Forecasting Recidivism: Mission Impossible

Cengiz Zopluoglu

Associate Professor

Educational Methodology, Policy, and Leadership

College of Education, University of Oregon

Author Note

Cengiz Zopluoglu https://orcid.org/0000-0002-9397-0262

Data, code, and additional online materials are publicly available at the project's GitHub repository (https://github.com/czopluoglu/nij-competition). This report is produced to respond to the National Institute of Justice's (NIJ) Recidivism Forecasting Challenge that aimed to increase public safety and improve the fair administration of justice across the United States. The models and procedures outlined and summarized in this report provided the 3rd best performance in the Challenge for predicting recidivism in Year 1 for male parolees, female parolees, and on average accuracy. They provided the best 5th performance in Year 2 for female parolees. All entries were submitted in the Large Team Category with a team name CrescentStar.

The author was awarded a total of $13,000 prize money in exchange for writing this final report. I have no other conflicts of interest to disclose. The findings and opinions reported and expressed in this report are solely my own and do not express the views or opinions of my employee, University of Oregon.

Correspondence concerning this article should be addressed to Cengiz Zopluoglu.

Email: cengiz.@uoregon.edu

Forecasting Recidivism: Mission Impossible

A. Introduction

The forecasting models become popular from time to time when there are limited resources for making decisions. National Institute of Justice's (NIJ) Recidivism Forecasting Challenge aims to improve the decision-making process by the community corrections officers with an appropriate balance of surveillance and support for persons on probation and parole. An unbiased forecasting model with adequate power may help reduce the caseloads of the community corrections officers by providing a nuanced view of the characteristics of persons most at risk for recidivating.

In this brief report, I outline the process of developing such a forecasting model. Two main algorithms were used for developing a prediction model in this attempt. The first type of model is a simple penalized regression model. The second type of model is Extreme Gradient Boosting, XGBoost, a highly effective gradient tree-boosting algorithm as has been demonstrated in many data science competitions. Both modeling approaches typically provide optimal solutions for rectangular data.

B. Datasets

B.1. Datasets provided by NIJ

The primary dataset was provided by NIJ and included observations from the State of Georgia about persons released from Georgia prisons on discretionary parole to the custody of the Georgia Department of Community Supervision (DCS) for post-incarceration supervision between January 1, 2013, and December 31, 2015. These datasets included a total of 48 predictor variables (e.g., gender, race, age at release) and three main binary outcome variables (0: not recidivated, 1: recidivated) in Year 1, Year 2, and Year 3. For Year 1 predictions, 33 predictor variables were available after excluding the supervision activities. For Year 2 and Year 3 predictions, all 48 predictors were available to use. A detailed description of available datasets can be found at this link (https://nij.ojp.gov/funding/recidivism-forecasting-challenge#ks8ofq). The detailed information about processing these variables before modeling will be given later under the Feature Engineering section.

B.2. Supplemental Datasets Compiled by the researcher

In addition to the datasets provided by NIJ, additional supplemental datasets were compiled. Most of the variables in these datasets were aggregated information about residential locations at release. NIJ provided 25 unique residence codes at release, and each unique residence code combined several US Census Bureau Public Use Microdata Area (PUMA). These 25 unique residence codes included a total of 72 PUMAs. First, the county names associated with each unique residence code were identified using the information at this link https://www2.census.gov/geo/maps/dc10map/PUMA_RefMap/st13_ga/. This link provides a PDF map for each PUMA code, and this map associates each code with a county name. Table

A1 in Appendix A provides a list of county names associated with each unique Residential Code provided by NIJ.

American Community Survey Public Use Microdata. Five-year estimates for 161 variables from the 2018 American Community Survey (ACS) were downloaded for 494,091 households in the 72 PUMAs from Georgia. ACS 2018 5-year estimates cover information from the 2013-2018 period. This period is selected so that the community data resembles the period for the data released by NIJ. All the variables in ACS were aggregated at the county level by taking the average across all households within a county.

This file can be found under the following link in the Github repository as an RData file.

https://github.com/czopluoglu/nij-competition/blob/main/data/supplemental%20data/geodata.RData

Crime Statistics. The crime statistics at the county level from 2013 to 2017 were compiled using the summary reports from the Uniform Crime Reporting (UCR) program by the Georgia Bureau of Investigation (https://gbi.georgia.gov/services/crime-statistics). These statistics included the number of crimes per 100,000 people for ten variables (murder, rape, robbery, assault, burglary, larceny, theft, arson, and total). The crime rates were aggregated by calculating the average crime rate across five years for each county.

The file that includes the county-level crime summary statistics can be found under the following link in the Github repository:

https://github.com/czopluoglu/nij-competition/blob/main/data/supplemental%20data/crime_summary.csv

Auxiliary Statistics. Other auxiliary information at the county level was compiled from the GeorgiaData initiative supported by the University of Georgia (https://georgiadata.org/data/data-tables). This information included county-level vital statistics, poverty data, lottery data, hospital data, unemployment data, voting data, public assistance data, population data, medicare data, sexually transmitted disease data, economic data, and agricultural data. The detailed information about these variables will be given later under the Feature Engineering section.

All supplemental data files that include the variables used in a model building can be found under the following link in the Github repository:

https://github.com/czopluoglu/nij-competition/tree/main/data/supplemental%20data

C. Feature Engineering (Variable Construction)

C.1. Processing variables in the original training and test datasets

The variables in the training and test datasets can be categorized as numeric, binary, ordinal, and nominal. For each variable with an ordinal nature, dummy variables were first constructed using a one-hot encoding approach. Then, additional variables representing polynomial contrasts were created. Also, if ordinal variables are presented as an interval, a numerical variable is created using the midpoint of each interval. For each binary variable, a single dummy variable was constructed. For each nominal variable, dummy variables were constructed using a one-hot encoding approach. Table A2 in Appendix A provides a list of all variables in the original dataset used in modeling, including the original nature of the variable, the process applied, and the constructed variables.

Below is an example of how each type is processed to construct new variables to represent the information in the original variable.

Ordinal Variables. Variable *Age_at_Release* in the original dataset had 7 categories presented as intervals: 18-22, 23-27, 28-32, 33-37, 38-42, 43-47, 48 or older. A total of 14 variables were constructed as following to represent the information in this variable.

	One-hot Encoding							Polynomial Contrast Coding						Numeric
	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14
18-22	1	0	0	0	0	0	0	-0.57	0.55	-0.41	0.24	-0.11	0.03	20
23-27	0	1	0	0	0	0	0	-0.38	0.00	0.41	-0.56	0.44	-0.20	25
28-32	0	0	1	0	0	0	0	-0.19	-0.33	0.41	0.08	-0.55	0.49	27
33-37	0	0	0	1	0	0	0	0.00	-0.44	0.00	0.48	0.00	-0.66	35
38-42	0	0	0	0	1	0	0	0.19	-0.33	-0.41	0.08	0.55	0.49	40
43-47	0	0	0	0	0	1	0	0.38	0.00	-0.41	-0.56	-0.44	-0.20	45
> 48	0	0	0	0	0	0	1	0.57	0.55	0.41	0.24	0.11	0.03	59

Nominal Variables. Variable *Prison Offence* type in the original dataset had five categories. A total of 5 variables were constructed to represent the information in this variable.

	One-hot encoding				
	V1	V2	V3	V4	V5
Drug	1	0	0	0	0
Property	0	1	0	0	0
Violent/Sex	0	0	1	0	0
Violent/Non-Sex	0	0	0	1	0
Other	0	0	0	0	1

Binary Variables. Variable *Gender* in the original dataset had two categories. A single dummy variable is constructed to represent the information in this variable.

	Dummy Coding
	V1
Female	0
Male	1

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.

After assigning PUMA level aggregated data to individuals based on their unique Residence Code, this supplemental dataset was merged with NIJ's original individual level training and test datasets.

Puma Level Data

PUMA	Variable X	Variable Y	Variable Z
1003	1	8	15
1008	2	9	16
1400	3	10	17
1500	4	11	18
1600	5	12	19
4300	6	13	20
4400	7	14	21

Individual Level Data

Subject	NIJ Residence Code	Associated PUMAs	Variable X	Variable Y	Variable Z
1	1	1003, 4400	(1 + 7)/2	(8+14)/2	(15+21)/2
2	2	1008, 4300	(2 + 6)/2	(9+13)/2	(16+20)/2
3	4	1400, 1500, 1600	(3+4+5)/3	(10+11+12)/3	(17+18+19)/3

C.3. Processing variables in the county-level crime statistics

In the county-level crime statistics, all variables were numerical variables indicating the crime rates per 100,000 people for ten variables (murder, rape, robbery, assault, burglary, larceny, theft, arson, and total). These variables were not processed. A similar procedure to the 2018 American Community Survey (5-year Estimates) was followed to assign these county-level crime rates to individual-level data. Each unique Residence Code consisted of one or more counties through assigned PUMAs (see Table A1 in Appendix A). The variables were aggregated by taking an average across all associated counties based on the assigned Residence Code for an individual and then merged to with the original individual-level training and test datasets.

County Level Data

County	Variable X	Variable Y	Variable Z
Fulton	1	8	15
Douglas	2	9	16
DeKalb	3	10	17
Newton	4	11	18
Rockdale	5	12	19
Clayton	6	13	20
Cobb	7	14	21

Individual Level Data

Subject	NIJ Residence Code	Associated PUMAs	Associated Counties	Variable X	Variable Y	Variable Z
1	1	1003, 4400	Fulton, Douglas	(1+2)/2	(8+9)/2	(15+16)/2
2	2	1008, 4300	DeKalb, Newton, Rockdale	(3+4+5)/3	(10+11+12)/3	(17+18+19)/3
3	9	5001, 6001, 6002	Clayton	6	13	20
4	11	1001, 3004, 4600	Fulton, Cobb	(1+7)/2	(8+14)/2	(15+21)/2

C.4. Processing variables in the county-level auxiliary statistics

In addition to county-level crime statistics, 233 county-level auxiliary variables were compiled in 17 different areas (poverty, voting, hospital, unemployment, public assistance, urban population, population age, Medicare, sexually transmitted diseases, money transfer, agriculture, income, juvenile court, bankruptcy, crime index, birth/death, lottery). These variables were all numeric, and no other process was applied. As described and demonstrated before, a similar procedure was followed to aggregate these variables and assign them to individuals based on the Residence Code provided by NIJ. A detailed list of 233 auxiliary variables can be found in Table A4 in Appendix A.

D. Model Building

D.1 Recidivism in Year 1

Five different XGBoost models and two linear regression models with ridge penalty were built using the original individual-level variables provided by NIJ and other aggregated PUMA- and county-level variables compiled by the author. There were a total of 644 variables available to use as features after processing all the variables. These models differed in the features being used. Below is a brief description of how each model is different than the others.

- **XGBoost1**: Only the processed feature variables provided by NIJ were used to develop a forecasting model.
- **XGBoost2**: All 644 variables, including the aggregated county- and PUMA-level variables, were used to develop a forecasting model. The learning rate is fixed to .05 while optimizing the rest of the tuning variables.
- **XGBoost3**: Correlation coefficients between the binary outcome variable and all 644 variables were calculated. Then, only the features with a correlation larger than .01 or smaller than -.01 were included to develop a forecasting model.

- **XGBoost4**: This is equivalent to XGBoost2 in terms of the set of features, and all features were used. The only difference the learning rate is fixed to .1 while optimizing the rest of the tuning parameters in the model.
- **XGBoost5**: The most important 50 predictors from XGBoost4 were identified, and a different XGBoost model was optimized using only these 50 most important predictors.
- **LR1**: A logistic regression model with both L1 penalty and L2 penalty was developed by including the main effects of all 644 variables.
- **LR2**: A logistic regression model with L2 penalty was developed by including the 56 significant main effects from LR1 and 1540 two-way interactions among these 56 variables.

When developing XGBoost models, parameters were optimized by first fixing the learning rate (0.05 or 0.1) and then tuning the rest of the parameters one by one in the following order: the number of trees, maximum depth of a tree and minimum child weight, gamma, maximum delta step, scale positive weight, lambda and alpha, subsample, and column subsample. After tuning all the parameters, the learning rate was recalibrated at the end. More information about the nature of these parameters can be found at the following link:

https://xgboost.readthedocs.io/en/latest/parameter.html

Table 1 below presents the final parameters used to train each XGBoost model.

Table 1. Final parameters used to train each XGBoost model

	XGBoost1	XGBoost2	XGBoost3	XGBoost4	XGBoost5
Eta	0.01	0.01	0.05	0.1	0.095
Number of Trees	2000	484	160	69	69
Max. Depth	4	5	4	4	4
Min. Child Weight	0.7	5.5	4	0.5	4.5
Gamma	0.12	0.03	0.96	0.51	0.74
Max. Delta Step	1.2	1.3	1.6	2.7	1.7
Scale Pos. Weight	1	1	1	1	1
Lambda	1	1.1	8.1	7.1	1
Alpha	0	0	0	1.5	0
Subsample (proportion)	0.45	0.5	0.7	1	0.6
Column subsample (proportion)	0.90	0.5	0.45	1	0.5

The training dataset provided by NIJ has a sample size of 18,023 for Year 1. A randomly selected 15,000 observations were used to optimize the model parameters with 10-fold cross-validation. The remaining 3,023 observations were used to evaluate the model performance. Once the model is finalized, the predicted values were obtained for the test dataset provided by NIJ and submitted through the project website.

D.2 Recidivism in Year 2 and Year 3

In Year 2 and Year 3 predictions, a single XGBoost model was trained by using the default parameters by fixing the learning rate to 0.01 and optimizing the number of trees. Below is a table that presents the final parameters used to train an XGBoost model for Year 2 and Year 3 predictions.

Before model building for Year 2, the individuals who recidivated in Year 1 were removed from the dataset, leaving a total of 12,651 observations in the training set. A randomly selected 11,000 observations were used to optimize the model parameters with 10-fold cross-validation. The remaining 1,651 observations were used to evaluate the model performance. Once the Year 2 model was finalized, the predicted values were obtained for the test dataset provided by NIJ and submitted through the project website.

Before model building for Year 3, the individuals who recidivated in Year 1 and Year 2 were removed from the dataset, leaving a total of 9,398 observations in the training set. A randomly selected 8,000 observations were used to optimize the model parameters with 10-fold cross-validation. The remaining 1,398 observations were used to evaluate the model performance. Once the Year 3 model was finalized, the predicted values were obtained for the test dataset provided by NIJ and submitted through the project website.

Table 2. Final parameters used to train the XGBoost model in Year 2 and Year 3

	Year 2	Year 3
Eta	0.01	0.01
Number of Trees	700	364
Max. Depth	6	6
Min. Child Weight	0	0
Gamma	0	0
Max. Delta Step	0	0
Scale Pos. Weight	1	1
Lambda	1	1
Alpha	0	0
Subsample (proportion)	1	1
Column subsample (proportion)	1	1

E. Results

For each year, a local test dataset was randomly selected from the available training dataset. The model performance was evaluated on this local test dataset before calculating and submitting the entries for the challenge test data. In Year 2, the individuals who recidivated in Year 1 were removed from the training dataset during model development. Similarly, in Year 3, the individuals who recidivated in Year 1 or Year 2 were removed from the training dataset during model development. Table 3 below presents the sample sizes available in the training dataset, local test dataset, and challenge test dataset for Year 1, Year 2, and Year 3. For all years, the model parameters were optimized using 10-fold cross-validation using the training dataset; then, the model performance was evaluated on the local test dataset. The best-performing model or stack of models was chosen to predict the outcome in the challenge dataset. Since the true outcome values in the challenge test datasets were released after the Challenge is over at the time of writing this report, I was also able to compare the performance of the models on the local test dataset and challenge test dataset.

Table 3. Number of observations in the training, local test, challenge test datasets

	Training Dataset	Local Test Dataset	Challenge Test Dataset
Year 1	15,000	3,028	7,807
Year 2	11,000	1,651	5,460
Year 3	8,000	1,398	4,146

E.1. Year 1 Predictions

The prediction accuracy measured by Brier Score for all seven models is presented in Table 4. All seven models provided a very similar performance on the local test data, and the differences among models were negligible. Fairness adjusted correctness index and AUC were also very similar across all these models.

A model averaging procedure was implemented to improve the prediction accuracy for the local test dataset. It was found that the model averaging of XGBoost1, XGBoost3, XGBoost4, and LR2 provided a slight improvement over the best performing model. Therefore, the predicted probabilities for the challenge test data were averaged across these four models for final submission.

The results from the challenge test data are presented in Table 5, and the bold entries at the bottom of Table 5 were placed in the 3rd position in this competition. The performance of the models on the Challenge test dataset resembled the performance observed in the local test dataset.

The most important 15 variables in these models and their brief descriptions are provided below in order of their importance as measured by XGBoost. The importance is calculated for a single tree by the amount that each feature improves the performance (as measured by gain weighted by the number of observations in each leaf) and then averaged across all trees within the model.

- **comp1**: a composite score obtained by averaging the variables Prior_Arrest_Episodes_Felony, Prior_Arrest_Episodes_Misdemeanor, Prior_Arrest_Episodes_Property, Xv1, and Xv2 (see the original codebook for more info about these variables).
- **age1c:** the contrast variable associated with the linear coefficient after polynomial contrast coding of variable Age_At_Release
- **gang**: a dummy variable of gang affiliation (0: no, 1: yes)
- **felony**: prior number of GCIC arrests with the most serious charge being felony (Prior_Arrest_Episodes_Felony)
- **age**: A numeric variable assigning the midpoint of original variable Age_At_Release
- **Xv1**: original numeric variable (Xv1)
- **risk1**: original Supervision_Risk_Score_First variable, a first parole supervision risk assessment score (1-10, 1=lowest risk)

- **prop**: prior number of GCIC arrests with the most serious charge being property (Prior_Arrest_Episodes_Property)
- **year1c:** the contrast variable associated with the linear coefficient after polynomial contrast coding of variable Prison_Years
- **misd**: prior number of GCIC arrests with the most serious charge being misdemeanor (Prior_Arrest_Episodes_Misdemeanor)
- **cmisd**: prior number of GCIC convictions with the most serious charge being misdemeanor (Prior_Conviction_Episodes_Misdemeanor)
- **age1**: a dummy variable indicating whether or not a parolee is between 18 and 22 years old
- **off3:** a dummy variable indicating whether or not the primary prison conviction offense group is property
- **gender**: a dummy variable indicating gender
- **mhsa**: a dummy variable indicating whether or not parole release condition is mental health or substance abuse

Table 4. Prediction Accuracy on the Local Test dataset (N=3,028)

	Brier Score			Fairness Adjusted Correctness Index		AUC
	Male	Female	All	Male	Female	
XGBoost1	0.1888	0.1561	0.1848	0.8045	0.7990	0.706
XGBoost2	0.1892	0.1559	0.1851	0.8106	0.8196	0.705
XGBoost3	0.1882	0.1555	0.1842	0.8106	0.8117	0.707
XGBoost4	0.1885	0.1566	0.1846	0.8046	0.8067	0.707
XGBoost5	0.1891	0.1576	0.1853	0.8059	0.8265	0.703
LR1	0.1892	0.1558	0.1851	0.8055	0.8239	0.703
LR2	0.1893	0.1548	0.1851	0.8003	0.8249	0.705
Final_Stacked	0.1881	0.1554	0.1841	0.8037	0.8162	0.708

Table 5. Prediction Accuracy on the Challenge Test dataset (N=7,807)

	Brier Score			Fairness Adjusted Correctness Index		AUC
	Male	Female	All	Male	Female	
XGBoost1	0.1909	0.1544	0.1864	0.7966	0.8365	0.708
XGBoost2	0.1923	0.1565	0.1879	0.8018	0.8345	0.702
XGBoost3	0.1915	0.1552	0.1870	0.8030	0.8325	0.706
XGBoost4	0.1923	0.1563	0.1879	0.8034	0.8367	0.702
XGBoost5	0.1923	0.1549	0.1876	0.7956	0.8258	0.703
LR1	0.1920	0.1554	0.1876	0.8016	0.8376	0.704
LR2	0.1916	0.1551	0.1872	0.7897	0.8411	0.705
Final_Stacked	**0.1910**	**0.1548**	**0.1866**	0.8001	0.8364	0.707

E.2. Year 2 Predictions

The prediction performance from a single XGBoost model trained mainly by default parameters is presented in Table 6 for both the local and Challenge test datasets. The bold entry at the bottom of Table 6 was placed in the 5th position in this competition. The most important 15 variables in this model and their brief descriptions are provided below in order of their importance as measured by XGBoost:

- **jobs**: jobs per year while on parole (Jobs_Per_Year)
- **pemployed**: % of days employed while on parole (Percent_Days_Employed)
- **avg_drug**: average days on parole between drug tests (Avg_Days_Per_DrugTest)
- **comp1**: a composite score obtained by averaging the variables Prior_Arrest_Episodes_Felony, Prior_Arrest_Episodes_Misdemeanor, Prior_Arrest_Episodes_Property, Xv1, and Xv2 (see the original codebook for more info about these variables).
- **age1c**: the contrast variable associated with the linear coefficient after polynomial contrast coding of variable Age_At_Release
- **risk1:** original Supervision_Risk_Score_First variable, a first parole supervision risk assessment score (1-10, 1=lowest risk)
- **felony**: prior number of GCIC arrests with the most serious charge being felony (Prior_Arrest_Episodes_Felony)
- **thc**: % drug tests positive for THC/Marijuana (DrugTests_THC_Positive)
- **gang**: a dummy variable of gang affiliation (0: no, 1:yes)
- **meth**: % drug tests positive for Methamphetamine (DrugTests_Meth_Positive)
- **misd**: prior number of GCIC arrests with the most serious charge being misdemeanor (Prior_Arrest_Episodes_Misdemeanor)
- **resch**: number of residence changes/moves (new zip codes) during parole (Residence_Changes)
- **prat:** number of program attendances
- **other**: % drug tests positive for other drug (DrugTests_Other_Positive)
- **cmisd**: prior number of GCIC convictions with the most serious charge being misdemeanor (Prior_Conviction_Episodes_Misdemeanor)

Table 6. Prediction Accuracy in Year 2

	Brier Score			Fairness Adjusted Correctness Index		AUC
	Male	Female	All	Male	Female	
Local Test Data (N=1,651)	0.1693	0.1362	0.1647	0.8492	0.8163	0.726
Challenge Test Data (N = 5,460)	0.1671	**0.1245**	0.1613	0.8287	0.8736	0.734

E.3. Year 3 Predictions

The prediction performance from a single XGBoost model trained by mainly default parameters is presented in Table 7 for both local test data and challenge test dataset. The most important 15 variables in this model and their brief descriptions are provided below in order of their importance as measured by XGBoost:

- **jobs:** jobs per year while on parole (Jobs_Per_Year)
- **comp1:** a composite score obtained by averaging the variables Prior_Arrest_Episodes_Felony, Prior_Arrest_Episodes_Misdemeanor, Prior_Arrest_Episodes_Property, Xv1, and Xv2 (see the original codebook for more info about these variables).
- **pemployed:** % of days employed while on parole (Percent_Days_Employed)
- **age1c:** the contrast variable associated with the linear coefficient after polynomial contrast coding of variable Age_At_Release
- **avg_drug:** average days on parole between drug tests (Avg_Days_Per_DrugTest)
- **thc:** % drug tests positive for THC/Marijuana (DrugTests_THC_Positive)
- **gang:** a dummy variable of gang affiliation (0: no, 1: yes)
- **meth:** % drug tests positive for Methamphetamine (DrugTests_Meth_Positive)
- **cmisd:** prior number of GCIC convictions with most serious charge being misdemeanor (Prior_Conviction_Episodes_Misdemeanor)
- **risk1:** original Supervision_Risk_Score_First variable, a first parole supervision risk assessment score (1-10, 1=lowest risk)
- **misd:** prior number of GCIC arrests with most serious charge being misdemeanor (Prior_Arrest_Episodes_Misdemeanor)
- **felony:** prior number of GCIC arrests with the most serious charge being felony (Prior_Arrest_Episodes_Felony)
- **educ1:** a dummy variable indicating whether or not an individual has at least some college education
- **age7:** a dummy variable indicating whether or not an individual 48 years old or older
- **gender:** a dummy variable indicating gender

Table 7. Prediction Accuracy in Year 3

	Brier Score			Fairness Adjusted Correctness Index		AUC
	Male	Female	All	Male	Female	
Local Test Data (N=1,398)	0.1385	0.1042	0.1336	0.8601	0.8958	0.707
Challenge Test Data (N = 4,146)	0.1529	0.1182	0.1478	0.8455	0.8818	0.693

F. Final Remarks and Future Considerations

This section will address some questions NIJ directed to the challenge participants in light of the findings provided in earlier sections.

Were variables added to the data set? If so, detail the variables.
What variables were constructed? How were the variables constructed?

In sections 2 and 3, detailed information about supplemental variables compiled by the author was given. Also, Appendix A provides some tables that list all variables used in modeling and how they were processed.

What type of model was used?

Two types of models were used: Extreme Gradient Boosting (XGBoost; Chen & Gu, 2016) and Penalized Logistic Regression with L1 penalty and L2 penalty (Tibshirani, 1996; Hoerl and Kennard, 1970).

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

I primarily considered the XGBoost models. By nature, XGBoost doesn't provide a statistical significance test for predictors in the model. Instead, it ranks the variables based on their importance using a specific metric. The variable importance is calculated for a single tree by the amount that each feature improves the performance (as measured by gain weighted by the number of observations in each leaf) and then averaged across all trees within the model. The most important 15 variables for the XGBoost models and their brief descriptions were provided in Section 5.

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

Yes, different XGBoost models were tried with different sets of input variables. Also, different penalized logistic regression models were tried. On the other hand, all these models performed very similarly in Year 1 predictions. The difference among the models was negligible.

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

I would suggest considering precision. Using the terminology and confusion matrix provided in the competition website, precision can be defined as A/(A+C). For those individuals whom the model predicted Fail, how many of them indeed failed?

		PREDICTED	
		Fail	Succeed
TRUE	Fail	A	B
	Succeed	C	D

Brier Score may not be the most helpful metric for this competition as one has to make a binary decision based on the model-predicted probability. This binary decision has to be made based on a threshold value (e.g., .5), and this decision is either correct or not correct. Optimizing a metric more relevant to the use of the predicted probabilities in practice may be more helpful.

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

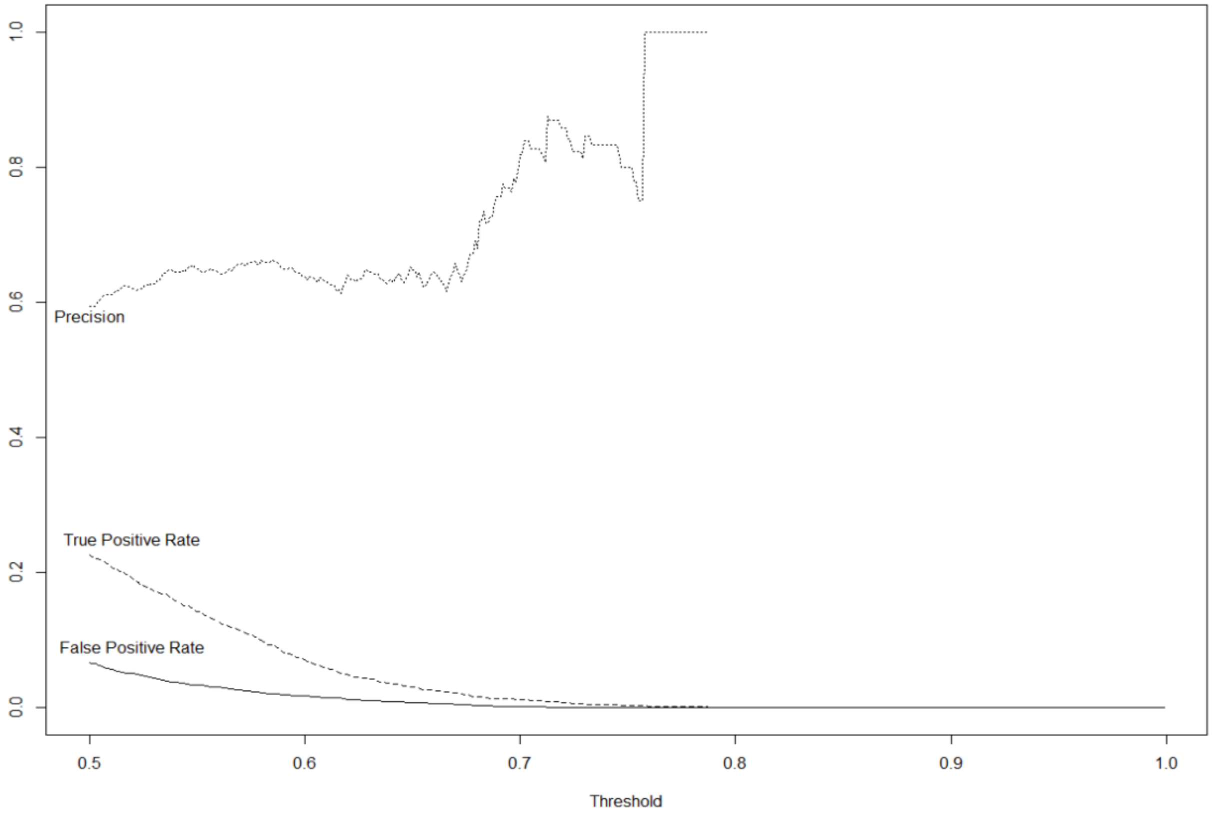
It depends on many factors when choosing a cut-off probability for making a binary decision and the type of model being used. In the plot below, I considered all numbers from 0.500 to 1.000 with increments of .001 when predicting recidivism in Year 1 using the XGBoost model. For every potential threshold value on the x-axis, I

converted the predicted probabilities to binary decisions for 7,807 individuals in the challenge test dataset,
created a confusion matrix based on the predicted binary outcome and true binary outcome for these 7,807 individuals,
then finally calculated True-Positive Rate, False-positive Rate, and Precision.

As you can see, precision (the proportion of correct decisions when the model predicted that an individual would be recidivated in Year 1) does not change when the cut-off value is between 0.5 and 0.7. At the same time, the true-positive rate significantly drops to the point that the model is not useful at all. Therefore, a cut-off value of 0.5 seems to be providing a good balance between the false-positive rate (0.066) and the true-positive rate (0.226), yielding a precision value of .59. It indicates about a 59% chance of making a correct decision when the model predicts recidivism (probability > 0.5). At the same time, the model would correctly identify about 22.6% of all recidivated individuals in Year 1.

The next optimal cut-off seems to be around 0.7, where the precision improves to .818; however, the true-positive rate drops to 0.012 while the false-positive rate is 0.001. It indicates an 82% chance of making a correct decision when the model predicts recidivism (probability > 0.7); however, the model would correctly identify only about 1.2% of all recidivated individuals in Year 1.

Therefore, it is all about the relative cost of making a false-positive and false-negative. If NIJ thinks making false-positive costs too much, then choosing a higher threshold is reasonable to improve precision at the expense of a low true-positive rate. On the other hand, if false-negative costs too much, then choosing a lower threshold is reasonable to improve true-positive rates at the expense of low precision.



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

I don't know.

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

This is not an area of my expertise; so, I can't make strong recommendations about practical/applied findings. However, I list a few insights based on these findings.

A simple regression model (penalized logistic regression) works as well as a fancy state-of-the-art algorithm (XGBoost).

I compiled hundreds of PUMA-level and county-level variables, but they made almost no difference in increasing the model's predictive performance. A model with individual-level variables provided by NIJ did as well as a model that includes hundreds of PUMA-level and county-level variables. It may be due to the noise in the data about the residence location of each individual. PUMA-level and county-level variables are aggregated weirdly due to the lack of precise information about individuals' location at release. Therefore, these variables may not be providing precise information. If NIJ compiles PUMA-level and county-level variables based on the exact location of these individuals at release, they may contribute better to the model performance.

The overall performance of these models can be considered mediocre at best, with AUC values around 0.69 – 0.74. It is challenging to justify the use of these models for high-stakes decisions about individual's lives. These models should NOT be used to make any decision about individual's lives. More work must be done to build more predictive models if these models are going to be used for making important decisions.

APPENDIX A

Table A1. County Names Associated with unique Residence Code provided by NIJ

Unique Residence Code (NIJ)	PUMA (combined)	Associated County Names
1	1003, 4400	Fulton, Douglas
2	1008, 4300	DeKalb,Newton, Rockdale
3	1200, 1300	Appling, Evans,Jeff, Davis, Montgomery, Tattnall, Telfair, Toombs,Wayne, Wheeler, Bleckley, Candler, Dodge, Emanuel, Johnson, Laurens, Treutlen, Wilcox
4	1400, 1500, 1600	Bibb, Houston, Pulaski, Baldwin, Crawford, Jones, Monroe, Peach,Putnam, Twiggs, Wilkinson
5	1700, 1800	Chattahoochee, Muscogee, Clay, Crisp, Dooly, Harris, Macon, Marion, Quitman, Randolph, Schley, Stewart
6	2001, 2002, 2003, 4005	DeKalb,Gwinnett
7	100, 200, 500	Camden, Glynn, McIntosh, Bryan, Liberty, Long,Atkinson, Bacon, Brantley, Charlton, Clinch
8	4000, 4100, 4200	Richmond, Columbia, Burke, Glascock, Hancock, Jefferson, Jenkins, Lincoln, McDuffle, Taliaferro, Warren, Washington
9	5001, 6001, 6002	Clayton
10	2400, 5002	Fayette,Clayton
11	1001, 3004, 4600	Fulton, Cobb
12	1002, 1005, 3300, 3400, 4001, 4002, 4006	Fulton, Forsyth,Hall, Gwinnett
13	3101, 3102	Cherokee
14	1900, 3900, 4003, 4004	Butts, Lamar, Pike, Spalding, Upson, Jasper, Morgan, Walton, Gwinnett
15	3001, 3002, 3003, 3005	Cobb
16	2500, 4500	Floyd, Haralson,Polk, Paulding
17	2800, 2900, 3200, 3500	Fannin, Gilmer, Gordon, Murray, Pickens, Bartow, Dawson, Lumpkin, Rabun, Towns, Union, White, Banks, Franklin, Habersham, Hart, Stephens
18	600, 700, 800	Lowndes, Ben Hill, Berrien, Brooks, Cook, Irwin, Tift, Turner, Colquitt, Thomas, Worth
19	900, 1100	Dougherty, Lee, Baker, Calhoun, Decatur, Early, Grady, Miller, Mitchell, Seminole, Terrell
20	300, 401, 402	Bulloch, Effingham, Screven, Chatham
21	1004, 2100	Fulton, Coweta
22	2200, 2300	Heard, Meriwether, Troup, Carroll
23	1006, 1007, 2004	Fulton, Dekalb
24	2600, 2700	Catoosa, Chattooga, Dade, Walker, Whitfield
25	3600, 3700, 3800	Clarke, Elbert, Greene, Madison, Oconee, Oglethorpe, Barrow, Jackson

Table A2. Process information for all variables in the training and test datasets provided by NIJ

	Variable Name	Type	Number of Categories	Process Applied				Number of Constructed Variables
				Dummy Coding	One-hot encoding	Polynomial contrast coding	Numerical Assignment	
1	Gender	Binary	2	x				1
2	Race	Binary	2	x				1
3	Age_at_Release	Ordinal	7		x	x	x	14
4	PUMAs	Nominal	25		x			25
5	Gang_Affiliated	Binary	2	x				1
6	Supervision_Risk_Score_First	Numeric					x	1
7	Supervision_Level_First	Ordinal	3		x	x		5
8	Education_Level	Ordinal	3		x	x	x	6
9	Dependents	Ordinal	4		x		x	5
10	Prison_Offence	Nominal	5		x			5
11	Prison_Years	Ordinal	4		x	x		7
12	Prior_Arrest_Episodes_Felony	Numeric		x			x	2
13	Prior_Arrest_Episodes_Misd	Numeric		x			x	2
14	Prior_Arrest_Episodes_Violent	Numeric		x			x	2
15	Prior_Arrest_Episodes_Property	Numeric		x			x	2
16	Prior_Arrest_Episodes_Drug	Numeric		x			x	2
17	Prior_Arrest_Episodes_DVCharges	Numeric		x				1
18	Prior_Arrest_Episodes_GunCharges	Numeric		x				1
19	Prior_Conviction_Episodes_Felony	Numeric		x			x	2
20	Prior_Conviction_Episodes_Misd	Numeric		x			x	2
21	Prior_Conviction_Episodes_Violent	Numeric		x				1
22	Prior_Conviction_Episodes_Property	Numeric		x			x	2
23	Prior_Conviction_Episodes_Drug	Numeric		x			x	2
24	X_v1	Numeric		x			x	2
25	X_v2	Binary	2	x				1
26	X_v3	Binary	2	x				1

27	X_v4	Binary	2	x		1
28	Prior_Revocations_Parole	Binary	2	x		1
29	Prior_Revocations_Probation	Binary	2	x		1
30	Condition_MH_SA	Binary	2	x		1
31	Condition_Cog_Ed	Binary	2	x		1
32	Condition_Other	Binary	2	x		1
33	Violations_ElectronicMonitorin	Binary	2	x		1
34	Violations_InstructionsNotFollowed	Binary	2	x		1
35	Violations_FailtoReport	Binary	2	x		1
36	Violations_MoveWithoutPermission	Binary	2	x		1
37	Delinquency_Reports	Numeric		x	x	2
38	Program_Attendances	Numeric		x	x	2
39	Program_UnexcusedAbsences	Numeric		x	x	2
40	Residence_Changes	Numeric		x	x	2
41	Avg_Days_per_DrugTest	Numeric			x	1
42	DrugTests_THC_Positive	Numeric			x	1
43	DrugTests_Cocaine_Positive	Numeric			x	1
44	DrugTests_Meth_Positive	Numeric			x	1
45	DrugTests_Other_Positive	Numeric			x	1
46	Percent_Days_Employed	Numeric			x	1
47	Jobs_Per_Year	Numeric			x	1
48	Employment_Exempt	Binary	2	x		1

Notes. The variables are listed in order they appear in the training dataset provided by NIJ. A total of 48 predictors are recoded into a total of 122 variables after processing all variables. In addition to these 122 variables, a Principal Component Analysis was run for crime related variables. PCA revealed that these variables can be grouped into four categories. Therefore, an additional four composite variables were created as basic sum score of the crime related variables in these four categories. The R code that is used to process these variables for more detailed information can be found at this link (https://github.com/czopluoglu/nij-competition/blob/main/R/03_data%20prep.r).

Table A3. List of variables aggregated at the PUMA level from 2018 American Community Survey (ACS) 5-year estimates

	Variable Name	Type	Number of Categories	Process Applied				Number of Constructed Variables
				Dummy Coding	One-hot encoding	Polynomial contrast coding	Numerical Assignment	
1	ACCESS	Nominal	3	x				1
2	ACR	Nominal	3		x			3
3	AGEP	Numeric					x	1
4	AGS	Ordinal	6		x			6
5	BATH	Binary	2	x				1
6	BLD	Nominal	10		x			10
7	BROADBND	Binary	2	x				1
8	BUS	Binary	2	x				1
9	CIT	Nominal	5		x			5
10	COMPOTHX	Binary	2	x				1
11	CONP (log transformed)	Numeric					x	1
12	COW	Nominal	9		x			9
13	DDRS	Binary	2	x				1
14	DEAR	Binary	2	x				1
15	DEYE	Binary	2	x				1
16	DIALUP	Binary	2	x				1
17	DIS	Binary	2	x				1
18	DOUT	Binary	2	x				1
19	DPHY	Binary	2	x				1
20	DRAT	Ordinal					x	2
21	DRATX	Binary	2	x				1
22	DREM	Binary	2	x				1
23	ELEFP	Nominal	3		x			3
24	ELEP	Numeric					x	1
25	ENG	Ordinal					x	1
26	FER	Binary	2	x				1
27	FES	Nominal	8		x			8
28	FINCP	Numeric					x	1
29	FPARC	Nominal	4		x			4

30	FS	Binary	2	x			1
31	FULP (log transformed)	Numeric				x	1
32	GASP (log transformed)	Numeric				x	1
33	GCL	Binary	2	x			1
34	GCM	Ordinal		x			5
35	GCR	Binary	2	x			1
36	GRNTP	Numeric				x	2
37	GRPIP	Numeric				x	1
38	HFL	Nominal	9		x		9
39	HHL	Nominal	5		x		5
40	HHT	Nominal	7		x		7
41	HINCP	Numeric				x	1
42	HINS1	Binary	2	x			1
43	HINS2	Binary	2	x			1
44	HINS3	Binary	2	x			1
45	HINS4	Binary	2	x			1
46	HINS5	Binary	2	x			1
47	HINS6	Binary	2	x			1
48	HINS7	Binary	2	x			1
49	HISPEED	Binary	2	x			1
50	HUGCL	Binary	2	x			1
51	HUPAC	Nominal	4		x		4
52	HUPAOC	Nominal	4		x		4
53	HUPARC	Nominal	4		x		4
54	INSP	Numeric		x			1
55	INTP	Numeric				x	1
56	JWMNP	Numeric				x	1
57	JWRIP	Numeric				x	1
58	JWTR	Nominal	12	x			6
59	KIT	Binary	2	x			1
60	LANX	Binary	2	x			1
61	LAPTOP	Binary	2	x			1
62	LNGI	Binary	2	x			1
63	MAR	Nominal	5		x		5
64	MARHD	Binary	2	x			1
65	MARHM	Binary	2	x			1

66	MARHT	Ordinal	3		x		3
67	MARHW	Binary	2	x			1
68	MHP	Numeric				x	1
69	MIG	Binary	2	x			1
70	MIL	Nominal	4		x		4
71	MLPA	Binary	2	x			1
72	MLPB	Binary	2	x			1
73	MLPCD	Binary	2	x			1
74	MLPE	Binary	2	x			1
75	MLPFG	Binary	2	x			1
76	MLPH	Binary	2	x			1
77	MLPI	Binary	2	x			1
78	MLPJ	Binary	2	x			1
79	MLPK	Binary	2	x			1
80	MRGI	Binary	2	x			1
81	MRGP	Numeric				x	1
82	MRGT	Binary	2	x			1
83	MRGX	Nominal	3		x		3
84	MSP	Nominal	6		x		6
85	MULTG	Binary	2	x			1
86	MV	Nominal	7		x		7
87	NATIVITY	Binary	2	x			1
88	NOC	Numeric				x	1
89	NP	Numeric				x	1
90	NPF	Numeric				x	1
91	NPP	Numeric				x	1
92	NR	Binary	2	x			1
93	NRC	Numeric				x	1
94	NWAB	Nominal	3		x		3
95	NWAV	Nominal	5		x		5
96	NWLA	Nominal	3		x		3
97	NWLK	Nominal	3		x		3
98	NWRE	Nominal	4		x		3
99	OC	Binary	2	x			1
100	OCPIP	Numeric				x	2
101	OIP	Numeric				x	1

102	OTHSVCEX	Binary	2	x			1
103	PAOC	Nominal	4		x		4
104	PAP	Numeric				x	1
105	PARTNER	Nominal	5		x		5
106	PERNP	Numeric				x	1
107	PINCP	Numeric				x	1
108	PLM	Binary	2	x			1
109	POVPIP	Numeric				x	1
110	PRIVCOV	Binary	2	x			1
111	PSF	Binary	2	x			1
112	PUBCOV	Binary	2	x			1
113	R18	Binary	2	x			1
114	R60	Ordinal	3		x		3
115	R65	Ordinal	3		x		3
116	RACAIAN	Binary	2	x			1
117	RACASN	Binary	2	x			1
118	RACBLK	Binary	2	x			1
119	RACWHT	Binary	2	x			1
120	RC	Binary	2	x			1
121	REFR	Binary	2	x			1
122	RETP	Numeric				x	1
123	RMSP	Numeric				x	1
124	RNTM	Binary	2	x			1
125	RWAT	Binary	2	x			1
126	SATELLITE	Binary	2	x			1
127	SCIENGP	Binary	2	x			1
128	SCIENGRLP	Binary	2	x			1
129	SEMP	Numeric				x	1
130	SINK	Binary	2	x			1
131	SMARTPHONE	Binary	2	x			1
132	SMOCP	Numeric				x	1
133	SMP	Numeric				x	1
134	SMX	Nominal	4		x		4
135	SRNT	Binary	2	x			1
136	SSIP	Numeric				x	1
137	SSMC	Ordinal	3	x			1

138	SSP	Numeric					x	1
139	STOV	Binary	2	x				1
140	SVAL	Binary	2	x				1
141	TABLET	Binary	2	x				1
142	TAXAMT	Numeric					x	1
143	TEL	Binary	2	x				1
144	TEN	Nominal	4		x			4
145	TOIL	Binary	2	x				1
146	VALP	Numeric					x	1
147	VEH	Numeric					x	1
148	WAGP	Numeric					x	1
149	WATFP	Nominal	3		x			3
150	WATP	Numeric					x	1
151	WGTP	Numeric					x	1
152	WIF	Ordinal	4		x			4
153	WKHP	Numeric					x	1
154	WKL	Nominal	3		x			3
155	WKW	Ordinal					x	1
156	WRK	Binary	2	x				1
157	YBL	Ordinal	22				x	1

Notes. The detailed codebook about these variables can be found in the 2014 – 2018 ACS PUMA Data Dictionary (https://www2.census.gov/programs-surveys/acs/tech_docs/pums/data_dict/PUMS_Data_Dictionary_2014-2018.pdf). A total of 157 predictors are recoded into a total of 295 variables after processing all variables. In addition, a Principal Component Analysis was run for all 295 variables. Standardized composite scores for the first four principal component were added to the dataset. The R code that is used to download the 2018 ACS database and process these variables can be found at this link https://github.com/czopluoglu/nij-competition/blob/main/R/02_geodata.r).

Table A4. List of auxiliary variables compiled at the county-level

Category	Variable
Vital Statistics (2018)	
1	Birth Rate per 1,000 Population
2	White Birth Rate per 1,000 Population
3	Black Birth Rate per 1,000 population
4	Hispanic Birth Rate per 1,000 population
5	White Low Weight Birth Rate per 100 Births
6	Black Low Weight Birth Rate per 100 Births
7	Hispanic Low Weight Birth Rate per 100 Births
8	Births to Unwed Mothers, All Ages, Rate per 100 Births
9	Births to Unwed Mothers, All Ages, White, Rate per 100 Births
10	Births to Unwed Mothers, All Ages, Black, Rate per 100 Births
11	Births to Unwed Mothers, All Ages, Hispanic, Rate per 100 Births
12	Births to Unwed Teen Mothers, Rate per 100 Births to Teen Mothers
13	Births to Unwed Teen Mothers, Rate per 100 Births
14	Births to Unwed Teen Mothers, Rate per 100 Births to Unwed Mothers
15	Death Rate per 1,000 Population
16	White Death Rate per 1,000 Population
17	Black Death Rate per 1,000 Population
18	Hispanic Death Rate per 1,000 Population
19	Major Cardiovascular Diseases Rate per 100,000 Population
20	Cancers Rate per 100,000 Population
21	Respiratory Diseases Rate per 100,000 Population
22	Nervous System Diseases Rate per 100,000 Population
23	Endocrine, Nutritional, and Metabolic Diseases Rate per 100,000 Population
24	Mental and Behavioral Disorders Rate per 100,000 Population
25	Reproductive and Urinary System Diseases Rate per 100,000 Population
26	Infectious and Parasitic Diseases Rate per 100,000 Population
27	External Causes Rate per 100,000 Population
28	External Causes, Suicide, Rate per 100,000 Population
29	External Causes, Homicide, Rate per 100,000 Population
Lottery Statistics	

	30	Gross Instant, Dollars
	31	Cash3/Cash4, Dollars
	32	Fantasy5 with CashMatch, Dollars
	33	Mega Millions, Dollars
	34	Keno, Dollars
	35	Powerball, Dollars
	36	Georgia Five, Dollars
	37	All Or Nothing, Dollars
	38	Jumbo Bucks Lotto with CashMatch, Dollars
	39	5 Card Cash, Dollars
	40	Cash 4 Life, Dollars
	41	Print n Play, Dollars
	42	Lottery Sales per Capita, Dollars
Poverty Statistics (2014-2018)		
	43	Total Persons in Poverty, Percent
	44	Children Under Age 18 in Poverty, Percent
	45	Related Children in Families, Age 5-17, in Poverty, Percent
	46	Persons Below Poverty Level, Percent
	47	Persons Below Poverty Level, White, Percent
	48	Persons Below Poverty Level, Black, Percent
	49	Persons Below Poverty Level, Hispanic, Percent
	50	Persons Age 65 and Over in Poverty, Percent
	51	Persons Less than High School Graduates in Poverty, Percent
	52	Persons with Bachelor's Degree or Higher in Poverty, Percent
	53	Families Below Poverty Level, Percent
	54	Families Below Poverty Level With a White Householder, Percent
	55	Families Below Poverty Level With a Black Householder, Percent
	56	Families Below Poverty Level, With an Hispanic Householder, Percent
	57	Families Below Poverty Level, Female Head of Household, No Husband Present, Percent
	58	Families Below Poverty Level With a White Householder, Female Head of Household, No Husband Present, Percent
	59	Families Below Poverty Level With a Black Householder, Female Head of Household, No Husband Present, Percent

	60	Families Below Poverty Level, With an Hispanic Householder, Female Head of Household, No Husband Present, Percent
Voting statistics (2016, 2018)		
	61	Votes Cast for President, Democratic Party, Percent
	62	Votes Cast for President, Republican Party, Percent
	63	Votes Cast for President, Libertarian Party, Percent
	64	Voting History
	65	Voter Turnout, Black Female, Percent
	66	Voter Turnout, Black Male, Percent
	67	Voter Turnout, White Female, Percent
	68	Voter Turnout, White Male, Percent
	69	Voter Turnout, Asian, Percent
	70	Voter Turnout, Hispanic, Percent
	71	Voter Turnout, Other Race, Percent
	72	Registered Voters, Black Female, Percent
	73	Registered Voters, Black Male, Percent
	74	Registered Voters, White Female, Percent
	75	Registered Voters, White Male, Percent
	76	Registered Voters, Asian, Percent
	77	Registered Voters, Hispanic, Percent
	78	Registered Voters, Other Race, Percent
Hospital Statistics (2018)		
	79	General Hospital Bed Capacity
	80	General Hospital Total Inpatient Days
	81	General Hospital Total Admissions
	82	General Hospital Occupancy Rate
	83	General Hospital Average Stay in Days
	84	General Hospital Total Emergency Department Visits
	85	General Hospital Emergency Department Inpatient Admissions
	86	General Hospital Total Admissions from Emergency Department, Percent
	87	Nursing Home Bed Capacity
	88	Nursing Home Total Days
	89	Nursing Home Average Occupancy, Percent

	90	Child Care Learning Centers
	91	Child Care Learning Center Capacity
	92	Family Child Care Learning Homes
	93	Uninsured Under Age 65, All Income Levels, Percent
	94	Uninsured Under Age 65, At or Below 200% Poverty, Percent
	95	Uninsured Under Age 19, All Income Levels, Percent
	96	Uninsured Under Age 19, At or Below 200% Poverty, Percent
Unemployment Statistics		
	97	2014 Unemployment Rate
	98	2015 Unemployment Rate
	99	2016 Unemployment Rate
	100	2017 Unemployment Rate
	101	2018 Unemployment Rate
	102	2018 Unemployment Insurance Initial Claims Monthly Average
Public Assistance Statistics (2018)		
	103	Persons Under Age 18 Receiving Benefits, Percent
	104	Persons Age 18-64 Receiving Benefits, Percent
	105	Persons Age 65 and Over Receiving Benefits, Percent
	106	Percent of SSI Recipients also Receiving OASDI
	107	Percent of Total Population Receiving SSI Benefits
	108	OASDI Beneficiaries Age 65 and Over (Percent)
	109	OASDI Beneficiaries as Percent of Total Population
Urban Population Statistics (2010)		
	110	Population Inside Urbanized Area, Percent
	111	Population Inside Urban Clusters, Percent
	112	Total Persons, Urban, Percent
	113	Urban Land, Percent
	114	Urban Area Population Density
	115	Rural Area Population Density
Population Age Statistics		
	116	2018 Median Age
	117	2018 Median Age, Male
	118	2018 Median Age, Female

119	2010 Median Age
120	2010 Median Age, Male
121	2010 Median Age, Female
122	2010 Median Age, White
123	2010 Median Age, White, Male
124	2010 Median Age, White, Female
125	2010 Median Age, Black
126	2010 Median Age, Black, Male
127	2010 Median Age, Black, Female
128	2010 Median Age, Hispanic
129	2010 Median Age, Hispanic, Male
130	2010 Median Age, Hispanic, Female
131	2018 Population 0-4 Years, Percent
132	2018 Population 5-9 Years, Percent
133	2018 Population 10-14 Years, Percent
134	2018 Population 15-19 Years, Percent
135	2018 Population 20-24 Years, Percent
136	2018 Population 25-29 Years, Percent
137	2018 Population 30-34 Years, Percent
138	2018 Population 35-39 Years, Percent
139	2018 Population 40-44 Years, Percent
140	2018 Population 45-49 Years, Percent
141	2018 Population 50-54 Years, Percent
142	2018 Population 55-59 Years, Percent
143	2018 Population 60-64 Years, Percent
144	2018 Population 65-69 Years, Percent
145	2018 Population 70-74 Years, Percent
146	2018 Population 75-79 Years, Percent
147	2018 Population 80-84 Years, Percent
148	2018 Population 85 and Over, Percent
149	2018 Population 18 and Over, Percent
150	2018 Population 65 and Over, Percent

Sexually Transmitted Disease Statistics (2018)		
	151	All Sexually Transmitted Diseases Reported Cases Rate per 100,000 Population
	152	Chlamydia Rate per 100,000 Population
	153	Gonorrhea Rate per 100,000 Population
	154	Syphilis (all Types Except Congenital) Rate per 100,000 Population
	155	Tuberculosis Rate per 100,000 Population
	156	HIV Prevalence Rate per 100,000 Population
Medicare Statistics (2018)		
	157	Hospital and/or Medical Enrollment
	158	Original Medicare Enrollment
	159	Prescription Drug Enrollment
	160	Prescription Drug Plans Enrollment
	161	Medicare Aged Total
	162	Medicare Disabled Total
	163	Physician Rate per 100,000 Population
Money Transfer Statistics		
	164	Personal Current Transfer Receipts, Dollars in Thousands
	165	Personal Current Transfer Receipts, Percent Change
	166	Personal Current Transfer Receipts, Percent Change
	167	Retirement/Disability Insurance Benefit Payments to Individuals as a Percentage of Total Transfer Receipts
	168	Medicare Payments to Individuals as a Percentage of Total Transfer Receipts
	169	Public Assistance Medical Care Benefit Payments to Individuals as a Percentage of Total Transfer Receipts
	170	Supplemental Security Income (SSI) Payments to Individuals as a Percentage of Total Transfer Receipts
	171	Earned Income Tax Credit (EITC) Payments to Individuals as a Percentage of Total Transfer Receipts
	172	Supplemental Nutrition Assistance Program (SNAP) Payments to Individuals as a Percentage of Total Transfer Receipts
	173	Other Income Maintenance Payments to Individuals as a Percentage of Total Transfer Receipts
	174	Unemployment Insurance Payments to Individuals as a Percentage of Total Transfer Receipts

	175	Veterans' Benefit Payments to Individuals as a Percentage of Total Transfer Receipts
	176	Percentage of Total Transfer Receipts to Non-Profit Institutions
	177	Retirement/Other Payments per Capita, Dollars
	178	Income Maintenance per Capita, Dollars
	179	Unemployment Insurance per Capita, Dollars
	180	Transfer Receipts per Capita, Dollars
	181	Transfer Receipts as a Percentage of Total Personal Income
Agricultural Statistics		
	182	Farms of 1-9 Acres, Percent
	183	Farms of 10-49 Acres, Percent
	184	Farms of 50-179 Acres, Percent
	185	Farms of 180-499 Acres, Percent
	186	Farms of 500-999 Acres, Percent
	187	Farms of 1000 or More Acres, Percent
	188	Farm Sales Below $2,500, Percent
	189	Farm Sales of $2,500-$9,999, Percent
	190	Farm Sales of $10,000-$49,999, Percent
	191	Farm Sales of $50,000-$99,999, Percent
	192	Farm Sales of $100,000 or More, Percent
	193	Estimated Market Value, Land and Buildings, Dollars per Acre
	194	Principle Producers, Black, proportion
	195	Principle Producers, Hispanic, proportion
	196	Principle Producers, Women, proportion
	197	Principle Producers Average Age
	198	Irrigated Acres
	199	Conservation Reserve Program Cumulative Enrollment, Acres
Juvenile Court Statistics		
	200	Juvenile Court Commitment Rate Per 1,000 at Risk
	201	Juvenile Court Commitments, White, Percent
	202	Juvenile Court Commitments, Black, Percent
	203	Juvenile Court Commitments, Male, Percent
	204	Regional Youth Detention Center Admission (Detention) Rate Per 1,000 at Risk
Income Statistics		

	205	Median Household Income, Dollars
	206	Median Household Income, White, Dollars
	207	Median Household Income, Black, Dollars
	208	Median Household Income, Hispanic, Dollars
	209	Households With Incomes Less than $10,000, Percent
	210	Households With Incomes Less than $10,000, White, Percent
	211	Households With Incomes Less than $10,000, Black, Percent
	212	Households With Incomes Less than $10,000, Hispanic, Percent
	213	Households With Incomes $10,000-$24,999, Percent
	214	Households With Incomes $10,000-$24,999, White, Percent
	215	Households With Incomes $10,000-$24,999, Black, Percent
	216	Households With Incomes $10,000-$24,999, Hispanic, Percent
	217	Households With Incomes $25,000-$49,999, Percent
	218	Households With Incomes $25,000-$49,999, White, Percent
	219	Households With Incomes $25,000-$49,999, Black, Percent
	220	Households With Incomes $25,000-$49,999, Hispanic, Percent
	221	Households With Incomes $50,000-99,999, Percent
	222	Households With Incomes $50,000-99,999, White, Percent
	223	Households With Incomes $50,000-99,999, Black, Percent
	224	Households With Incomes $50,000-99,999, Hispanic, Percent
	225	Households With Incomes $100,000 or More, Percent
	226	Households With Incomes $100,000 or More, White, Percent
	227	Households With Incomes $100,000 or More, Black, Percent
	228	Households With Incomes $100,000 or More, Hispanic, Percent
Bankruptcy Statistics		
	229	Deposits of all FDIC-Insured Institutions, Dollars in Thousands, 2014-2018 average
	230	Percent Change in Deposits of all FDIC-Insured Institutions, 2014-2019
	231	Bankruptcy Filings Rate per 1,000 population,2018
Crime Index		
	232	2017 Crime Index

Notes. The related county-level data on these variables are available from the GeorgiaData initiative supported by the University of Georgia, https://georgiadata.org/data/data-tables.