Accounting for Racial Bias in Recidivism Forecasting, Year 3 Male Parolees Report

National Institute of Justice Recidivism 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.

Accounting for Racial Bias in Recidivism Forecasting, Year 3 Male Parolees Report

National Institute of Justice Recidivism Challenge

August 31, 2021

Submitted by:

SAS Institute Inc. Team:
Mary Beth Carroll, Rodney Carson, Mike Clark, Adam Cottrell, Jim Georges, Tyler Nelson, Hiwot Tesfaye, Halil Toros, Sree Vuthaluru

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.

Contents

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.

SAS Recidivism Forecasting Challenge

Introduction

The National Institute of Justice (NIJ) recently sponsored a Recidivism Forecasting Challenge that "aims to improve the ability to forecast recidivism using person and place-based variables with the goal of improving outcomes for those serving a community supervision sentence." Participants were challenged to analytically understand variables impacting recidivism. NIJ encouraged merging supplemental data with correctional data to increase accuracy of risk predictions for individuals under supervision. NIJ anticipated outcomes of the challenge could lead to factors for evaluating risk and highlight the collection and analysis of data that contribute to reincarceration, as well as provide specific strategies to account for racial bias.

*SAS was again recognized in 2020 as a Leader in the Gartner Magic Quadrant for **Data Science and Machine Learning Platforms** and by the **Forrester Wave, Multimodal Predictive Analytics and Machine Learning Solutions***

Participating in the NIJ Recidivism Challenge allowed SAS to showcase the functionality of our analytics as well as our analytical subject matter expertise. As a result, SAS continues to be driven to effect positive change in justice and public safety by helping to identify the key components of recidivism. As a leader in machine learning technology, SAS can help analyze recidivism metrics to answer key questions that relate to the who, what, why, and when behind recidivism, ultimately improving policy and practices. SAS also can help proactively identify gaps related to the collection of data from disparate data sources to ensure this data can help improve longitudinal views as they relate to recidivism.

In this report, SAS describes the process taken to account for racial bias in recidivism forecasting, year 3 male parolees, addressing NIJ's specific questions in *blue italics*. SAS built models using a data-mining approach with both SAS and open-source software. SAS trained several machine learning algorithms to

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.

derive measures of variable importance. SAS found leakage in the data, which impacted our ability to make significant conclusions. In addition to our analysis, our report includes specific recommendations to reduce the gaps and limitations of the data provided for analysis to improve future Challenges.

Variables

Overall, *age*, *drug use*, *employment status*, *prior arrests*, *prior convictions*, and *gang affiliation* repeatedly appeared as important predictors of recidivism. However, SAS learned the supervisory variables directly leaked information about arrest status, so there are serious limitations in the generalizability of insights gained from the studied variables.

Data

The NIJ presented a clean and comprehensive data set. SAS' preprocessing efforts included label encoding ordinal variables and imputing missing values. Eleven (11) variables presented missing values. We show the extent of their missingness as well as our imputation strategy in Table 1.

Table 1. Missing Values with Imputed Values

Variable	Percent Missing (n=9,398)	Imputed Value
Case Information		
Gang Affiliation	**14.85**	**'Missing'**
Prison Offense	12.75	'Missing'
Supervision Level	**6.16**	**-1**
Supervision Risk Score	1.62	-1
Supervision Activities		
Drug Tests – Days	**21.97**	**-1**
Drug Test – THC	15.48	-1
Drug Test – Cocaine	**15.48**	**-1**
Drug Test – Meth	15.48	-1
Drug Test – Other	**15.48**	**-1**
Jobs per Year	5.68	-1
% Days Employed	**3.27**	**-1**

Of note, missing values for gang affiliation were exactly tied to the gender variable. All parolees missing gang affiliation were female, and no male parolees were missing the gang affiliation variable.

Were variables added to the data set? If so, detail the variables.

We explored a set of variables connected to PUMA groups representing income and household

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.

characteristics. A complete listing of these variables can be found in Appendix 1. We gathered these variables from the American Community Survey Public Use Microdata Sample data. As seen, these additional PUMA-based variables were not significant. The issue here is granularity. Because of the 'roll-up' of PUMA zones which was implemented to protect privacy, any demographic variable merged to the provided PUMA variable cannot provide any additional information specific to a parolee.

What variables were constructed? How were the variables constructed?

Dimension Reduction

As we noticed several clusters of correlated variables in the NIJ dataset, we applied principal component analysis (PCA) to mitigate multi-collinearity and reduce overall dimensionality. We took several approaches to grouping variables before applying PCA:

- All prior arrest and conviction variables
- Prior arrest and conviction variables by type of offense
- All supervisory variables
- Supervisory variables by type (drug, employment, program attendance, violations)

The first component of the all-priors approach was used as an input to our champion model. Other elements did not prove useful.

Leakage

We also constructed a set of 'leakage' variables from employment and drug test data. In this sense, 'leakage' refers to artifacts of data gathering/preparation, which are unrelated to the subject matter domain (i.e., recidivism), that may inadvertently clue data miners into deterministic data regions that can be exploited to artificially improve model performance. As more fully explained in Appendix 2, the leakage variables allowed us to identify a subset of parolees we knew for certain did not recidivate due to the way the data was prepared (i.e., it allowed us to identify outcomes based on mistakenly provided "future data").

Which variables were statistically significant?

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.

Several methods exist to determine the relevance and predictive power of variables. We share our findings of variable importance here, but for reasons discussed below, place no reliability on these estimates.

Our first view of variable importance derives from a standardized lasso-regularized logistic regression model, which provides both measures of significance (p-values) and magnitude (changes in odds). We created 20 instances of the model using the NIJ provided inputs and a different 60% training-validation split on each run. Table 2 provides information for the 10 variables which most often appeared significant ($p < 0.05$). Because the prior variables are highly correlated, they seem to 'take turns' entering the model, and thus no single prior variable is strongly significant or appears consistently across model instances.

Table 2. Variable Significance Derived from Logistic Regression.

Variable	Appearances	Mean P-Value ± STD	% Change Odds ± STD
Age at Release	**20**	**<0.001 ± <0.001**	**-31.36 ± 2.02**
Percent Days Employed	18	0.028 ± 0.089	-11.24 ± 2.30
Gang Affiliated – True	**18**	**0.009 ± 0.016**	**-42.47 ± 10.10**
Gang Affiliated – Missing	15	0.042 ± 0.041	-22.57 ± 5.48
Prior Arrest – Property	**14**	**0.043 ± 0.034**	**13.46 ± 2.47**
Jobs Per Year	12	0.107 ± 0.207	9.30 ± 3.48
Avg Days Between Drug Tests	**11**	**0.070 ± 0.070**	**7.35 ± 2.29**
Parole Violations – Instructions	10	0.118 ± 0.152	7.08 ± 2.56
Prior Convictions – Misdemeanors	**9**	**0.130 ± 0.183**	**10.89 ± 4.04**
Prison Years	9	0.133 ± 0.179	-7.37 ± 2.69

Our second view comes from tree-based approaches to modeling. Figure 1 on the next page displays an example of a variable importance chart generated from a random forest model. Both these views suggest *age*, *employment data*, *drug test data*, and having *prior arrests* as being predictive of recidivism. It is SAS' contention, however, that this assessment of predictive power is being clouded by target leakage.

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.

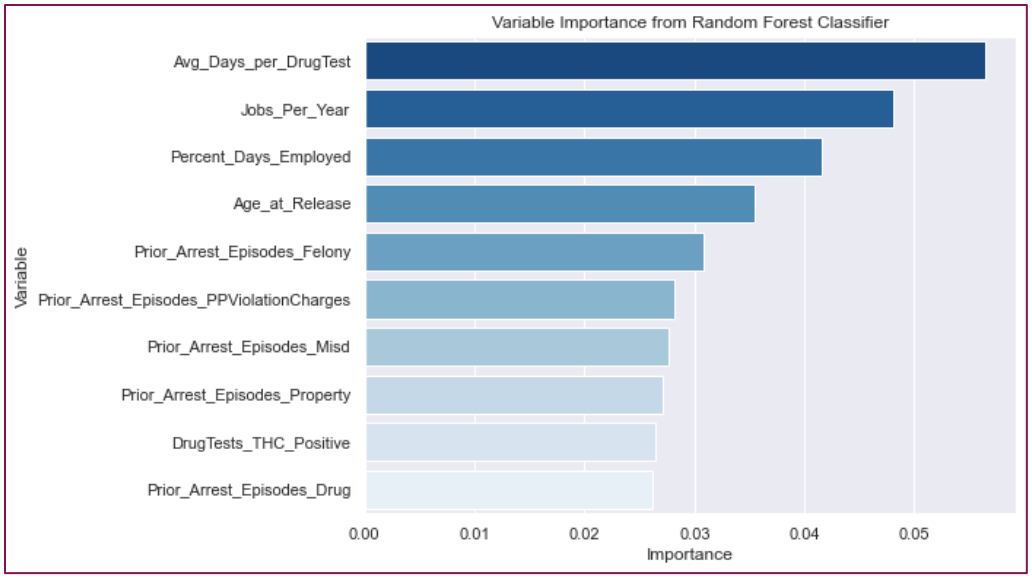


Figure 1. Variable Importance as determined by random forest classifier.

A third view of variable importance comes from our champion model and helps us understand the impact of the leakages. As mentioned, a set of variables was derived to explicitly indicate cases with deterministic (leaked) target values. A model was constructed from these indicators and all other available variables *to predict the champion model's predicted probabilities*. This *surrogate model*, in essence, reveals which factors (leakage or otherwise) are most responsible for driving predictions. Table 3 shows us the relative impact each group of variables has on affecting the predicted probabilities.

Table 3. Champion Model Variable Importance Summary

Variable Category	Importance
Target Value Leakage Indicators	**72%**
Prior Arrest and Conviction	7%
Age	**6%**
Supervision Activity	4%
Drug Use	**3%**
Employment	3%
Gang	**3%**
Other	1%

In this approach, each variable takes a turn being held at fixed values and the variability in the resulting predictions is recorded. The variables are then grouped into categories and the overall variability is aggregated.

The preeminence of leakage is apparent, and this is the primary reason for our skepticism of any insights

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.

gained from submitted models. It is SAS' conviction that any top-tier model in the Challenge will explicitly or implicitly owe its predictive prowess to the leaked information. As the leakages can deterministically delineate re-arrests in a significant number of cases, no model can accurately estimate the role other variables play in affecting recidivism. Given this, we do not recommend the field rely on any Challenge models for future decision making.

What variables were not statistically significant? How was this handled? For example, were they dropped from the overall model?

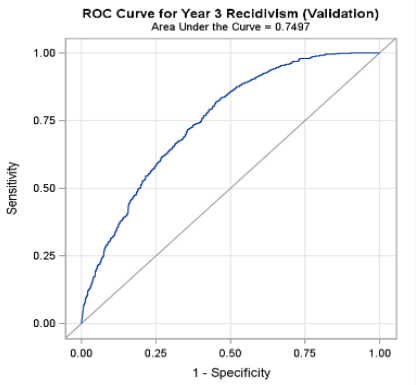
In terms of handling insignificant variables, we allowed our chosen algorithms to perform automatic selection. In our logistic regression, non-predictive variables were excluded through regularization. The more advanced 'black-box' algorithms inherently underweighted or ignored variables that did not improve predictive power.

Models

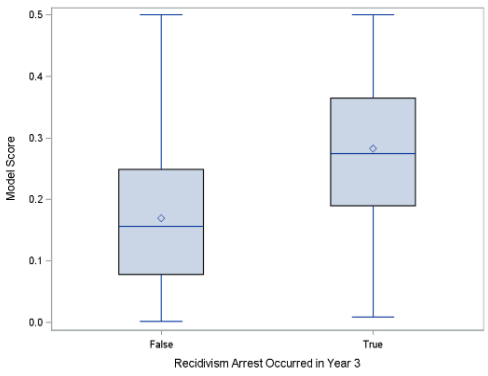
Using both open-source and SAS proprietary platforms, our team undertook multiple modeling approaches. We fit all models to a 60% training partition of the NIJ's training dataset and reported fit statistics against the remaining 40% validation partition.

What type of model was used?

Our champion model was an ensemble of neural networks, discussed in more detail in Appendix 3. Performance metrics against validation data for this model are shown in a., b., c., and d. in Figure 2.

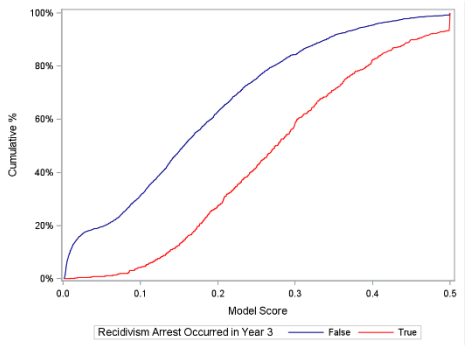


a. ROC Curve with AUC

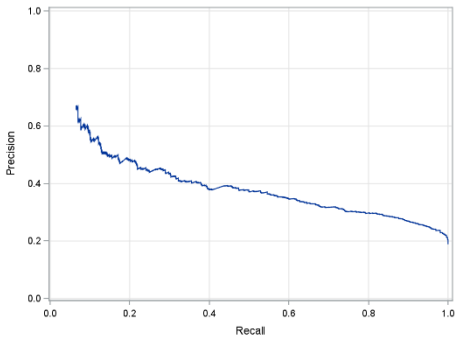


b. Predicted Probability by Target

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.



c. Cumulative Score Distribution by Target



d. Precision vs Recall

Figure 2. Panel of Performance Summary for Champion Model

Did you try other models? Were they close in performance? Not at all close?

Due to interaction effects, the 'black-box' machine-learning models generally outperformed regression-based techniques incorporating only main effects, however, the differences were small. We present a comparison of model performance in Table 4. The Brier score serves as a measure of predictive accuracy, while the Fair and Accurate score penalizes the Brier score based on the disparity between false positive rates for white and black parolees. As seen, most models performed similarly. In fact, so long as each model used an optimized set of hyperparameters, we believe the observed differences in performance were insignificant. Differences resulted from randomness in the validation data and/or the randomness used by the modeling algorithms themselves.

Table 4. Model Comparison by Brier Score on Validation Partition.

Model	Brier Score	Fair and Accurate Score
LASSO Logistic Regression	0.1481	0.8519
Random Forest Classifier	0.1471	0.8529
Gradient Boosting Classifier	0.1461	0.8539
Ensembled Neural Network	0.1444	0.8556

What other evaluation metrics should have been considered/used for this Challenge? For example, using false negatives in the penalty function.

A metric that incorporates both false positives and negatives seems appropriate for this task. F_β-scores or Youden's J-statistic are both common measures in this regard. Alternatively, the NIJ could define a custom loss-function to weight false positives and negatives in a way that reflects their domain knowledge (i.e., a false positive is 'worth' more or less than a false negative depending on the perceived

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.

severity of one type of error versus another).

A second suggestion for evaluating submissions would be to calculate a team's score as the average of performance on several bootstrapped samples of the test data. Our own explorations with cross-validation methods suggest category winners were determined by small and non-meaningful variations in scores. Differences on the order of ~5% seem possible simply due to randomness in the ML algorithms themselves and/or fluctuating proportions of events to non-events in different samples.

On the topic of assessing bias, future considerations should include more than one type of bias assessment (false positive rate). NIJ may consider these assessments in the future:

- Group fairness: assess whether there are differences in model accuracy by group (e.g., race, ethnicity, educational attainment, gender, etc.). Each group fairness metric may reveal differences in the model's accuracy or allocation of resources, in this case, by race. Metrics could include, but are not limited to, the following:
  - Equalized odds
  - False positive/negative parity
  - True positive/negative parity
  - Demographic parity
  - F1 score parity
  - AUC parity
- Individual fairness: assess whether nearly identical individuals receive similar scores.
- Counterfactual fairness: assess whether changes in certain inputs result in unexpected changes in the predicted outcome.
- Proxy variables: assess whether there are other variables in the data that are significantly associated with sensitive variables (e.g., race, gender) as they may unintentionally introduce bias to the model.
- Moving beyond known sensitive variables:

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.

- Assessing disparities should not be constrained to known sensitive variables.
- Error Analysis: Modelers should consider assessing cohorts within the dataset that have higher error rates than the average/global model error rate.
- Clustering: Modelers should consider identifying clusters within the data, characterizing clusters, and assessing performance, allocation of resource and quality of service disparities across those clusters.

Did the 0.5 threshold affect anything? Would your team recommend a different threshold?

Unfortunately, any exploration of potential bias in our models was stymied by the pre-set probability threshold of 0.5. While overall recidivism is high (57.80% of parolees), the year-over-year recidivism rate is low (29.83%, 25.71%, and 19.06%). Yearly models, therefore, will tend to assign positive event probabilities that are close to the year-over-year rates. As a result, our champion model estimated very few predicted probabilities over 0.5 for any parolee.

An approach for creating a meaningful bias-assessment threshold is to use the marginal recidivism rate for a given year as a classification cut-off value (~30% for year one, ~26% for year two and ~19% for year three). The practical interpretation of these threshold values is to classify a case as "risky," when there is a higher-than-average chance to recidivate. Using these lower thresholds guarantees a significant number of cases for assessing model biases and having the desirable theoretical properties related to balancing model sensitivity and specificity.

Did the fact that the fairness penalty only considered false positives affect your submission?

Given the construction of the fairness penalty, we surmised that no gain in Brier score could mitigate an increase in false positive rate. To maximize the Fair and Accurate score, we limited all predictions to a maximum of 0.4999, and thus our final predictions included no false positives. As measured by the evaluation criteria, our model was unbiased, although it should be noted that this is a somewhat artificial perception (i.e., a different cut-off threshold could yield a false positive rate difference across various

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.

demographic groups).

Future Considerations

Are there practical/applied findings that could help the field based on your work? If yes, what are they?

It is difficult to derive valuable practical findings from a short-term contest using publicly available data. We believe practical/applied findings could be achieved in the future with more granular, detailed data and a more rigorous analytic process. We are hesitant to accept any insights gained from submitted models and question the reliability of their performance. We would also discourage the use of any submitted models in live environments.

For future Challenges, what should NIJ consider changing to improve Challenges? For example, more/less time, different topic, or data issues (missing data)? What should NIJ have considered changing (other than metrics) to improve this Challenge?

We believe that the 'supervisory activity' variables leaked information about the targets. While the leakages were important to the outcome of this Challenge, it is equally important to know how they arose. A detailed description from those who prepared the data would help avoid similar issues in the future. This challenge would have been greatly improved had the supervision variables been structured to provide an accurate 'snapshot' of a parolee's information for the relevant time-period, thereby eliminating the leakage of future information into the training data. Finally, selection of a different positive-event threshold would have enabled a much better discussion about the biases present in contestants' submissions.

- NIJ should provide more detailed data, in a format that allows integration with supplemental external data sets. In general, supplemental data will be required for more accurate and more applicable models, and the NIJ should consider what could be provided while still protecting PII. A date-of-release variable is one particularly relevant feature that could be added, as additional data that could be joined to a parolee most likely has a temporal aspect (i.e., demographic or economic data).
- NIJ should consider collaborating with researchers studying the intersection of equity and recidivism or organizations that allow the NIJ to tap into the voices and experiences of previously incarcerated

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.

citizens. Participants could be vetted and held under NDAs, which would allow teams to have access to parolees' valuable supplemental PII data. The results of such work would be imminently more applicable to real-world situations.

- Requiring a model card/fact sheet along with the submission of any model would help promote transparency of the modeling process and provide clear guidance to end users on how to use the model. The model card/fact sheet also defines the intended use of the model, capture ethical considerations, known risks of misuse, risk mitigation techniques applied/should be considered, and overall limitations of the model.
- Finally, future challenges could benefit if models could be developed in a secure, CJIS compliant environment with the use of the full spectrum of variables available in CJIS data.

Conclusion

SAS concludes the following with respect to stated Recidivism Challenge Goals and offers recommendations for consideration:

NIJ Recidivism Challenge Goal #1: Enhance recidivism forecasting using person- and place-based factors.

SAS Conclusions:

1. *Age at release*, *drug use*, *employment status*, *gang affiliation*, and *priors* repeatedly appeared as important predictors of recidivism. However, it seems that a form of look-ahead bias introduced by the supervisory variables has directly leaked information about arrest status, so there are serious limitations in the generalizability of insights gained from the studied variables.
2. 'Leakage variables' created from *percent days employed*, *jobs per year*, *average days between drug tests*, and the *percent drug tests positive* variables give away arrest status for 18.4% (n=1,726) of the parolees in the Round 3 training data. These features are **not** powerful predictors – they are actually proxies for the target variable.

SAS Recommendations:

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.

1. SAS recommends restructuring the supervision variables to provide an accurate 'snapshot' of a parolee's information for the relevant time-period, thereby eliminating the leakage of future information into the training data.
2. A detailed description from those who prepared the data would help avoid similar issues in the future.
3. As SAS observed impacts of most variables are either inflated or masked due to the presence of these leakages, the models generated through participation in this challenge are not reliable and should not be used in practice.

NIJ Recidivism Challenge Goal #2: Increase the accuracy of risk predictions for all individuals under Correction Custody supervision to provide the most accurate risk and needs assessments available.

SAS Conclusions:

1. We cannot deliver reliable estimates of variable importance or place confidence in the accuracy of predictions due to the dominant effect of the leakage variables.
2. Research shows advanced modeling techniques can make use of increasing amounts of data because the models are able to detect complicated interactions between inputs and outcomes. The limited granularity of the data provided in this challenge prevents submitted models from realizing these improvements in predictive accuracy.
3. The choice of threshold and resulting actions of competing teams may cause submitted models, including our own, to artificially exhibit low levels of bias.

SAS Recommendations:

1. Provide more granular and thorough data. Organizers could create enriched features from data traditionally considered too complex to analyze. For example, a social network analysis of a parolee's relationships with other inmates or community contacts may provide strong evidence for their likelihood to recidivate.

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. Consider requiring a model card/fact sheet along with the submission of any model to promote transparency of the modeling process and provide clear guidance to end users on how to use the model. These cards also help to define the intended use of the model as well as document ethical considerations, known risks of misuse, risk mitigation techniques applied/should be considered, and overall limitations of the model.
3. Regarding assessing bias, more than one type of bias assessment (false positive rate) should be considered. The SAS Data Ethics Practice discussed later can help with these assessments.

NIJ Recidivism Challenge Goal #3: **Understand how supplemental data can be integrated with official records to enhance precision and increase accuracy of the models.**

SAS Conclusion:

1. SAS analyzed the available supplemental data and determined the PUMA data offered the only method to join extra data to the analysis. But because of the 'roll-up' of PUMA zones which was implemented to protect privacy, any demographic variable merged to the provided PUMA variable cannot provide any additional information specific to a parolee. Accordingly, SAS concluded the other supplemental data would not be more useful than the PUMA group itself.

SAS Recommendation:

1. SAS agrees adding supplemental data enhances precision and increases accuracy. However, a model must be able to join it to existing data for analytical purposes for it to be useful. SAS recommends collaborating with NIJ in the development of models with the use of the full spectrum of available CJIS data variables under a CJIS Security Addendum approved by the US Attorney General.

In conclusion, SAS reiterates our gratitude for the opportunity to participate in the NIJ's Recidivism Forecasting Challenge. As the worldwide leader in analytics, SAS understands the value of data and its power to positively impact our communities. There is no question that the challenge provided a unique

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.

opportunity for us to expand our knowledge of analytics as applied to supervised populations, however, the results of the challenge should be closely scrutinized due to significant data quality concerns. It is of our utmost concern that the analytical processes implemented throughout the justice system meet rigorous standards that account for reliability, validity, and bias.

SAS takes the ethics and equity of such models seriously and has a range of options for empowering transparency of models. SAS acknowledges that technology alone cannot solve for equity. Robust and proactive governance and supporting processes to reduce the risk of impact to society is needed. As such, the SAS Data Ethics Practice (DEP) supports SAS, our customers, and communities with human-centered best practices and solutions that promote responsible, equitable judgments where SAS technologies are deployed, with a particular emphasis on minimizing harm to vulnerable populations. The DEP institutionalizes norms – within SAS and commercialized offerings for customers – guided by a set of human centered principles. The DEP also provides thought leadership on minimizing risks to vulnerable populations and the public interest while ensuring SAS' approach to data ethics is consistent and coordinated. SAS understands the bias and ethical challenges historically tied to criminal justice data and is committed to promoting the responsible use of data in this space and welcomes collaboration with NIJ on these efforts (alongside researchers and other subject matter experts).

About SAS

Governments increase oversight, reduce losses and pinpoint risk when they use automation and artificial intelligence to monitor all activities, not just a small sample. The digital transformation of justice involved population data drives better decisions that improve outcomes and enable evidence-based re-entry services to connect individuals with treatment and support services. As the global leader in analytics, SAS enables you to integrate, authenticate and standardize data. By harnessing quantitative and qualitative data, governments can better identify risks, analyze effectiveness of interventions, evaluate program and policies, identify gaps, and improve outcomes for justice involved populations.

SAS has been at the forefront of tackling the world's hardest problems with data and analytics for over 45 years. SAS software and services are used across industries and government agencies to assess equitable practices such as policy simulations, disparate impact analysis, and adherence to fair lending laws. SAS has been a pioneer in the advanced analytics industry, including in the areas of AI and ML. SAS is recognized as a Leader in 2020 Gartner Magic Quadrant for Data Science and Machine Learning Platforms. We have extensive industry and subject matter expertise, as well as our commitment to the core concept of "Responsible AI" - AI that is governed, transparent, interpretable, and ethical.

Whether managing financial, public safety, environmental or regulatory risks, governments need a comprehensive and proactive approach to monitor agency performance. SAS provides the analytical capabilities needed to identify and assess risks by integrating data from disparate sources, identifying complex linkages and generating better oversight.

The SAS platform allows a variety of users to build and expand upon sophisticated models to get accurate results – all in a single, collaborative environment. In this specific instance, our expert modelers and data scientists were able to access SAS capabilities from their preferred coding environment – Python, R, Java or Lua. The comprehensive pipeline view allowed for data preparation, model building, model comparison, and model selection with easily interpretable results.

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.

Agencies can reduce the time spent on ingesting and manipulating data by 50% or more. This allows individuals to move more quickly through their workloads. SAS also delivers the risk-based alerting and behavior analysis needed to focus efforts for the most effective results. Recent, relevant projects include:

- Indiana Department of Corrections (DOC): Reduce and Mitigate Violence: Using SAS, the Indiana DOC started aggregating data from various systems including its offender information management systems to detect and communicate when offenders were most volatile and systematically reduce assaults on officers and fellow offenders. The tool needed to do three things: incorporate all relevant data sources, more accurately predict violence, and clearly communicate this insight to facility staff.
- North Carolina: Criminal Justice Law Enforcement Automated Data System (CJLEADS) is an on-demand, web-based application created and hosted by SAS. It integrates criminal offender data to provide courts, law enforcement, probation and parole agencies with a complete view of a criminal offender. The system also includes a watch list that allows officials to monitor the change of any offender's status, such as arrests, future court appearances or a release from custody.
- Wake County, NC: SAS provided entity resolution to match data across systems of service to identify the population of individuals with recurring interactions with costly county services (e.g., lack of stable housing and its effects on chronic cycling between hospitals/jails). By better understanding the "familiar faces," Wake County can ensure they are proactively targeting programs, such as subsidized housing, to the appropriate at-risk individuals.
- Riverside County, CA: Riverside County partnered with SAS to reduce reincarcerations for probationers and emergency room visits via screenings and referrals to targeted interventions and programs. With SAS, Riverside can evaluate its Whole Person Care program, combining data from multiple county departments, including the public hospital, behavioral health, jail health and probation. This program serves the probationary population with some early outreach and engagement pre-release in jail.

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.

- SAS helps a **State Department of Education** achieve its goal of providing students with equitable access to excellent educators by analyzing whether certain student groups are disproportionately placed with teachers who are identified as less effective in terms of their students' growth. Users can focus on students identified as low-achieving, economically disadvantaged, English learners, special education, or students of color. The analysis leverages existing teacher value-added reporting with student-level information to report whether equity gaps exist in districts and schools. This reporting allows the State to put concrete information and resources in the hands of State, district and school administrators to add to their understanding of equitable access, thus increasing education opportunities for students.

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.

Relevant Literature

In the last few decades, much research has been devoted to evaluating the accuracy and fairness of using predictive analytics to create risk assessment instruments, or RAIs. Two commonly used tools are the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) and the Level of Service Inventory Revised (LSI-R). Both tools generate an offender's recidivism likelihood based on several static and dynamic risk factors such as demographics, criminal history, and personal characteristics. SAS has provided a separate attachment, **Recidivism Forecasting Relevant Literature**, discussing the relevant literature and issues regarding predictive analytics and RAIs.

For the NIJ Challenge, rather than replicate existing tools and research, SAS followed a data-mining approach to model building using SAS and open-source software, training several machine learning algorithms to derive measures of variable importance.

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.

Appendix 1 - Additional PUMA Variables

Table 5. Additional Puma Variables

Description	Value Key
18-65 Adult Population	Integer
18-22 Age Population	Integer
18-30 Age Population	Integer
Not Worked Last 12 Months	Integer
Not Worked Last Week	Integer
Adults Not Married	Integer
Adults with Annual Earnings < 1000	Integer
Income to Poverty Ratio < 50%	Integer
Income to Poverty Ratio < 125%	Integer
Median Earnings	Integer
Households Received SNAP	Integer
Adult Households	Integer
Median Property Value	Integer
Median Household Income	Integer
Urban Description	Large metro, Non metro, Small metro
Estimated PUMA Count: Drug Offenses	Integer
Estimated PUMA Count: Other Offenses	Integer
Estimated PUMA Count: Property Offenses	Integer
Estimated PUMA Count: Sex Offenses	Integer
Estimated PUMA Count: Violent Offenses	Integer
Estimated PUMA Count per 10K: Drug Offenses	Real number
Estimated PUMA Count per 10K: Other Offenses	Real number
Estimated PUMA Count per 10K: Property Offenses	Real number
Estimated PUMA Count per 10K: Sex Offenses	Real number
Estimated PUMA Count per 10K: Violent Offenses	Real number
PUMA Count: Less than HS	Integer
PUMA Count: HS Graduate	Integer
PUMA Count: At Least Some College	Integer

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.

Appendix 2 - Explanation of Leakage Variables

We should note that while the leakages described here are important, just as important is knowing how they arose. A detailed description from those who prepared the data would help avoid similar issues in the future.

In our analysis, we noticed that certain combinations of variables created regions of a single class. Figure 3 below shows such an occurrence, plotting percent days employed against jobs per year.

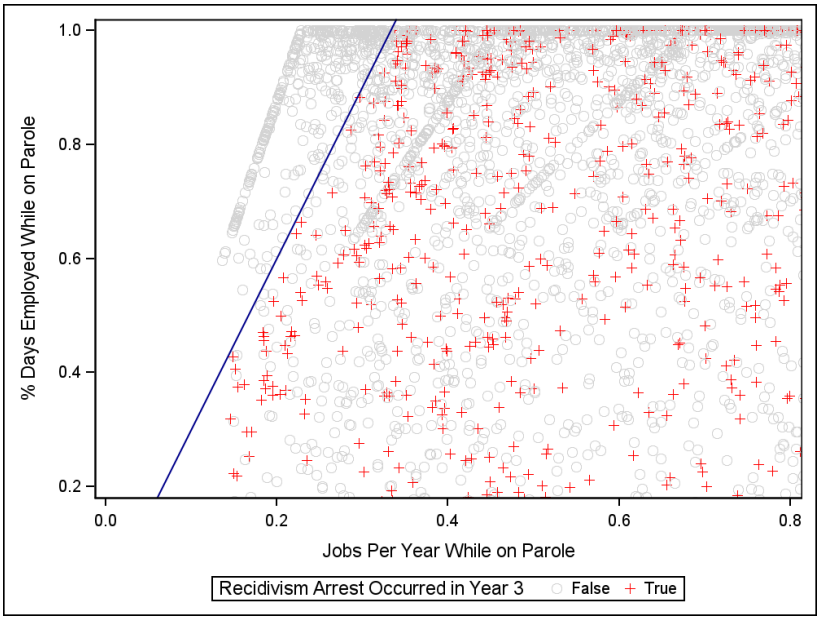


Figure 3. Deterministic Determination of Recidivism

Any parolee above the blue boundary line is guaranteed to have not recidivated. We also note a series of ray-like structures of parolees who did not recidivate. We derived mathematical relationships (see equations 1 and 2 below) from these variables which determined, independently of the model, which parolees recidivated.

$$\%\ Days\ Employed > 3 \times (Jobs\ per\ Year) \quad (1)$$

$$mod\left(\frac{Jobs\ per\ Year}{\%\ Days\ Employed} \times \frac{3}{0.684837}\right) = n, where\ n = 1, 2, 3, \ldots \quad (2)$$

We did not develop a satisfactory explanation for the existence of these relationships, but we believe it is a form of look-ahead bias introduced by the supervisory variables. Our assumption is that the values for

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.

the supervisory variables reflect their status at the time the dataset was created, i.e., sometime in 2020, instead of relative to a parolee's release date. In some cases, we may be getting 5, 6, or even 7 years of supervisory data as opposed to the expected two years' worth, indicating that the parolee was indeed not rearrested within 3 years.

We should note that in a small number of cases (n=7), our leakage variables classify parolees incorrectly. Figure 4 on the following page shows these instances. We again are unsure why this occurs but believe it may have to do with a parolee being rearrested before the end of 3 years, but being released in time to accrue further supervisory data.

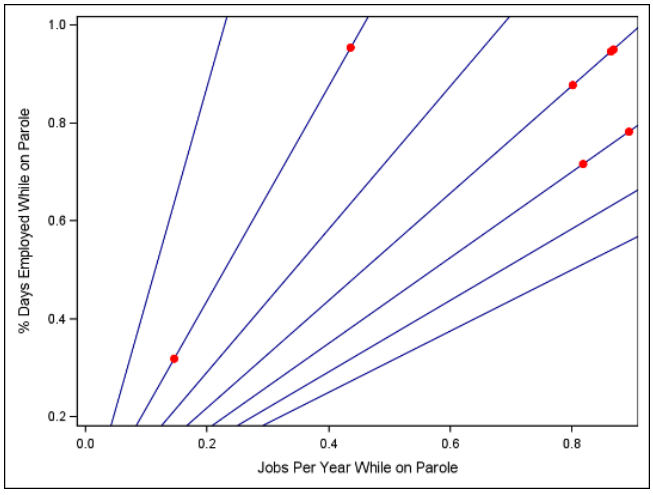


Figure 4. Leakage Variable Errors

A second class of leakage variables was derived from drug test data. In some cases, parolees were missing time between test data, but nonetheless had information showing they returned positive tests. In these cases, parolees could also deterministically be classified as not having recidivated.

$$(Avg\ Days\ Between\ Tests\ is\ Missing) \cap (Any\ Drug\ Test\ \%\ Positive\ is\ not\ Missing) \cap (Any\ Drug\ Test\ \%\ Positive > 0) \rightarrow Non\ Arrest \quad (3)$$

We were unable to derive a plausible explanation for the relationship among the drug test variables.

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.

Appendix 3 - Champion Model Description

The champion model was an ensemble of multiple neural networks. We allowed the model to make use of all available NIJ inputs, our additional PUMA variables, and the first principal component of the priors variables. By nature, neural networks reduce the impact of weak predictors by assigning low edge weights to those inputs.

The actual model consisted of the following elements:

- Ensemble 1:
  - Averaged prediction of 30 neural networks
  - Leaked cases excluded from model training
  - Predicted probabilities for leaked rows set to zero
- Ensemble 2:
  - Averaged prediction of 30 neural networks
  - Leaked cases not excluded from model training
  - Allowed models to predict probabilities for leaked cases
- Aggregated predicted probabilities equaled the average of the prediction from each ensemble
- Final predicted probabilities were capped at 0.4999

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.

To contact your local SAS office, please visit: sas.com/offices

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.