Crime in Communities:

Predicting Violent Crime by County

Jacob Titcomb

Spring 2024

Contents

Introduction

Motivations

Crime, especially violent crime, holds a unique position within the American psyche. Fear mongering and finger pointing can obfuscate the true problems and solutions to better manage the issue of crime. It is important to first consider *where* crime is happening and *how* crime can be addressed meaningfully. This project aims to predict violent crime rates based off demographic data, with the goal of locating where violent crime is more prevalent and determining which geographic and socioeconomic factors can be considered to minimize perpetuated violence. These models could assist policy makers and advisers, informing future policies and government spending.

According to the Pew Research Center, the critical perception of crime has increased recently, with 58% of Americans considering crime reduction as a top political priority, compared to the 47% at the beginning of 2021 (Gramlich, 2024). Yet crime as a whole has gone down substantially from the early 1990s, when there was a spike across most crime categories, violent included. Violent crime is a category by the Federal Bureau of Investigation (FBI) which is comprised of robberies, aggravated assault, murder/nonnegligent manslaughter, and rape ("Violent Crime", 2019). Based on FBI data, rates for all but the latter have decreased between 1993 and 2022—rape is excluded because how the offence was counted changed in 2013 (Gramlich, 2021). With regards to murder rates, the U.S. has seen a downward trend since the 1970s (Gramlich, 2024); whether that is an effect of "tough on crime" policies, increased community resources, and/or other social phenomena is difficult to determine. But by and large, Americans are concerned for public safety and have an interest in the de-escalation of crime in general.

In the pursuit of better policing, Artificial intelligence (A.I.) has already been implemented to assist law enforcement in many forms, chief among them facial recognition technology and "predictive policing," which anticipates future criminal activity ("Predictive Policing Explained", 2020). Implicit and explicit biases exist within these existing technologies, and addressing them is beyond the scope of this project, but are worth considering ("Artificial Intelligence in Predictive Policing Issue Brief", 2024). Some believe that A.I. is capable of contributing to community safety, but that goal faces two particularly large hurdles: the issues of equity and racial bias. Equitable policing by humans is already difficult and polarizing; with the current state of A.I. still being in development, there is room to improve before further reliance on A.I. is reasonable. Racial bias in A.I. is a prevalent issue (Ferrer et al., 2021), and implementation in policing could exacerbate existing problems in law enforcement today. Compared to those other A.I. models, the use of machine learning in this project serves as a policy-informing tool for communities, rather than a tool to measure an individuals propensity for crime.

Determining the number of violent crimes has broad social and political implications. Decreasing crime generally increases overall community safety; however, improper policing, arrests, and convictions could damage communities, particularly communities of color and marginalized groups. Law enforcement is one of the few facets through which the government directly interacts with the people it represents. Managing and de-escalating crime expends resources, a difficult issue to address efficiently and effectively for governments. While crime has been on the decline, government at all levels still contends with the issue of public safety, so determining the rate of crime and factors associated with violent crime would inform decision making. Findings could influence the allocation of funds and political effort towards policing, community support programs, and other programs. Of course, all these policies come with a price, both social and economic.

Both violent crimes and the means of addressing them have economic effects. From the crime itself, an individual or property are harmed, resulting in a loss of value of future economic contribution (particularly significant in the case of death). Their families are also affected, having to deal with the material and emotional damage caused by the crime, leading to loss of productivity and a potentially negative economic shock to those close to the victim(s) ("Violence & Socioeconomic Status"). Yet similar conclusions can be said to those close to the perpetrator—these factors include the emotional and economic toll of the perpetrator being imprisoned, namely the loss of income, legal fees, and lost productivity. The impact of false convictions are even more severe: the same effects inflicted upon someone innocent. Thus the issue of violent crime has widespread economic implications, from the personal level to the federal level.

The aim of this project is to predict the number of violent crimes per 100,000 based off demographic data for the year 1995. The data for this project was sourced from the 1990 U.S. Census and the 1995 FBI Uniform Crime Report (UCR), via the UC Irvine Machine Learning Repository (Redmond, 2011). Each observation represents a different county, with the variables being demographic information for each county. Some counties—mostly in the Midwest—noted issues with counting rapes; those counties were excluded from this project.

The Features

As mentioned earlier, the target variable we aim to predict is the number of violent crimes per 100,000, which we transformed with a log transformation. For the predictors, there were 124 numeric variables and 1 categorical (the county's state). There were 22 columns with over 80% of the values missing, likely due to issues with how the variable was recorded at the county-level. Those 22 columns were therefore removed, leaving 103 predictors.

For ease of use, we one-hot encoded the categorical variable (`State`), creating an indicator variable for each state.

Now with a dataset with 152 predictors, we wanted to reduce the number of predictors. In order to do so, we performed principal component analysis (PCA) and found that 70 principal components account for over 95% of the variance. Then among those 70 principal components, we found that 94 features contribute to over 80% of the summed absolute loadings, so those 94 features will be the variables we keep in the model. Of the 94 features, 49 are numeric, and the other 45 are categories.

The final change to the features was that we scaled and centered the numeric variables, such that they have mean 0 and a standard deviation of 1.

Methods

In this project, we constructed the 10 following models:

1. Ordinary least squares (OLS)
2. Least Absolute Shrinkage and Selection Operator (LASSO) regression
3. Ridge regression
4. Elastic net regression
5. Principal component regression (PCR)
6. Piece-wise polynomial
7. Multivariate Adaptive Regression Splines (MARS)
8. Generalized Additive Model (GAM)
9. Gaussian Process Regression (GPR)
10. Bayesian ridge regression

When training the models, we used 10-fold cross validation, repeated 5 times. To assess out-of-sample performance, we also constructed an 80-20 train-test split. Our performance metrics will be the RMSE, R-squared, and MSE.

In addition, we fitted a poisson regression with the raw (rather than logged) target variable, but it performed so poorly that it was not worth including in the model.

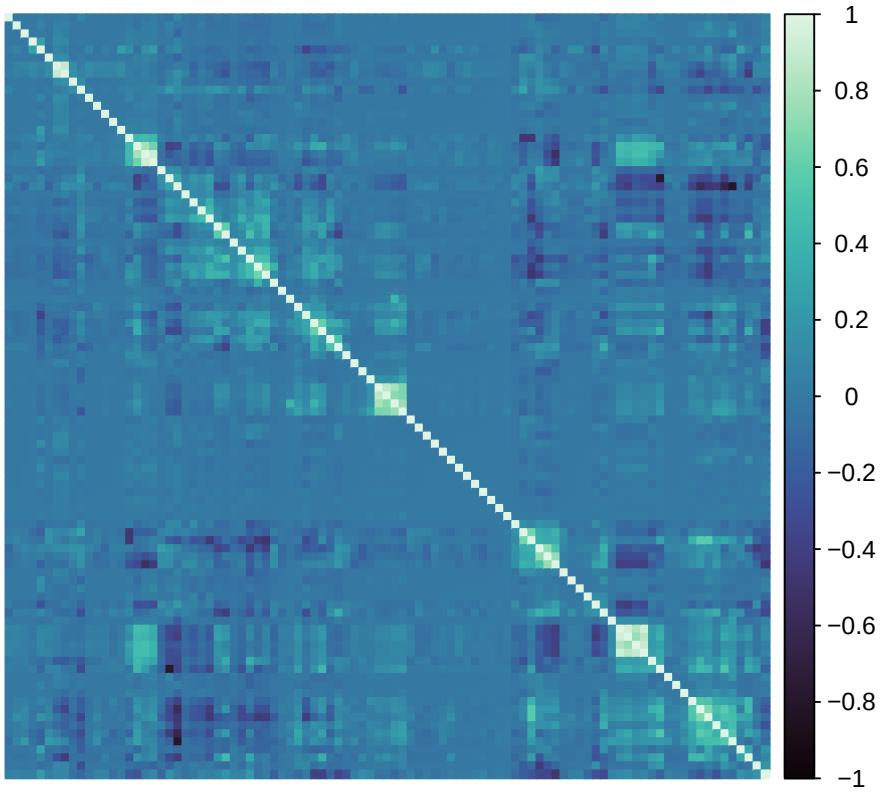
For the spline regression, we fitted both a traditional spline model and a MARS model. Since the MARS model outperformed the traditional spline by all measures, we opted to include the MARS model as representing the general classs of "spline regressions."

Exploratory Data Analysis

Since this data has many predictors (roughly 150), we will not be granting them an extensive preliminary study. We did observe that most of the features relating to economic information (e.g., median house price, mean wages) were very right-skewed, as is to be expected from socioeconomic statistics.

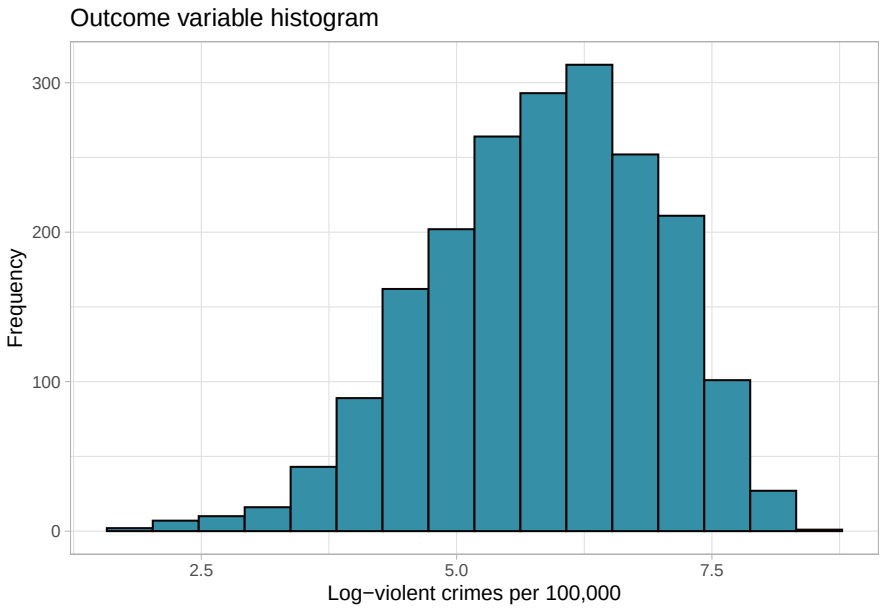
When we performed an initial dimension reduction using PCA, we found that over 95% of the variance was captured in just 70 principal components. Of the around 90 we selected based off of PCA, about half (45) were one-hot encoded states while the other 49 were continuous variables.

In the correlation plot below, there are many swaths of features with practically no correlation. Most of those pairings are states with other states, which is to be expected. There are some areas of higher positive correlation, mostly the socioeconomic measures. Features which constitute percentages (e.g., percent white or black in a county) show some negative correlation, as the percentages have a ceiling of 100 to their sum.



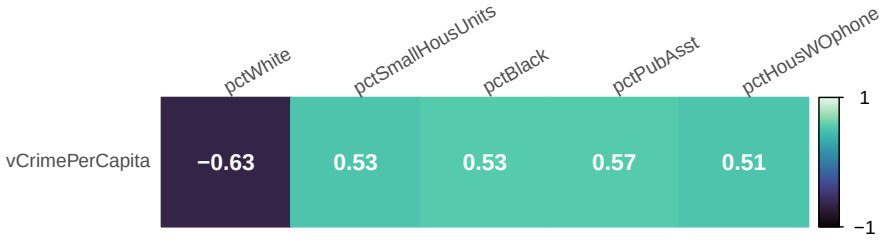
The Target Variable

As mentioned, the outcome variable of study is the number of violent crimes per capita, log transformed. The variable has a mean of 5.853, a standard deviation of 1.108497, and a range of around 7. We can see the distribution in the histogram below.



With a p-value of 0.03069, the above distribution is *not* approximately normal, based on a Kolmogorov-Smirnov test.

Below, we have the features with the highest correlation with the target variable. Of course, it is worth noting that these only indicate some linear relationship; non-linear effects are not represented, so we cannot extract any conclusions thusfar.



Model Fitting

Model 1: Ordinary least squares

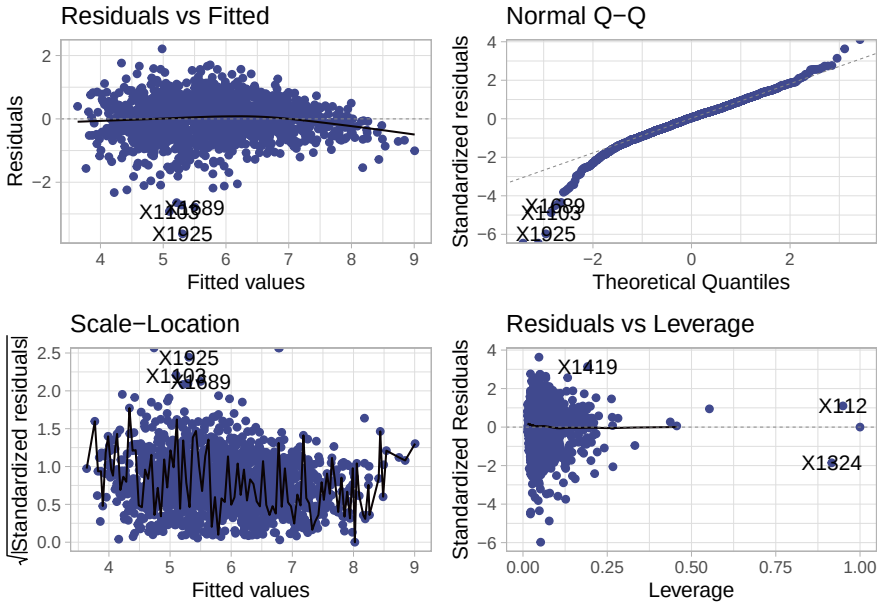
First we have the performance measures for a standard multiple linear regression.

CV.RMSE	RMSE	Rsquared	MAE
0.6627	0.7247	0.5582	0.5211

And the 6 features with the highest importance according to the model.

Variable	Importance
pctUrban	100.00000
pctOfficDrugUnit	94.53754
pctSmallHousUnits	90.87587
pctHousOccup	88.48184
pctPubAsst	78.57790
pctEmployMfg	73.25859

Lastly, we look at the residual plots below. There appears to be a pattern in the residual plot and a decrease in the scale-location plot, both indicating non-constant variance. Thus this model is not valid.

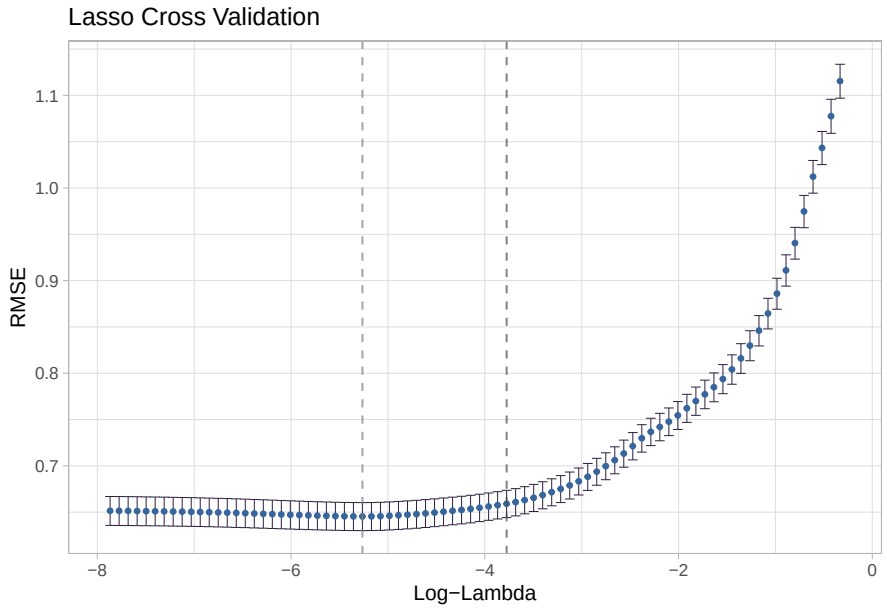


Model 2: LASSO

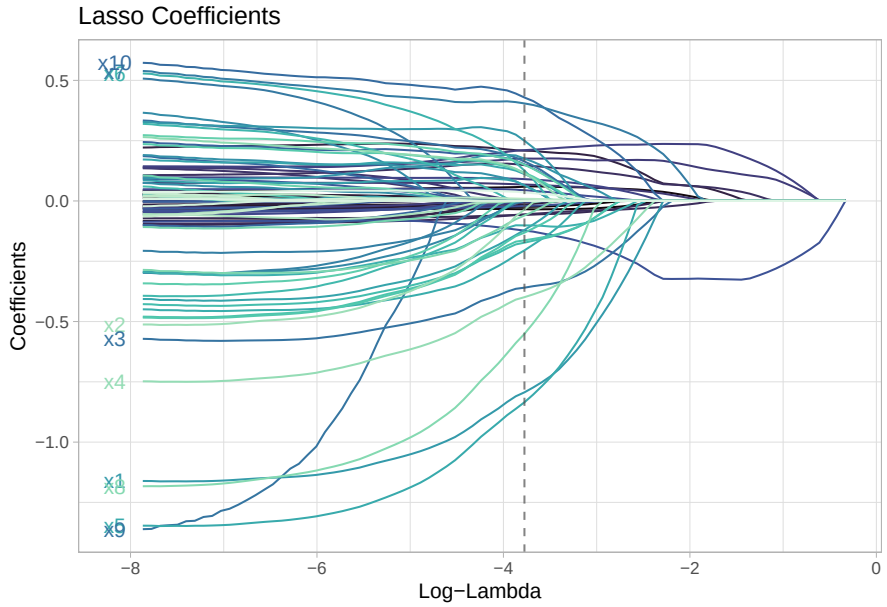
For the LASSO model, we performed repeated 10-fold cross validation manually, since the original `train()` function in `caret` was having difficulty with the high dimensionality of the data.

Using cross validation, we found the optimal regularization parameter to be $\lambda = 0.0229433$. We are using the highest λ within one standard deviation of the minimizing λ so as to favor parsimony in the model. The final number of features included (not reduced to 0) was 48.

Below is the cross validation showing how the regularization parameter was chosen.



And the below plot shows the survival of the coefficients under LASSO.



Below are the performance measures.

CV.RMSE	RMSE	Rsquared	MAE
0.659	0.7343	0.5403	0.5351

Lastly, here are the 6 features with the highest importance in the model.

Variable	Importance	Sign
StateND	100.00000	NEG
StateME	95.01640	NEG

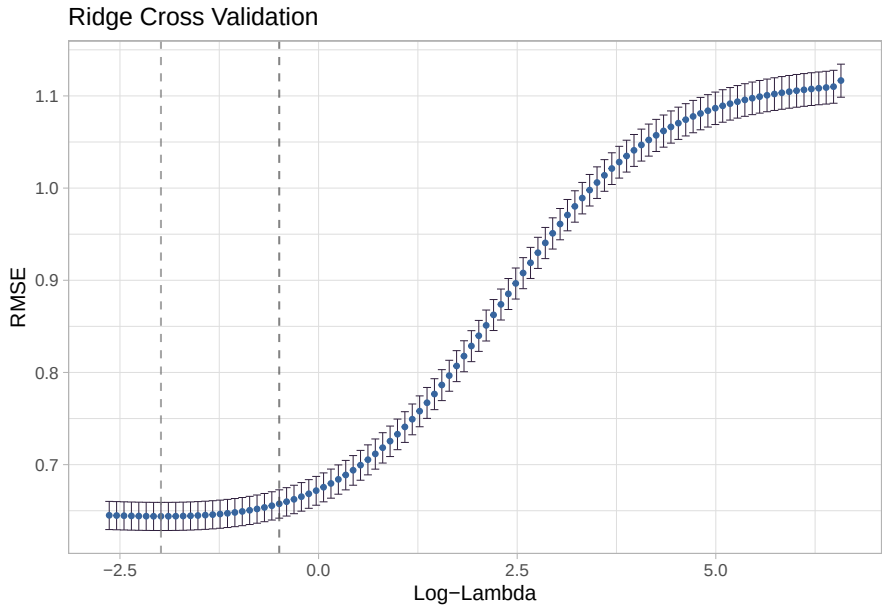
Variable	Importance	Sign
StateVT	65.97737	NEG
StateCA	51.79100	POS
StateFL	48.56833	POS
StateWI	47.70682	NEG

Model 3: Ridge

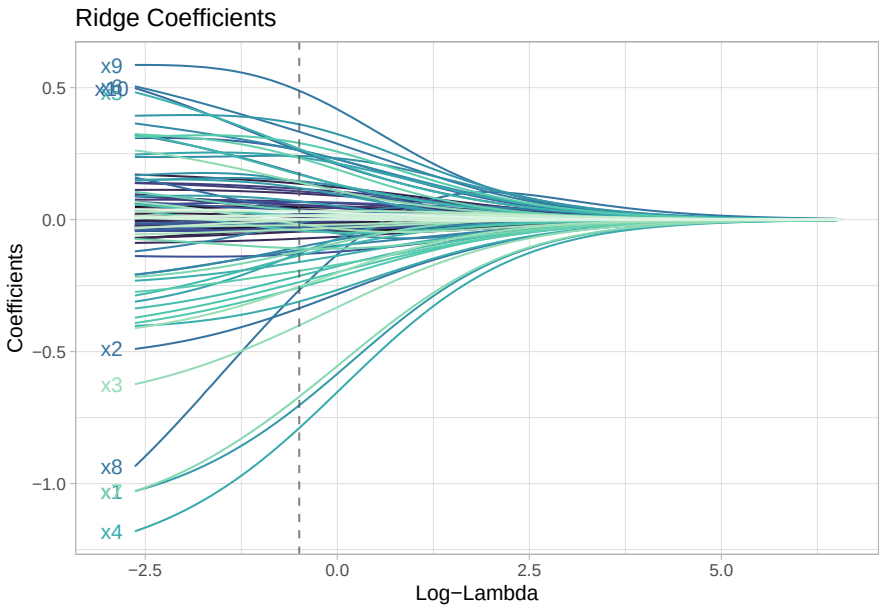
Similar to the LASSO model, we performed repeated 10-fold cross validation manually for the ridge regression, since the original `train()` function in `caret` was having difficulty with the high dimensionality of the data.

Using cross validation, we found the optimal regularization parameter to be $\lambda = 0.609393$. We are using the highest λ within one standard deviation of the minimizing λ so as to favor parsimony in the model.

Below is the cross validation showing how the regularization parameter was chosen.



And here is the survival plot of the coefficients.



Below are the model performance measures.

CV.RMSE	RMSE	Rsquared	MAE
0.6575	0.7243	0.5491	0.5243

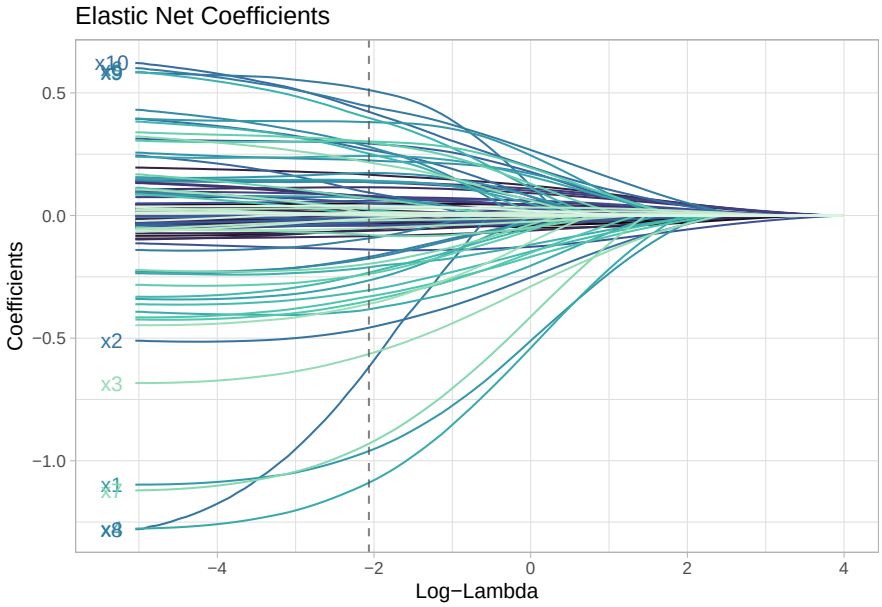
Lastly, here are the 6 features with the highest importance.

Variable	Importance	Sign
StateND	100.00000	NEG
StateME	89.13088	NEG
StateVT	84.98771	NEG
StateDE	61.93025	POS
StateWI	50.78152	NEG
StateMD	45.78063	POS

Model 4: Elastic net

Using cross validation, we found the optimal regularization parameter to be $\lambda = 0.1268456$. We also determined the optimal $\alpha = 0.0134228$, which is clearly close to 0. Since both the λ and α are close to those of the ridge regression ($\alpha = 0$ in that case), these two models are very similar.

Below is the survival plot, showing behavior similar to that of the ridge regression survival plot.



In the following table, we have the performance measures for the model.

CV.RMSE	RMSE	Rsquared	MAE
0.642	0.7176	0.562	0.5169

And lastly, here are the 6 features with the highest importance in the elastic net model.

Variable	Importance
StateND	100.00000
StateME	88.13954
StateVT	85.39664
StateDC	56.70187
StateWI	51.82111
StateDE	46.98794

Model 5: PCR

For the principal component regression, we found the optimal number of principal components to be 82. This number is very close to the full number of features of 90, indicating that most of the included features already contribute substantially the variation observed.

Below are the performance measures for this model.

CV.RMSE	RMSE	Rsquared	MAE
0.6449	0.7185	0.5656	0.517

And here are the 6 most important features, according to the model.

Variable	Importance
pctWhite	100.00000

Variable	Importance
pctPubAsst	81.20643
pctBlack	79.96451
pctHousWOphone	73.58767
pctSmallHousUnits	63.62770
houseVacant	52.56120

Model 6: Piece-wise polynomial

For the piece-wise polynomial model, we ended up constructing this model from scratch. Using cross validation, we determined the optimal order of the numeric variables to be 2.

Below are the performance measures for the model.

CV.RMSE	RMSE	Rsquared	MAE
0.7653	0.8121	0.4921	0.618

We chose to not include the variable importance because they did not provide much information, as most of the top performing variables were variable subsets that were virtually unintelligible.

Model 7: MARS

Though we did fit a traditional spline model, we chose to include the MARS model to represent the "family" of spline regression models, since it outperformed the B-spline and natural cubic spline models by every metric. For this model, we found the optimal number of degrees of interaction to be 1, and the optimal number of terms in the pruned model to be 24.

Below is the table of MARS performance measures.

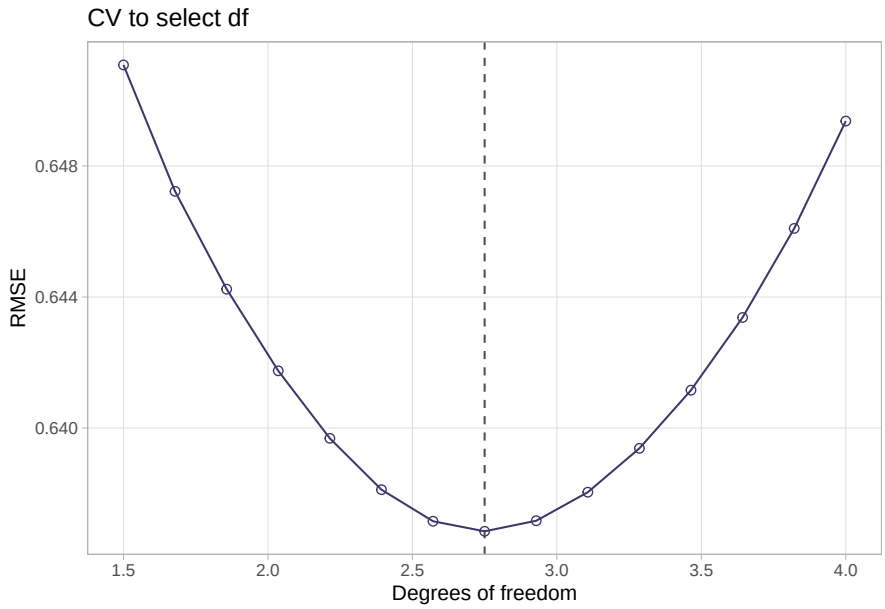
CV.RMSE	RMSE	Rsquared	MAE
0.6344	0.7322	0.558	0.5176

And the 6 features with the highest importance.

Variable	Importance
pctWhite	100.00000
pctHousWOphone	52.63968
houseVacant	45.69627
StateFL	31.89371
pctPubAsst	29.34801
StateMA	29.01005

Model 8: GAM

Below is the plot of the cross validation used to determine the optimal number of degrees of freedom for the GAM model. We found the optimal $df = 2.75$.



Below are the performance measures.

CV.RMSE	RMSE	Rsquared	MAE
0.6368	0.716	0.5641	0.5084

And lastly, the variable importance for the 6 most important.

Variable	Importance
medNumBedrm	100.00000
StateCA	32.40801
StateNJ	24.27825
StatePA	22.92012
StateOH	20.73016
pctWhite	11.48156

Model 9: GPR

As part of the data transformation process, we ended up transforming many variables using a logarithmic or inverse hyperbolic sine transformation (if the data was not strictly positive). As a result, many of the numerical predictors followed somewhat normal distributions. In particular, the outcome variable was almost approximately normal. To take advantage of this approximate normality, we opted to fit a Gaussian process regression (GPR), a non-parametric regression method that uses a Gaussian kernel to learn a distribution which is likely to have generated the outcome variable.

Below is the table of performance metrics.

CV.RMSE	RMSE	Rsquared	MAE
0.7034	0.7104	0.5746	0.5065

And here is the variable importance for the 6 most important in the model.

Variable	Importance
pctWhite	100.00000
pctPubAsst	81.38615
pctBlack	80.03136
pctHousWOphone	73.30396
pctSmallHousUnits	63.64465
houseVacant	58.58766

Model 10: Bayesian ridge

Lastly, we wanted to try a modified ridge regression which learns the regularization parameter through a Bayesian framework. We would expect this method to learn a slightly better model than the traditional ridge regression from earlier.

Below is the performance metrics

CV.RMSE	RMSE	Rsquared	MAE
0.6465	0.7187	0.5631	0.5184

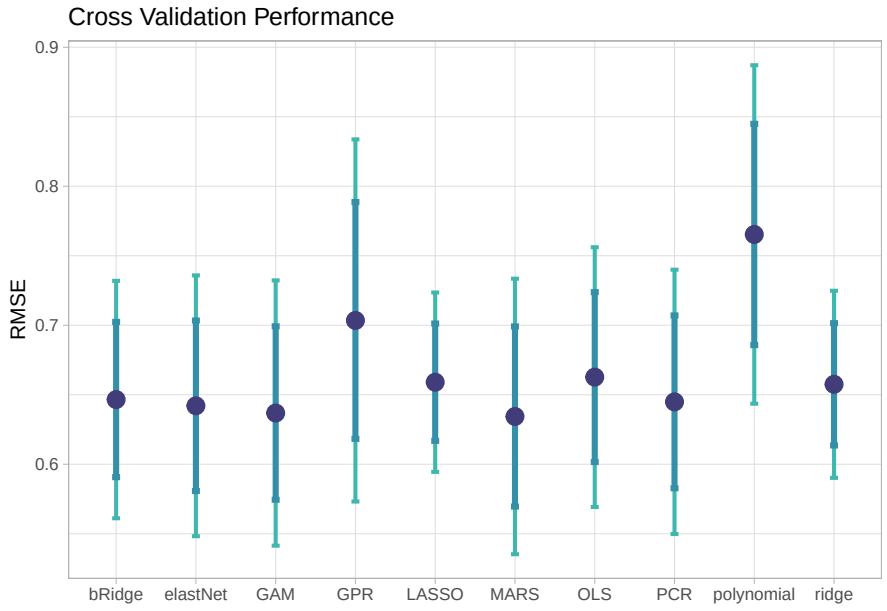
And of course, the 6 most important features.

Variable	Importance
pctWhite	100.00000
pctPubAsst	81.20643
pctBlack	79.96451
pctHousWOphone	73.58767
pctSmallHousUnits	63.62770
houseVacant	52.56120

Conclusions

Comparing Models

Below we have the cross validation RMSE measures for each model. While most of the models have similar distributions, there are three notable observations. Most apparent is the Gaussian process regression, which has a higher than average RMSE with a larger variance, yet future comparisons will show this to be a well-performing model. Secondly, the piece-wise polynomial performs markedly worse compared to the other models. Our third observation is that both the LASSO and ridge regressions have slightly smaller variance compared to the other models. This last note makes sense, since their higher bias would result in more consistent predicted values.



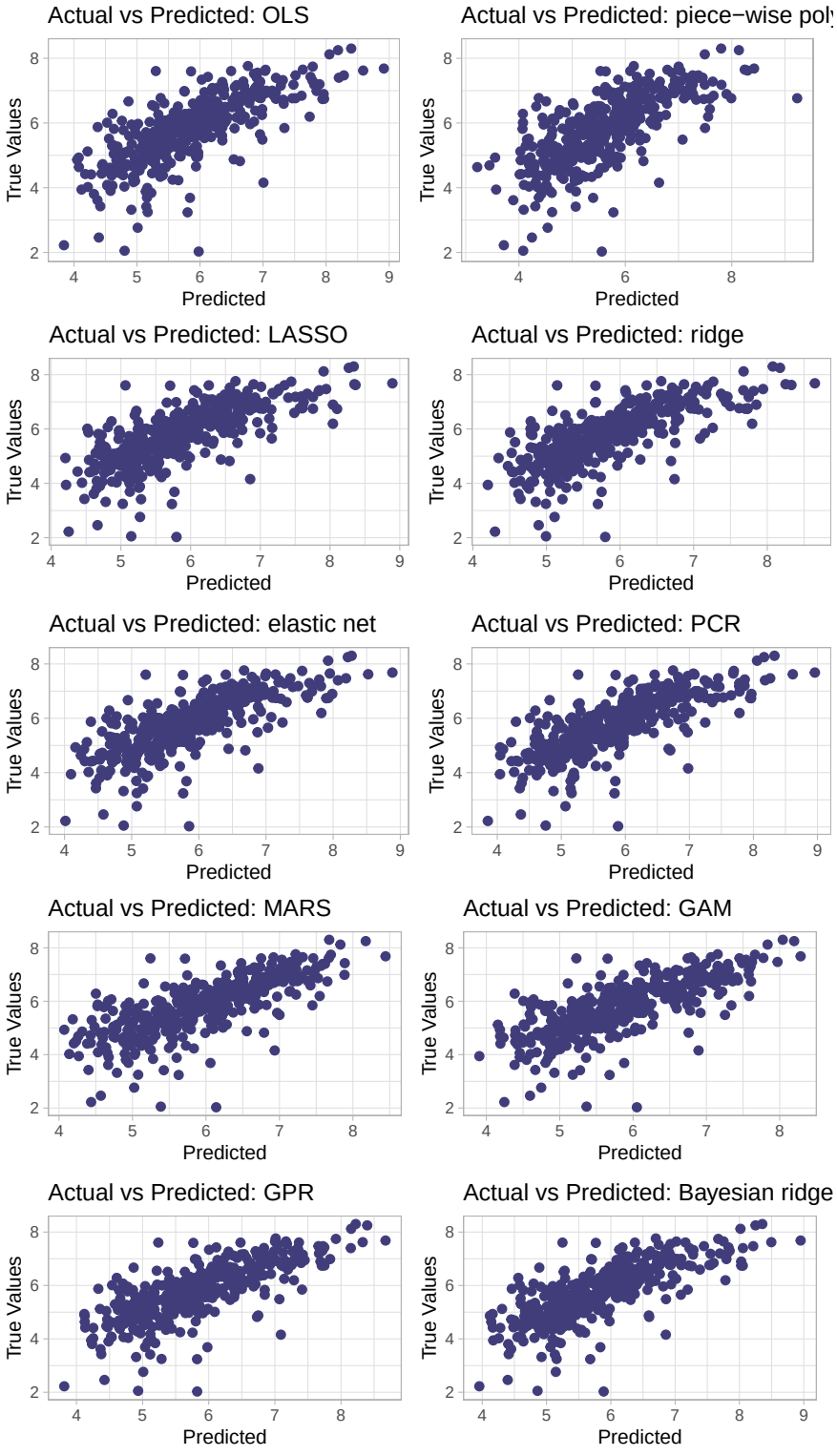
For the main driver of our analysis, we will compare the out-of-sample performance measures below. Ranked by increasing RMSE, it is apparent that the Gaussian process regression performs the best for by each metric, with the GAM and elastic net regression performing second and third best, respectively. As expected from the cross validation performance, the piece-wise polynomial performs the most poorly.

Notably, these RMSE values are all very close, indicating that these models tend to perform fairly similarly. Furthermore, compared to the standard deviation of 1.108497, all the models differ in their estimates by, on average, less than one standard deviation. Thus we would consider these models to be generally successful in their goals.

	RMSE	R.squared	MAE
GPR	0.7104	0.5746	0.5065
GAM	0.7160	0.5641	0.5084
elastic net	0.7176	0.5620	0.5169
PCR	0.7185	0.5656	0.5170
Bayesian ridge	0.7187	0.5631	0.5184
ridge	0.7243	0.5491	0.5243
OLS	0.7247	0.5582	0.5211
MARS	0.7322	0.5580	0.5176
LASSO	0.7343	0.5403	0.5351
p-w polynomial	0.8121	0.4921	0.6180

Overall, we would prefer the GPR model and the GAM model, favoring non-linear models over linear models.

This page contains graphs of true versus predicted values. As expected, the distributions for the GPR and GAM model have the most linear trend, indicating better fit.



Economic Relevance

Since the RMSE performance measures are close, differing in their estimates by less than 1 standard deviation. In that sense, the models hold similar predictive power, though each tells a different story of the data. The most economic interpretability can come from the feature importance for the models, from which we can construct an idea of the most relevant features to the issue of violent crime.

Among the regularized models (not including Bayesian ridge), all three had indicator variables for states as their most important features. In particular, North Dakota, Maine, and Vermont were the top 3 features for all three of the models.

From the other important variables, there seems to be an association between violent crime and aspects of socioeconomic class. For example, the percentage of people on public assistance, the percent of houses without phones, the number of vacant homes, and the number of small house units are all traits relating to socioeconomic class, and they showed up as important across multiple models. Most interesting is the percentage of houses without phones, an unlikely feature to be considered important. Further study might connect the feature to underlying trends.

With socioeconomic variables showing relative importance in the models, policy makers might consider redistributing funds within counties with high poverty rates. Directing economic support towards social programs and programs aimed at people below the poverty line might better mitigate violent crime within those affected communities, compared to increased investment in law enforcement and policing. With counties and states limited by tax revenues and apportioned funding, these models could aid in structuring government spending schemes, improving efficiency with regard to community safety.

Lastly, we must mention that race, specifically the percent white and the percent black, consistently showed up as important features. Though these might be related to violent crime, these features dealing with race do not imply any sense of causality. It is precisely for results like these that addressing bias within AI is very important, especially in the realm of policing. The models in this project serve to better understand groups of locally organized people and *not* individuals; in that way, they are tools for understanding social phenomena and policy, rather than making statements on the choices people make.

Limitations & Future Work

The most glaring issue of the data is the difficulty of tracking both demographics and reports of crimes. Marginalized groups such as undocumented immigrants might not be willing to provide their demographic information to the government ("What Immigrants Need to Know about Census"), and including incarcerated persons in the counts for a county artificially modifies the county's demographic makeup. On the crime reporting side, rapes and incidents of sexual violence tend to be underreported ("Statistics about Sexual Violence", 2015), which also affects how violent crimes are tracked.

Also within the data, we only looked at one slice of what would be a large panel dataset, tracking crime and demographics over time. Therefore the scope of this project is really only applicable to U.S. counties in 1995.

As for the analysis, our ability to mitigate bias, both implicit and explicit, was very limited by the simplicity of the tools used. For example, deciding whether race should be used to predict violent crime inherently implies bias within the model. Proper AI research is aimed at these issues, particularly in the realm of predictive policing, but we were incapable of adopting those methods at this time in this project. Future work could potentially aim to build models impervious to these biases, or could isolate one state to actually look at the community level. Both possible projects would be taking a more nuanced view of the issue of violent crime than we were capable of in this project.

At the end of the day, the purpose of this project was to give a high-level overview of a social phenomenon whose roots and influences go beyond the limited data used. The causes and effects permeate every strata of society and vary by each individual's circumstance. Crime, especially violent crime, does not happen in a vacuum, so the simplistic models of this project simply cannot model the individual complexities of those who commit crimes.

Bibliography

"Artificial Intelligence in Predictive Policing Issue Brief." *NAACP*, February 15, 2024. https://naacp.org/resources/artificial-intelligence-predictive-policing-issue-brief#:~:text=Bias%20and%20Discrimination%3A%20AI%20models,policing%20and%20resources%20are%20made.

Black, Tawanna, Cailean Kok, Jennifer S. Vey, Anthony F. Pipa, Gary Geiler, Andre M. Perry, Hanna Love, and Jenny Schuetz. "Want to Reduce Violence? Invest in Place." *Brookings Institute*, October 17, 2023. https://www.brookings.edu/articles/want-to-reduce-violence-invest-in-place/.

Ferrer, Xavier, Tom van Nuenen, Jose M. Such, Mark Cote, and Natalia Criado. "Bias and Discrimination in AI: A Cross-Disciplinary Perspective." *IEEE Technology and Society Magazine* 40, no. 2 (June 2021): 72–80. https://doi.org/10.1109/mts.2021.3056293.

Gramlich, John. "What the Data Says about Crime in the U.S." *Pew Research Center*, April 24, 2024. https://www.pewresearch.org/short-reads/2024/04/24/what-the-data-says-about-crime-in-the-us/#:~:text=A%20growing%20share%20of%20Americans,Joe%20Biden's%20presidency%20in%202021.

Gramlich, John. "What We Know about the Increase in U.S. Murders in 2020." *Pew Research Center*, October 27, 2021. https://www.pewresearch.org/short-reads/2021/10/27/what-we-know-about-the-increase-in-u-s-murders-in-2020/.

"Predictive Policing Explained." *Brennan Center for Justice*, April 1, 2020. https://www.brennancenter.org/our-work/research-reports/predictive-policing-explained.

Redmond, Michael. "Communities and Crime Unnormalized." *UCI Machine Learning Repository*, March 1, 2011. https://archive.ics.uci.edu/dataset/211/communities+and+crime+unnormalized.

"Statistics about Sexual Violence." *National Sexual Violence Resource Center*, 2015. https://www.nsvrc.org/sites/default/files/publications_nsvrc_factsheet_media-packet_statistics-about-sexual-violence_0.pdf.

"Violence & Socioeconomic Status." *American Psychological Association*, 2010. https://www.apa.org/pi/ses/resources/publications/violence#:~:text=Community%20level%20risk%20factors%20for,2016%3B%20McMahon%20et%20al.%2C.

"Violent Crime." *The U.S. Federal Bureau of Investigation*, September 13, 2019. https://ucr.fbi.gov/crime-in-the-u.s/2019/crime-in-the-u.s.-2019/topic-pages/violent-crime.

"What Immigrants Need to Know about Census." *The Southeast Asia Resource Action Center*. Accessed May 10, 2024. https://www.searac.org/wp-content/uploads/2019/08/SEAA-Immigrants-Census-Confidentiality-and-the-Citizenship-Question_FINAL-FINAL.pdf.