National Institute of Justice Recidivism Challenge Report

Team: Aurors

Introduction

Forecasting algorithms have been used in a wide range of applications, from marketing purposes to credit applications. The fast advancement of forecasting techniques has made them more accurate and reliable, making them suitable for implementation in the criminal justice system. In the United States (US), law enforcement agencies have leveraged these techniques for predictive policing. According to Perry (2013), predictive policing consists of applying quantitative methods based on statistical forecasting to prevent crime, among other purposes. The National Institute of Justice (NIJ), following the priorities set by the US Department of Justice, is leading the research, development, and evaluation of methods to increase public safety and the fair administration of justice. For that purpose, the NIJ funded the Recidivism Forecasting Challenge, aiming to improve the ability to forecast recidivism.

The following report is the result of this challenge. This work uses individual and geographic-specific databases to design and calibrate forecasting models that help predict recidivism. The database was built using data from Georgia about people in parole supervision between 2013 and 2015, provided as a part of the NIJ Challenge, and data from the US Census Bureau about socio-economic characteristics at a geographic level. The forecasting models include regression analysis methods (binary logit and LASSO regressions) and machine learning techniques (random forest), combined through a model averaging procedure. The output consists of the percent likelihood of individuals recidivating within one, two, or three years from release. While this report explains the database construction process and the modeling approach, the results focus on the section of female paroles recidivating within three years, since that is the category where the team came in second place.

The report is structured as follows. Section 2 presents an overview of the relevant literature. Section 3 discusses the construction of the variables. Section 4 contains a discussion of the modeling process. Section 5 presents a summary of the findings. Section 6 focuses on future considerations. The report is finalized with conclusions and references.

Relevant Literature

This section provides a brief revision of the literature associated with prediction in criminology. Early research in the criminology field started with predicting the success or failure of parole (Burgess, 1928; Tibbitts, 1931). Using past records of the Illinois Parole System, Tibbitts (1931) predicted the observance or violation of parole with statistical methods. Lately, more sophisticated techniques have been used for prediction in criminology. Logistic regression (LR) was used to forecast recidivism in mentally-ill offenders (Gagliardi et al., 2004) and older adults (Rakes et al., 2018) released from prison.

Palocsay et al. (2000) predicted criminal recidivism using neural networks (NN). Through numerical experiments, they demonstrated that NN models have a higher classification accuracy for criminal recidivism compared with logistic regressions. The authors also highlighted the importance of selecting the network topology and the training methodologies in prediction models. Wadsworth et al. (2018) used an adversarially-trained NN model to predict recidivism and reduce racial bias. Wang et al. (2010) presented a general framework to predict criminal recidivism with support vector machines (SVM). The authors compared the performance of the NN, SVM, and LR models. As a result, in some cases, NN and SVM models outperform LR models; furthermore, combining the three models presents a higher performance than each single model. Tollenaar and Heijden (2012) questioned the best method to predict recidivism comparing statistical, machine learning, and data mining models. The authors demonstrated that traditional methods, e.g., LR, perform equally good or better than modern models.

Zeng et al. (2015), concerned about recidivism models that are accurate and easy to use and interpret, used a Supersparse Linear Integer Model (SLIM) that met these criteria. Ozkan (2017) predicted recidivism in the U.S. criminal justice system through machine learning techniques. The author compared several machine learning models and found that XGBoost and neural networks outperformed the other models considered. Overall, the revision of the literature shows the continuing need to find methods that improve the overall performance and the reduction of bias of predictive models in criminology.

Variables

The dataset was constructed from two sources: (1) the data provided by NIJ, which consist of individual-specific variables of people in parole supervision between one to three years (this dataset from Georgia ranged from January 1st, 2013, to December 31st, 2015); and (2) data from the Census Bureau's American Community Survey Public Use Microdata Sample (ACS PUMS). More specifically, we used geographic-specific data by PUMA area tied to the 25 regions (aggregated PUMAs) of the first dataset. Among the variables from the ACS PUMS database, we selected those that help better describe the socio-economic conditions of the PUMA area, based on our judgment. The chosen variables were the following ones:

- Code of aggregated PUMAs in our dataset
- Average age
- Percentage of unemployment
- Average family income
- Percentage of households where grandparents are responsible for children
- Percentage of people with no health insurance coverage
- Percentage of houses with size less than 1 acre
- Percentage of houses with size less than 10 acres
- Percentage of commuters that do not use car
- Percentage of workers that work from home
- Average travel time of commute
- Percentage of households using food stamps
- Percentage of people born in a foreign country

- Income to poverty ratio
- Percentage of people with no public health insurance
- Average monthly rent
- Percentage of females
- Percentage of houses that are rented
- Percentage of houses that are not owned but the residents don't pay
- Average property value
- Percentage of households with zero cars, with one car, and with two cars

These variables do not include the base variables, e.g., "percentage of households with more than two cars" is not included since that would be the complement of the included variables on car ownership.

Acknowledging that there would be multicollinearity among the geographic-specific variables, since they did not have variation across individuals belonging to the same PUMA region, a Principal Component Analysis (PCA) was conducted to reduce the dimensionality of geographic-specific dataset while retaining the maximum amount of variability among them. The first three components were used, retaining together approximately 79% of the total variance, as can be seen in Figure 1.

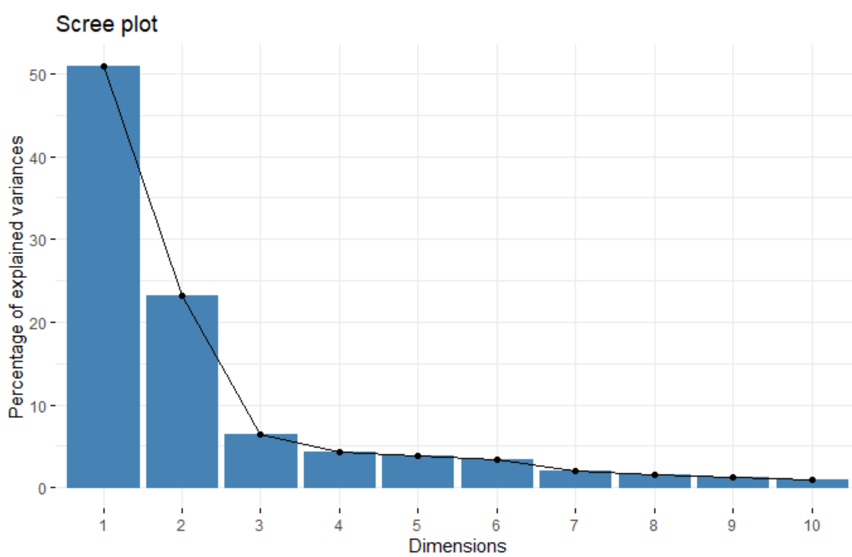

Figure 1. Percent of variance explained by each principal component (in decreasing order)

Regarding the data provided by NIJ, the authors converted the categorical variables into dummy variables. For the modeling process, the lowest category was used as the base; e.g., for the variable of age, the base case is the group of individuals between 18-22 years old.

Model

A brief description of the various model types used in the analysis is presented in this section. Particular attention is given to the results obtained after their application to predict the recidivism rates for females during the third year.

All models mentioned in this section were trained and tested to check their individual accuracy using an 80%-20% training-testing split of the training dataset available (without considering the testing dataset), i.e., 80% of the individual training records were randomly chosen to conform the training dataset and then used to estimate the different model structures. With the estimated models, a validation of their accuracy, separately or combined, was made using the remaining 20% of the data records.

Binary logit model

The authors used a binary logit model to forecast the recidivism for females within the third year. This model was very suitable for this scenario because it considers that the convict has the choice to come back to jail and predicts this probability using a logistic function. The specification of the model is shown in Equation 1 (Greene, 2003):

$$y_i^* = \beta' x_i + \varepsilon_i \quad (1)$$

Where y_i^* is the binary outcome variable of interest for individual i, x_i are the explanatory factors, e.g., the supervision and prison case information, prior Georgia criminal and community supervision history, the Georgia board of pardons and paroles conditions of supervision, and supervision activities. ε_i is the error term.

whitev
Can you describe this a little more. Were eight variables selected 1,000 times and for each of the 1,000 selections one of the eight variables were randomly chosen to start the branching process?

started the challenge by conducting a literature review to have an informed decision about selecting suitable models for the problem at hand. We then started looking for variables that can provide additional insights on recidivism, and then conducted PCA to select the three components that can explain the most variance while decreasing the dimension of the dataset.

We then split the training dataset to 80% training and 20% testing, we used different models (Binary logit, LASSO regression, Random Forest, SVM) along with a model averaging technique to select the best predictive model using the 20% testing dataset. These models provided good insights with relatively small differences among them individually, ranging from a Brier score of 0.1474 for Binary logit to 0.151 for SVM. Combing these models via model averaging procedure provided better results with a Brier score of 0.14724. The final submitted model was a combination of Binary logit model (with a coefficient of 0.75), the LASSO regression model (with a coefficient of 0.13) and the RF model (coefficient of 0.12).

Looking at the Binary logit model and the LASSO regression, it was found that the indicator of percentage of drug tests positive for methamphetamine was the most prominent variable in positively predicting recidivism, while age was found to have an inverse relation with the probability to recidivate. These findings can provide decision makers with a clear idea about the target of their policies to reduce recidivism rates by targeting younger criminals with drug problems especially those with gang affiliation.

References

- Berkson, J. (1944). "Application of the Logistic Function to Bio-Assay." Journal of the American Statistical Association **39**(227): 357365.
- Burgess E.W. (1928). “Factors Determining Success or Failure on Parole.” In: The workings of the indeterminate sentence law and the parole system in Illinois. Springfield, IL: Illinois State Board of Parole.

- Breiman, L., & Spector, P. (1992). Submodel Selection and Evaluation in Regression. The X-Random Case. International Statistical Review / Revue Internationale De Statistique, 60(3), 291-319. doi:10.2307/1403680
- Gagliardi, G.J., D. Lovell and P.D. Peterson (2004). "Forecasting Recidivism in Mentally Ill Offenders Released from Prison." Law Human Behavior 2: 133–155. https://doi.org/10.1023/B:LAHU.0000022319.03637.45
- Greene, W. H. (2003). Econometric Analysis, Pearson Education, Inc.
- Kohavi, R. (1995) A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. Proceedings of the 14th International Joint Conference on Artificial Intelligence, Vol. 2, Montreal, 20-25 August 1995, 1137-1145.
- Ozkan, T. (2017). "Predicting Recidivism Through Machine Learning." PhD. Dissertation.
- Palocsay, S.W., P. Wang and R.G. Brookshire (2000). "Predicting Criminal Recidivism Using Neural Networks." Socio-Economic Planning Sciences, Elsevier 34(4): 271-284, December.
- Perry, W. L. (2013). "Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations". Rand Corporation.
- Rakes, S., S. Grace and S.J. Tripodi (2018). "Recidivism Among Older Adults: Correlates of Prison Re-entry." Justice Policy Journal 15:1-15
- Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological), 58(1), 267-288.
- Tibbitts, C. (1931). "Success or Failure on Parole Can be Predicted: A Study of the Records of 3,000 Youths Paroled from the Illinois State Reformatory.". Journal of Criminal Law & Criminology, 22: 11-50.
- Tollenaar, N. and P.G.M. Heijden (2012). "Which Method Predicts Recidivism Best? A Comparison of Statistical, Machine learning, and Data Mining Predictive Models." Journal of the Royal Statistical Society: Series A (Statistics in Society). 176: 565-584

- Wadsworth, C., F. Vera and C. Piech (2018). “Achieving Fairness through Adversarial Learning: an Application to Recidivism Prediction.” ArXiv: 1-5, abs/1807.00199.
- Wang, P., R. Mathieu, J. Ke and H. J. Cai (2010). "Predicting Criminal Recidivism with Support Vector Machine." 2010 International Conference on Management and Service Science: 1-9, doi: 10.1109/ICMSS.2010.5575352.
- Zeng, J., B. Ustun and C. Rudin (2015). “Interpretable Classification Models for Recidivism Prediction.” Journal of the Royal Statistical Society: Series A (Statistics in Society). 180. 10.1111/rssa.12227.