Crime in US Communities

Team 51 Final Report

JENNY CHEN

JONATHAN CASCIOLI

MAKRAND KALYANKAR

MEIR WEINER

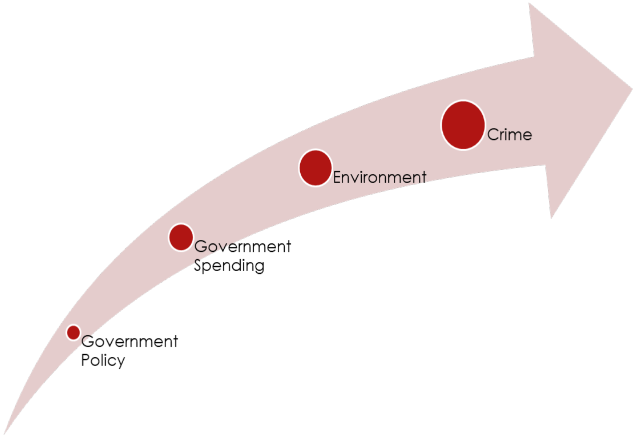
XIAOLU SU

Overview

Our project focused on the problem of violent crime in US communities. We began this project with a hypothesized life cycle. We held the belief that government policy served as the origin point for the most influential variables that contribute to violent crime. Therefore, posing the idea, that different stances on government policy, thereby dictating the allocation of government funding, directly create the overall environment that makes up a community.

These environmental variables serve as the seeds that grow a community's socio-economic climate. Healthy seeds, dictated by informed policy, create a prosperous community. Whereas unhealthy seeds lead to a hostile environment, thereby breeding crime.

The overall goal of our project was to uncover the key factors of policy that a government (local or federal) could influence. We gathered data and used machine learning methodologies, with the intent of creating a tool that would inform policy makers on the optimum stance to exercise on various issues. With an emphasis on understanding the relationship of these issues and crime(inference), as opposed to one of predicting the rate of crime under circumstantial input



Data Selection and Cleaning:

Data Sources:

As part of the project, we started out with crime data (Schema) as our base dataset. We intended to enrich the data with features that communities deploy that may affect the number of crime incidents occurring in the community. This required us to collect features available through various agencies and publicly available datasets.

Planned Parenthood:

We were unable to find planned parenthood data from the specific year of our main crime dataset (1995). We gathered the data by scraping the planned parenthood website. This marks an area where our project, as mentioned in the proposal, must dive into the scope of being theoretical. More on this later in the challenges section. A simple binary variable (0,1) made up our planned parenthood data.

Education Data:

Our Educational data source was gathered from the National Center for Education Statistics (NCES) from 1995-1996 and contained information on:

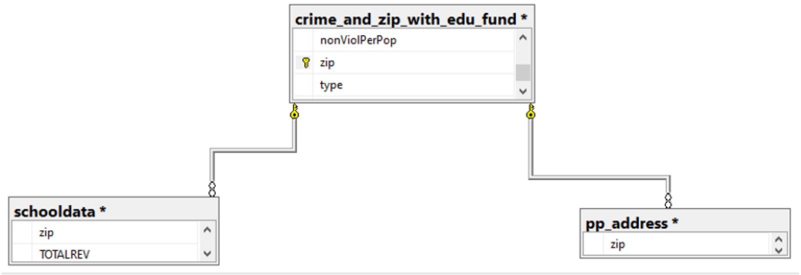
School Information	*Finance*	*State Level Sum:*	*School Counts*	*Grade & Grad counts*
School District	Financial Aid	Financial State Summary	Dropouts	Grade (1-12) Counts
School Codes	Revenue		Ethnicity Counts	Teacher Counts
Student Counts	Taxes		School Staffing	Librarians
etc	Employee Salaries			School Admins Etc
	Employee Benefits			

	etc			

This marks an area where due to time constraints we were unable to incorporate all the data we thought would be valuable. We joined financial ratios of federal, state, and local aid contributions into our main dataset as we thought these held the most promise and hypothesized would have the greatest explanatory strength.

Data Cleaning:

Our additional data sets were reported on the zip code level or some sort of government statistical code (NCES). The base crime data was reported at community level with non-standardize community name. A significant amount of time was spent just enabling the data to be joined to our main dataset. The team enriched this main dataset with Zip code information based on the existing State-Community combination. Once Zip codes were associated with communities, we could experiment with relevant feature data.



We followed industry best practices in cleaning the data:

- We leveraged RDBMS inherent ability to enforce domain constraints i.e. column data types. This guarantees that the data for the field is of data types we define it to be. Any data not meeting the criteria was type casted.
- Removed any duplicate rows originating from join condition
- Normalized missing data to NULL instead of mixture of NA, “empty string”, NULL so there is a consistent way to deal with them

Exploratory Data Analysis:

Once the team joined the core raw dataset with 2 additional sources, we started an investigation in python as part of our exploratory data analysis. This led to further data preparation. We discovered there were almost 200 features, and 11,516 rows with lots of missing values.

Filtering Features:

First, we went through and manually chose all the values that were relevant to our business problem and ethical to use. For example, we removed all columns related to race. Our remaining explanatory variables made up several categories:

1. Age Demographics
2. Social program / assistance
3. Poverty indicators
4. Education History
5. Education Spending

6. Employment
7. Family Demographics
8. Housing Occupancy
9. Policing

With Violent Crime Rates Per population unit as our response variable, as we had originally posed in our proposal.

We did some additional manipulations and aggregations to get the data to be unique at the community level. Discovering that when we incorporated our zip code data to our original dataset, it created duplication of rows. After the aggregation and cleaning to this point, we held 1886 unique communities / observations.

Feature Standardization:

From here we started a further null value analysis null value, removing additional features that contained greater than 30% null values.

After the final cleaning and standardization process, we landed on 1,414 observations with 38 explanatory variables.

Statistical Exploration:

With our final features, we performed some light exploratory data analysis. First, we noticed that the distribution of the response variable was right-skewed (fig 1), implying that there are only a few communities with high crime rates.

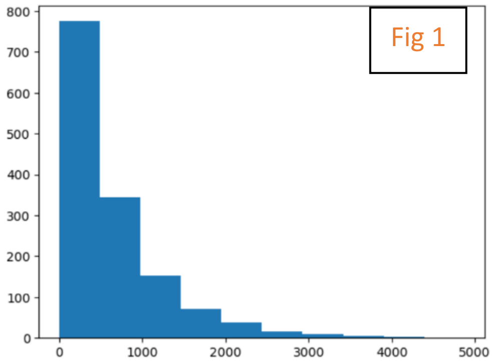


Fig 1

We then looked at some of the explanatory variables that we hypothesized to be most important and compared them to our response variable (Violent Crimes per Population Unit). These consisted of:

- % with less than 9th Grade Education (fig 2)
- Access to Planned Parenthood
- (fig 3)
- % of population using public assistance (fig 4)
- % of homes that are non-vacant (fig 5)
- % of population that is unemployed (fig 6)
- % of education funding by local government (as opposed to federal or state) (fig 7)

**Note: for each of the following figures the format is:*

- *explanatory variable – x axis vs*
- *response variable Violent Crimes per Population Unit – Y axis*

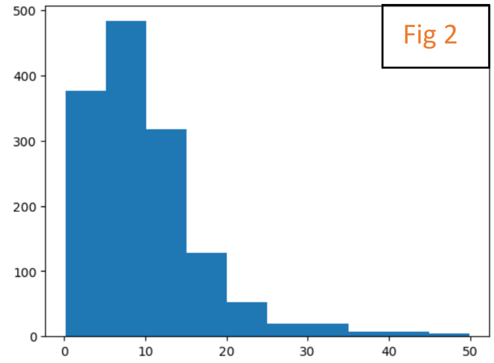


Fig 2

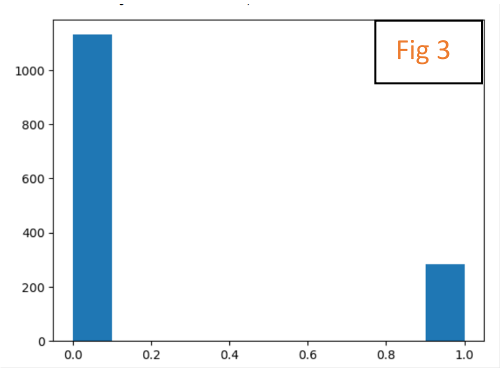


Fig 3

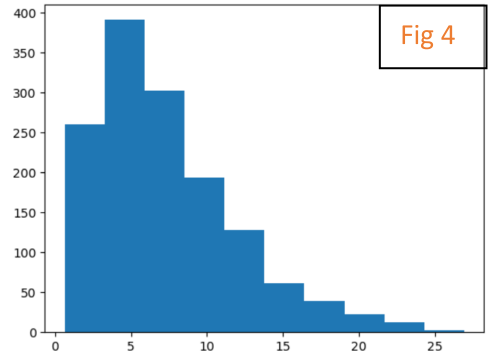


Fig 4

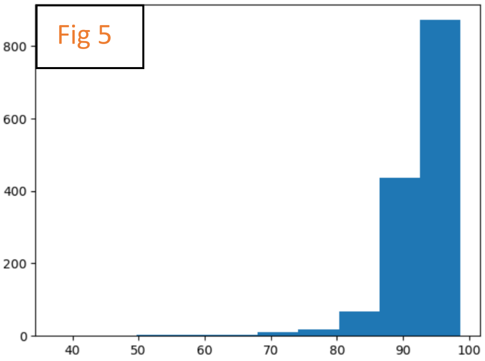


Fig 5

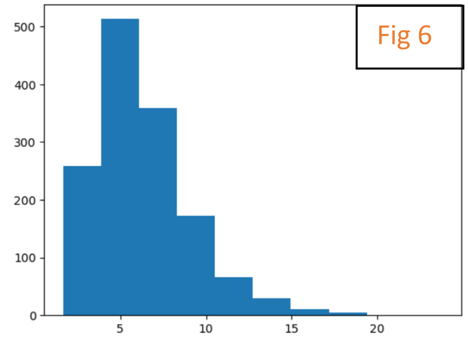


Fig 6

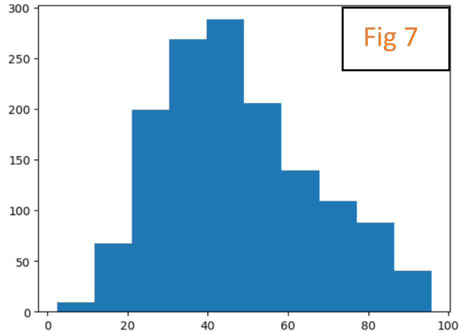


Fig 7

Here, our 2 key findings were that (1) many of the features do not have linear or linear- transformable relationships with the response, and (2) there is a decent amount of correlation amongst the explanatory variables themselves. Both findings suggest that a multivariate linear regression may not be the best model for our data.

Modeling: Linear Regression (with Greedy Algorithm applied)

Data Split:

Before beginning our modeling training process, we left out the variable "state" because it was an unalterable location factor that would not be helpful for the study of local crime rate. In order to evaluate the models based on the dataset we have, we split the data into 70% of the train set and 30% of the test set. In other words, out of the 1,414 observations in our final dataset, there will be still around 420 datapoints set aside as the test set for the model validation.

For Linear Regression model, we calculated the mean squared error (MSE) of the predicted response values in the train set MSE(train), and then we compared MSE to that the MSE of the predicted response values calculated using the test set MSE(test).

VIF Analysis:

To better understand the dataset, we ran a VIF analysis on the variables. As a rule of thumb, if VIF > 5, it indicates a presence of multicollinearity. Our VIF analysis showed that many variables were highly correlated with each other, but we were unable to know in what way they were related. Therefore, we decided to apply Greedy Algorithm, specifically Forward Feature Selection and Backward Feature Elimination. For forward selection, it is good at dealing with data with collinearity because it builds up the empty model by adding whatever feature that can improve the model. Backward elimination shares the similar goal to only improve the model performance, but instead of adding features to the model, it removes features from the model step by step.

Approaches:

The first type of model we chose was Linear Regression, which has been mostly used to solve regression problems. It has simple estimation procedures, and the results are easy to interpret. We used three different approaches to produce our best linear regression model:

1. Linear Regression
2. Forward Feature Selection and Linear Regression
3. Backward Feature Elimination and Linear Regression.

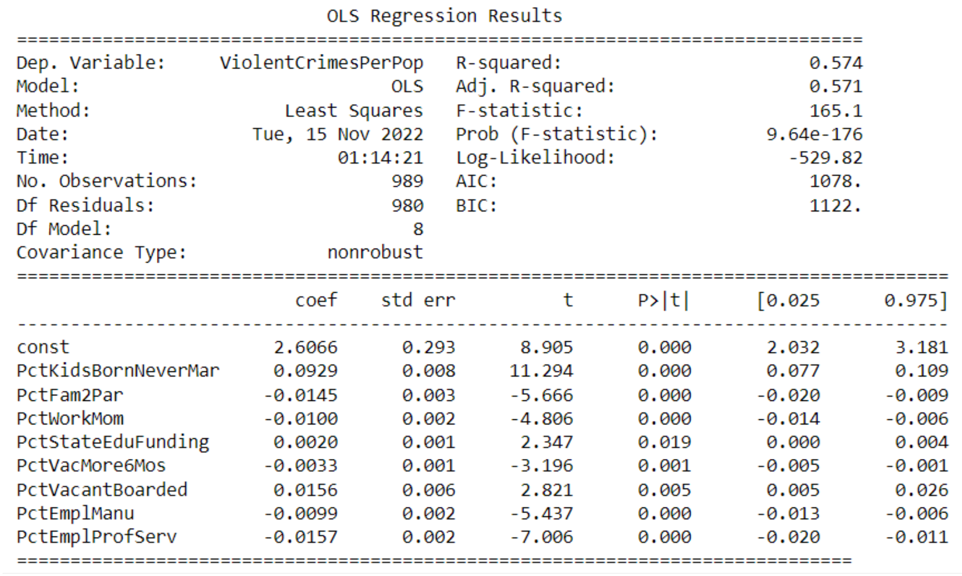
Approach 1: Linear Regression

First, we trained the linear regression model including all independent variables to get a general understanding of the data. The model had an R-squared of 0.596 and adjusted R-squared of 0.585, which means the model was able to explain around 59% of the data variance. Even though the percentage was not high in theory, it was still considered strong in practice. We were surprised by the model's MSE(train) = 0.162 and MSE(test)= 0.134. Both were exceptionally low before any feature selection. However, as we expected, the model included too many unimportant variables based on their high p-values. After we set the threshold of 95% Confidence Interval, 8 variables remained, which was a large decrease. But choosing the variables from the premature model was not enough to prove their importance, so we needed to train the linear model only on those selected variables.

We trained the linear regression model for a second time on those 8 predictors:

ViolentCrimesPerPop = 2.6066 + 0.0929 * PctKidsBornNeverMar – 0.0145 * PctFam2Par – 0.01 * PctWorkMom + 0.002 * PctStateEduFunding – 0.0033 * PctVacMore6Mos + 0.0156 * PctVacantBoarded – 0.0099 * PctEmplManu – 0.0157 * PctEmplProfServ

As there were less predictors to train on, we were not surprised that R-squared would decrease. But given that the R-squared was reduced to 0.574 and adjusted R-squared to 0.571, it was a relatively small reduction compared to the number of features excluded from the previous model trained. This time, MSE(train) = 0.171 and MSE(test) = 0.138. While the MSE's showed slight increases, we traded a bit more errors for the emphasis on the significant factors and better interpretation of the prediction model.

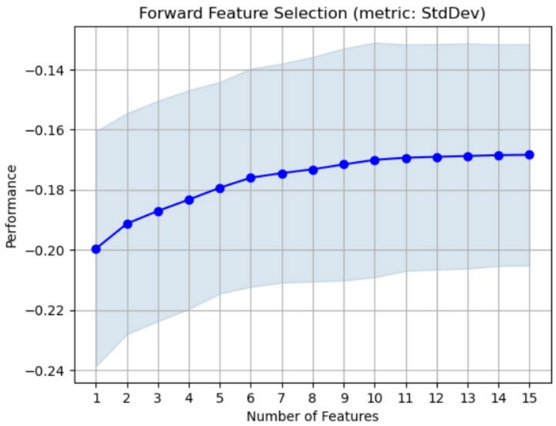


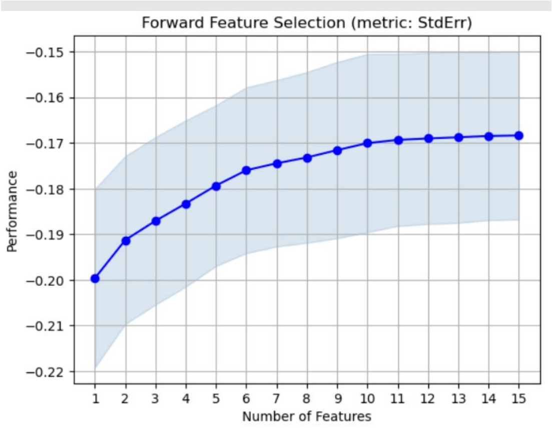
```
                            OLS Regression Results
==============================================================================
Dep. Variable:     ViolentCrimesPerPop   R-squared:                       0.574
Model:                             OLS   Adj. R-squared:                  0.571
Method:                  Least Squares   F-statistic:                     165.1
Date:                 Tue, 15 Nov 2022   Prob (F-statistic):          9.64e-176
Time:                         01:14:21   Log-Likelihood:                -529.82
No. Observations:                  989   AIC:                             1078.
Df Residuals:                      980   BIC:                             1122.
Df Model:                            8
Covariance Type:             nonrobust
========================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------
const                   2.6066      0.293      8.905      0.000       2.032       3.181
PctKidsBornNeverMar     0.0929      0.008     11.294      0.000       0.077       0.109
PctFam2Par             -0.0145      0.003     -5.666      0.000      -0.020      -0.009
PctWorkMom             -0.0100      0.002     -4.806      0.000      -0.014      -0.006
PctStateEduFunding      0.0020      0.001      2.347      0.019       0.000       0.004
PctVacMore6Mos         -0.0033      0.001     -3.196      0.001      -0.005      -0.001
PctVacantBoarded        0.0156      0.006      2.821      0.005       0.005       0.026
PctEmplManu            -0.0099      0.002     -5.437      0.000      -0.013      -0.006
PctEmplProfServ        -0.0157      0.002     -7.006      0.000      -0.020      -0.011
==============================================================================
```

According to the OLS Regression Report, all predictors are percentages of the different community factors. The coefficients told us how each factor would influence local violent crimes. “Const” was the base estimation of the number of violent crimes per 100,000 people with zero percent of any other factors. For the other coefficients, on one hand, if the coefficients were positive, such as PctKidsBornNeverMar, PctStateEducFunding, PctVacantBoarded, the higher percentage of these factors would result in more violent crimes. On the other hand, if the coefficients were negative, like PctFam2Par, PctWorkMom, PctVacMore6Mos, PctEmplManu, PctEmplProfServ, the higher the percentage of these factors, the less violent crimes there were predicted to be. “ Based on this prediction model, the most significant factor was the PctKidsBornNeverMar. While other factors remained the same, if the percentage of kids born in the family where the parents never married increased by 1%, the violent crime would increase by 0.093 cases. This model suggested that local violent crimes could be affected by parenting, education funding, house vacancy, and employment.

Approach 2: Forward Feature Selection and Linear Regression

We started this approach by using one type of the Greedy Algorithm, Forward Feature Selection, with 5-fold cross validation to pick the top 15 key features. The outputs of standard deviations and standard errors showed that when the number of features k = 10, the selection would have the biggest improvement from the model with k-1 number of features. After k = 10, adding more features would barely have any influence on the performance. Therefore, we selected the top 10 significant features from the forward selection with which to train a linear regression model.

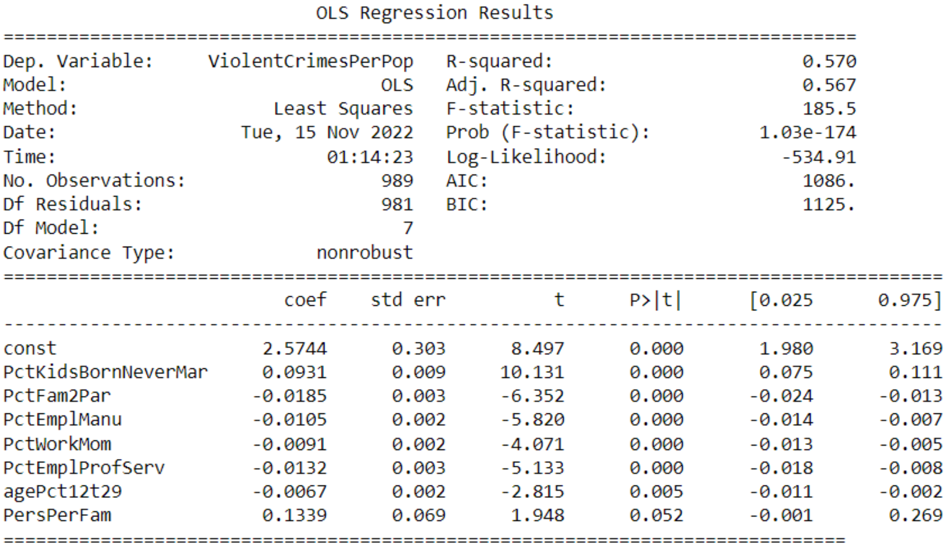




Similar to Approach 1, we trained the first linear regression model to find the features that fell within the 95% Confidence Interval, and we reduced the number of features to the top 7 that met the requirement. Then we trained the linear regression model again with only those 7 variables:

ViolentCrimesPerPop = 2.8589 + 0.1022 * PctKidsBornNeverMar – 0.0155 * PctFam2Par – 0.0103 * PctEmplManu – 0.0149 * PctEmplProfServ – 0.0108 * PctWorkMom – 0.0052 * agePct12t29

The model achieved the MSE(train) = 0.173 and MSE(test) = 0.142, which were both a little higher than the MSE's from the Approach 1. The following OLS Regression Results:

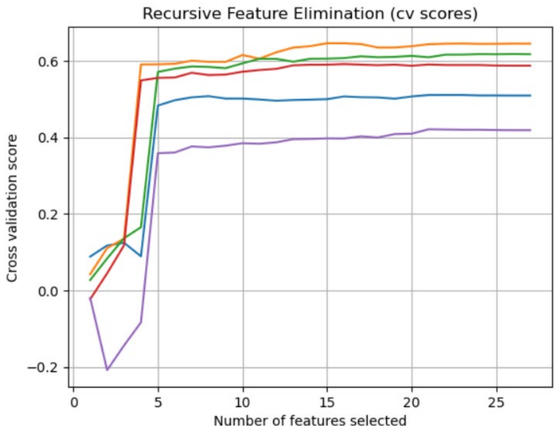


```
                            OLS Regression Results
==============================================================================
Dep. Variable:     ViolentCrimesPerPop   R-squared:                       0.570
Model:                             OLS   Adj. R-squared:                  0.567
Method:                  Least Squares   F-statistic:                     185.5
Date:                 Tue, 15 Nov 2022   Prob (F-statistic):         1.03e-174
Time:                         01:14:23   Log-Likelihood:                -534.91
No. Observations:                  989   AIC:                             1086.
Df Residuals:                      981   BIC:                             1125.
Df Model:                            7
Covariance Type:             nonrobust
=======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
const                   2.5744      0.303      8.497      0.000       1.980       3.169
PctKidsBornNeverMar     0.0931      0.009     10.131      0.000       0.075       0.111
PctFam2Par             -0.0185      0.003     -6.352      0.000      -0.024      -0.013
PctEmplManu            -0.0105      0.002     -5.820      0.000      -0.014      -0.007
PctWorkMom             -0.0091      0.002     -4.071      0.000      -0.013      -0.005
PctEmplProfServ        -0.0132      0.003     -5.133      0.000      -0.018      -0.008
agePct12t29            -0.0067      0.002     -2.815      0.005      -0.011      -0.002
PersPerFam              0.1339      0.069      1.948      0.052      -0.001       0.269
==============================================================================
```

The R-squared and adjusted R-squared are both around 0.57, which was slightly lower than the ones of Approach 1. Besides, they shared most of the significant features. To be specific, both models found PctKidsBornNeverMar, PctFam2Par, PctWorkMom, PctEmplManu, PctEmplProfServ to have similar positive or negative significant influence on violent crimes. This model reflected that parenting, employment, and age could most significantly affect local violent crimes. It narrowed down the categories but incurred more errors.

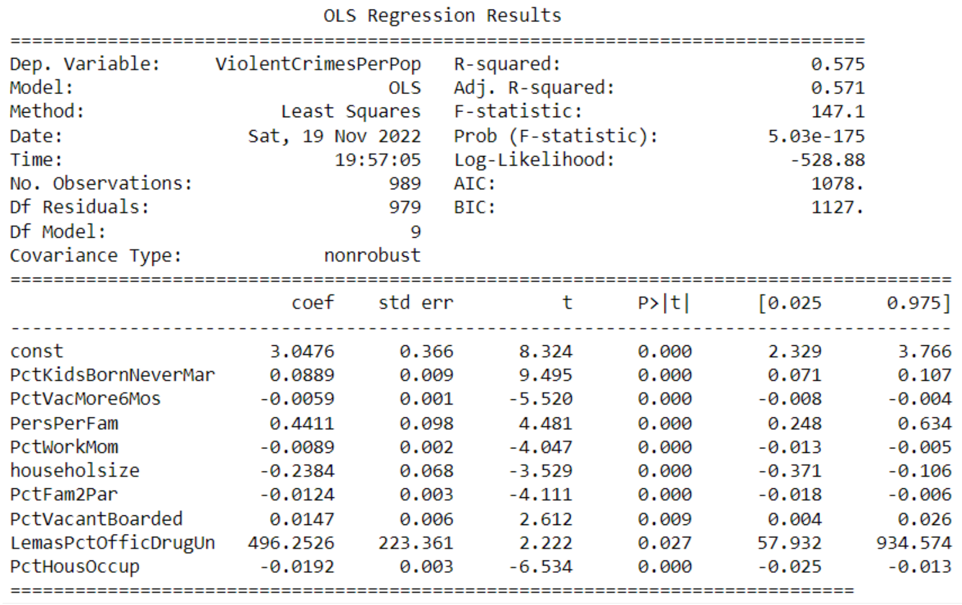
Approach 3: Backward Feature Elimination and Linear Regression

This approach replaced Forward Selection by Backward Feature Elimination also with 5-fold cross validation without specifying the number of features we wanted. The cross-validation scores suggested that when the number of features selected k=12, the models' performances would be stabilized. Thus, we selected the top 12 significant features found by the backward elimination and trained a linear regression model with them. We also expected that some features would not fit the 95% Confidence Interval, so we removed those features and trained the model again with the remaining 9 variables.



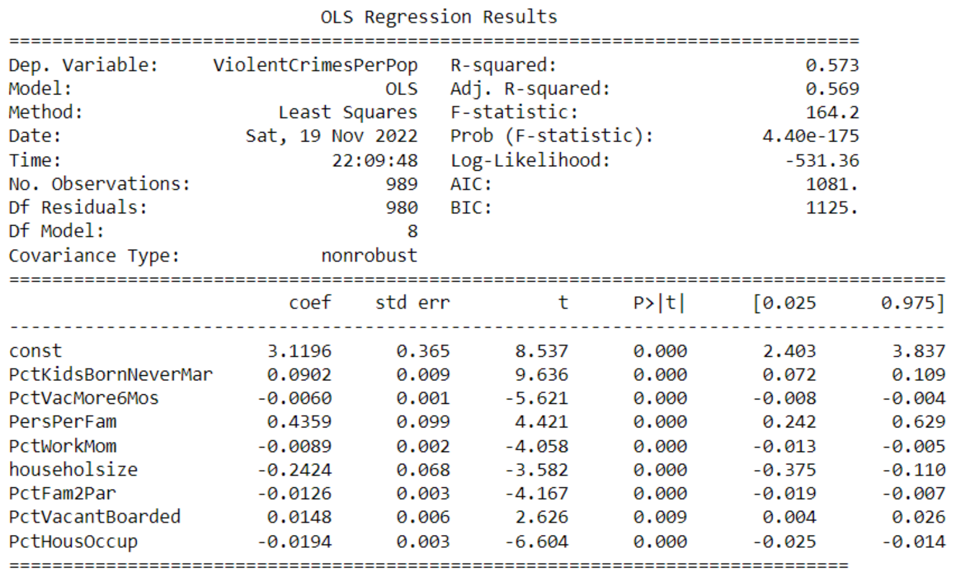
The OLS Regression Results of this approach showed almost the same R-squared and adjusted R-squared as Approach 1. The MSE(train) = 0.171 and MSE(test) = 0.141 were slightly higher than the Linear Regression alone. The model in this approach also indicated that some important features found in the previous approaches have significant impact on local violent crimes, such as PctKidsBornNeverMar, PctFam2Par, PctWorkMom. Nevertheless, the findings in this approach agreed more with Approach 1, because they were both aware of the importance of the house vacancy issues (PctVacMore6Mos, PctVacantBoarded, PctHousOccup) aside from other community factors. This approach has an obvious drawback though. We noticed that the coefficient of LemasPctOfficDrugUn was particularly high due to the correlation between this factor and the response. Since it simply suggested that the number of violent crimes and percentage of officers assigned to drug units were positively correlated, we could only learn that one of the categories of violent crimes was drug related, we decided to remove it from the features and trained the linear regression model once again:

ViolentCrimesPerPop = 3.1196 + 0.0902 * PctKidsBornNeverMar – 0.0126 * PctFam2Par – 0.0089 * PctWorkMom – 0.006 * PctVacMore6Mos + 0.0148 * PctVacantBoarded – 0.2424 * householsize – 0.0194 * PctHousOccup + 0.4359 * PersPerFam



```
                            OLS Regression Results
==============================================================================
Dep. Variable:     ViolentCrimesPerPop   R-squared:                       0.575
Model:                             OLS   Adj. R-squared:                  0.571
Method:                  Least Squares   F-statistic:                     147.1
Date:                 Sat, 19 Nov 2022   Prob (F-statistic):          5.03e-175
Time:                         19:57:05   Log-Likelihood:                -528.88
No. Observations:                  989   AIC:                             1078.
Df Residuals:                      979   BIC:                             1127.
Df Model:                            9
Covariance Type:             nonrobust
=========================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
const                   3.0476      0.366      8.324      0.000       2.329       3.766
PctKidsBornNeverMar     0.0889      0.009      9.495      0.000       0.071       0.107
PctVacMore6Mos         -0.0059      0.001     -5.520      0.000      -0.008      -0.004
PersPerFam              0.4411      0.098      4.481      0.000       0.248       0.634
PctWorkMom             -0.0089      0.002     -4.047      0.000      -0.013      -0.005
householsize           -0.2384      0.068     -3.529      0.000      -0.371      -0.106
PctFam2Par             -0.0124      0.003     -4.111      0.000      -0.018      -0.006
PctVacantBoarded        0.0147      0.006      2.612      0.009       0.004       0.026
LemasPctOfficDrugUn   496.2526    223.361      2.222      0.027      57.932     934.574
PctHousOccup           -0.0192      0.003     -6.534      0.000      -0.025      -0.013
==============================================================================
```

This time, we found that the only difference between this model and the model in Approach 1 was that it replaced PctStateEducFunding with PctHousOccup, and not only did this model explain slightly less data variance than Approach 1, the MSE(train) = 0.171 and MSE(test) = 0.140 also showed that it was subject to more errors. Besides, PersPerFam and householsize were highly correlated, so were PctHousOccup, PctVacantboarded and PctVacMore6Mos. They would affect the accuracy of the linear regression model. So we determined that Approach 1 to be the winner in this Linear Regression part.

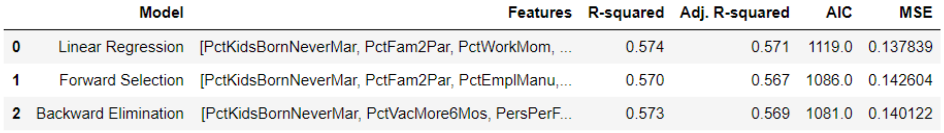


```
                            OLS Regression Results
==============================================================================
Dep. Variable:     ViolentCrimesPerPop   R-squared:                       0.573
Model:                             OLS   Adj. R-squared:                  0.569
Method:                  Least Squares   F-statistic:                     164.2
Date:                 Sat, 19 Nov 2022   Prob (F-statistic):          4.40e-175
Time:                         22:09:48   Log-Likelihood:                -531.36
No. Observations:                  989   AIC:                             1081.
Df Residuals:                      980   BIC:                             1125.
Df Model:                            8
Covariance Type:             nonrobust
=======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
const                   3.1196      0.365      8.537      0.000       2.403       3.837
PctKidsBornNeverMar     0.0902      0.009      9.636      0.000       0.072       0.109
PctVacMore6Mos         -0.0060      0.001     -5.621      0.000      -0.008      -0.004
PersPerFam              0.4359      0.099      4.421      0.000       0.242       0.629
PctWorkMom             -0.0089      0.002     -4.058      0.000      -0.013      -0.005
householsize           -0.2424      0.068     -3.582      0.000      -0.375      -0.110
PctFam2Par             -0.0126      0.003     -4.167      0.000      -0.019      -0.007
PctVacantBoarded        0.0148      0.006      2.626      0.009       0.004       0.026
PctHousOccup           -0.0194      0.003     -6.604      0.000      -0.025      -0.014
==============================================================================
```

Summary and Touch-up:

To sum up on the first modeling part, all approaches shared a few common important features, but Approach 1: Linear Regression and Approach 3: Backward Elimination and Linear Regression were more wholesome as they were able to identify certain important community factors like vacancy, education funding, and employment status that Approach 2 omitted. We chose Approach 1: Linear Regression without applying Greedy Algorithm to be the best model among the three for the following reasons:

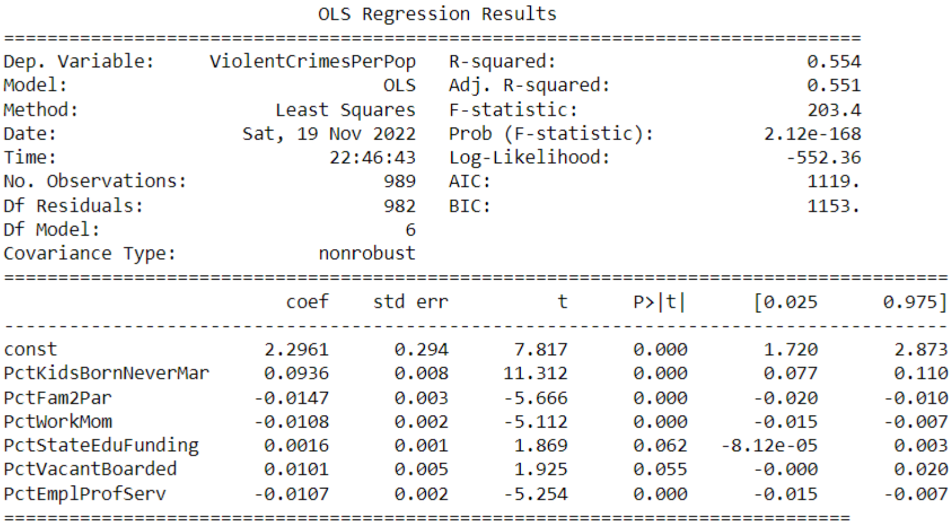
1. MSE was the metric we mainly used to compare across models
2. Other metrics collected were only slightly different
3. The features found in Approach 1 made the best, reasonable interpretation



	Model	Features	R-squared	Adj. R-squared	AIC	MSE
0	Linear Regression	[PctKidsBornNeverMar, PctFam2Par, PctWorkMom, ...	0.574	0.571	1119.0	0.137839
1	Forward Selection	[PctKidsBornNeverMar, PctFam2Par, PctEmplManu,...	0.570	0.567	1086.0	0.142604
2	Backward Elimination	[PctKidsBornNeverMar, PctVacMore6Mos, PersPerF...	0.573	0.569	1081.0	0.140122

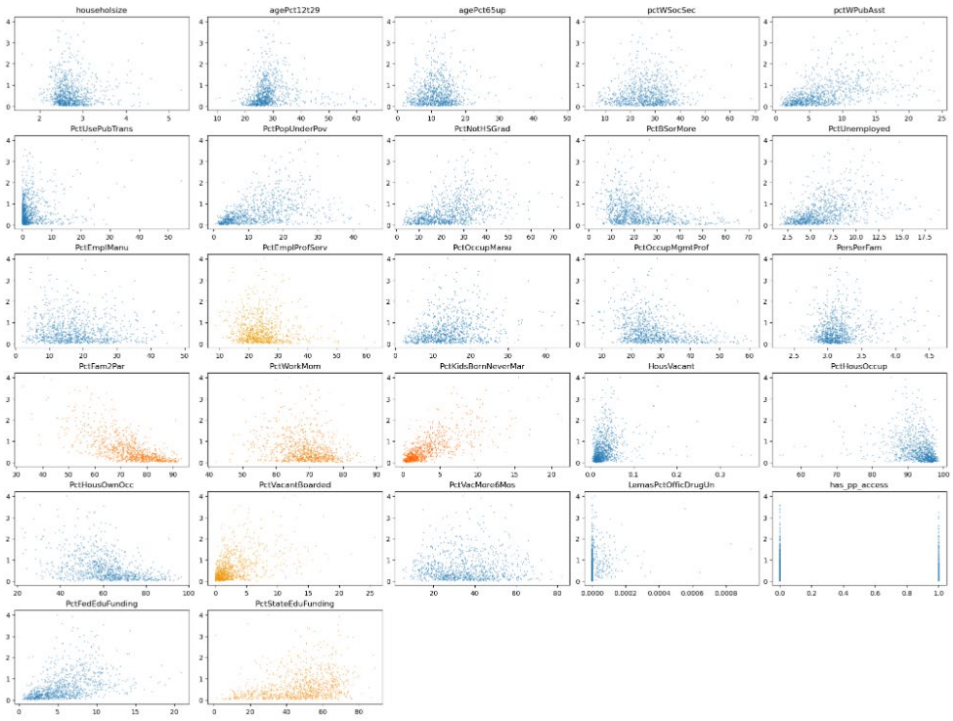
Nevertheless, we considered the feature PctVacMore6Mos found in Approach 1 and 3 to be controversial. This feature could be debated given it had a different coefficient sign than the PctVacantBoarded. We understood that the more vacant boarded houses might lead to more crimes, but we could not properly explain the reason for the reduction in crimes due to house vacancies longer than 6 months. And because it was a feature found in both Approach 1 and 3, and in Approach 1, there was another vacancy factor PctVacantBoarded, we decided we should remove PctVacMore6Mos from Approach 1 and refine the model. Another feature we debated over was PctEmplManu, which was highly related to `PctEmplProfServ`, because whoever was working in manufacturing was very unlikely to be working in professional services. Therefore, we decided to remove this feature as well. After the adjustment, we trained our linear regression model one last time:

ViolentCrimesPerPop = 2.2961 + 0.0936 * PctKidsBornNeverMar – 0.0147 * PctFam2Par – 0.0108 * PctWorkMom + 0.0016 * PctStateEduFnding + 0.0101 * PctVacantBoarded – 0.0107 * PctEmplProfServ



```
                            OLS Regression Results
==============================================================================
Dep. Variable:     ViolentCrimesPerPop   R-squared:                       0.554
Model:                             OLS   Adj. R-squared:                  0.551
Method:                  Least Squares   F-statistic:                     203.4
Date:                 Sat, 19 Nov 2022   Prob (F-statistic):          2.12e-168
Time:                         22:46:43   Log-Likelihood:                -552.36
No. Observations:                  989   AIC:                             1119.
Df Residuals:                      982   BIC:                             1153.
Df Model:                            6
Covariance Type:             nonrobust
=======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
const                   2.2961      0.294      7.817      0.000       1.720       2.873
PctKidsBornNeverMar     0.0936      0.008     11.312      0.000       0.077       0.110
PctFam2Par             -0.0147      0.003     -5.666      0.000      -0.020      -0.010
PctWorkMom             -0.0108      0.002     -5.112      0.000      -0.015      -0.007
PctStateEduFunding      0.0016      0.001      1.869      0.062   -8.12e-05       0.003
PctVacantBoarded        0.0101      0.005      1.925      0.055      -0.000       0.020
PctEmplProfServ        -0.0107      0.002     -5.254      0.000      -0.015      -0.007
==============================================================================
```

It was understandable that the R-squared and Adjusted R-squared would decrease, because we were feeding the model with less variables to train with. We noticed that `PctStateEduFunding` and `PctVacantBoarded` were slightly weighted less as in the feature importance, but they did not fall too far from the 95% Confidence Interval. We also anticipated that the MSE(train) = 0.179 and MSE(test) = 0.145 would be higher than the previous models, but in this case, we found it necessary to have minor human interference on the model to avoid confusion in interpretation. The scatterplots slow helped visualize the general distribution of all variables and how some of the significant features selected were different from the others. The features highlighted with orange gradient were the ones selected from this part of modeling. The darker the orange, the more important the feature was predicted to be.



Modeling: Decision Tree

Rationale:

On top of a multivariable linear regression, we wanted to try a model that could potentially improve the fit to our data. Two of our key concerns with applying the linear regression were (1) non-linear relationships and (2) collinearity. A decision tree is more effective at handling both cases and generally has better prediction accuracy.

Additionally, we went with decision tree over random forest or boosted tree because overall, interpretability was more important to us than prediction accuracy. Although the results of our modeling can certainly be used to predict future crime rates, the inference from feature importance is what would drive our recommendation to the business.

Approach:

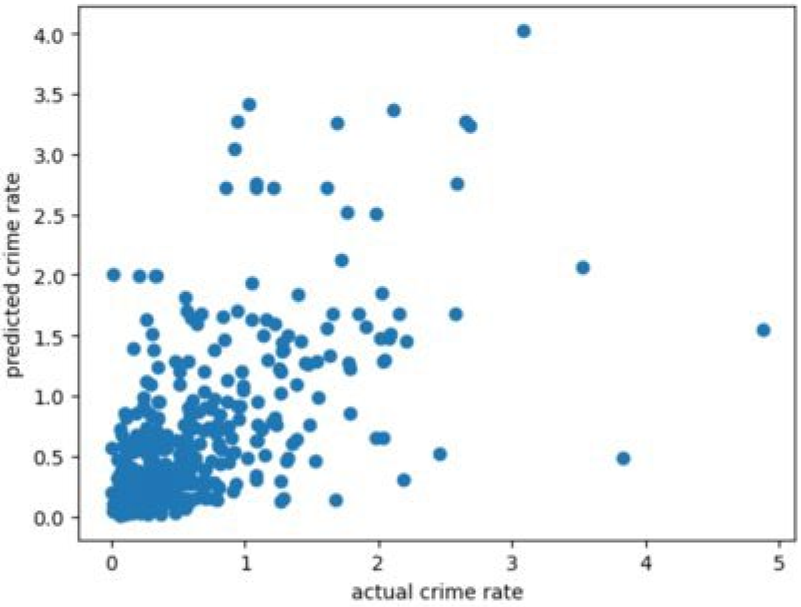
We tried 3 different models with the following specs:

1. All features, default hyperparameters
2. Feature selection method, default hyperparameters
3. All features, tuned hyperparameters

Note: for each of the below models, we used the same train/test split as the linear regression.

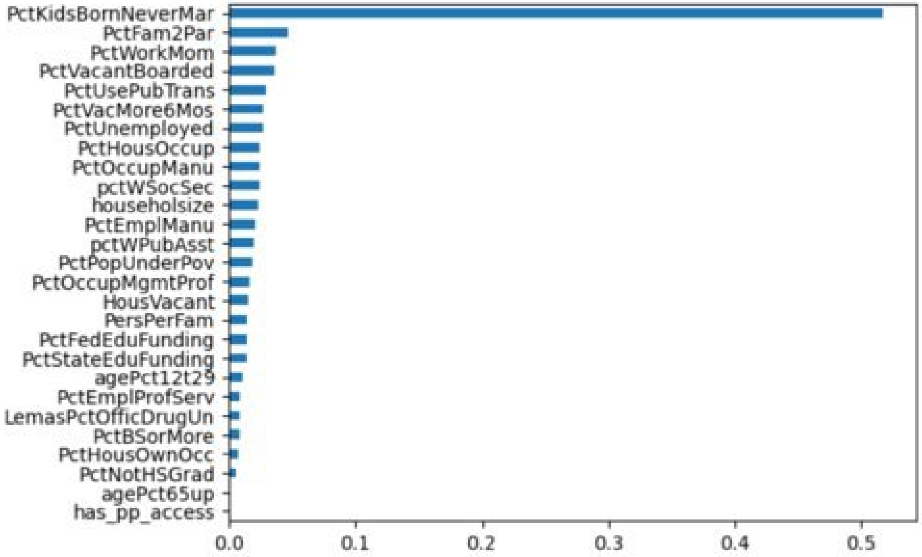
Approach 1: All features, default hyperparameters

The first “base” regression tree we modeled included all features from the cleaned dataset. After fitting the model, we analyzed the differences between the predicted and actual response in the test set, shown below:



There is not a strong linear relationship in the above graph, with a calculated mean squared error of 0.34.

From this first model, we also gleamed some preliminary insights from the feature importance output:



From the above it appears that the PctKidsBornNeverMar feature is the primary variable on which this tree is making splits. This is in-line with some of the results we saw from the linear regressions.

The final output tree from this model is incredibly complex with many splits and leaves, making it impractical to show to a stakeholder audience.

Approach 2: Feature selection, default hyperparameters

For our second tree model, we tried out a built-in feature selection function in the hope that it would make our tree more digestible by choosing fewer features. We leveraged the ExtraTreesRegressor class, typically used as a type of ensemble learning technique, as a quick feature selection tool.

We chose the top 5 features from the feature importance output and fit a tree (default hyperparameters) with similar accuracy results as our first model. The MSE only decreased by about 0.03, and there was not much visible cause to suggest this model was superior to the first.

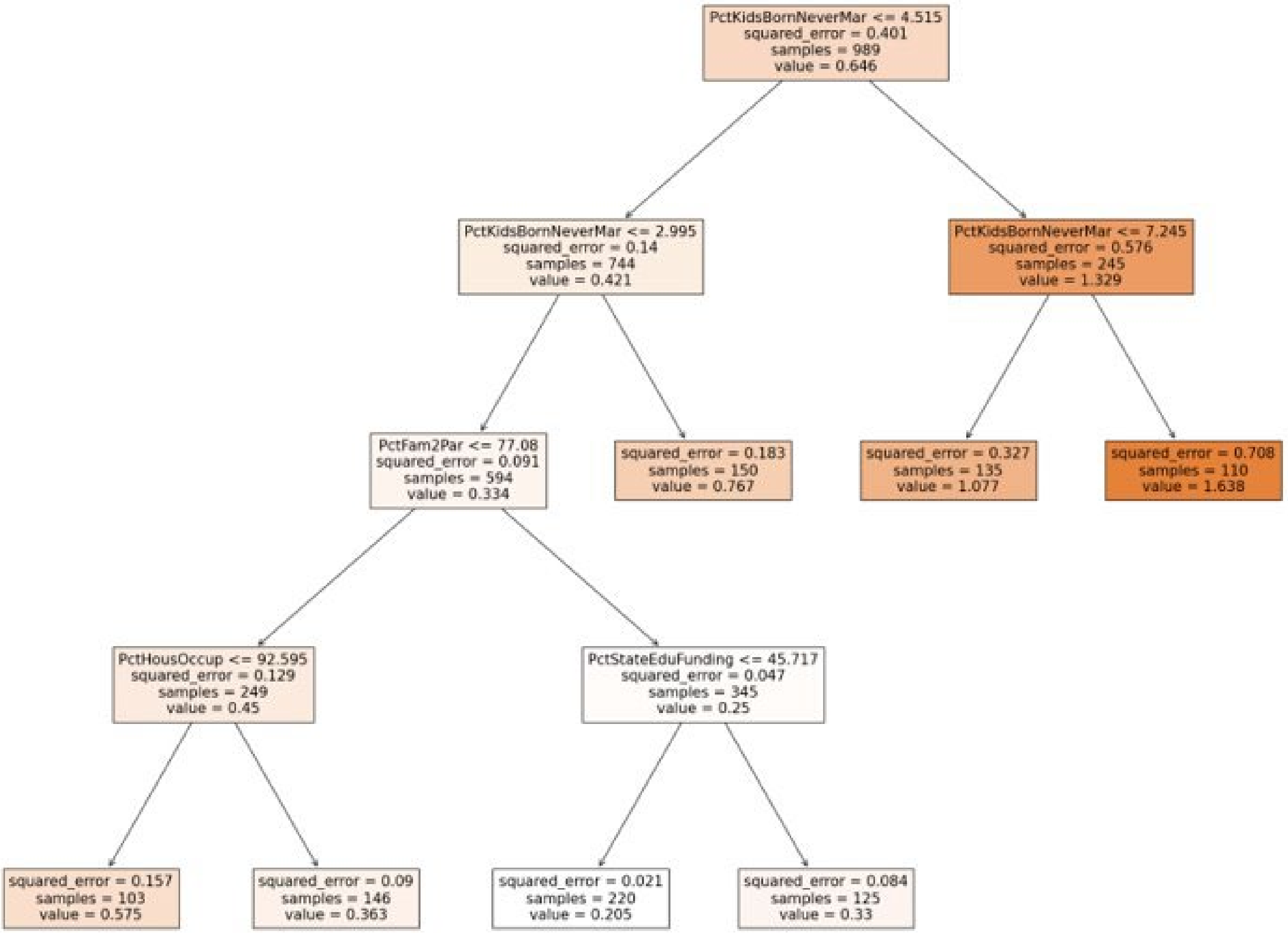
Approach 3: All features, with hyperparameter optimization

For our third and final approach, we wanted to tune some of the hyperparameters manually to output a tree that had higher prediction accuracy and would be easy to explain to our stakeholders.

We initialized our hyperparameter ranges keeping in mind our business use case and stakeholders. For example, we wanted a low max depth and low max leaf nodes for simplicity's sake. We also wanted to ensure a minimum amount of samples in each leaf node to avoid overfitting.

We then performed a grid search to find the optimal hyperparameters using negative mean squared error as our metric. Using this set of hyperparameters, we fit a final model, training on the full feature set.

Unsurprisingly, this tuned model had the best performance with regard to prediction accuracy, with a MSE of 0.19. See below for the splits:



We can see that the two most important features are still PctKidsBornNeverMar (% of kids born to parents that were not married) and PctFam2Par (pct of families with both parents), which are the top two splits the tree makes. The results make sense to us because we'd expect that family stability plays a key role in fostering healthy and safe communities, which both of these variables confirm.

The other two splits are on PctHouseOccup (percent of housing units that are occupied) and PctStateEduFunding (percent of education funded by the state, as opposed to local or federal). The tree indicates that the lower the occupancy rate (and thus the higher the vacancy rate), the lower the crime. This seems counterintuitive as we would expect that more vacant units would increase crime. However, this variable could make sense because a high number of vacant units does not necessarily mean run-down or abandoned units – perhaps this number reflects a high number of developments being built that do not have residents yet, or the city has a stricter code for what they deem as safe for a citizen to occupy. We would have to further investigate and maybe break down this feature to better understand the relationship.

If we had to choose an optimal tree model, we would go with Model 3 to take to our stakeholders. It is worth mentioning that had we allowed for a deeper tree or fewer minimum observations in the leaves, we could have increased our prediction accuracy even more.

Final Model Selection

For our final model, we selected based on using mean-squared error on the test set* as the evaluation metric. It was a linear regression model that seemed to have the best fit:

ViolentCrimesPerPop = 2.2961 + 0.0936 * **PctKidsBornNeverMar** – 0.0147 * **PctFam2Par** – 0.0108 * **PctWorkMom** + 0.0016 * **PctStateEduFnding** + 0.0101 * **PctVacantBoarded** – 0.0107 * **PctEmplProfServ**

*Note: ideally we would have split our data into train, test, and validation sets and used the validation set performance to choose between linear regression and decision tree. However, we made the decision to forgo this split as we felt this would have reduced our datapoints in each set to too few.

On top of having the best robust performance on MSE, the linear regression model has the benefit of being very easy to interpret, explain, and predict. We can gleam lot of insights simply by looking at the direction and magnitude of the coefficients, and they are easy for non-technical audiences to understand as well. Additionally, as a simple linear equation is output, predictions on future data are quite easy to make with a simple spreadsheet or calculator. Governments often lack the more advanced technology necessary to predict on more complex models.

There are of course a few drawbacks to our final model. One, it is hard to resolve the issue of multicollinearity when explanatory variables exhibit correlation amongst themselves. Also, a linear regression is very sensitive to outliers and prone to overfitting, especially when there are fewer observations. We could improve on this model by adding a regularization term, such as lasso, ridge, or ElasticNet.

Business Recommendations:

Disclaimers:

A few important notes must be discussed before we outline our contextual interpretation of these variables.

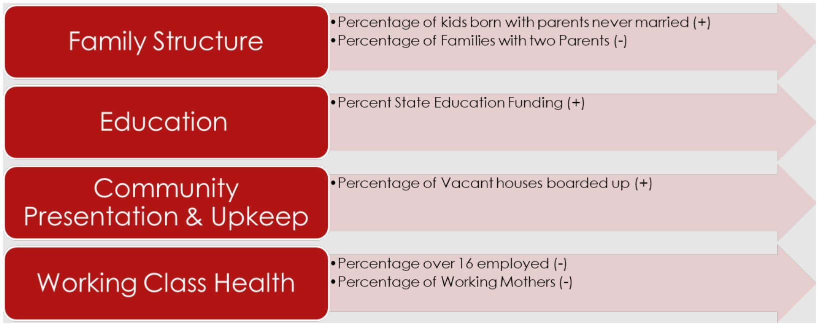
We must emphasize the word "Inform", understanding the relationships of influence between explanatory variables can only help you choose a policy or program based on the desired effects. The study of the influence of different policies and programs is outside of the scope of our project, though it would serve as interesting additional research.

Correlation does not equal causation. Our project is not trying to evaluate causation. Therefore, our business suggestions do not imply that.

Professional understanding and context are required to accurately interpret the model and come up with the best solutions. Though our group has performed extensive research, context can only fully be understood by an expert in the field of government policy, political science and social reform. No action should be taken without expert consultation and further research.

Interpretation:

As stated previously, our focus of this project was on creating a tool for governments to use to inform their policy making on factors that could influence communities and thereby crime rates. Based on our final model, we designated six variables to be the most important determinants of crime. We broke down our six variables into four broader categories that we believe policy should be focused on.



Family Structure
- Percentage of kids born with parents never married (+)
- Percentage of Families with two Parents (-)

Education
- Percent State Education Funding (+)

Community Presentation & Upkeep
- Percentage of Vacant houses boarded up (+)

Working Class Health
- Percentage over 16 employed (-)
- Percentage of Working Mothers (-)

Family Structure

While the government can't mandate family structure it can influence it, especially at the local level. Here we would recommend increasing funds into planned parenthood, and sexual education. Why? The idea has a simple premise, reducing early life pregnancies, reduces the likelihood that children are born into a single parent, or joint custody situation. Delaying the age at which an individuals have children also increases the odds of the individual being financially stable and being better able to support a child, “Teenagers, unmarried women, and the economically disadvantaged … are those most at risk to give birth to children who would engage in criminal activity”(Donohue & Levitt 2020).

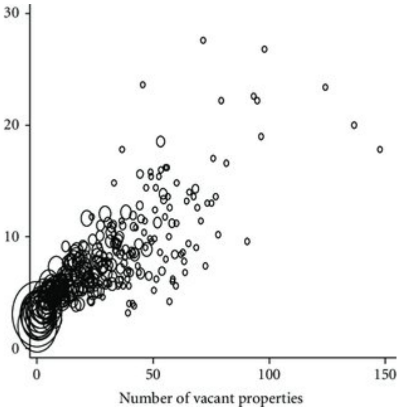
Education

As mentioned, correlation does not equal causation. We notice that our model has a positive relationship between State Education Funding and Crime. Areas with high fraction of school funding coming from state and federal resources most likely already have high crime rates, as Noghanibehambari discusses, “Several states-initiated school finance reforms ...the primary purpose of providing adequate funding for low-income school districts” (2022). Our suggestion is for local governments to emphasize programs that try to increase public buy-in to the community. Creating the idea of investing in themselves and establishing the desire to give back and build stronger bonds amongst their local peers. In affluent communities you will often notice high levels of private funding, the goal is to augment state funding to further the impact.

> “I find that exposure to reform reduces the juvenile arrest rates, increases the likelihood of high school graduation, increases the time spent on educational activities, and reduces risky behaviors at schools. A 10%% rise in real per-pupil spending is associated with 7.4 fewer arrests per 1,000 in the population aged 15–19”(Noghanibehambari,2022).

Community Presentation & Upkeep

Our model found the housing vacancy rates had a positive relationship with violent crime. Housing vacancies could contribute negatively to a community in multiple ways. A study performed by Case Western Reserve University, in Cleveland, discusses some of these affects, ”... Clusters of vacant homes can sometimes become an unguarded location for illicit activity, or a signal of social disorder and vulnerability to potential criminals”(2017) Further research confirmed this relationship, in a study done in Philadelphia, “Between block groups, the risk of aggravated assault increased 18% for every category shift of vacant properties (IRR 1.18, 95% CI: 1.12, 1.25, P < 0.001)”(Branas 2013). We propose emphasizing policy to clean up the presentation of communities and re-purpose vacant spaces.



Scatter plot showing the Unadjusted association between vacant properties and aggravated assaults across 1816 block groups in Philadelphia County between 2002-2006 * . * Area of circles proportional to the number of block groups with level of vacant properties. (Branas 2013).

Working Class Health

Both explanatory variables we modeled in this category had a negative relationship with violent crime rates. However, we found that historically, the subject of maternal employment has been a heavily debated subject, even since the great depression. Ven & Cullen, discussing some of controversial aspects of this topic, state, “Indeed, the belief that maternal employment is related to youth crime is fairly pervasive... (a study) found that 40% of their survey respondents believed youth were more likely to suffer some sort of negative consequences, including delinquent involvement, if their mothers worked.” (2004). This paper did a detailed analysis however, reviewing historical studies and performing multiple regression on an assortment of data. Their findings were intriguing, concluding that maternal employment in the paid workforce had no impact on youth crime, regardless of hours (Ven & Cullen 2004). A fascinating conclusion they drew speaks to the complexity of the question of the origins and influences on crime. Explaining these complexities, Ven & Cullen pose, “It appears that it may be coercive, unsatisfying, and low-paying maternal employment, and not employment per se, that may be criminogenic” (2004). Based on these findings, and our own, we recommend emphasizing policy and programs that create work opportunities and training for single mothers, and young teenagers. Thereby stimulating the local economy, creating role models in the family structure and empowering women, and eliminating downtime of youth. As the study discusses, it is imperative to make these opportunities intriguing and meaningful to have positive impacts.

Challenges Faced

Preliminary Data Processing:

Though much of the data that we were interested in was publicly available, our team was overly ambitious in what we wanted to accomplish and include. Though we accomplished a significant amount of quality work, learning about the problem and gathering additional data sources; the data was quite messy. This led to us spending substantial time cleaning and processing the data. Specifically joining the data and normalization. Despite the project instructions specifically stating to emphasize focus on the modeling, we found the conventional criterion of “80% of data scientists time is spent on data preparation and cleaning” to hold true and unfortunately be necessary. Working with older data compounded this problem.

Theoretical Limitations:

As mentioned above, we were unable to gather overlapping data from the 1995 timeframe. We cannot quantify the significance of this in our output. However, it must be acknowledged. Specifically, this affected us in terms of zip-codes and planned parenthood data. Our zip-code dataset used to clean and join our data together was modern. We discovered in our research, as one would logically conclude, that overtime zip-codes have been added or decommissioned. Our planned parenthood data was modern as well. As we unveiled in our research, to understand the effect of social programs on a specific period, you must sometimes understand the state of social programs and policy at a times significantly before the period in which you are analyzing (Donohue & Levitt 2020). This can be due to the impact that such social programs might have had on individuals when they were children, or an earlier point in their life when they were exposed to the program.

Recommended Augmentation:

Many tertiary subjects and interesting questions arose during our project. In order to get full context and make an informed decision on policy surrounding this issue, one should consider doing further research in the following areas:

- Methods for establishing strong community bonds
- Factors influencing the likelihood of a couple having a child getting married
- Impact of vacant homes on property value
- Highschool Work Study & Apprenticeship Programs Effectiveness
- Creating meaningful working opportunities and training programs in low-income areas

Conclusion:

We encountered many different challenges throughout the project. It revealed to the team the real level of detail and effort that must be put into such a complex problem of this nature. Our project can only serve as a template for another to dedicate more time with extensive research and data vetting to create an effective, informed approach to political policy decisions.

Research & Work Cited

Our group attempted to conduct background research on our hypothesized and conclusive variables to better understand the intricacies of the problem. Some sources were not referenced and therefore not cited but studied for context and framing of the problem.

Abdul Jalil, M. @, Mohd, F., & Mohamad Noor, N. M. (2017). A comparative study to evaluate filtering methods for crime data feature selection. Procedia Computer Science, 116, 113–120. https://doi.org/10.1016/j.procs.2017.10.018

Branas, Charles & Rubin, David & Guo, Wensheng. (2013). Vacant Properties and Violence in Neighborhoods. ISRN public health. 2012. 246142. 10.5402/2012/246142.

Center on Urban Poverty and Community Development. (2017, June 5). *Exploring the relationship between vacant and distressed properties and ...* Case Western Reserve University. Retrieved November 18, 2022, from https://case.edu/socialwork/healthcounseling/sites/case.edu.socialwork/files/2018-10/vacant_distressed_props_comm_health_safety.pdf

Durlauf, S. N., Navarro, S., & Rivers, D. A. (2010). Understanding aggregate crime regressions. Journal of Econometrics, 158(2), 306–317. https://doi.org/10.1016/j.jeconom.2010.01.003

John J. Donohue, Steven D. Levitt, The Impact of Legalized Abortion on Crime over the Last Two Decades, American Law and Economics Review (2020)

MARKOWITZ, F.E. (2006), PSYCHIATRIC HOSPITAL CAPACITY, HOMELESSNESS, AND CRIME AND ARREST RATES. Criminology, 44: 45-72. https://doi.org/10.1111/j.1745-9125.2006.00042.x

Hamid Noghanibehambari, School Finance Reforms and Juvenile Crime, American Law and Economics Review, Volume 24, Issue 1, Spring 2022, Pages 1–86, https://doi.org/10.1093/aler/ahac001

Stephen Machin, Olivier Marie, Sunčica Vujić, The Crime Reducing Effect of Education, The Economic Journal, Volume 121, Issue 552, May 2011, Pages 463–484, https://doi.org/10.1111/j.1468-0297.2011.02430.x

Ven, T. V., & Cullen, F. T. (2004). The Impact of Maternal Employment Serious Youth Crime: Does the Quality of Working Conditions Matter? Crime & Delinquency, 50(2), 272–291. https://doi.org/10.1177/0011128703253165