Numeric Variables. Variable *Prior_Arrest_Episodes_Violent* was a numerical variable with values 0, 1, 2, 3+.

	Numerical Assignment	
	V1	V2
0	0	0
1	1	0
2	2	0
3 or more	3	1

Note that two variables were constructed for the numerical variables including a value such as "X or more," where X is a number. Otherwise, only one variable was constructed for the numerical variables.

Principal Components. In addition, a Principal Component Analysis (PCA) was conducted for 16 crime-related variables reporting the frequency of prior arrest and convictions (https://nij.ojp.gov/funding/recidivism-forecasting-challenge#prior-georgia-criminal-history). PCA revealed that these 16 crime-related variables could be grouped into four categories. Therefore, four composite variables representing these categories were constructed using a simple sum score from variables within each category.

Missing values. Two primary models were used in the model building process: Extreme Gradient Boosting (XGBoost) and Logistic Regression with Ridge Penalty. Since XGBoost doesn't require anything about missing values and can handle datasets with missing values, no action was taken, and missing values were left as missing when building the XGBoost models. For Logistic Regression, missing values were filled with the median value for each feature variable.

C.2. Processing variables from the 2018 American Community Survey (5-year Estimates)

A total of 157 variables were pulled from the 2018 American Community Survey (5-year Estimates). A similar approach as described earlier for numeric, binary, ordinal, and nominal variables were used to recode these variables. A list of these variables and the process applied to each variable is given in Table A3 in Appendix A. After processing these 157 variables, 295 predictor variables were constructed for use in subsequent modeling. In addition, a Principal Component Analysis (PCA) was run for all 295 variables; standardized composite scores for the first four principal components were added to the dataset. As the last step, the household level data were aggregated by taking the average of each variable across all households within a PUMA. So, a total of 299 features at the PUMA level were derived from ACS.

Since the forecasting is at the individual level, the PUMA level features had to be assigned to each individual based on the unique Residence Code assigned by NIJ. Each unique Residence Code consisted of two or more PUMAs (see Table A1 in Appendix A); therefore, the variables were aggregated by taking an average across all PUMAs assigned to the unique Residence codes for an individual assignment. Below is a sample that demonstrates how this procedure was done for some hypothetical variables.