Crime Patterns and Predictors in California Counties

Erdős Institute Data Science Project

Contributors: Deepesh Singhal
Yuxin Lin
Feride Kose
Leonard Afeke

Mentor: David Ken

Background

- Crime rates depend on a variety of socio-economic and demographic factors.
- Identifying the strongest drivers can help policymakers allocate resources more effectively.

Goal

- We aim to understand **which factors are the most predictive of high or low crime rates** in California counties 

- **Specifically, we ask:**

- What are the most important **socioeconomic factors** linked to crime rates?
  (e.g., welfare spending, education, unemployment etc.)

- Do these factors **differ across urban, suburban, and rural counties?**  

Data Collection

We collected data in the following categories:

- Crime statistics: violent-crime counts and clearance rates.

- Demographics: population, median age, religious composition.

- Socioeconomic indicators: income, poverty, unemployment, housing.

- Government expenditure: spending on education, health, policing, etc.

- Education: student dropout rates and public-school enrollment.

Data Source

Category	Source
Crime data	CA Department of Justice
Demographic	U.S. Census / CA DOF / ARDA
Economic indicators	U.S. Census /BLS
Government spending	State Controller's Office
Education & Health	U.S. Census ACS

Data imputation and Feature Engineering

Missing Values: Except for the crime stats which are available from 1985, most of the data are available from 1990 or from a later year. We consider two strategies for dealing with missing data:

- Row deletion for missing values, thus using data from 2010 onwards (works best for Urban/Rural).
- Time-series imputation: Fit a simple linear regression (feature vs. year) per county and fill missing values (works best for Suburban).

Inflation & per-capita adjustment:

- Divide all monetary features by that year's CPI
- Normalize by county population
- For example:
  - Adjusted_income = median household income / CPI index
  - Adjusted_police_budget = Police budget / (Population * CPI index)

Feature Selection

We start with 30 engineered features across Demographics, Economics, Housing, Education, Health and Expenditure.

To avoid overfitting and retain interpretability, we design a feature selection function that recursively drops features via a DFS search that maximizes our out-of-sample R^2 score. We thus arrive at a subset of the features that performs well.

For example in the urban model we pick a set of 9 features using DFS.

Evaluation Metrics

1. Mean Squared Error (*MSE*)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

2. R^2 score:

$$R^2 = 1 - \frac{MSE(model)}{MSE(y_{train})}$$

Why out-of-sample R^2 score ?

$$R^2 = 1 - \frac{MSE(model)}{MSE(baseline)}$$

- Crime rates for different years within the same county are close to each other.

Baseline $= \overline{y_{test}}$ → Predicting $\overline{y_{test}}$ is not a valid baseline model. Get negative R^2 score!!

Baseline $= \overline{y_{train}}$ → The R^2 score measures the ability of the model in predicting the crime rate in a new county.

Validation method

We use the following three cross-validation methods:

- **5-Fold Cross-Validation**
- **Leave-One-County-Out Validation**
  - For each iteration, one county is held out as the test set. The model is trained on all **other counties**.
- **Time-Series Cross-Validation**
  - Trained on data **up to 2018** and tested on data from **2019 onward**.

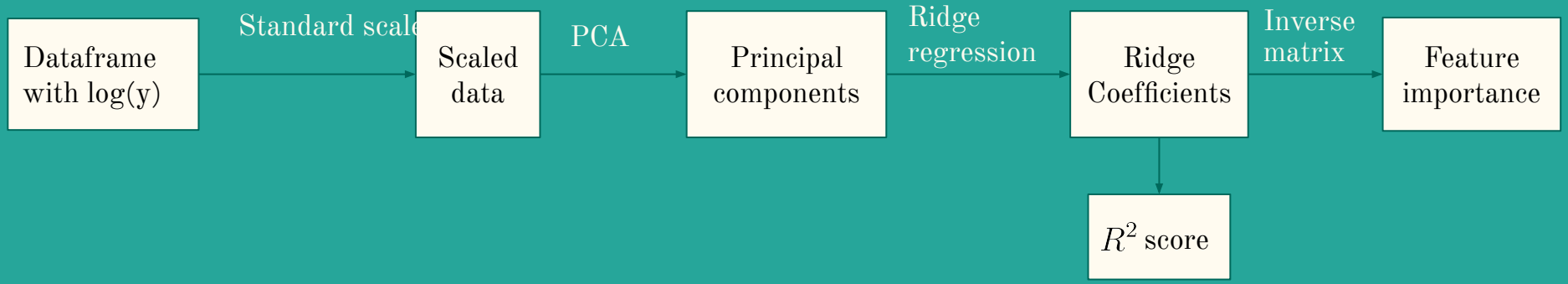
Models considered: **Ridge regression**, **Random Forests** and **XGBoost**.

Modeling Pipeline

- We predict the log crime rate with Ridge regression:

$$log(y) = Ridge(Features)$$

- We train three separate models for: **Urban, Rural, Suburban** counties**.**
- We use Principal component analysis and Ridge regularization to reduce overfitting.
- The ridge pipeline:



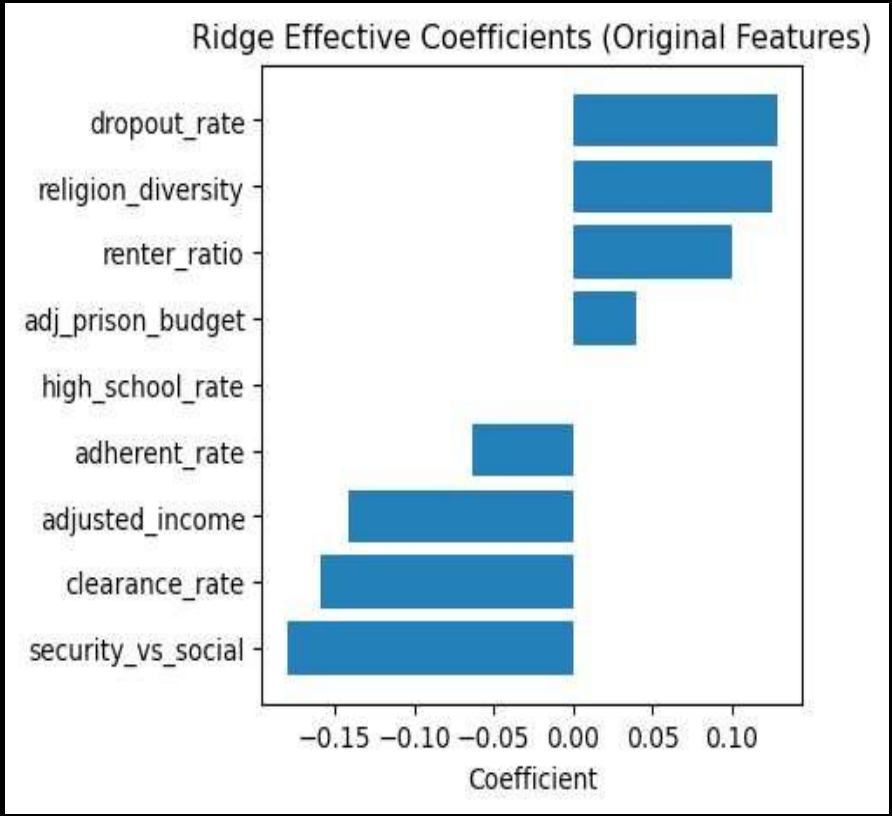
Final Result

The MSE and R^2 of the urban model:

Model	Type	MSE	MR2
ridge	Regular train	6.394766464749542e-07	0.8297497349280686
	Regular val	7.241765287692242e-07	0.825614088429796
	County train	6.087137264289414e-07	0.8351574335910459
	County val	8.185807957824113e-07	0.7175272572112615
	Time train	5.290472022998657e-07	0.8495139146143015
	Time val	1.3650382926719547e-06	0.691964027584281
	Test	5.397108884719842e-07	0.8161354855674486

We see that our model predicts the crime rate of the urban counties successfully.

Feature importance of the Urban model



The Top 4 important features are:

1. Security spending/social spending
2. Clearance rate
3. Adjusted income
4. Drop-out rate

Conclusion

1. Urban counties: Ridge model generalizes well across counties and years.

2. Key levers to reduce the crime are:

- Increase the security expenditure (e.g. police budget, prison budget).
- Effectively solve the crimes to increase the clearance rate.
- Increase resident's incomes.
- Reduce the dropout rate for teenagers.

For the Suburban and Rural counties...

- We carry out a similar analysis on the suburban & rural counties.
- The cross-county R2 score are around 0.4 and 0.3, respectively.
- We suspect that the data quality for suburban and rural counties is not as accurate.

Thank you