# 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/0000-0001-6905-1832) <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](https://nij.ojp.gov/funding/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, and 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 of 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

# Introduction

# Datasets

# 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 the purpose of 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](https://nij.ojp.gov/funding/recidivism-forecasting-challenge%23ks8ofq)). The detailed information about processing these variables before modeling will be given later under the Feature Engineering section.

# 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 actual county names associated by each unique residence code was 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 1 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 a total of 161 variables from 2018 American Community Survey (ACS) was downloaded for 494,091 households in the 72 PUMAs from Georgia. ACS 2018 5-year estimates covers information from 2013-2018 period. This period is selected so that the community data resembles as much as the time period for the 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 crime rates per 100,000 people for 10 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 were 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 model building can be found under the following link in the Github repository:

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

# Feature Engineering (Variable Construction)

* 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 also 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. Below provides an example about how each type is processed to construct new variables to represent the information in the original variable.

Table 2 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.

***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 that includes 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 can be grouped into four categories. Therefore, a total of four composite variables representing these categories was constructed using a simple sum score from variables within each category.

***Missing values.*** In the model building process, two primary models were used: 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 median value for each feature variable.

* 1. **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 3. 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; and, standardized composite scores for the first four principal components were added to the dataset. As a 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 1), therefore the variables were aggregated by taking an average across all PUMAs assigned to the unique Residence codes for individual assignment. Below provides a sample demonstration 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 the original individual level training and test datasets provided by NIJ.

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 |

* 1. **Processing variables in the county-level crime statistics**

All variables in the county-level crime statistics were numerical variables indicating the crime rates per 100,000 people for 10 variables (murder, rape, robbery, assault, burglary, larceny, theft, arson, and total). These were variables were not processed. A similar procedure to 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 one or more counties through assigned PUMAs (see Table 1). 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 |

* 1. **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 additional process was applied. A similar procedure as described and demonstrated before were 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 4.

# Model Building

* 1. **Predicting Recidivism in Year 1**

Five different XGBoost model and two linear regression models with ridge penalty was 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 644 variables available to use as features after processing all the variables. These models differed in the features being used. Below provides 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 of 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 was identified, and a different XGBoost model was optimized by using only these 50 most important predictors.

**LR1**: A logistic regression model with a ridge penalty was developed by including the main effects of all 644 variables.

**LR2**: A logistic regression model with a ridge penalty was developed by including the main effects of all 644 variables and all two-way interactions (N = 207,046) among these 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: 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 these parameters can be found at the following link:

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

Below is a table that presents the 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. For all models, a randomly selected 15,000 observations were used to optimize the model parameters with a 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.

* 1. **Predicting 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 bulding for Year 2, the individuals 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 a 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 bulding for Year 3, the individuals 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 a 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.

|  |  |  |
| --- | --- | --- |
|  | 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 |

# Results

# Future Considerations and Final Remarks

* Were variables added to the data set? If so, detail the variables.
* What variables were constructed? How were the variables constructed?
* Which variables were statistically significant?
* What variables were not statistically significant? How was this handled? For example, were they dropped from the overall model?
* What type of model was used?
* Did you try other models? Were they close in performance? Not at all close?
* What other evaluation metrics should have been considered/used for this Challenge? For example, using false negatives in the penalty function.
* Did the 0.5 threshold affect anything? Would your team recommend a different threshold?
* Did the fact that the fairness penalty only considered false positives affect your submission?
* Are there practical/applied findings that could help the field based on your work? If yes, what are they?
* What should NIJ have considered changing (other than metrics) to improve this Challenge?
* For future Challenges, what should NIJ consider changing to improve Challenges? For example, more/less time, different topic, or data issues (missing data)?