Math 156: Socioeconomic Predictors of Crime via Classical and Deep Learning Methods

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

March 2026

1 Introduction

Understanding the socioeconomic determinants of crime is a longstanding problem at the intersection of criminology, economics, sociology, and public policy. Numerous studies suggest that exogenous variables such as poverty [8], income inequality [3], unemployment [6], racial or gender composition [4], education levels [5], population density [10], and housing features [9] are associated with crime rates. Others focus on variables endogenous to law enforcement systems, such as crime detection rates and punishment severity [1]. Nonetheless, the relationship between these variables and crime is complex; they could be nonlinear, have spatial dependencies between neighboring regions, and vary across time. Better understanding of these relationships could help with allocating law enforcement resources and engineering effective policies in reducing crime levels.

The goal of this project is to model and predict violent crime rates across U.S. communities using demographic and socioeconomic indicators using selected supervised machine learning methods [2], thereby seeking to interpret determinants in a predictive (rather than strictly causal) sense. Specifically, we aim to compare classical regression-based approaches with neural network models in modeling socially complex phenomena such as crime.

Finally, we compare the predictive performance of the different models. Our results show that the classical linear regression model already achieves strong predictive performance, with only marginal improvements from polynomial regression and slightly worse performance from neural networks due to overfitting. These findings suggest that relatively simple models are sufficient to capture most of the predictive signal in the dataset.

2 Data

Our dataset of interest is the **Communities and Crime** dataset from the UCI Machine Learning Repository [7]. The dataset contains $n = 1994$ observations (”communities”), with 122 predictive features. The problem we aim to solve is a **regression problem**, with the continuous target output variable *ViolentCrimesPerPop*, representing a normalized rate of violent crimes computed from population counts and reported incidents of murder, rape, robbery, and assault. As a sidenote, official reporting of severity of crimes is a controversial economic and criminology issue as well; many communities which had missing values for rape instances were omitted from this dataset because of this.

The Communities and Crime dataset features many socioeconomic predictive features. To highlight some of the main categories:

- Demographic: *population*: population for community (numeric - decimal); *householdsize*: mean people per household (numeric - decimal); *racepct[RACE]*: percentage of population that is [RACE], where [RACE] is ”black”, ”White”, ”Asian” or ”Hisp” (numeric - decimal); *agePct[AGE1]t[AGE2]* percentage of population that is [AGE1] to [AGE2] in age, where ages are either 12-21, 12-29, 16-24, or 65 and up (numeric - decimal); *NumImmig*: total number of people known to be foreign born (numeric - decimal)

- Economic: *medIncome*: median houshold income (numeric - decimal); *perCapInc*: per capita income (numeric - decimal); *[RACE]PerCap*: per capita income for [RACE], where [RACE] is ”black”, ”indian”, ”Asian”, ”white”, ”Hisp”, or ”Other”; *PctPopUnderPov*: percentage of people under the poverty line; *PctUnemployed*: percentage of people 16 and over, in the labor force, and unemployed (numeric - decimal)

- Education: *PctLess9thGrade*: percentage of people 25 and over with less than a 9th grade education (numeric - decimal); *PctNotHSGrad*: percentage of people 25 and over that are not high school graduates (numeric - decimal); *PctBSorMore*: percentage of people 25 and over with a bachelors degree or higher education (numeric - decimal)

- Household: *TotalPctDiv*: percentage of population who are divorced (numeric - decimal); *PctKids2Par*: percentage of kids in family housing with two parents (numeric - decimal)

None of the above features (as well as any we will use) have any missing values and will have minimal to no pre-processing required. However, numeric variables have been normalized to lie within the interval $[0, 1]$, while maintaining their distribution and skew. This does not affect how our models predict crime counts from this dataset, but does affect the interpretability of coefficients in nonlinear models we use.

- **Preprocessing:** Non-predictive identifiers were removed and only numerical predictors were retained. Since the dataset already provides normalized variables in the interval $[0, 1]$, additional scaling was not required. The dataset was randomly divided into training (80%) and testing (20%) subsets in order to evaluate the generalization performance of each model.

3 Methodology

We explore both classical statistical learning approaches and modern deep learning models. Linear regression provides a simple and interpretable baseline, polynomial regression allows modeling nonlinear relationships between predictors, and neural networks provide a flexible framework capable of learning complex interactions between socioeconomic variables. We formulate the task as a supervised regression problem. Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ denote the matrix of socioeconomic predictors (the input feature vectors) and $\mathbf{y} \in \mathbb{R}^n$ denote the target variable *ViolentCrimesPerPop* (the labels), representing the number of violent crimes per 100,000 population. We randomly divide our initial data ($n = 1994$) into training and test sets via an 80% / 20% split. Our objective is to learn a decision algorithm $f : \mathbb{R}^p \to \mathbb{R}$ on the training data that minimizes prediction error on unseen community test data.

Evaluation Metrics Our primary evaluation metric will be the mean-squared error (MSE), defined as

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

for some function f parameterized by $\mathbf{w}$. We will compute MSE on the test data to evaluate performance and on the training data to evaluate generalization.

Feature Selection / Dimensionality Reduction Communities and Crime comes with 122 predictive features and a comparatively low 1994 data points. This can quickly explode model complexity and also lend itself to multicollinearity in linear regression models if not controlled. We control this via two strategies:

- L2-regularization: to stabilize regressions and models, we attach an L2 regularization term to our loss function
$$\mathcal{L}_{\text{MSE}}(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - f_{\mathbf{w}}(\mathbf{x}_i))^2 + \lambda \|\mathbf{w}\|_2^2$$
to eliminate unnecessary complexity and control correlation between predictors. We vary the parameter λ across tests.

- F-test feature selection: we also employ an F-test ranking and select only the top k features most correlated with the target variable before fitting our models. We vary the number k across tests.

4 Polynomial Regression with L2-Regularization

We aim to capture nonlinear predictive relationships for each feature, but for multivariate regression, we also need to capture feature interactions among variables. To keep the model from being over-parameterized, we restrict our model to a degree-2 polynomial featuring cross-terms:

$$\hat{y}(\mathbf{x}; \mathbf{w}) = w_0 + \sum_{j=1}^{p} w_{1,j} x_j + \sum_{j=1}^{p} w_{2,j} x_j^2 + \sum_{j=1}^{p} \sum_{k=j}^{p} w_{jk} x_j x_k,$$

where $\mathbf{W} = [w_0, w_{1,j}, w_{2,j}, w_{jk}]$ is our parameter vector of coefficients. To further control over-parametrization and reduce overfitting, particularly if we have a lot of features, we apply L2 regularization to control model complexity. $\mathbf{w}$ is estimated by minimizing the MSE and we can find the optimal $\mathbf{w}$ through standard linear regression.

4.1 Preliminary: Ordinary Linear Regression

As a first baseline, we perform standard linear regression, equivalent to fixing cross-term and degree-two term coefficients to zero. This is the simplest model and also has the highest interpretability, due to how the variables were normalized. Table 1 shows the performance of the baseline model using all 122 predictors, which achieves an relatively low test MSE of 0.01863 and $R^2 = 0.6307$, indicating the features display a strong linear association with crime counts. Figure 1 plots the predictions against the actual crime counts, showing relatively good accuracy, indicated by the 45 degree line.

Model	Train MSE	Test MSE	R^2
Linear Regression	0.01643	**0.01863**	0.6307

Table 1: Performance of the baseline linear regression model.

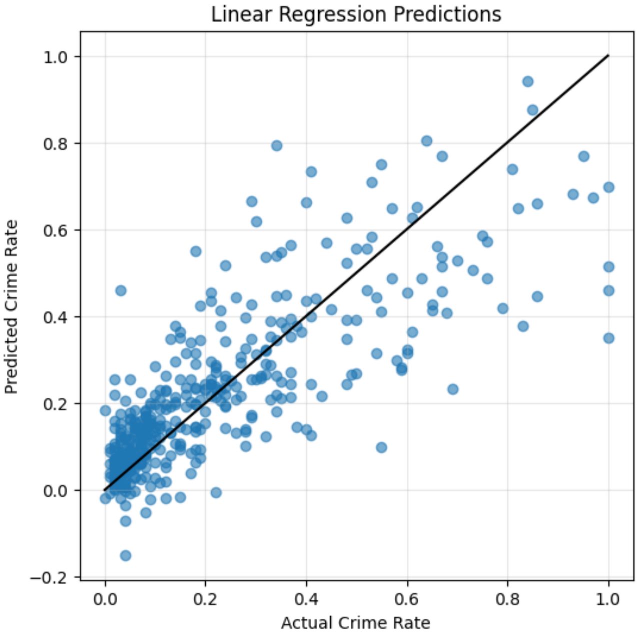

Figure 1: Linear Regression Predictions vs. Actual.

Table 2 shows some of the top predictors in the linear model, ranked by absolute coefficient magnitude. While again, these coefficients should be taken with a grain of salt due to multicollinearity, from the top five coefficients alone, it is clear housing and domestic features play a large part in predicting crime. For example, a larger percent of people in owner occupied homes (*PctPersOwnOccup*) lead to a lower crime prediction, while increases in the median rent lead to a higher crime count prediction.

Feature	Coefficient
PctPersOwnOccup	-0.1504
PctHousOwnOcc	0.1244
PersPerOccupHous	0.0991
PctLargHouseOccup	-0.0901
MedRent	0.0842
⋮	⋮

Table 2: Top predictors in the linear regression model ranked by absolute coefficient magnitude.

Using the F-test to select features most correlated with crime counts, we also test various numbers of predictors in Figure 2. The model's predictive performance on unseen data stabilizes past around $50 - 60$ top predictors. While this is relatively crude to control for multicollinearity compared to VIF or correlation matrix based tests and offers very little causal substance, these results provide a good comparison in terms of predictive power for our more complex models later.

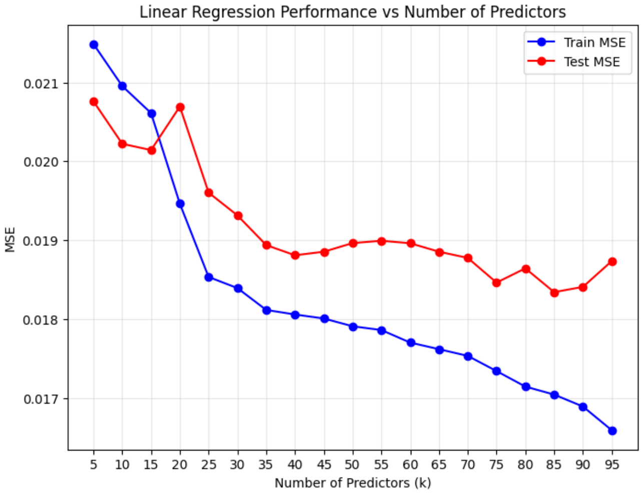

Figure 2: Linear Regression Performance using top k predictors.

4.2 Polynomial Regression with L2-Regularization

The baseline linear regression model demonstrates strong predictive power, but we now go back to our full polynomial regression model to see if we can increase the predictive power by enabling the model to learn potentially nonlinear relationships between socioeconomic phenomena and crime.

However, using even a degree-two polynomial dramatically increases the number of parameters in the model. For p predictors, the number of polynomial features grows on the order of $O(p^2)$, which can easily lead to overfitting, particularly given the moderate sample size of $n = 1994$ observations. To mitigate this, we incorporate L2-regularization (ridge regression) with level λ, which penalizes large coefficient magnitudes and stabilizes estimation in the presence of correlated predictors. We also again use F-test feature selection to retain the k predictors most correlated with the target variable.

Figure 3 summarizes the training and test mean-squared errors as a function of the number of selected predictors k for several values of the regularization parameter λ. Each subplot corresponds to a different regularization strength, from $\lambda = 0, 1, 5, 10$.

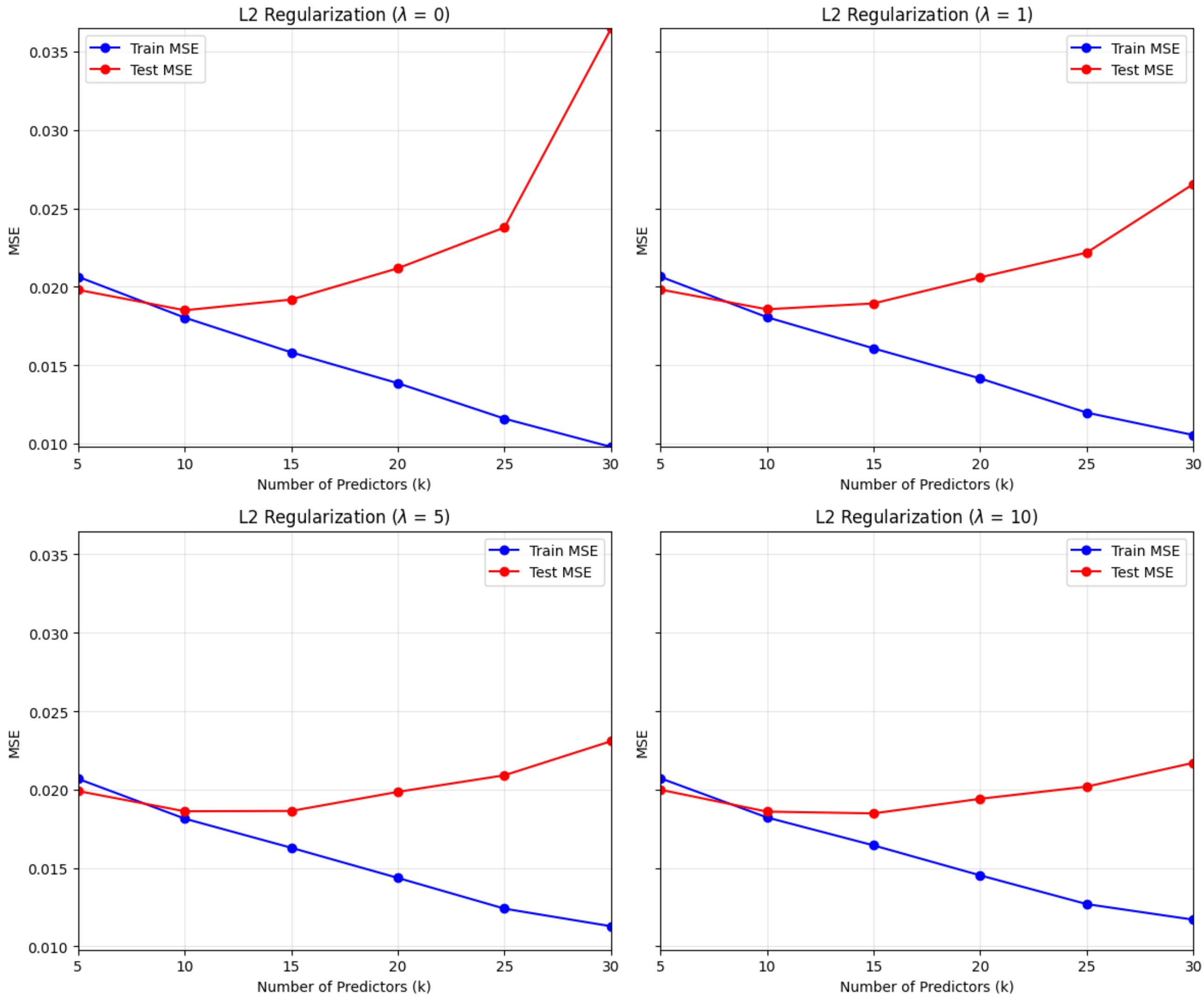

Figure 3: Polynomial regression performance as a function of the number of predictors k under different L2-regularization strengths.

We make some observations. First, increasing the number of predictors generally decreases the training error, while the test error explodes for large values of k, indicating overfitting as model complexity grows. Second, increasing the level of L2 regularization generally helped decrease the amount of overfitting for large values of k, but not much. Using top k selection, we find that the optimal number of predictors to use is around 10. Among this range, the effect of L2 regularization is similar; all models perform relatively equivalently on the test data.

Table 3 reports the best-performing polynomial regression models. Overall, the degree-2 polynomial regression model yields marginal improvements over the baseline linear regression (test MSE: 0.01863) in certain configurations, particularly when select features are used and moderate regularization is applied. However, the gains are relatively small, suggesting that a large portion of the predictive structure in the dataset may already be captured by linear relationships. Figure 4 displays the predictions of the polynomial regression ($\lambda = 0, k = 10$) against the true crime counts.

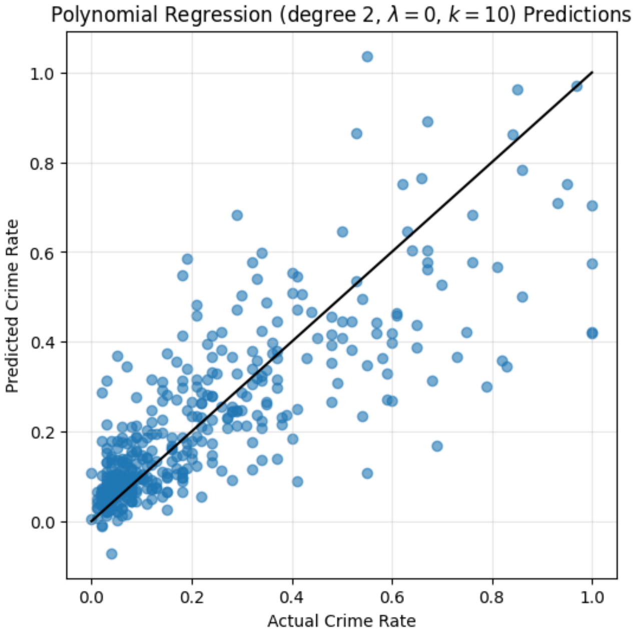

Figure 4: Polynomial Regressions Predictions vs. Actual.

Number of Predictors (k)	λ	Train MSE	Test MSE
10	0	0.01802	**0.01850**
10	1	0.01806	**0.01856**

Table 3: Selected polynomial regression configurations and their predictive performance.

5 Feedforward Neural Network

While our results from polynomial regression hinted that the linear model already captured most predictive relationships, we again try to capture potentially more complex nonlinear relationships between socioeconomic variables and crime rates by implementing a fully connected feedforward neural network (FFNN). Neural networks are capable of approximating highly nonlinear functions and learning complex interactions between predictors without explicitly constructing polynomial features.

The neural network architecture used in this study consists of:

- An input layer of dimension p (equal to the number of predictors).
- Two hidden layers with ReLU activation functions.
- A single linear output neuron to produce the predicted crime rate.

The network parameters $\mathbf{w} = [\mathbf{W}_1, \mathbf{W}_2, (\mathbf{W}_3)]$ are optimized by using backpropagation (to compute gradients) and stochastic gradient descent (Adam optimizer) on the mean-squared loss with L2 regularization. To keep the model from being too complex, our baseline neural network architecture contains two hidden layers with 32 and 16 neurons respectively. ReLU activation functions are used in each hidden layer due to their strong empirical performance in deep learning models. The final output layer is linear to accommodate the regression setting.

Training Procedure The network is trained for up to 500 iterations with an early stopping rule. As with previous experiments, we evaluate performance using mean-squared error on both the training and test datasets for various regularization strengths ($\lambda = 0, 0.001, 0.01, 0.1$) and top k features. Figure 5 displays the results of this experiment. We can see that similar to the polynomial regressions, the neural network experiences large overfitting as the number of predictors increases, due to the increasing complexity of the model. L2 regularization with strength $\lambda = 0.1$ seems to stabilize the neural network the best, achieving the lowest observed test error across all models. Table 4 shows the network configuration that achieves the best performance on the test data with a MSE of 0.01912 ($k = 10$, $\lambda = 0.1$), and Figure 6 shows the predictions by this model plotted against the actual crime count values.

Model	Number of Predictors (k)	λ	Train MSE	Test MSE
Feedforward Neural Network	10	0.1	0.01886	**0.01912**

Table 4: Performance of the highest performing FFNN.

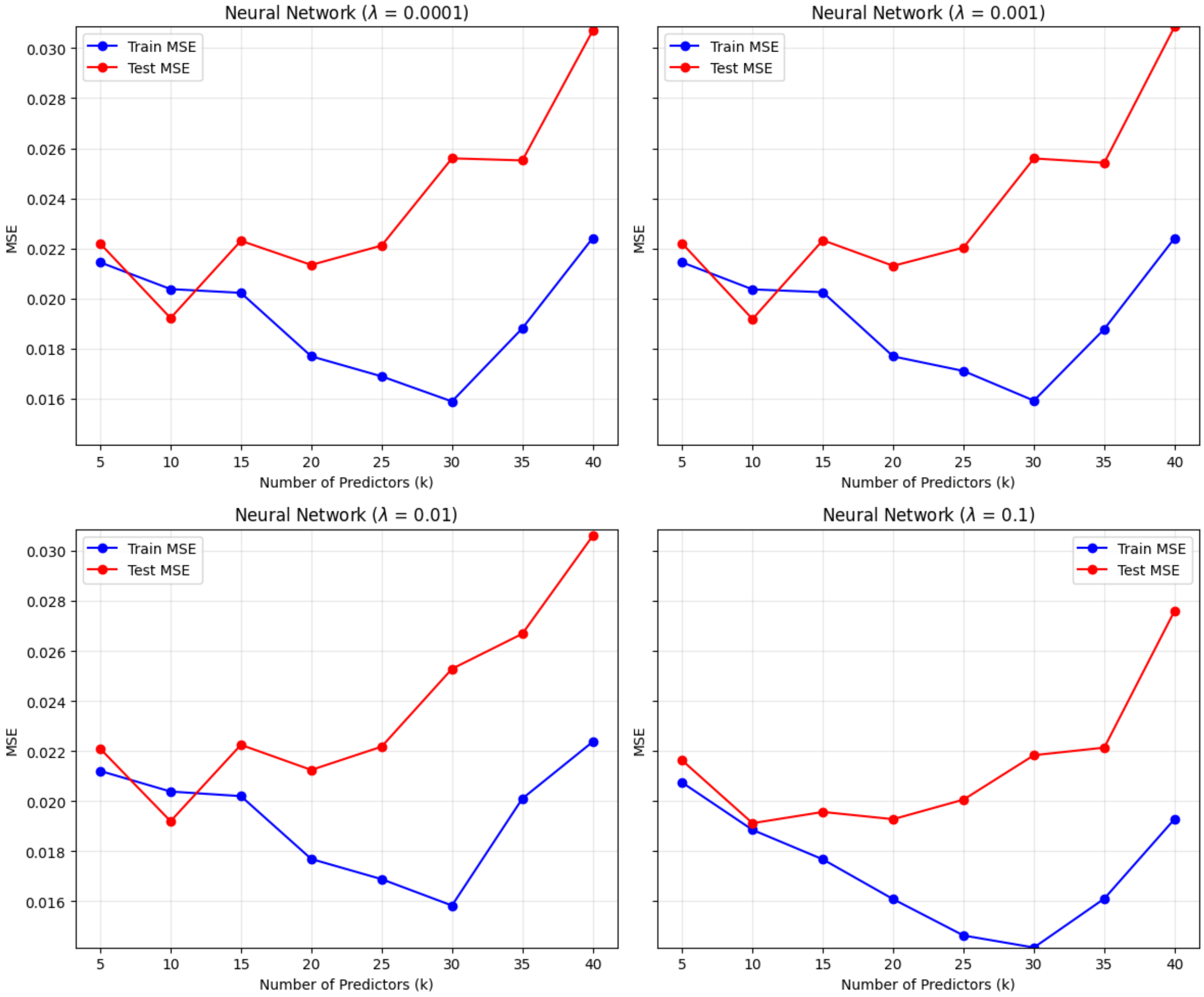

Figure 5: FFNN performance as a function of the number of predictors k under different L2-regularization strengths.

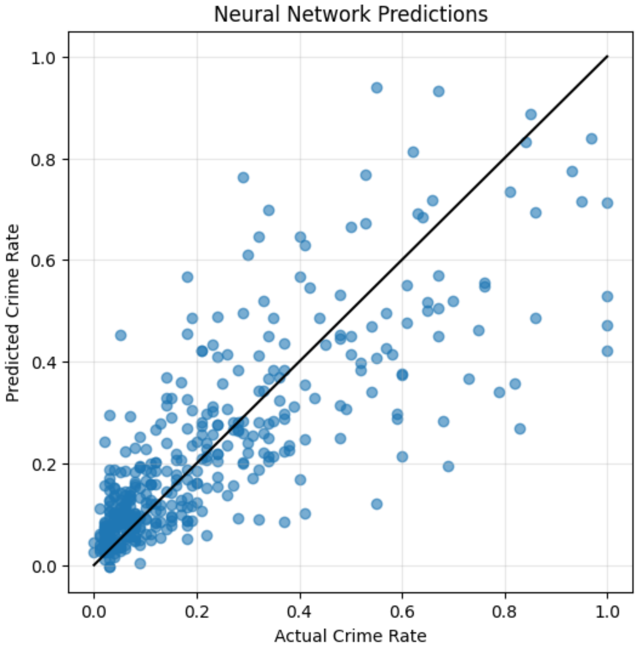

Figure 6: Predictions vs actual for the Feedforward Neural Network.

Overall, we see that the fully-connected feedforward neural network performs comparably to, and in most configurations *slightly worse than*, the classical regression models. This is likely due to the complexity of the neural network framework, which tends to overfit our small dataset and not generalize well.

6 Final Approach and Results

Model	Predictors (k)	Regularization (λ)	Train MSE	Test MSE
Linear Regression	122	0	0.01643	0.01863
Polynomial Regression (deg = 2)	10	0	0.01802	**0.01850**
Polynomial Regression (deg = 2)	10	1	0.01806	0.01856
Feedforward Neural Network	10	0.1	0.01886	0.01912

Table 5: Comparison of predictive performance across all models tested.

Table 5 summarizes the predictive performance of all models considered in this project. 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. The small sample size of the dataset limits the effectiveness of these deep learning models; classical regression techniques can perform competitively when the number of observations is moderate and predictors are selected appropriately. Overall, these results suggest that much of the predictive structure in the dataset can already be captured by relatively linear relationships, and we endorse the classical linear regression approach for this problem for its simplicity, performance, and overall interpretability.

7 Interpretability

7.1 Association vs. Causation

At face value, it seems as though our models identify multiple socioeconomic factors that are seemingly strongly associated with crime; however, this should not be viewed as causal. It is important to note that the dataset is observational, meaning that there was no deliberate experimental control and it is obviously apparent by inspection of the dataset that many factors are heavily correlated. Looking at these variables, they span domains such as income, poverty, education, housing and family structure which are often heavily linked due to underlying causes. Thus, even if one variable seemingly displays a strong predictive relationship with crime, it cannot be concluded that the variable is independently causal. Instead of treating our models as an analysis of what factors cause crime, rather, it should be perceived purely as a predictive tool to gauge crime based on a variety of input factors.

For example, our earlier top-k feature analysis identified home ownership related variables as strong predictors. Rather than suggesting that home ownership itself determines crime rates, this more likely indicates that such features act as proxies for broader neighborhood characteristics correlated with crime, like income or education. Our models should therefore be understood as learning patterns that are useful for forecasting crime rates across communities, as opposed to identifying the true causal determinants of crime.

7.2 Redundant Predictors

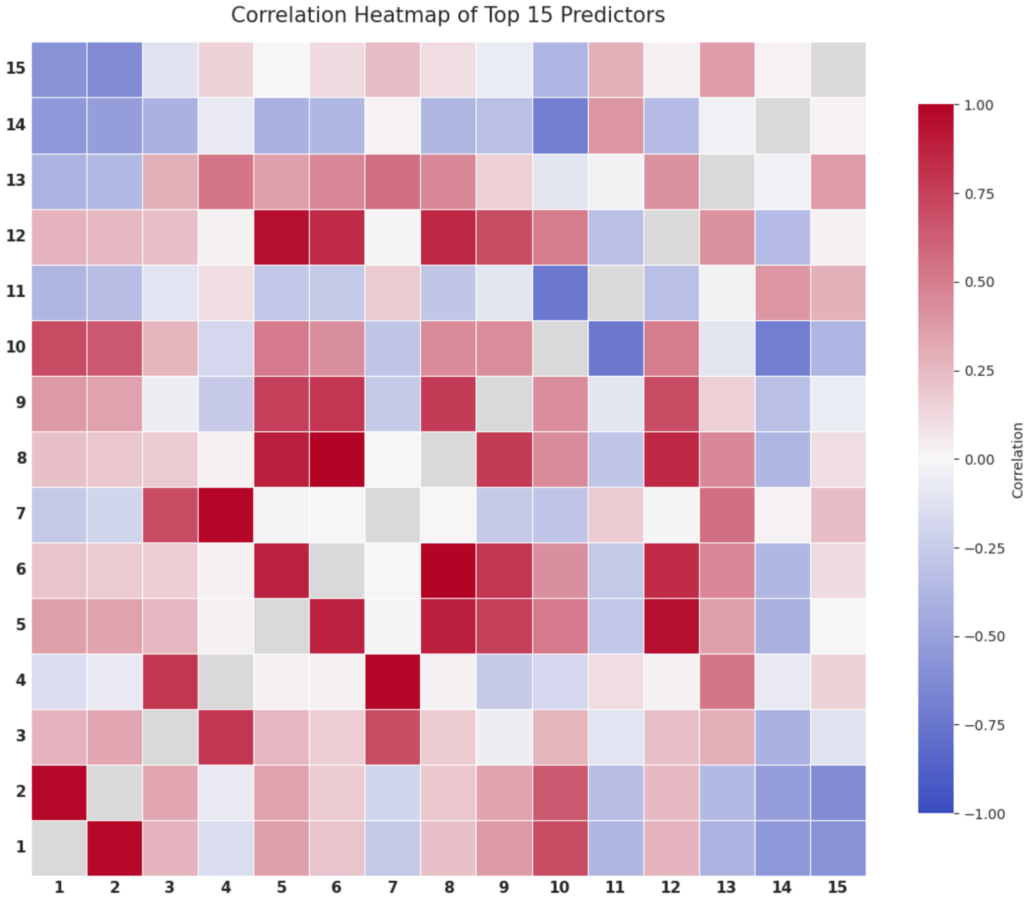

Figure 7: Correlation heatmap of the top 15 predictors, showing substantial redundancy among the most important features.

The diagram above depicts that, even when filtering all the input variables to the top 15 predictors, there is still heavy correlation between many of them, and that the variables identified as important by our models are not acting independently. However, this is to be expected if one looks at the dataset, as variables that span socioeconomic, demographic, housing, and family conditions are naturally correlated. In fact, there are multiple pairs of variables within the top 15 predictors that have near perfect correlation.

This redundancy in variables is important to point out as, along with the small size of our dataset, is a key contributing factor to why our model's performance does not seem particularly strong - one would anticipate that with so many input features, the model would be highly accurate. However, as the heatmap shows, some predictors may as well be interchangeable, so the true amount of information within the input variables is a lot less than it seems.

7.3 Dimensionality Reduction - PCA

Broadly, our takeaway from performing PCA suggests that a big portion of the predictive signal in the dataset is concentrated in a minority of latent dimensions. Generally, for all three models, the test error falls off sharply as we go from 1 up to around 8-12 components, after which the gains become minimal if not zero. Our linear regression model shows this pattern very clearly, where the MSE (test) rapidly decreases and then levels off after 8-12 components, with very minimal improvements after. This means that the primary predictive structure can still be extracted despite heavy compression of the original dataset. In context, the raw socioeconomic predictors are rather redundant, and PCA can preserve the most vital contents with far fewer variables than our input data.

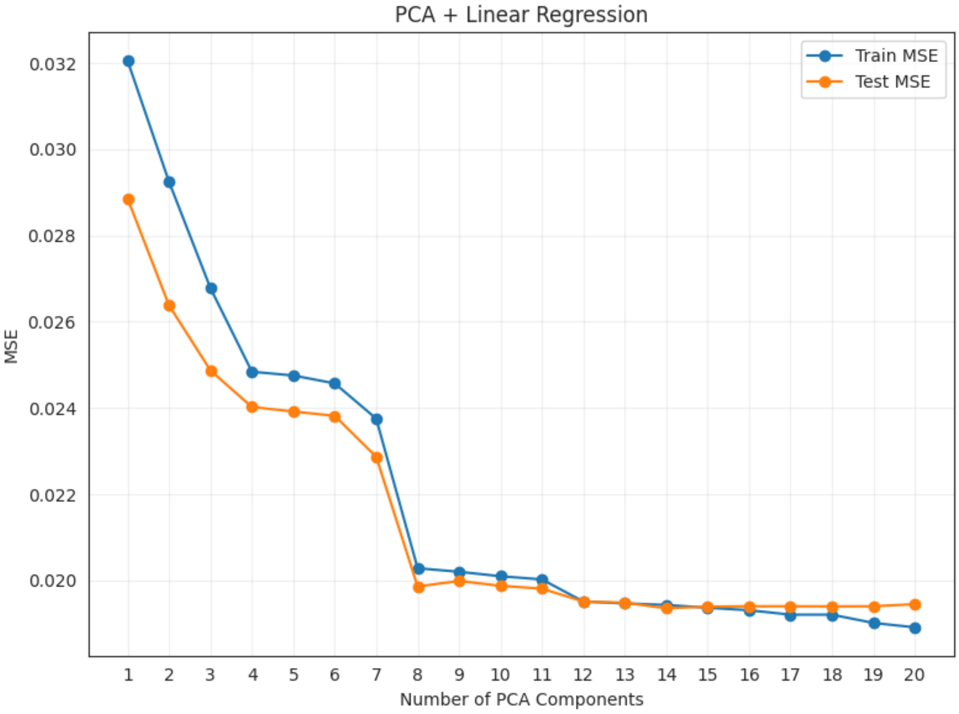

Figure 8: PCA + linear regression. Train and test MSE are shown as functions of the number of retained principal components.

Between the three models, linear regression has by far the most smooth and stable result after applying PCA. The test MSE steadily decreases when adding more components up until around 8-12 components, at which point the curve flattens and there are no additional benefits to adding additional components. This suggests that the broadest principal directions already contain most of the predictive signal that a linear model can take advantage of, and additional components barely contribute.

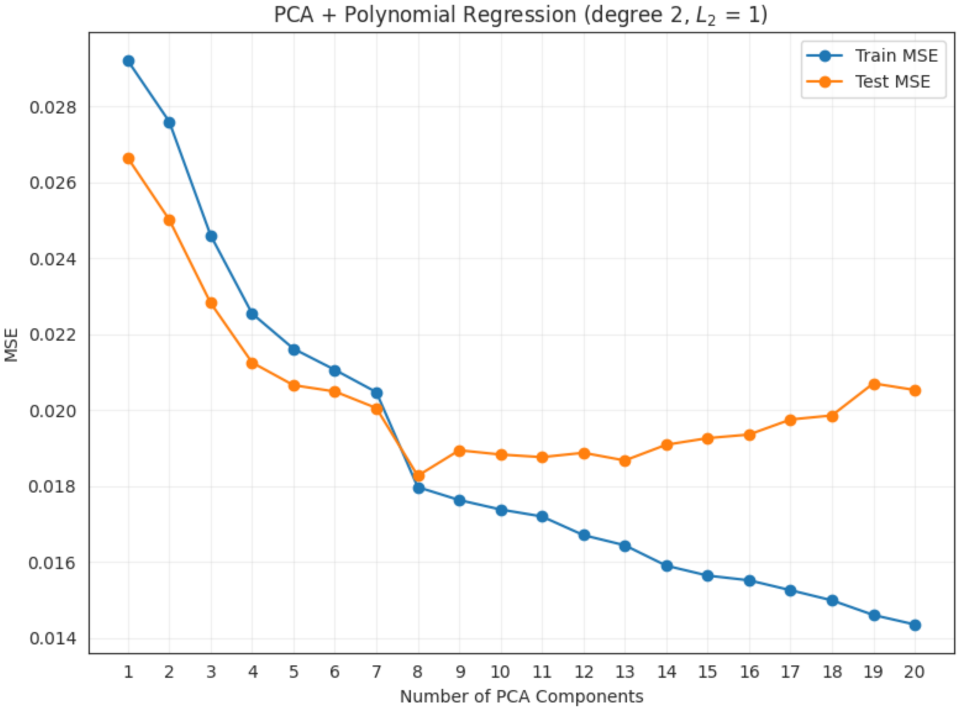

Figure 9: PCA + polynomial regression with degree 2 and $L_2 = 1$. Train and test MSE are shown as functions of the number of retained principal components.

Polynomial regression shows a very similar result to linear regression, although it is more pronounced. Similar to linear regression, polynomial regression benefits greatly up to 8 components, but afterwards, unlike linear regression, performance drops with additional components. This is an important result, since it signifies that the quadratic model is able to exploit components up to a certain point effectively, but afterwards starts underperforming as potentially noisier directions are added.

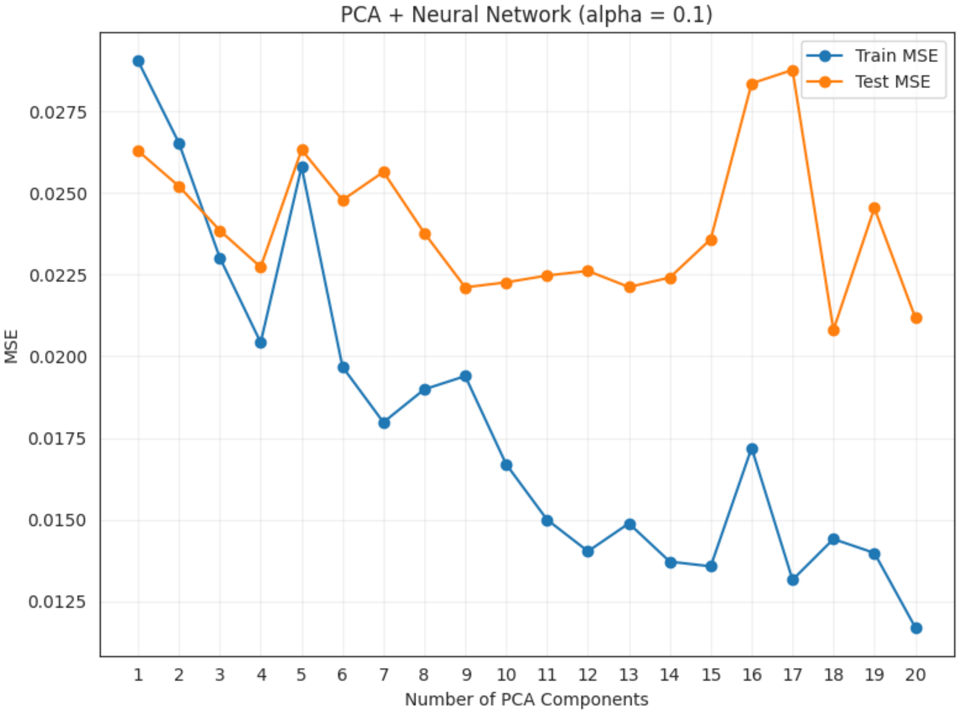

Figure 10: PCA + neural network with $\alpha = 0.1$. Train and test MSE are shown as functions of the number of retained principal components.

Notably, the neural network behaves very differently than the other two simpler models. Its test MSE improves quickly for the first few dimensions added, but the overall curve has significantly more variance and is less smooth. Performance seems to drop noticeably between 4-8 dimensions, then improves quite a bit, retaining good results between 9 and 14 components, after which test error once again increases and decreases seemingly sporadically. In the neural network, we also markedly see a widening gap between training and test result which signifies that our neural network is more sensitive to the additional noisy components.

All together, the three graphs lead us to conclude that crime prediction as a task is largely driven by a relatively small number of broad, latent socioeconomic factors rather than the large full set of predictors independently that we saw in the dataset. Overall, it is also a significant result that both linear regression and polynomial regression are able to reach near optimal results with merely 8-12 principal components. This suggests that not only is the input data able to be compressed significantly while still retaining predictive power, it is also the fact that a simple linear model can show good results, which the predictive signals are strong enough to be learned by weaker models. Finally, the polynomial regression model along with the neural network showed that too many components can also be detrimental, as performance drops seemingly due to overfitting / noise being learned. To conclude, PCA reaffirms our suspicion that most of the predictive power lies within a few principal components.

8 Conclusion and Discussion

In this project, we examined the predictive relationship between socioeconomic indicators and violent crime rates across U.S. communities using several supervised learning approaches. We started with a baseline OLS model and explored increasingly flexible models including polynomial regression with L2 regularization and a feedforward neural network. The baseline linear regression model already demonstrated strong predictive performance, achieving a test MSE of 0.01863 and an R^2 value of approximately 0.63. These results already suggest that a substantial portion of the predictive signal for crime counts can be explained by linear relationships. The difference in performance over the linear baseline was relatively small for both polynomial regression and feedforward neural networks. Even with regularization and feature selection, the best polynomial models only marginally improved test MSE. This suggests that while nonlinear interactions exist, they do not dramatically alter predictive accuracy relative to the linear model.

We acknowledge several limitations. First, the dataset contains only 1994 observations (with some omitted variables and data points), which restricts the complexity of models that can be reliably estimated without overfitting. Second, the predictors are highly correlated with one another, which complicates interpretation of individual coefficients. Third, the dataset has been normalized to the interval $[0, 1]$, which preserves distributions within features but reduces the interpretability of nonlinear interaction terms.

Future work could extend this analysis in several directions. One promising avenue would be to incorporate spatial information into the modeling process. Crime rates often exhibit geographic dependence, and incorporating spatial correlations could improve predictive accuracy. One potential extension would be to analyze more granular datasets, such as census tract or neighborhood level crime data. Larger datasets would allow models such as neural networks to be trained more effectively. In particular, convolutional neural networks could be applied to spatially structured crime data to learn crime count trends across geographic regions. Finally, future work could draw on previous econometric and causal analyses on crime to identify variables that are strongly associated with crime rates, which could help reduce the dimensionality of the features (alternatively to F-tests or PCA) while preserving predictive power.

9 Credits

- Kenny: Implementation of OLS, polynomial regression, and FFNN. Produced figures and wrote report.
- Jinghui: Data acquisition and data processing. Wrote proposal, report, and slides.
- Kievan: Contributed to model analysis, interpretation of results, and report writing.
- Rishauv: Interpretability section (code, report/analysis, diagrams, presentation)

References

[1] G. S. Becker. Crime and punishment: An economic approach. *Journal of political economy*, 76(2):169–217, 1968.

[2] C. M. Bishop and N. M. Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[3] J. Choe. Income inequality and crime in the united states. *Economics Letters*, 101(1):31–33, 2008.

[4] P. Leighton, G. Barak, A. Cotton, C. L. Buist, and K. S. León. *Class, race, gender, and crime: The social realities of justice in America*. Bloomsbury Publishing PLC, 2024.

[5] L. Lochner. Education and crime. In *The economics of education*, pages 109–117. Elsevier, 2020.

[6] S. Raphael and R. Winter-Ebmer. Identifying the effect of unemployment on crime. *The journal of law and economics*, 44(1):259–283, 2001.

[7] M. Redmond. Communities and Crime. UCI Machine Learning Repository, 2002. DOI: https://doi.org/10.24432/C53W3X.

[8] P. Sharkey, M. Besbris, and M. Friedson. Poverty and crime. 2016.

[9] M. S. Tillyer and R. J. Walter. Low-income housing and crime: The influence of housing development and neighborhood characteristics. *Crime & Delinquency*, 65(7):969–993, 2019.

[10] R. E. Watts. The influence of population density on crime. *Journal of the American Statistical Association*, 26(173):11–20, 1931.