon; Christian Haas, "Fairness in Machine Learning: A Survey," Cornell University, 2020. https://arxiv.org/abs/2010.04053, Accessed on September 16, 2021.
[7] https://www.kaggle.com/c/inclusive-images-challenge, Accessed on September 16, 2021.
[8] Link to go live and be inserted here after incorporating feedback from competition organizers to improve the paper.
[9] https://towardsdatascience.com/introduction-to-gradient-boosting-on-decision-trees-with-catboost-d511a9ccbd14, Accessed on September 15, 2021.

data matched to the categorical variable Residence_PUMA as another strategy for modeling the categorical variables.

For the one-year forecast, we first considered variables with numerical values and a category for all values over a certain value (e.g., Prior_Arrest_Episodes_Felony). We removed the variable Prior_Arrest_Episodes_Drug because its relationship to the dependent variable was not monotonic. We included all other variables provided for the one-year test set. We also converted each of these features to numeric values:

- Prison_Years
- Prior_Arrest_Episodes_Felony
- Prior_Arrest_Episodes_Misd
- Prior_Arrest_Episodes_Violent
- Prior_Arrest_Episodes_Property
- Prior_Arrest_Episodes_PPViolationCharges
- Prior_Conviction_Episodes_Felony
- Prior_Conviction_Episodes_Misd
- Prior_Conviction_Episodes_Prop
- Prior_Conviction_Episodes_Drug
- Dependents
-

Although gradient boosting algorithms allow for interactions with multiple variables, model performance may improve if relevant variables are combined through arithmetic operations to explicitly model these interactions with other variables. The models do not need to discover interactions through branching on trees. We created three additional variables for the one-year model. First, we combined several arrest variables into a total arrest variable: Total Arrests = Prior_Arrest_Episodes_Felony+ Prior_Arrest_Episodes_Misd + Prior_Arrest_Episodes_Property + Prior_Arrest_Episodes_Violent.

Next, we combine several variables related to convictions:
Total Convictions = Prior_Conviction_Episodes_Felony+Prior_Conviction_Episodes_Misd+ Prior_Conviction_Episodes_Prop+Prior_Conviction_Episodes_Drug.

Through model iteration and development, we also combined several variables that had a high relative influence in earlier iterations of the model. A higher value of this variable implies a higher likelihood of recidivism. We subtracted dependents to be consistent with this relationship since parolees with more dependents have a lower average recidivism rate:
Total Significant Variables = Prior_Arrest_Episodes_Felony+ Prior_Arrest_Episodes_Misd + Prior_Arrest_Episodes_Property + Prior_Arrest_Episodes_Violent+ Prior_Arrest_Episodes_PPViolationCharges + Prior_Conviction_Episodes_Felony + Prior_Conviction_Episodes_Misd+ Prior_Conviction_Episodes_Prop+ Prior_Conviction_Episodes_Drug +(3- Dependents)

These were our initial variables in the one-year forecasting model. Shortly before the deadline, we attempted an enhancement. We used these variables to also predict two-year recidivism and three-year recidivism. We developed out-of-sample forecasts for each of these variables. Then we predicted these variables for the test set as well. This is a process called stacking. We then used these forecasts for two-year and three-year recidivism along with the variables discussed previously to predict one-year recidivism. Our attempt with this multi-stage model was to add information about two-year and three-year recidivisms to increase the sample size of data available to fit the model. Since this model performed slightly better for a short validation exercise before the deadline, we used the forecast with stacking. But we chose to use the concept behind the simpler model for the second and third rounds of the competition after further validation analysis. Table 1 shows the performance for the one-year forecasts on the test set for the model used in the submission and the simpler model.

Table 1. Brier score of simple model and stacked model for one-year forecast

	Male	Female
Simple Model	0.191185	0.155625
Stacked Model (Submission)	0.191306	0.155214

A measure of importance for variables in a gradient boosting model is the relative contribution of each variable in the model. These values are calculated by measuring the decrease in model performance by removing variables and normalizing the sum to 100%. Table 2 shows the relative importance of variables in the submission for the one-year forecast. We see age and gang affiliation are the two most influential variables in the model. The prediction for two-year recidivism is the next most significant variable, but this variable is derived as a prediction from other variables in the model.

Table 2. Relative importance of variables for the one-year forecast for the submission.

Variable	Relative Importance
Age_at_Release	14.9%
Gang_Affiliated	9.7%
Probability Recidivism Two Years	8.1%
Total Significant Variables	6.6%
Prison_Years	5.0%
Supervision_Risk_Score_First	4.7%
Prior_Arrest_Episodes_Property	4.5%
Prison_Offense	4.1%
Residence_PUMA	4.0%
Total Arrests	3.7%
Prior_Arrest_Episodes_PPViolationCharges	3.6%
Education_Level	3.5%
Prior_Arrest_Episodes_Felony	3.1%
Supervision_Level_First	3.1%
Condition_MH_SA	2.5%
Probability Recidivism Three Years	2.1%
Prior_Conviction_Episodes_Misd	1.7%
Prior_Revocations_Parole	1.5%
Prior_Conviction_Episodes_Drug	1.5%
Prior_Conviction_Episodes_Prop	1.4%
Total Convictions	1.4%
Prior_Arrest_Episodes_Misd	1.4%
Race	1.3%
Dependents	1.0%
Condition_Cog_Ed	0.9%
Prior_Arrest_Episodes_Violent	0.9%
Gender	0.9%
Prior_Conviction_Episodes_Felony	0.6%
Prior_Conviction_Episodes_PPViolationCharges	0.5%
Prior_Arrest_Episodes_GunCharges	0.4%
Prior_Revocations_Probation	0.3%
Condition_Other	0.3%
Prior_Conviction_Episodes_Viol	0.3%
Prior_Arrest_Episodes_DVCharges	0.3%
Prior_Conviction_Episodes_DomesticViolenceCharges	0.2%
Prior_Conviction_Episodes_GunCharges	0.1%

Table 3 shows the confusion matrix for the one-year forecast for the submission. This confusion matrix is based on a threshold of 0.5. Since the model is optimizing squared error and the portion of parolees who are recidivists is less than 0.5, the number of predicted recidivists is low relative to those predicted not to be recidivists.

Table 3. Confusion matrix for one-year submission

	Predicted: Not Recidivist	Predicted: Recidivist
Actual: Not Recidivist	5176	284
Actual: Recidivist	1890	457

The training set for both the two-year and three-year forecasts included additional information on parolees. We used the approach of the simple model for one-year with additional variables for both the two-year and three-year forecasts. The same model variables were used for both forecasts, but we trained the model on the data specific for that forecast. We used the same set of numeric variables in common with the one-year model, but we included additional numerical variables:

- Program_Attendances
- Program_UnexcusedAbsences
- Avg_Days_per_DrugTest

We also engineered additional features such as job turnover for a given time working:
Jobs per Days Employed = Jobs_Per_Year/(0.1+ Percent_Days_Employed)

We also engineered some variables to approximate the number of positive drug tests:
Num Drug Positive = 1089/(Avg_Days_per_DrugTest)*(DrugTests_THC_Positive+ DrugTests_Cocaine_Positive+ DrugTests_Meth_Positive+ DrugTests_Other_Positive)

Num Drug Positive THC = 1089/(Avg_Days_per_DrugTest)*(DrugTests_THC_Positive)

Num Drug Positive Meth = 1089/(Avg_Days_per_DrugTest)*(DrugTests_Meth_Positive)

We also estimated the percent of absences to normalize across different levels of attendance. Someone missing one class out of 10 is different than someone missing one class out of 2:
Percent Absence = Program_UnexcusedAbsences/(1+ Program_Attendances)

Table 4 shows the relative importance of variables for the two-year model. Variables related to employment status are the most influential in the model. These variables exceed the influence of the most influential variables from the one-year model in age and gang affiliation. Table 5 shows the confusion matrix for the two-year forecast. Similarly, Tables 6 and 7 show these results for the three-year forecast. Table 7 shows that the number of forecasted recidivists in the confusion matrix becomes small given the threshold of 0.5 and the decreasing portion of recidivists as we move to a three-year horizon.

Table 4. Relative importance of variables for the two-year forecast.

Variable	Relative Importance	Variable	Relative Importance
Percent_Days_Employed	13.4%	Violations_ElectronicMonitoring	0.5%
Jobs per Days Employed	11.4%	Race	0.5%
Jobs_Per_Year	5.3%	Gender	0.5%
Age_at_Release	4.9%	Prior_Conviction_Episodes_Prop	0.4%
Gang_Affiliated	4.7%	Dependents	0.4%
Num Drug Positive	3.6%	Violations_FailToReport	0.3%
Avg_Days_per_DrugTest	3.5%	Condition_Other	0.2%
Prior_Arrest_Episodes PPViolationCharges	3.1%	Violations_Instruction	0.2%
Total Arrests	3.0%	Prior_Arrest_Episodes_DVCharges	0.1%
Delinquency_Reports	3.0%	Prior_Arrest_Episodes_GunCharges	0.1%
DrugTests_THC_Positive	2.9%	Prior_Conviction_Episodes_GunCharges	0.1%
Total Significant Variables	2.7%	Violations_MoveWithoutPermission	0.1%
Supervision_Risk_Score_First	2.6%	Prior_Conviction_Episodes_Viol	0.1%
Education_Level	2.5%	Employment_Exempt	0.1%
Residence_PUMA	2.3%	Prior_Conviction_Episodes DomesticViolenceCharges	0.1%
Percent Absence	2.1%	Condition_Cog_Ed	0.0%
Num Drug Positive Meth	2.1%		
Prison_Years	2.0%		
Residence_Changes	1.9%		
Num Drug PositiveTHC	1.8%		
DrugTests_Meth_Positive	1.4%		
Prior_Revocations_Parole	1.4%		
Supervision_Level_First	1.3%		
Prior_Arrest_Episodes_Misd	1.3%		
Prior_Arrest_Episodes_Felony	1.3%		
Condition_MH_SA	1.2%		
Prior_Conviction_Episodes_Misd	1.2%		
Prior_Arrest_Episodes_Violent	1.0%		
Program_UnexcusedAbsences	1.0%		
Prison_Offense	0.9%		
Prior_Arrest_Episodes_Property	0.9%		
DrugTests_Other_Positive	0.8%		
Total Convictions	0.7%		
Prior_Conviction_Episodes PPViolationCharges	0.6%		
Prior_Conviction_Episodes_Drug	0.6%		
Prior_Revocations_Probation	0.6%		
Prior_Conviction_Episodes_Felony	0.6%		
DrugTests_Cocaine_Positive	0.6%		

Table 5. Confusion matrix for two-year forecast

	Predicted: Not Recidivist	Predicted: Recidivist
Actual: Not Recidivist	4021	125
Actual: Recidivist	1159	155

Table 6. Relative importance of variables for the three-year forecast

Variable	**Relative Importance**	**Variable**	**Relative Importance**
Jobs per Days Employed	9.5%	Prior_Conviction_Episodes_Drug	1.0%
Percent_Days_Employed	8.0%	Prior_Arrest_Episodes_Property	1.0%
Jobs_Per_Year	5.6%	Violations_Instruction	1.0%
Gang_Affiliated	4.6%	DrugTests_Other_Positive	0.9%
Age_at_Release	4.5%	Violations_MoveWithoutPermission	0.8%
Total Arrests	4.4%	Prior_Revocations_Parole	0.8%
Avg_Days_per_DrugTest	3.8%	Prior_Arrest_Episodes_GunCharges	0.8%
Num Drug PositiveTHC	3.8%	Prior_Conviction_Episodes DomesticViolenceCharges	0.6%
Total Significant Variables	3.6%	Gender	0.6%
Residence_Changes	3.2%	Prior_Arrest_Episodes_Violent	0.6%
Supervision_Risk_Score_First	3.2%	Prior_Conviction_Episodes_Prop	0.5%
Prior_Arrest_Episodes PPViolationCharges	3.0%	Condition_MH_SA	0.4%
Num Drug Positive	2.8%	Prior_Conviction_Episodes_Felony	0.4%
Education_Level	2.6%	Violations_ElectronicMonitoring	0.4%
Residence_PUMA	2.2%	Prior_Conviction_Episodes_Viol	0.3%
Prior_Conviction_Episodes_Misd	2.1%	Prior_Conviction_Episodes_GunCharges	0.3%
DrugTests_THC_Positive	2.0%	Condition_Cog_Ed	0.2%
Delinquency_Reports	1.9%	Condition_Other	0.1%
Num Drug Positive Meth	1.7%	Prior_Arrest_Episodes_DVCharges	0.1%
Dependents	1.7%	Prior_Revocations_Probation	0.1%
Prison_Offense	1.7%	Employment_Exempt	0.1%
Supervision_Level_First	1.7%	Race	0.1%
DrugTests_Meth_Positive	1.7%	Prior_Conviction_Episodes PPViolationCharges	0.1%
Percent Absence	1.6%	Violations_FailToReport	0.1%
Program_UnexcusedAbsences	1.5%		
Prison_Years	1.4%		
Prior_Arrest_Episodes_Misd	1.3%		
Prior_Arrest_Episodes_Felony	1.1%		
Total Convictions	1.1%		
DrugTests_Cocaine_Positive	1.1%		

Table 7. Confusion matrix for three-year forecast

	Predicted: Not Recidivist	Predicted: Recidivist
Actual: Not Recidivist	3314	10
Actual: Recidivist	817	5

Racial Biases

Racial and other biases are important ethical issues to consider when using machine learning algorithms to develop predictions. Before considering models and predictions, it is important to recognize that a possible source of bias is the system or process itself. The recidivism rates could be different based on biases inherent in the existing system. The underlying reasons for differences in recidivism across races are extremely complex issues; rigorous study is required to fully understand these differences and how to appropriately address them.

Another potential area of bias is that the algorithms themselves could be biased in predicting recidivism. For example, the algorithm could perform poorly for a particular sub-population if fewer observations for that group exist in the training data. The algorithm could systematically overestimate or underestimate recidivism for a population because of this input data. Similarly, an image classifier model that labels wedding images is likely to perform poorly for images of wedding traditions in underrepresented parts of the world.[10]

For the competition, the racial bias metrics include a penalty for differences in false positive rates across races. We attempted to address potential bias by explicitly including race and interactions with race to adjust for any biases. If a model is over- or underestimating recidivism for a particular race, including race explicitly will adjust for this bias. Ideally, relative contribution of race to the models would be minimal. This would imply that the model predictions are robust across racial groups. Furthermore, it would not at all be appropriate to apply consequences differently to someone based on their race because the model predicts someone of a particular race is more likely to be a recidivist. While this is our approach to develop the most accurate predictions across races, ethical issues remain in how to use the model results. We discuss these concerns in the practical application section.

Results and Conclusions

Table 8 shows a summary of our models across categories where we placed in the top five. Our models placed in the top five in 9 of the 15 categories which we competed in. Our forecasts were consistent in the variables that were most influential. Job related variables were the most influential when those variables were available. In all three forecasts, age and gang affiliation were among the variables with the most relative influence on the model.

[10] https://ai.googleblog.com/2018/09/introducing-inclusive-images-competition.html, Accessed September 15, 2021.

Table 8. Categories with top five finishes

Category	Place	Team Name	Year	Population	Brier Score
Large Team	4	KMG BQR	1	Male	0.1913
Large Team	5	KMG BQR	1	Female	0.1552
Large Team	4	KMG BQR	1	Average	0.1733
Large Team	4	KMG BQR	2	Male	0.1631
Large Team	4	KMG BQR	2	Female	0.1224
Large Team	4	KMG BQR	2	Average	0.1428
Large Team	5	KMG BQR	3	Male	0.1507
Racial Fairness	2	KMG BQR	2	Male	0.8356
Racial Fairness	5	KMG BQR	3	Male	0.8481

Practical Implications

The desired outcome related to this challenge is lower recidivism. We think these results add value in identifying correlations between recidivism and other factors. These correlations could be further analyzed with empirical techniques to determine causal relationships and evaluate appropriate policy changes. For example, we observed that variables related to jobs have a high relative influence in our model. Further research could be performed to understand potential interventions about employment to reduce recidivism.

While we identified machine learning algorithms as the best tool to predict recidivism, the potential for bias from these algorithms raises ethical concerns. We do not think it is appropriate to use only the machine learning predictions to target interventions without further analysis. For example, longer prison sentences should not be applied to some parolees simply because a model predicts they have a higher recidivism rate. First, our model explicitly includes variables for race and it would not be appropriate to give longer sentences to people based on their race. Furthermore, people should be able to clearly understand their potential length of sentence or other consequences for criminal acts. Their anticipated consequences create deterrence. It would be better to apply consequences in a more transparent way than using predictions from a machine learning model.

Additionally, if consequences are applied differently to different groups of people, the corresponding policy should be debated in the open. For example, minors are treated different than adults, but everyone is aware of that policy difference. But applying consequences differently based on predictions from machine learning algorithms would not be transparent to allow for debate about biases in treatment.

Future Considerations

The competition hosts requested that participants share thoughts on future competitions. We think the incentives for the competition encouraged participants to optimize their forecasts with the Brier score since the majority of the prizes were awarded for accuracy with this metric. Our approach was not affected by the fairness penalty only considering false positives. By minimizing the Brier score, we also minimized false positives to some degree and applying the same methodology across racial groups. If we had weighted the predictions to place more emphasis on minimizing false positives, the estimates would have been biased. These biased estimates would not have performed as well for the Brier score. Similarly, if we had adjusted our model for the threshold of 0.5 or a different level, that would have biased our estimates and increased the Brier score. We considered race and interactions with race explicitly in the model, which improves both the overall Brier score and the racial fairness metric. Although it did not affect our approach, the threshold of 0.5 did cause the racial fairness metric to be heavily influenced by on the tail of the distribution, particularly for the three-year forecasts. If only considering the racial fairness threshold, people may be incentivized to forecast under the threshold.[11] Another approach to the racial fairness aspect of a competition could be award separate prizes for forecasts for different racial groups similar to the separate awards by gender.

Several other options exist for measuring the performance of classification models. The Brier score used in the competition measures accuracy. Another example is the receiver operating characteristic (ROC) area under the curve (AUC) that measures the relative ranking of predictions. The choice of metrics is dependent on the desired outcomes. For example, NIJ selected false positives due to the meaning of the error. We recommend aligning both metrics and incentives with desired outcomes.

Another interesting topic for a future challenge would be to forecast hotspots or levels of crime in hotspots similar to the real-time crime forecasting challenge in 2017. As discussed in the practical implications section, the results in our forecasts for the recidivism competition would benefit from further empirical research to understand the reasons for recidivism. Forecasting hotspots with machine learning algorithms could have more direct practical implications for prepositioning resources, surveillance, and sting operations. We do acknowledge that unreported crime is still an issue that needs to be considered to develop a comprehensive approach to allocating law enforcement resources. Empirical techniques can be used to estimate this unobserved crime, but predictions of observed crime are useful in a comprehensive approach to law enforcement. In the real-time crime forecasting challenge in 2016, a key element of the competition was designing the shapes of the cells. Perhaps it may be more beneficial for practical applications to focus efforts on forecasting if the grid of cells is already provided. It may also be interesting to forecast levels of crime and consider a larger geographic region to try to gain more insights about the correlations of factors with crime rates. The out of sample forecast for the future from the real-time forecasting challenge could also be a good feature for a future competition. The barrier to entry for the real-time crime analytics challenge was higher given the ArcGIS modeling required. The data set provided for the recidivism challenge is straight forward for people from different research backgrounds to

[11] Mohler, George; Michael D. Porter, "A note on the multiplicative fairness score in the NIJ recidivism forecasting challenge," Crime Science volume 10, Article number: 17 (2021). https://crimesciencejournal.biomedcentral.com/articles/10.1186/s40163-021-00152-x, Accessed on September 16, 2021.

understand and use while researchers had the flexibility to expand this data set with outside data to allow for more creativity. The time for the recidivism challenge was reasonable.