Executive Summary: Crime Patterns and Predictors in California (1985–2023)

Deepesh Singhal, Yuxin Lin, Feride Kose, Leonard Afeke

GitHub: https://github.com/git-loa/CrimePatternsProject.git

1. Introduction

High violent crime rates can be related to many factors, such as low income, high unemployment rate, or high dropout rate. It can also be influenced by insufficient government funding for security and social services. With limited public resources, identifying the key drivers of high crime rate can help governments allocate resources more effectively to reduce the crime rate. In this project, we investigate long-term violent crime patterns across California's 58 counties from 1985 to 2023, aiming to identify the socioeconomic and demographic factors most predictive of crime rates.

2. Dataset

The target variable is the crime rate, which is computed by the number of violent crimes divided by the population of the county. We obtained 30 predictor features from 14 datasets. The predictor variables fall into five categories: **crime statistics** (e.g., violent crime rate and clearance rate), **demographics** (e.g., population, age, religious diversity), **socioeconomic indicators** (e.g., median income, poverty rate, unemployment), **government expenditures** (e.g., government spending in education, health, policing) and **education** (e.g., drop-out rate, high school rate). The majority of our datasets are from the official website of the CA Department of Justice, U.S. Census, and State Controller's Office. While the crime statistics are available from the year 1985 to 2023, many of the other features are unavailable before the year 2010.

3. Data Engineering

We used two approaches to handle missing values. First is to apply linear interpolation across years within each county to impute missing data. Second is to simply drop rows with missing values, resulting in the loss of data prior to 2010 and all data for San Francisco County. We prepare both the datasets and choose the one that performs better.

Monetary values such as household income and home value are adjusted for inflation using the Consumer Price Index (CPI). Count-based features (e.g. number of unemployed individuals, religious adherents) are normalized by the total population to produce per capita measures. After the adjustment, we obtain 30 features. We also divide the dataset based on the type of the county (14 urban, 17 suburban and 27 rural) and build separate models for them.

We designed a feature selection function to reduce the number of features. For each model, the function identifies and retains only those features that improve the cross-county R^2 score. The function uses depth first search to find such features.

4. Model Processing and final results

We tried four models: **linear regression, ridge regression, XGBoost** and **Random Forest.** For each of the models, we first applied the feature selection function to select the best

performing features, then run the selected features through a pipeline of standard scalar, principal component analysis and the corresponding model.
We chose Ridge regression as our final model since:

- It is the most interpretable, which helps with policy insights;
- It has comparable out-of-sample performance;
- It has stronger generalization across counties and over time.

Our final model is the **log ridge regression model**: $log(y) = Ridge(features)$. We use principal component analysis and ridge regularization to prevent overfitting. This model achieved cross-county R^2 scores of 0.7 (urban), 0.4 (suburban), and 0.3 (rural).
For the **urban** counties, our model generalizes well across counties and across years. It suggests that the most important factors in determining the crime rates are:

1. The ratio security spending/social spending;
2. clearance rate (the proportion of cases that are solved by the police);
3. adjusted income (median household income adjusted for inflation);
4. drop-out rate (The proportion of 15 to 17 year olds not attending school).

For the **suburban** and **rural** model our model is less accurate. We observed several features (e.g. clearance rate) with the following phenomenon: Our models perform poorly for the counties with high variance in these features. The rural counties have higher variance for all these features when compared to the urban counties. We suspect that this feature variability may result from data collection errors, and it might explain the low R2 score for rural counties.
For future work, we hope to collect more accurate data for suburban and rural counties and come up with better models for these counties.

5. Reference

Hawinkel, S., Waegeman, W., & Maere, S. (2023). Out-of-Sample *R*2: Estimation and Inference. *The American Statistician*, *78*(1), 15–25.
https://doi.org/10.1080/00031305.2023.2216252