Socioeconomic Predictors of Crime via Classical and Deep Learning Methods

Kenny Guo, Jinghui Chen, Keivan Bolouri, Rishauv Kar-Roy

MATH 156: Machine Learning

March 11, 2026

Agenda

- Introduction: Problem & Goal
- Dataset Overview
- Methods & Models
- Results & Interpretability
- Limitations

Motivation

- Crime affects community safety & policy decisions
- Complex relationship: poverty, inequality, unemployment, education, housing, law enforcement institutional behavior
- Better understanding helps allocate law enforcement resources efficiently and facilitate effective crime reduction policies

Problem

- Question: Can we predict violent crime rates in United States communities using socioeconomic data?
- Model Effectiveness: How can we use ML models and which are the most effective in predicting crime rates?

Goal

This project aims to build **regression models** to estimate violent crime rates across U.S. communities.

Data Overview

Source: Communities and Crime Dataset (UCI)

Property	**Value**
Observations	1,994 communities
Features	122 predictors
Target	`ViolentCrimesPerPop` (numerical)
Preprocessing	Normalized to $[0, 1]$, some missing values

Feature Categories:

- **Demographic**: population, race, age, immigrants
- **Economic**: income, poverty, unemployment
- **Education**: high school, college
- **Household**: rent, divorce, kids with two parents

Challenge: 122 features, but only 1994 samples; risk of overfitting

Methodology

- **Problem type**: Supervised regression
- **Evaluation**: Mean Squared Error (MSE)

$$\mathcal{L}_{\mathrm{MSE}}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f_{\mathbf{w}}(\mathbf{x}_i))^2$$

- **Train / Test**: 80% / 20% split

Preventing Overfitting / Dimensionality Reduction

Two strategies:

1. **L2 regularization**: penalizes large coefficients
2. **F-test feature selection**: keep only top k features most correlated with output
3. **PCA**: preserves information

Baseline: Simple Linear Regression

- All predictors used

Performance

- Test MSE: **0.01863**
- R^2: **0.63**

Linear model already captures 63% of variance!

Largest Coefficients

Feature	Coef
PctPersOwnOccup	-0.150
PctHousOwnOccup	+0.124
PersPerOccupHous	+0.099
PctLargHouseOccup	-0.090
MedRent	+0.084

- *Predict less crime when more owner-occupied homes*
- *Predict more crime when higher median rents*

Baseline: Simple Linear Regression

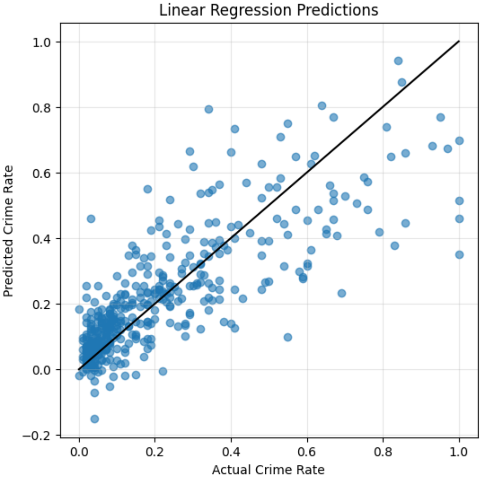


Figure 1: Linear Regression, Predictions vs. Actual Crime Counts

Linear + Feature Selection

Using only top k features (F-test)

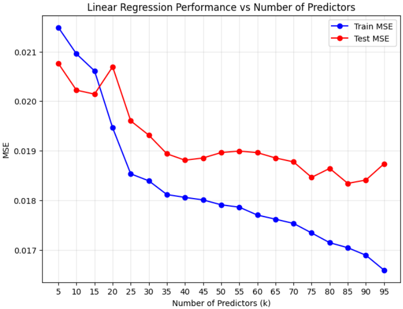


Figure 2: Linear Regression Performance using top *k* predictors.

Performance stabilizes after 50 features.

Likely multicollinearity and redundancy is present.

Polynomial Regression

Why?

Capture non-linear effects & interaction effects:

$$\hat{y} = w_0 + \sum w_{1,j} x_j + \sum w_{2,j} x_j^2 + \sum \sum w_{jk} x_j x_k$$

The Problem

For p features, there can be $O(p^2)$ parameters, overfitting
Solution: L2 regularization, F-test (keep only top k predictors)

Polynomial Regression Results

Test MSE vs. Number of Predictors (k)

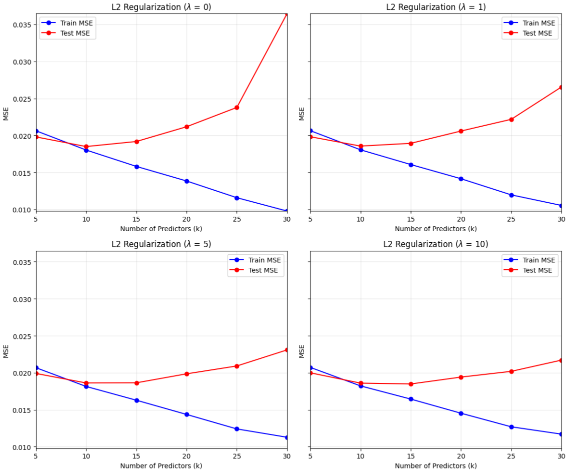


Figure 3: Polynomial regression performance across different k and regularization strengths.

Polynomial Regression Accuracy

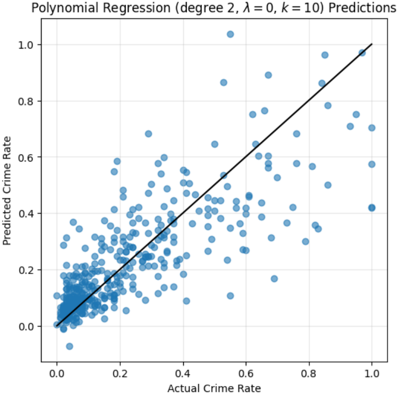


Figure 4: Polynomial Regression ($k = 10, \lambda = 0$ Predictions vs. Actual Crime Counts

Best at $k = 10$: Test MSE = **0.01850**

Marginal improvement over linear (0.01863)

Neural Network (FFNN)

Architecture

- Input layer: top k features
- Hidden layer 1: 32 neurons, ReLU
- Hidden layer 2: 16 neurons, ReLU
- Output: 1 neuron (linear)

Regularization

- L2 penalty (λ)
- Early stopping

Neural Network Results

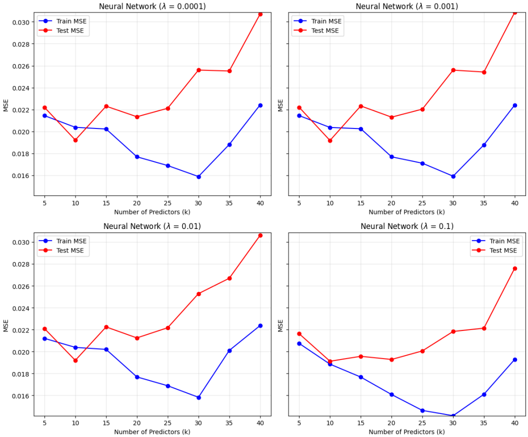


Figure 5: FFNN performance as a function of the number of predictors k under different L2-regularization strengths.

Best: $k = 10, \lambda = 0.1 \rightarrow$ Test MSE $= \mathbf{0.01912}$

Neural networks overfit on small datasets.

Final Model Comparison

Model	k	λ	Test MSE
Linear Regression	122	0	0.01863
Polynomial (deg 2)	10	0	**0.01850**
Polynomial (deg 2)	10	1	0.01856
Neural Network	10	0.1	0.01912

Polynomial wins

While polynomial regression achieves the lowest observed test MSE, the improvement over the baseline linear regression model is marginal. The feedforward neural network, despite its greater flexibility, performs slightly worse on the test data due to overfitting.

Key Result

Predictive signal is mostly linear.

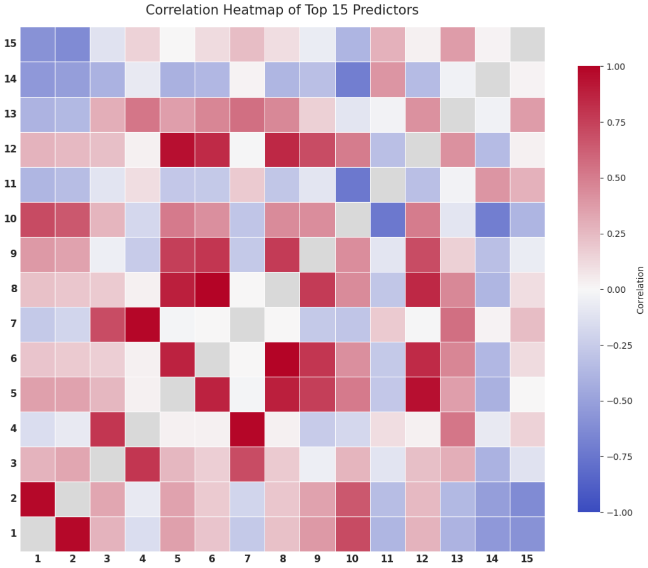
- Complex models (polynomial, NN) don't help much
- Small dataset ($n = 1994$) limits complex models
- For policy: simple + interpretable > tiny accuracy gain

Association vs. Causation

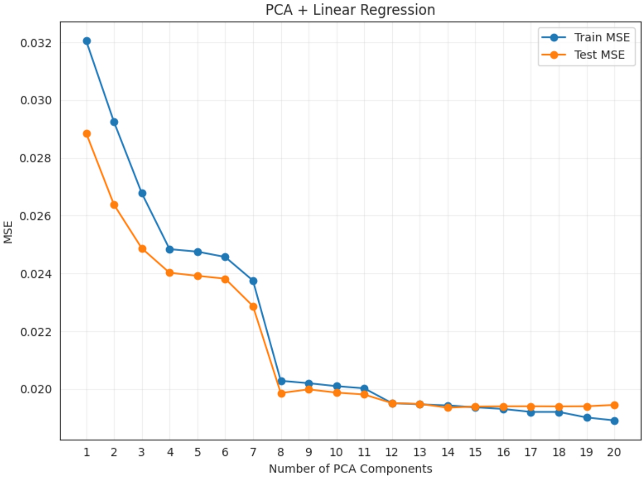
Largest Coefficients

Feature	Coef
PctPersOwnOccup	-0.150
PctHousOwnOccup	+0.124
PersPerOccupHous	+0.099
PctLargHouseOccup	-0.090
MedRent	+0.084

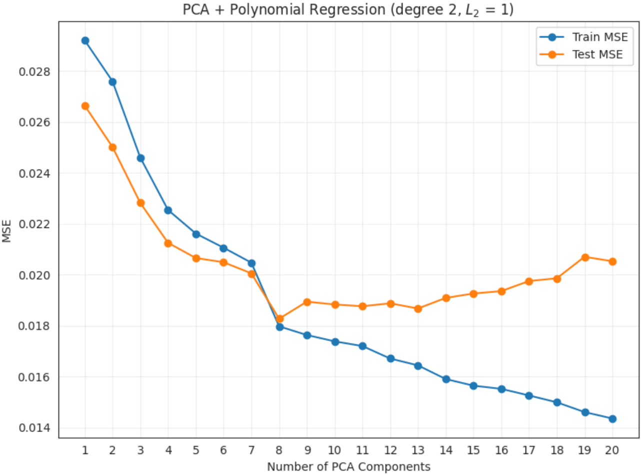
Redundant Predictors



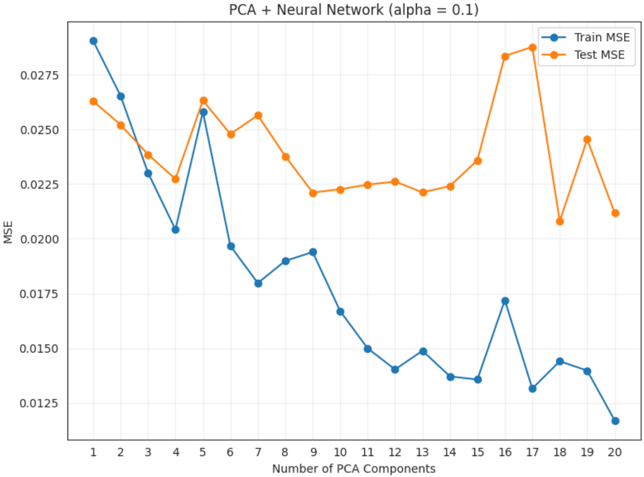
PCA - Linear Regression



PCA - Polynomial Regression



PCA - Neural Network



Limitations & Future Work

Limitations

- Small sample size, data collection issues, normalization issues
- Highly correlated features affects interpretability and true effects of variables
- No spatial info

Future Directions

- Add spatial dependence
- Use census tract data (larger n)
- CNN on spatial crime maps
- Use domain knowledge for feature selection

Thanks! Any Questions?