# Predicting Recidivism with Neural Network Models

Sara Debus-Sherrill, Ph.D. and Colin Sherrill

August 2021

## Introduction and Relevant Literature

Each year, millions of people leave jails and prisons to re-enter communities around the United States (ASPE, n.d.). Unfortunately, the majority (62-68%) will reoffend within three years (Durose & Antenangeli, 2021; Durose, Cooper, & Snyder, 2014). These communities are faced with the challenge of how to support these individuals in their reintegration while also ensuring public safety is upheld. To assist with both of these goals, risk-need assessment instruments have grown in popularity to predict an individual's risk for reoffending and identify targets for behavioral intervention expected to influence recidivism risk. These instruments use information about the individual, gathered from existing records and interviews with the person, to make these assessments.

Risk-need instruments have been both praised and criticized, in turn. While some argue that these instruments have greater accuracy than unstructured human decision-making and serve to reduce bias in the justice system (Goel et al., 2021; Grove et al., 2000; Jung et al., 2020; Lin et al., 2020), others claim the imperfect accuracy and remaining bias raise ethical questions about whether and how to use these tools (Angwin et al., 2016; PJI, 2020). Given these ethical concerns and the implications for individuals' personal liberties based on these tools' results, it is critical to carefully examine these tools and ensure they are as accurate and fair as possible.

The National Institute of Justice (NIJ) created the Recidivism Forecasting Challenge to identify promising approaches to predicting recidivism. Competition participants used existing data provided by NIJ to create models to predict recidivism for a sample of people on parole in Georgia. The performance metrics for the competition prioritized both accuracy and reduced



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.

racial bias for awarding winners. This paper details the approach used by "Team Sherrill" to predict recidivism, which resulted in awards in the Small Teams category for predicting recidivism for females under parole supervision in the first year after release (2nd place) and in the third year after release (1st place).

## Methodological Overview

*Sample*. NIJ provided de-identified data of adults released to parole supervision in Georgia from January 1, 2013 through December 31, 2015. Individuals were excluded who were a race/ethnicity other than White or Black, under the age of 18, and/or missing critical pieces of data. The data was split into a training dataset (70% of the data, n=18,028) and test data set (30% of the data, n=7,807), so that models could be built with the training data and then applied to people in the test data. The test data sets for years 2 and 3 excluded those individuals from the test data population who had already recidivated in a prior year.

*Predictor Variables*. For the first round predicting recidivism in the first year after release, the predictor variables in the test data included basic demographics (e.g., gender, race, age, education, number of dependents), an approximate geographic location indicator (PUMA group), criminal history (e.g., offense, years in prison, prior arrests and convictions, prior revocations, gang affiliation), supervision risk score and level, and release conditions. For years 2 and 3, additional predictor variables were included in the test datasets, such as supervision violations, program attendance, resident changes, drug test outcomes, and post-release employment. These variables were not included in the year 1 test data to mimic the reality that supervision agencies do not yet know these interim, short-term reentry outcomes at the start of supervision. Participants could also use supplemental data to build their models if they liked

(e.g., data about geographic locations, data about Georgia's supervision populations or policies). However, our team did not use any supplemental data.

*Outcome Variables*. The outcome variables of interest were whether or not the individual was arrested for a new felony or misdemeanor crime in the first year of release, second year of release, or third year of release. These outcome variables were only included in the training dataset and had to be predicted for the test datasets. This was done through a sequential series of submissions where competition participants provided predictive probabilities for each individual in the test data for whether or not they recidivated within each year-long time period. Our team provided predictions for all three time periods.

*Analytic Method*. To prepare the data, we imputed missing values with the mean value for that variable, stratified by gender and race. The Gang Relatedness variable was not available for women in the sample, and was therefore imputed without gender stratification. Categorical variables were each expanded into a series of dichotomous or "dummy" variables.

We used deep neural network (DNN) machine learning to build models to predict recidivism with the provided data. We selected this approach because neural networks learn from the data provided as opposed to being limited to our own existing knowledge, can be trained to minimize cross-entropy and other selected loss functions, are able to model complex and non-linear relationships, and are well suited for building models with interaction effects such as those related to gender and racial differences. We used the Keras library in Python to train and apply DNN models using TensorFlow 2.

For model development, we further split the training data into two sub-sets of training data (80% of the data) and test data (20% of the data). This allowed us to examine model performance and assess over-fitting for various versions of our models at the expense of not

using all of the data for training. Once the parameters were finalized based on these earlier iterations, the entire training dataset was then used to train a final model to be applied to NIJ's test data.

*Model Performance Metrics*. In order to assess model performance, we used a few different metrics. First, we used the metrics designated by NIJ to assess validity and racial bias. Validity was measured using the Brier score, as defined below:

$$Brier\ Score = \frac{1}{n}\sum_{t=1}^{n}(f_t - A_t)^2$$

Racial bias was assessed through a "fair and accurate" measure which accounted for disparate false positive rates between races, as well as overall model accuracy. This measure was defined as:

$$Fair\ and\ Accurate = (1 - BS)(FP)$$

where FP is defined as:

$$FP = 1 - |FP_{Black} - FP_{White}|$$

We also examined the raw rates of false positives by race, as well as the AUC. To calculate the AUC, both sensitivity (True Positive rate) and specificity (True Negative rate) metrics are calculated for a specified threshold (e.g., 0.5) of test data predictions. A "receiver operating characteristic curve" (ROC curve) considers all thresholds and plots each (sensitivity, specificity) as a point along that curve. The AUC is the area under the curve, where a value of 1.0 indicates a perfect model.

For each model we attempted, we generated these metrics for the female sample, male sample, and combined gender sample. We selected our model to use with the final test data set

based on which model was performing best on these metrics with the training data. More detailed information about the variables used and models employed, along with their findings, are shown below in the requested sections.

## Variables

Because we were using a DNN approach which can handle larger numbers of variables and learns how to discard and transform variables based on their predictive power, we included all variables for model-building. One exception to this was we did not include "Residence PUMA" due to the high cardinality[1] of that variable. Since this is a "black box" approach, it is difficult to interpret which variables were used or how they were applied in the model.

When we received the additional variables for the second round of predictions, we attempted to combine and/or drop some employment variables which we believed could be consolidated or were less useful. However, the model performed better without these manipulations, so we re-included the variables as originally structured. No variables were added to the dataset from outside sources.

## Models

To build our DNN models, we relied on the Keras Python library to interface with TensorFlow 2. We experimented with a few different approaches to building our model with the training data, comparing the metrics each time, before settling on the final models we used with the test data. We trained the model to optimize cross-entropy as opposed to mean squared error,

[1] There are 25 different values for census zone, each of which would become its own dichotomous variable. This dramatically increased the number of variables, without providing additional accuracy based on our metrics.

because optimizing based on mean squared error caused the model to become overly conservative[2] and predict most people's probability of recidivating to be right around 0.5.

We provided three sets of predictions, one for each post-release year period. Our models varied slightly for each post-release year period. This was partially due to customizations for what we were predicting and partially due to experimenting and trying different things during each round.

For the first round predicting whether individuals recidivated in the first year after release, we built our model with the training data, experimenting with different neural network parameters and ways to structure the outcome variable. In the end, we used an outcome of 0 = no recidivism, (1/3) = recidivated in the third year, (1/2) = recidivated in the second year, and 1 = recidivated in the first year. We included all of the variables available for the first year except for the Residence PUMA, as explained above.

For the second round, we included the newly provided variables in the model. We again tried different ways of structuring the outcome variable for building our model with the training data, but ultimately used an outcome variable that was coded as: 0 = no recidivism, (1/2) = recidivated in year 3, and 1 = recidivated in year 1 or 2. We also adjusted the parameters of our neural network based on experimentation, removing layer 3 and changing the number of nodes in layer 2.

In this second round, we attempted to perform manual adjustments of the DNN results to reduce racial bias in the predictions once the model was finalized. We tried various linear transformations to reduce Black prediction values by varying factors, depending on how far from the 0.5 mid-point the prediction value was. However, the resulting predictions did not perform as

[2] In this context, "conservative" indicates that the model preferred to make indecisive predictions (likelihoods around 0.5) for all individuals.

well as the model without these adjustments, based on NIJ's selected metrics. Therefore, this post-processing step was abandoned.

For the third round predicting recidivism in year 3, we identified an issue where our models were suffering from a calibration error where the average prediction was too large. Since people who recidivated on year 1 and 2 were included in the training data (but not the test data), this was distorting the model's application to the test data. To address this, in the third round, we applied weighting to more heavily weight non-recidivators in the training data, so this would more closely mirror the test data's removal of those who recidivated in earlier years. This fixed the calibration error, so that the average model prediction matched the average outcome variable.

We did not apply any post-model corrections to reduce racial bias for the third round, because there were so few people predicted to recidivate, resulting in a low false positive rate already. Therefore, any corrections would have very minimal impact on the fairness metrics. Given this and the fact that these types of corrections did not provide an overall benefit in round 2, it was not fruitful to use these for round 3.

We considered a couple different approaches to structuring the outcome model for the third round. We first tried creating a model with the training data where the outcome variable was structured as: 0 = no recidivism, (1/3) = recidivated in year 1, (1/2) = recidivated in year 2, and 1 = recidivated in year 3. This model predicted that no one would recidivate in the third year based on the 0.5 threshold. We then tried modeling the training data where the outcome variable was structured as: 0 = no recidivism and 1 = recidivism at any point during the 3 years after release. This model, in contrast, did predict a small number of people to recidivate.

We debated which model to use. Ultimately, we decided to use the latter model for predicting recidivism with the test data, because we felt it was more in the spirit of the

competition to attempt to predict individuals who would recidivate, as opposed to predicting no one recidivates. However, the data proves that it is exceptionally difficult to accurately predict individuals who will recidivate in the third year after release if they have not yet recidivated in the first two years. Therefore, in a real life setting, it would be of limited utility and therefore unethical to predict individuals recidivating for the first time *specifically* in their third year of release.

## Conclusions and Future Considerations

The NIJ Recidivism Forecasting Challenge was a unique opportunity to grow knowledge around recidivism prediction models. The competition format allowed us to creatively explore model-building techniques we might not otherwise use due to their "black box" nature. In the end, this freedom led us to models which were particularly beneficial for predicting recidivism among women. This may be due to neural networks' ability to handle multiple interactions simultaneously. Since women's pathways to offending are known to be different from men (Brennan et al., 2012; Dehart et al., 2014; Simpson, Yahner, & Dugan, 2008), there may be many relevant interaction effects that other modeling approaches using combined gender samples may not capture.

The success of our models in the NIJ competition indicate that DNN models may be useful for predicting recidivism in the real world. However, a number of limitations exist which need to be fully considered prior to implementation. To understand if DNN or other machine learning models are appropriate for predicting risk of recidivism in real-life settings, future empirical work should examine how these types of models compare to more simple algorithms and determine whether the benefits outweigh the limitations. If the accuracy and fairness benefits are

not sufficiently large to warrant the trade-offs listed below, a simpler, more transparent approach may still be preferable.

*Limitations*. Unfortunately, one drawback of DNN models is that is difficult to identify how the model is handling different variables in combination with each other and what variables or interaction effects are significantly contributing to the model. This is a serious drawback of such "black box" models where transparency is traded for computational sophistication. In addition to how this limits the field's knowledge base of what predicts reoffending, this has been a critique in the ongoing debate about racial bias in risk-need assessments since it is unknown how the models may be introducing or perpetuating bias.

Moreover, DNN models are non-deterministic, so results will change slightly each time the model is trained, even if nothing entered into the model changes. This poses another challenge for real-life application, as we would need to contend with the ethical implications of the same model and sample predicting someone as being likely to recidivate in one instance of training the model, but predicting them not to recidivate based on another instance with an identical setup.

At this point, it is important to note that any modeling using criminal history variables is prone to systemic biases influencing the outcomes of that model. While this concern has been raised in recent arguments related to artificial intelligence or machine learning, it is also true of any sort of modeling using these types of criminal history variables (or other variables potentially linked to structural racism). We attempted to use post-model corrections, as described above, to address some of these concerns and reduce racial bias in our models. While this did not result in better performance according to NIJ's metrics for this competition, we still believe this is an avenue worthy of further experimentation to see if this approach can reduce racial bias in

other circumstances. That said, even without modifications, the DNN models' fairness penalties were greater than 0.96 for nearly all of our modeling attempts.

Other limitations are that the variables provided were limited and focused primarily on demographics, criminal history, risk, release conditions, and supervision performance. Other data which are important predictors of recidivism (e.g., criminogenic needs) were not provided. The sample was also a sub-selection of re-entering individuals. It only included individuals on parole and excluded races other than White or Black race (Latino/a/x ethnicity was categorized as its own racial category and was therefore excluded). It is unknown if the models produced under this competition would be generalizable to all justice-involved individuals who do not meet these same criteria.

*Future Considerations*. We have a few suggestions to consider for future data competitions similar to this one. NIJ tasked participants with predicting recidivism within a particular year after release. While this created an interesting computational challenge, it may be unnecessary to constrain predictions in this narrow way. Predicting who is likely to recidivate in a two- or three-year period may be more practical for real-life settings and may result in more accurate models. It was a notable finding, however, to learn that it is exceedingly difficult to determine who will recidivate in their third year of release if they have not yet done so in the two prior years. In fact, it may be more ethical in a real-life setting to act as if no one will recidivate if they have not recidivated by this point and assume the risk of false negatives, as opposed to imposing liberty restrictions on someone in anticipation of them recidivating in the third year.

Secondly, using a threshold measure might not be the best way to capture bias, especially for the Year 3 competition where there are low base-rates of recidivism. It's possible for all predictions to be less than 0.5, but still be racially biased. As opposed to a threshold, it might be

more useful to examine the AUC by race. AUC spans the full range of prediction values and is generally regarded as a good performance indicator for modeling. Equivalent AUCs would indicate that you could make similar tradeoffs for sensitivity and specificity, by race.

It would be interesting for future competitions to include more variables for inclusion in models, such as additional criminogenic needs and stability factors. These would likely improve the accuracy of models, making them more ethical for potential use. Including other states and races/ethnicities would also be helpful for knowledge-building and generalizability. Finally, we hope NIJ will continue to support competitions such as this one, as they are likely to elicit creative approaches and interdisciplinary contributions which can help enhance understanding for some of the field's most pressing issues.

# REFERENCES


Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Assistant Secretary of Planning and Evaluation. (n.d.). Incarceration and reentry. https://aspe.hhs.gov/topics/human-services/incarceration-reentry-0

Brennan, T., Breitenbach, M., Dieterich, W., Salisbury, E. J., & Van Voorhis, P. (2012). Women's pathways to serious and habitual crime: A person-centered analysis incorporating gender responsive factors. Criminal Justice and Behavior, 39(11), 1481-1508. https://doi.org/10.1177/0093854812456777

Durose, M.R., & Antenangeli, L. (2021). Recidivism of prisoners released in 34 states in 2012: A 5-Year follow-up period (2012-2017). U.S. Department of Justice, Bureau of Justice Statistics. https://bjs.ojp.gov/sites/g/files/xyckuh236/files/media/document/rpr34s125yfup1217.pdf

Durose, M.R., Cooper, A.D., & Snyder, H.N. (2014). Recidivism of prisoners released in 30 states in 2005: Patterns from 2005 to 2010- update. U.S. Department of Justice, Bureau of Justice Statistics. https://bjs.ojp.gov/content/pub/pdf/rprts05p0510.pdf

DeHart, D., Lynch, S., Belknap, J., Dass-Brailsford, P., & Green, B. (2014). Life history models of female offending: The roles of serious mental illness and trauma in women's pathways to jail. Psychology of Women Quarterly, 38(1), 138-151. https://doi-org.mutex.gmu.edu/10.1177/0361684313494357

Goel, S., Shroff, R., Skeem, J., Slobogin, C. (2021). The accuracy, equity, and jurisprudence of criminal risk assessment. In R. Vogl (Ed.), Research Handbook on Big Data Law, (pp.9-28). http://dx.doi.org/10.2139/ssrn.3306723

Grove, W.M., Zald, D.H., Lebow, B.S., Snitz, B.E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. Psychological Assessment, 12(1), 19-30.

Jung, J., Concannon, C., Shroff, R., Goel, S., & Goldstein, D.G. (2020). Simple rules to guide expert classifications. Statistics in Society, 183(3), 771-800. https://doi.org/10.1111/rssa.12576

Lin, Z.J., Jung, J., Goel, S., & Skeem, J. (2020). The limits of human predictions of recidivism. Science Advances, 6(7), 1-8. DOI: 10.1126/sciadv.aaz0652

Pretrial Justice Institute. (2020). The case against pretrial risk assessment instruments. https://university.pretrial.org/viewdocument/the-case-against-pretrial-risk-asse

Simpson, S. S., Yahner, J. L., & Dugan, L. (2008). Understanding women's pathways to jail: Analysing the lives of incarcerated women. Australian & New Zealand Journal of Criminology, 41(1), 84-108. https://doi.org/10.1375/acri.41.1.84