# 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 openly 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 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 49 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 49 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 code 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. The detailed information about these variables, and how these variables were processed before modeling will be given later under the Feature Engineering section.

***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.

***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>). These 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, economical 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 variabled 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)

# Model Building

# 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)?