Introduction to Data Science for Public Policy *March 29, 2017*
Prof. Jeff Chen and Prof. Dan Hammer
McCourt School of Public Policy, Georgetown University

Predicting Recidivism in the U.S. Prison Population

- Seth Taylor, Shashank Rai, and Viola Hilbert -

Hypothesis/Research Question: Our goal is to analyze correlations between recidivism and socio-economic characteristics, mental health, family background, drug use, type of prior offenses and convictions, etc. to predict the probability of recidivism among prisoners. We intend to make predictions regarding important risk factors and to draw policy conclusions as to how the risk of recidivism can be lowered for different types of (former) prisoners. For instance, identifying certain mental health patterns or drug abuse behaviors that significantly increase the risk of recidivism would help in setting priorities in health care and/or social work care for released prisoners.

Data Experiment Design: Recidivism is our target variable, where we want to predict recidivism as a binary variable taking the values 0 or 1. This variable is created from a survey variable on Criminal History that accounts for different cases of first time offenders and recidivists. In the 2004 survey data, both for federal and state prisons, one-third of prisoners are first time offenders (Y=0) while two-thirds are recidivists (Y=1). Thus, both Y=0 and Y=1 have at least n=100 and are represented in well more than 10% of the sample.

Data: Our data comes from the 2004 Survey of Inmates in State Correctional Facilities (SISCF) and the 2004 Survey of Inmates in Federal Correctional Facilities (SIFCF). These surveys, collectively referred to as the 2004 Survey of Inmates in State and Federal Correctional Facilities (SISFCF), provide nationally representative data on inmates held in state prisons and federal prisons. Collected through personal interviews conducted from October 2003 through May 2004, the data captures information about prisoners' current offense and sentence, criminal history, family background, socio-economic characteristics, prior drug and alcohol use and treatment programs, gun possession and use, as well as prison activities, programs, and services. While the data is well sampled and representative of the US nation, it is less certain that we can generalize on the prison population in 2017. Sentencing and time served are likely to affect recidivism rates. Thus, important changes in legislation after data collection, such as the Fair Sentencing Act of 2010, might influence recidivism risks for the current prison population.

Proposed Processing Methods: First, we will generate the binary target variable from a categorical recidivism variable in the survey that distinguishes between current and prior minor and violent offenses. Only 88 observations out of a total of 18,185 observations in the dataset are missing our target variable recidivism. However, the data suffers from missing data for independent variables such as prior drug use, where in almost 65% of cases the convicts refused to answer (or the data is not available for different reasons). Such a high non-response rate is likely to cause some bias, because those who refused to answer are likely to have certain motivations for not disclosing that kind of information.

Proposed Analytical Methods: We plan to train different prediction models using k-folds cross validation. We believe that there are sufficient observations in our data for both target variable outcomes to render data upsampling unnecessary. We would like to be able to go beyond simple predictions of binary outcomes for recidivism. More specifically, we would like to go in depth as to the predictive variables, i.e. the risk factors inclining a former prisoner to commit a new offense. For logistic prediction models, this would entail analyzing marginal effects, i.e. how a change in behavior or characteristics will impact the likelihood of recidivism. For other prediction models where marginal effects are difficult to identify, we could attempt to find a "tipping point" at which our prediction would change from non-recidivist to recidivist. Finally, since we have data for both federal and state prisons, we would want to examine whether recidivism rates and risk factors differ between federal crimes and state crimes. If so, it will be interesting to analyze whether there is an interaction effect with socio-economic background, criminal history, and other predictive variables. This will be rather straightforward in a logistic prediction model, but might turn out to pose important computation and interpretation challenges for other supervised learning methods.

We may want to try applying our prediction models trained with the 2004 SISFCF data to two different datasets on Monitoring of Federal Criminal Sentences in 2007 and 2015. The latter do not measure recidivism but share a range of predictive variables with the 2004 survey. While we understand that data sampling as well as the measurements of specific variables differ between the 2004 SISFCF data and the 2007/2015 Federal Criminal Sentences data, so our predictions must be interpreted very cautiously, this could be an interesting way of testing the models we built.

Ethical Considerations: The data we are working with is individual-level data that covers an inherently vulnerable population (inmates). While the ID's are anonymized, it contains sensitive information such as history of past abuse, family criminal history, drug use, etc. Therefore, we should be thorough in ensuring that any potentially identifiable information (PII) has been removed or masked within the dataset. The dataset has been pre-processed by the National Archive of Criminal Justice Data to be suitable for public use. Additionally, we are aware that any predictive model of recidivism risk could be misused as a justification to deny parole or early-release for prisoners. This is contrary to our intentions, because our goal is to improve the former inmates' chances of remaining compliant with the law by understanding the factors that affect recidivism and by identifying support structures for high-risk released prisoners which improve their living conditions and help them reintegrate into society.