**Crime and Punishment: Predicting Recidivism with Machine Learning**

**Motivation**

Recidivism is one of the most fundamental concepts in criminal justice. Defined as a person's relapse into criminal behavior, often after receiving sanctions for a previous crime, recidivism is measured by criminal acts that resulted in re-arrest, reconviction, or return to prison during a three-year period following the prisoner's release. [[1]](#footnote-1) Bureau of Justice Statistics studies have found high rates of recidivism among released prisoners. One study that tracked 404,638 prisoners in 30 states after their release from prison in 2005. found that within three years of release, about two-thirds (67.8 percent) of released prisoners were rearrested.[[2]](#footnote-2) In light of such phenomena, we believe that predictive algorithms using machine learning can contribute much to help judges, police officers, and reformative educators, to assess a criminal defendant’s likelihood of becoming a recidivist restructure reformative education for predicted high likelihood recidivists, and thus lower their chances of actually reoffending. Moreover, an analysis of recidivism datasets will uncover correlations between features and recidivism risk, thereby enabling the law enforcement ecosystem to design targeted responses for inmates who express those features.

Moreover, machine learning can also help to vastly reduce, if not eliminate human bias. For example, studies have shown that black defendants were far more likely than white defendants to be incorrectly judged to be at a higher risk of recidivism, while white defendants were more likely than black defendants to be incorrectly flagged as low risk.[[3]](#footnote-3) To the extent that such biases can be identified and rectified, the processes of justice can be much improved when it is backed up by data instead of being heavily influenced by personal feelings and biases. Deemed fitting for the duration of this project, as well as datasets available, we have chosen to train our model on 7215 criminal defendants in Broward County, Florida, with their recidivism rates with the rate that occurred over a two-year period.

**Method**

Linear Regression

**Preliminary experiments**

Dumb baseline

**Next steps:**

Given your preliminary results, what are the next steps that you're considering?

**Contributions**

Peicun took charge of data pre-processing. After understanding and making sense of our dataset in different files, he removed what was deemed unnecessary and combined what was useful for analysis. He then processed the data, through encoding and documentation, so it can be analysed by our machine learning models. He also wrote a fair part of this report.

Meilinda analysed the dataset to provide an overall picture of the data which will inform how we build our models (race, gender, age breakdown; type of offense; recidivism outcome). She sliced and diced the data in order to find correlations between some features and recidivism outcome (i.e. whether the previous offense was violent). Another purpose of analysing the data was to identify whether any features were underrepresented, because having too few training examples demonstrating certain features can hurt the model. Below is a summary of her findings:

=== Overview ===

Number of people in data set: 7214

=== AGE ===

Below 25: 1529 ( 21 %)

Between 25-45: 4109 ( 57 %)

Above 45: 1576 ( 22 %)

=== RACE ===

Caucasian: 2454 ( 34 %)

African American: 3696 ( 51 %)

Hispanic: 637 ( 9 %)

Asian: 32 ( 0.444 %)

Native American: 18 ( 0.250 %)

Other: 377 ( 5.226 %)

=== GENDER ===

Male: 5819 ( 81 %)

Female: 1395 ( 19 %)

=== OVERALL RECIDIVISM ===

Recidivists: 3471 ( 48 %)

Non-recidivists: 3743 ( 52 %)

=== TWO-YEAR RECIDIVISM ===

Two-year recidivists: 3251 ( 45 %)

Non-two-year recidivists: 3963 ( 55 %)

=== CHARGE DEGREE ===

Misdemeanor: 2548 ( 35 %)

Felony: 4666 ( 65 %)

=== VIOLENCE AMONG RECIDIVISTS ===

Total # or recidivists: 3471

Recidivists who were violent: 819 ( 24 %)

Recidivists who were nonviolent: 2652 ( 76 %)

=== PRISON DURATION AMONG RECIDIVISTS ===

NOTE: Some outliers in duration skewed metrics (i.e. several people were imprisoned for 40 000+ days, whereas most were imprisoned for several hundred days

NOTE: Sentence duration data unavailable for non-recidivists

Mean: 1010.708

Standard deviation:5536.666

A screenshot of a cell phone

Description automatically generated

**Works Cited**

Recidivism. (2014, June 17). Retrieved from https://www.nij.gov/topics/corrections/recidivism/Pages/welcome.aspx

Durose, Matthew R., Alexia D. Cooper, and Howard N. Snyder, [Recidivism of Prisoners Released in 30 States in 2005: Patterns from 2005 to 2010 (pdf, 31 pages)](http://www.bjs.gov/content/pub/pdf/rprts05p0510.pdf), Bureau of Justice Statistics Special Report, April 2014.

Larson, J., Angwin, J., Kirchner, L., & Mattu, S. (2016, May 23). How We Analyzed the COMPAS Recidivism Algorithm.

1. Recidivism, 2014 [↑](#footnote-ref-1)
2. See Durose et al. 2014 [↑](#footnote-ref-2)
3. See Larson et al. 2016 [↑](#footnote-ref-3)