Predicting Recidivism in Johnson County

Ran Bi
Yanning Cui
Ruochen Qiu

Jun 2, 2017

1. Background and Introduction

Recidivism is an important measure of the criminal justice system's effectiveness in reducing crime. It is usually measured by criminal acts that resulted in return to prison with or without a new sentence during a three-year period following the prisoner's release. According to a recent report by the Pew Center on the United States, more than 40% offenders return to state prison within three years of their release despite a massive increase in state spending on prisons.[1]

While a lot of resources are devoted at the front end of the justice system, by sending offenders into jail, it is argued that the tail end of the justice system has not been paid equal attention to. Without education, job skills, and other basic services, offenders are likely to repeat the same steps that brought them to jail in the first place. And recidivism is affecting almost everyone in the community: the formerly incarcerated person, their family members, the victims, and those who work in the justice system. Therefore, policymakers should also take efforts to decrease the likelihood of reoffending through certain interventions. Our group is motivated to help local governments, policymakers, and practitioners find effective strategies to reduce recidivism rates, increase public safety, save money spent on incarceration, and cut correctional costs.

The goal of our project is to identify top 200 individuals who are most likely to return to jail during a 12-month period following his/her release. We hope to build a model to predict these high-risk individuals and find out what factors might contribute to the recidivism. If the local government can identify these individuals as early as possible and execute customized interventions based on their personal characteristics and circumstances, they can receive proper rehabilitation during the incarceration or better fit back in with 'normal' life after they are released back into the community. (i.e. targeted rehabilitation, transition planning, community correctional programs)

2. Related work

Statistics from the Bureau of Justice tracked a sample of 404,638 former prison inmates from 30 states for five years following their release in 2005 and discovered that: 1) within three years of release, about 67.8 percent of released prisoners were rearrested; 2) within five years of release, about 76.6 percent of released prisoners were rearrested; 3) property offenders were the most likely to be rearrested, with 82.1 percent of released property offenders arrested for a new crime compared with 76.9 percent of drug offenders, 73.6 percent of public order offenders and 71.3 percent of violent offenders.[2] Other studies also support the statistics that drug offenders are the most likely to return to jail. They commit crimes either under the influence of drugs, or in order to get money to buy drugs.[3]

[1] Pew Center on the States. "State of recidivism: The revolving door of America's prisons." (2011). Link: http://www.pewtrusts.org/en/research-and-analysis/reports/0001/01/01/state-of-recidivism

[2] https://www.nij.gov/topics/corrections/recidivism/Pages/welcome.aspx

[3] Austin-Ketch, Tammy L., et al. "Addictions and the criminal justice system, what happens on the other side? Post-traumatic stress symptoms and cortisol measures in a police cohort." Journal of Addictions Nursing 23.1 (2012): 22-29.

The study from Pew Center offers three successful example of attacking recidivism: Oregon, Michigan and Missouri. Inmates receive risk and needs assessments at intake, and targeted case management during incarceration, along with detailed transition planning before release. Released offenders are equipped with tools to succeed in the community, through individualized correctional programs based on the individual's risk, needs, and strengths.[4]

Recently, scholars start to use machine learning tools to predict recidivism, which allows agencies to base their personnel and policy decisions on a scientifically proven method. Philadelphia uses random forest modeling to identify probationers likely to reoffend within two years of returning to the community.[5] DSSG (Data Science for Social Good) at the University of Chicago worked with Johnson County, Kansas, and used machine learning to prioritize outreach to individuals most at risk of being booked into jail within the next year. They argued that mental health problems might contribute in some individuals to an increased risk of recommitting crimes. Their model predicted jail booking in the following year with 51% accuracy, which outperformed both random selection and simple heuristics at identifying people with histories of mental illness and incarceration at risk of returning to jail. Their work provided a data-driven framework for Johnson County as well as many other jurisdictions to develop mental health and social service interventions to avoid incarceration. [6] Inspired by DSSG, our work also utilizes machine learning to identify the high-risk individuals that are most likely to return to jail.

3. Problem formulation and Overview of the solution

Our study estimates the recidivism patterns of 22,110 unique persons released from 2011 to 2016 (booked from 2010 to 2015), from Johnson County jail. The crime data detailed the arrest, release, adjudication, and incarceration experiences of these former inmates. Our project used a 5-year follow-up period, which offered supplementary information for policymakers and practitioners on the officially recognized criminal behavior of released prisoners. The longer recidivism period also provides a more complete assessment of the number and types of crimes committed by released persons in the years following their release. We also use American Community Survey 5-Year Data (2009-2015), which is available via API from the United States Census Bureau.[7] The time range of this dataset corresponds with the crime data, and the multiyear estimates provide the increased statistical reliability for small population subgroups (i.e. zip code).

We trained our model with data from jail system (booking, charge, and demographic information of individuals) as well as information of income, marital status, education level, and health insurance coverage from ACS. We hope that our model can be translated into practice to help Johnson County illuminate customized intervention strategies for the top 200 individuals who are most at risk, and help cut reoffending and corrections costs.

4. Data Description

4.1 Criminal Justice

The primary data sources come from Johnson county Criminal Justice department. We utilized three datasets to predict the probability of recidivism within one year: *booking*, *current charges* and *person.* The Johnson County jail and court system has an integrated justice management system that spans interactions from booking through probation. A total of 28,579 jail bookings are recorded in this dataset. There are 22,110 unique individuals - 16,107 males and

[4] http://www.pewtrusts.org/~/media/legacy/uploadedfiles/pcs_assets/2011/pewstateofrecidivismpdf.pdf

[5] Ritter, Nancy. "Predicting recidivism risk: New tool in Philadelphia shows great promise." *National Institute of Justice Journal* 271.February (2013): 4-13.

[6] Salomon, Erika, et al. "Reducing Incarceration through Prioritized Interventions." (2017).

[7] https://www.census.gov/data/developers/data-sets/acs-5year.html

6,002 females - since January 1, 2010. Among them, 15,902 inmates in total are residents in Kansas State.

Table *booking* provide information on dates of entry and exit, bail type, bail amounts, crime type and outcome per person per case. Table *current charges* provides detailed charge information per case, including charges level and descriptions, drug offense, severity, charge position, found guilty or not, and trial disposition. Table *person* records demographic information such as race, gender, address, zip code, date of birth, Johnson County residency, etc.

4.2 American Community Survey

The external data source is American Community Survey 5-Year Data (2009-2015) available via API from United States Census Bureau. The American Community Survey (ACS) is an ongoing survey that covers a broad range of topics about social, economic, demographic, and housing characteristics of the U.S. population. We chose 5-Year data over 1-Year data to ensure higher statistical reliability of the data for less populated areas and small population subgroups.

4.3 Cleaning and combining datasets

The DSSG project group has done the record linkage and cleaning process on the criminal justice datasets. In the cleaned datasets, each individual is represented by a unique *mni_no* based on hashed SSN and each case is represented by a unique *case_no*. We used *booking* table as the base table for feature engineering. Each row represents a booking record for an *mni_no – case_no* combination. After generating features based on individual tables, we joined ACS data to *person* table on *zip code*. Then *person* table was joined to *booking* table on *mni_no* while *charges* table was joined to *booking* table on *case_no*.

5. Methods

The goal of the project is to identify individuals who are at high risk of returning to jail within one year. The policy goal is translated into a binary classification problem - to label whether an individual will re-enter the jail within one year after release. As Johnson County is capable of intervening 200 high-risk individuals annually, our model scores each inmate's risk of re-entering jail within the next year and generates a list of 200 top ranked individuals to intervene. We generated 154 features, fit 6 types of classifiers and used temporal validation to select a model that performed well in terms of precision at 200 score and was relatively stable over the five-year valuation window.

5.1 Features

Features were first generated based on available datasets, feature engineering strategies covered in class and consultation of crime literature. We created four groups of features: demographics, jail booking statistics, case specific information, community characteristics. *Table 1* shows example of each group of features. Feature engineering strategies we adopted include categorical to binary, categorical histogram, features for missing values, discretization, date/time features, and aggregation over time periods.

Demographics
Gender Race Age when arrested / buckets Joco resident City
Jail booking statistics
Bail amount (current booking, sum last year/month/week, average last year/month/week) Bail amount buckets Bail type Booking count last year/month In jail length Arrest agent Crime type
Case Specific Information
Trial count Severity count Guilty count Charge position count Crime class count Charge description count

Community Characteristics
Income buckets
Health * Age buckets interaction
Education level
Marriage status

Table 1. Examples Features: demographics, jail booking statistics, case specific information, community characteristics.

5.2 Model Fitting

Models were fit using the scikit-learn package in Python and validated by temporal holdout strategy. *Table 2* (please see the appendix) exhibits the model and hyperparameter combinations we trained and the rationale of choosing them respectively. For temporal validation process, we held out jail bookings in 2015 as final test set. We trained and validated models on jail bookings between 2010 and 2014 in two ways. The first method was to train model on features and label of jail bookings in one year, and use the model to generate predictions for outcomes from the following year (e.g. the model trained by 2010 data was used to predict recidivism in 2011 data, and so on). The second method was to train model on features and label of jail bookings since 2010 up to certain year, and validate the model by the following year data (e.g. the model trained by 2010-2013 data was used to predict recidivism in 2014).

5.3 Model Selection and Evaluation Methodology

Since Johnson County is capable of reaching out to at max 200 individuals, we most cared about whether the 200 individuals of highest predicted risk score actually re-entered the jail within one year after release. Therefore, we used *precision at 200*, i.e. true positive out of all positive predictions at the threshold of top 200 predictions, to select the best performing model out of the temporal validation process. We also checked whether the model performed consistently well in each of the prediction years. Once we selected the best performing model and corresponding training window (i.e. train by the previous year data or train by data since 2010 up to the previous year), we applied the model to jail bookings of 2015 and evaluated its prediction results.

For baseline comparison, the best performing model is compared with selection at random, simple heuristics and expert heuristics. We used simple decision tree of depth 1 and 2 to simulate how human identify individuals at risk based on one or two rules.

6. Results

The final model selected based on data from 2011 to 2014 was a Gradient Boosting classifier with 10 boosting stages to perform, 0.001 in learning rate, full samples used for fitting the individual base learners, and a maximum depth of 5. The corresponding training data was jail booking since 2010 up to start of the prediction year. Among the 200 individuals the model identified as highest risk in our holdout year (2015), 84 were actually booked into jail – a precision of 42%.

6.1 Baseline Comparison

The simple heuristics baseline is a decision stump which classified individuals as at risk if he/she was booked to jail in the last year. Among 200 individuals sampled from such group in 2015 data, 21% were actually re-booked to jail within one year after release. Our model was twice as precise as the simple heuristics baseline in terms of precision at 200.

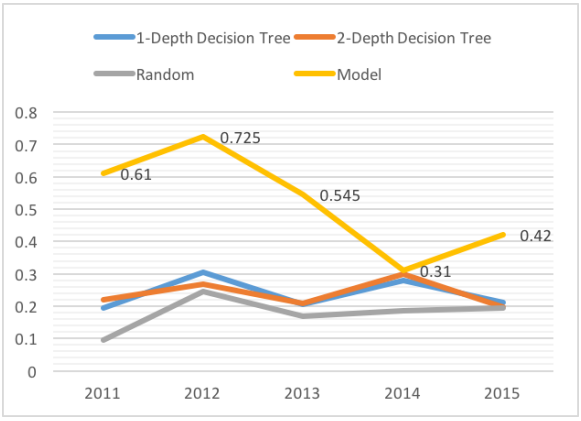
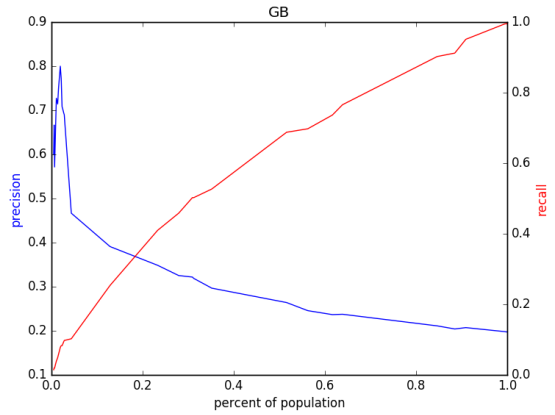


Figure 1 Baseline Comparison: Precision for the top 200 individuals at risk

The expert heuristics baseline is a two-depth decision tree which classified individuals as at risk if he/she was booked to jail in the last year, and had more than 1 charges in his/her crime case. The precision at 200 for 2015 jail booking data is even slightly lower than simple heuristics. *Figure 1* shows the Gradient Boosting model consistently perform better than baselines over the five years.

Figure 2 is the precision-recall curve for the final model in 2015 when the threshold for the positive label is set at the top x% of the risk score ranking. Although the precision remains around 30% after more than 30% population are labelled as at risk, the precision is quite high if we aim at a small percentage of population (precision is as high as 80% at around 5% threshold). Taking intervention capacity into consideration, the model should work well for Johnson County to target at less than 200 top risk individuals.

Overall, our model performs better than random selection, simple heuristics and expert heuristics in terms of precision at 200 score in the five-year time span. Moreover, its high precision at small threshold fits Johnson County's intervention capacity.



Figures 2 Precision and recall of the final model in 2015

6.2 Exploring the (Predicted) High Risk Individuals

The average age of the top 200 high risk individuals is substantially younger than the whole population (29 years old compared with 32), with 14 as the minimum for their first arrest and 61 as the maximum. The top 200 high risk individuals' average incarceration is around 27 days whereas the whole population averaged nearly a month (31 days). In the criminal justice system, the higher risk group's average bail amount is $3520 as compared to $4113 for the whole population. Meanwhile, although most criminal justice contacts involve men, the proportion within the top 200 was higher (89% compared to 76.34%) and white people are more highly represented (79%) in the top 200.

We also found those who had multiple charges per case are more likely to return *(Figure 3, please see the appendix),* while compared with domestic violence and juveniles, the individuals with a criminal case type are more likely to return. Most individuals spent less than one day in jail, and those who stay shorter periods have higher risk to return. *(Figure 4, please see the appendix)* We also identified the "high risky city" since the high risky recidivists are mostly from Olathe, Overland Park, Kansas City and Shawnee. In addition, the individuals who are bailed out through surety bond are more likely to return.

7. Policy Recommendations

Based on the result discussed above, it may necessary for Johnson County to implement an innovative Juvenile Support Program to deal with the high risk of recidivism among juveniles. Research[8] has shown that prevention and early intervention are more effective to prevent the juvenile delinquent behavior than remediating visible and/or longstanding disruptive behavior. The early intervention program can be cooperating with the local school and community college and focus on bullying prevention, classroom and behavior management and afterschool recreation activities to meet both the physical and emotional needs of the juveniles.

For the adult inmates with high recidivism probability, some community support programs

[8] Loeber, R., Farrington, D. P., & Petechuk, D. (2003). Child Delinquency: Early Intervention and Prevention.

including conflict resolution and violence prevention curriculums, job oriented skill building and anger control programs can be implement to assist the inmate during the detention and after release. More efforts should be put on those high risk cities identified through the research. Future researches such as new model generation to distinguish the detailed attribute of the recidivism might be considered to better support this policy. It would be more efficient by directing the inmates to the appropriate services like housing assistance, mental health triage, or substance abuse counseling base on their priority needs.

In fact, surety bonds may lower the crime cost and induce recidivism. Local criminal justice systems may put limitations on individuals who tend to be bailed out through surety bonds. Some specific factors related to the inmate's background should also be considered when implementing this policy. For instance, if an inmate had more than two recidivism record within a year and been involved in drug trafficking or gang activities, some constraint on bail out method or amount can be considered to raise their "cost" on recidivism.

8. Limitations and Improvement

There are less than 30,000 data entries in total for this study, which may constrain the training and prediction capability of features and models. In addition, the ACS data (income, marital status, education level, health insurance coverage, etc.) fetched for the project only goes down to zip code level, and it limits the precision to measure each individual's circumstances. Since our goal is to provide additional support to individuals who may be at risk of jail interactions, it would be important to lower the probability of missing individuals due to the vague data.

To further improve the research work, other advanced models, such as isolation forest, can be adopted since the datasets we have are relatively imbalanced. Meanwhile, more specified data for each inmate such as income, mental health data, medical recorded, marital and employment status can also be helpful to developing more detailed features and assist the model perform better. Another important area of future work is to involve the notions of bias and discrimination in the model predictions. The historical data collection in criminal justice may inevitably biased in the process and the model generated based on those data may lead to reinforcing that bias.

9. Future Experimental Design

A well designed experiment can be considered to continue to improve our models and to understand the generality of this method. To future testify the prediction and adapt the model, we plan to randomly divide the the top 200 individuals with high risk of recidivism into two groups. There will be no customized intervention for the control group, whereas for the treatment group, two steps intervention will be adopted to test whether the policy suggestion mentioned above can help to reduce the recidivism rate:

Phase 1: Provide inmates with targeted case management and services during incarceration, to get them prepared for fitting back in with normal life (ex. job skill training)

Phase 2: Provide inmates with up to 12 months of supportive services in the community after they are released.

One-year recidivism rates from both groups will be collected for further analysis and comparison to examine the accuracy of the prediction and the capability of the intervention works. Considering the potential cost issue, the group size can be reduced to 50 in each group by randomly picking while the implementations remain the same.

10. Conclusion

Using criminal justice data and American Community Survey data, we built a machine learning model that predicts an individual's risk of recidivism within one year after release and generates a list of top 200 individual at risk per year for Johnson County. The model is selected by temporal validation method and consistently performs better than random selection or heuristics baseline in terms of precision at 200. Based on analysis of those predicted high risk individuals, we recommend Johnson County to

implement an innovative Juvenile Support Program and community support programs, and put limitations on individuals who tend to be bailed out through surety bonds. To further improve our models and to understand the generality of this method, an experimental design is also recommended.

Appendix

Classifiers -Hyperparameters	**Rationale**
Decision Tree criterion: gini, entropy max_depth: 1, 2 min_samples_split: 2, 5, 10 min_samples_leaf: 1, 200	Simulate expert heuristics as evaluation baseline (Simple 1 or 2 deep Decision Tree).
K Nearest Neighbors n_neighbors: 1,5,10,25,50,100 weights: uniform, distance algorithm: auto, ball_tree, kd_tree	Robust to noisy data. Generally fit for classification questions.
Logistic Regression Penalty: l1, l2 C: 0.00001,0.0001,0.001,0.01,0.1,1,10	Mostly used by existing recidivism risk assessment tools.
AdaBoost algorithm: SAMME, SAMME.R n_estimators: 1,10,100	Requires little tweaking of parameters or settings. Versatile with textual data.
Random Forest n_estimators: 1,10,100 max_features: sqrt, log2 max_depth: 1,5,10,20,50,100 min_samples_split: 2,5,10	Considers the nonlinear effects of a large number of variables with complex interactions.
Gradient Boosting n_estimators: 1,10,100 learning_rate: 0.001,0.01,0.05,0.1,0.5 subsample: 0.1,0.5,1.0 max_depth: 1,3,5,10,20,50,100	Ensemble of weak prediction model. Reduce overfitting issues.

Table 2. Grid Search Parameters for Model Selection

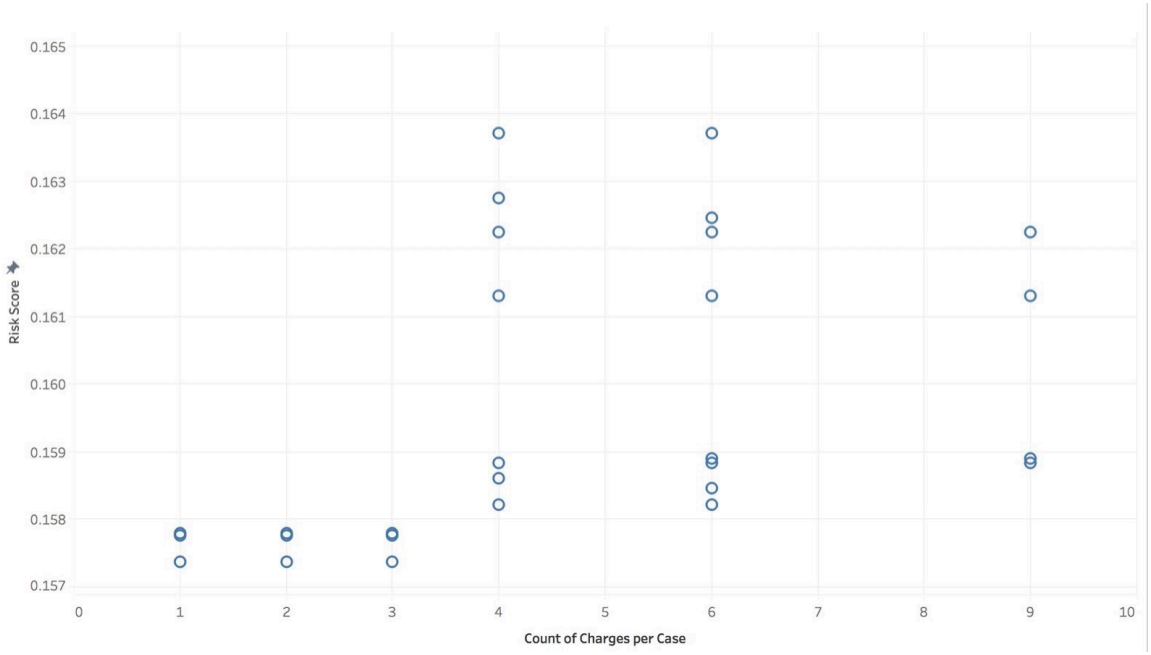


Figure 3 those who had multiple charges (4 or 6) per case are more likely to return

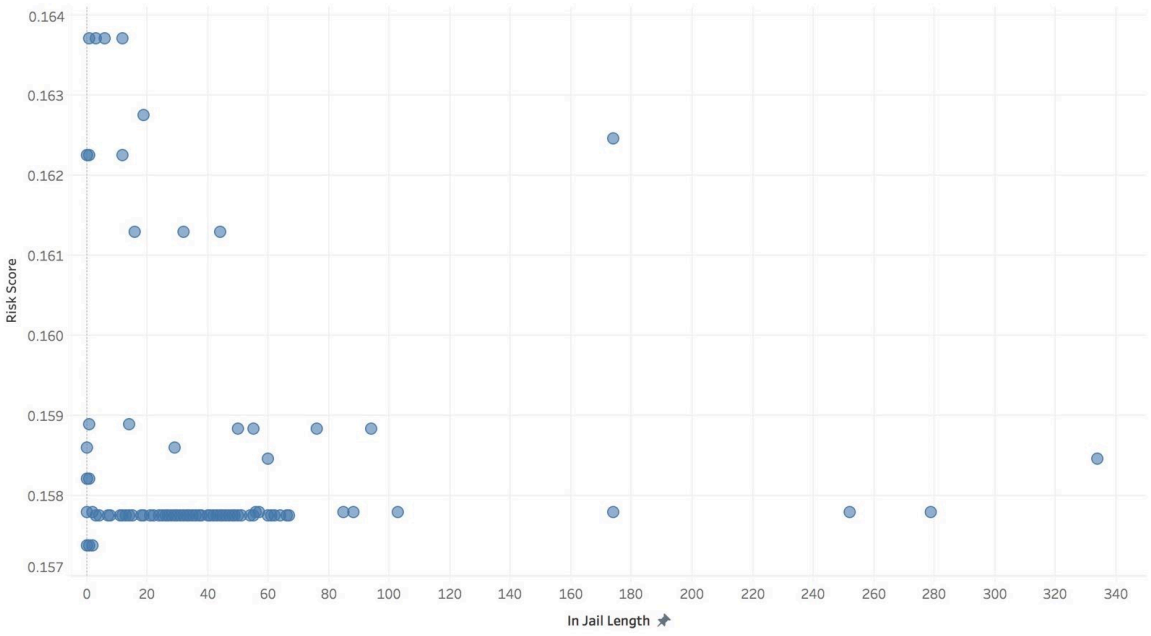


Figure 4 those who stay shorter periods have higher risk to return