Socioeconomic Influences On Gun Violence

Stat 471 Final Project

Alex Rand and Patrick McNally

https://github.com/alexrand18/stat-471-final-project

December 19, 2021

Executive Summary

Gun violence is a national public health epidemic that exacts a substantial toll on U.S. society. Despite increasing efforts by the U.S. government and local policymakers, 2021 is on pace to be the worst year for gun violence in decades, surpassing even the high levels last year.[1] Therefore we decided to look into various socioeconomic factors across US counties and analyze which variables were most predictive of fatal gun incidence rates from 2013 to 2018.

Our dataset pulled data from two sources, the first of which being a GitHub repository containing all violent gun incidents from January 1, 2013 until March 3, 2018. Our second data source was a dataset called `county_complete` from the R package `tidycensus` that consists of 188 socioeconomic and demographic variables for each of the 3142 counties in the United States. We realized that many of the rows in the gun violence dataset did not use the name of the county as the location. Instead, we used the latitude and longitude coordinates in each row to fetch the FIPS code for the county in which the incident happened, which required the use of the government's publicly available geolocation API. Our explanatory variables span various measures of socioeconomic status including the age, ethnicity, and income level of each county among other variables. Our primary response variable of interest is the cumulative annual violent gun incident rate per capita, calculated as the average number of violent gun crimes per year from 2014 until 2017 divided by the population of the county.

Before exploring our data or creating any models, we split our data into a training dataset and a test dataset, using the test dataset to assess and compare the performance of our models. We then explored our data to evaluate the correlations between our explanatory variables as well as between variables and the response. To create the model with optimal predictive performance, we then built eight different cross-validated models: intercept-only, ordinary least squares, ridge regression, LASSO regression, elastic net regression, regression tree, random forest, and boosting. Of the regression models, ordinary least squares had the lowest test error and of the tree-based models, the random forest model had the lowest test error. The random forest had the lowest test error overall as well.

Unsurprisingly, we found that the ordinary least squares regression and random forest both signified similar variables as the strongest predictors of violent gun incidence rates. Our random forest model denoted that the variables related to the poverty level of a county as well as the percentage of the population that was African American were the most significant predictors. We hope our analysis can inform and improve social and economic policies aimed at improving the gun violence epidemic in the country today.

[1] https://www.cnn.com/2021/09/19/politics/gun-violence-spike-2021-explainer/index.html

Introduction

More so than any other industrialized country in the world, the United States faces a severe gun violence problem in major cities and rural suburbs. In this country there are 120.5 firearms per 100 people, more than twice the amount of the second highest country in the world for the same statistic (Yemen)[2]. Incidents like Columbine, Parkland, and most recently Oxford, Michigan have grabbed the nation's attention, spurring political discourse about the need for the Second Amendment and whether or not gun laws should be made stricter. In 2020 alone, there were 40 "active shooter" cases in the United States, which is 37 more than the respective count at the turn of the century[3]. While most citizens are aware of the gun violence epidemic and how costly it is both fiscally ($229 billion or 1.2% of GDP) and emotionally for the nation, they aren't as aware about how the social fabric and different racial, demographical, and economic factors affect violent gun rates[4].

Every day, of the 316 people shot in the United States, black people are 2 times more likely than the next highest race to be involved in a fatal gun incident. What about these communities make African-Americans more likely to be involved in a violent firearm encounter? There aren't many studies that go beyond the statistics and surface-level measurements relating gun crimes and demographics/socioeconomic features, so our investigation aims to uncover a deeper relationship between such factors and gun crime[5]. More specifically, we will analyze which socioeconomic and racial features have the highest impact on violent gun incident rates, calculated as the total number of violent gun incidents annually per capita in a county. Furthermore, from using only socioeconomic and demographic measurements at the county level, we will aim to predict the gun violence rates in such areas. To measure success, we will compare our models against simply guessing the average violent gun incident rate for each county and extract which features our models used with the most significance to predict gun violence rates.

By analyzing which features are most predictive of gun violence, we hope to shed more light on why some communities are more adversely affected by this epidemic. Our results should help suggest ways for local governments to decrease crime by illustrating which factors are negatively correlated with gun violence and where tax dollars should be allocated (i.e education, homeownership) to minimize intemperate firearm usage.

[2] https://www.bbc.com/news/world-us-canada-41488081
[3] https://abcnews.go.com/US/america-gun-violence-problem/story?id=79222948
[4] https://health.ucdavis.edu/what-you-can-do/facts.html
[5] https://phys.org/news/2020-01-gun-violence-socioeconomic-root.html

Data Description And Exploration

Data Sources

Our datasets were pulled from two independent sources, the first of which being a GitHub repository containing all violent gun incidents from January 1, 2013 until March 3, 2018[6]. To construct their dataset, Github user jamesqo used Python scripts to query the Gun Violence Archive API and fetch all incidents, as the data itself was not published in a single location. The user notes that the data from 2013 is incomplete, as only 279 incidents were cataloged during this year. In addition, the infamous Las Vegas Massacre was omitted from this dataset, as the information about this shooting was incomplete and could not be directly imported into the larger dataset.

We merged the data from this source with a dataset called `county_complete` from the R package `tidycensus`. This dataframe consists of 188 socioeconomic and demographic variables for each of the 3142 counties in the United States. Per the documentation, this information was originally compiled by the USDA, Economic Research Service, Bureau of Labor Statistics, and the Census Bureau. The measurements themselves are annual estimates primarily from the years 2010 until 2019, so some of the data is obsolete from the perspective of the gun violence dataset. We downloaded both datasets in the form of csv files from the respective sources.

Data Cleaning Process

To clean the violent gun incident dataset, we first had to remove rows with incomplete data about the location of the event. Each row had the name of the location in which the incident occurred, but we realized that this data was insufficient to merge with the county statistics dataset because many of the rows did not use the name of the county as the location. Instead, we used the latitude and longitude coordinates in each row to fetch the FIPS code for the county in which the incident happened, which required the use of the government's publicly available geolocation API. We discovered that some of the latitude and longitudes were cataloged incorrectly (some were completely outside of the United States), so we had to drop these additional rows. Since there were over 230,000 violent gun incidents in the dataset, fetching the FIPS codes took many hours to complete. Therefore, we broke the data into smaller chunks, and stored the intermediate results in smaller csv files before recombining the data after fetching all county codes. Finally, we extracted only the incidents from 2014 until 2017, as the data from 2013 was incomplete and that data from 2018 only spanned until March.

Cleaning the county level dataset was a much quicker process. Since the incidents from the gun dataset were pulled from 2014 to 2017 primarily, we extracted the variables for each measurement that were calculated closest to our desired time window. For example, there were measurements for per capita income in 2010 and 2017, so we only kept the latter value. Additionally, to get an estimate for the population in each community, we averaged the

[6] https://github.com/jamesqo/gun-violence-data

population estimates from 2014 to 2017. Some of the columns that needed to be a per capita measurement, such as converting total number of employed residents to an employment rate, were then normalized using the population estimate. After dropping rows from the county data with NA values, we merged the two datasets on the FIPS code, grouped the gun resulting dataframe by FIPS code, and calculated the total number of gun incidents per county and the resulting gun incident rate per capita.

Data Description

Observations

There are 2813 observations in our final dataset corresponding to each of the counties that had non null measurements for each of the 17 socioeconomic/demographic variables that we included. In total, there are 18 columns in our dataset, one for our response variable and 17 explanatory variables.

Response Variable

Our response variable is the cumulative annual violent gun incident rate per capita, calculated as the average number of violent gun crimes per year from 2014 until 2017 divided by the population of the county. We calculated this continuous metric using a simple mutate call.

Explanatory Variables

Using the data compiled from the Census Bureau, USDA, Bureau of Labor Statistics, and the Economic Research Service, we extracted 17 explanatory variables describing the demographic and socioeconomic status of each county. The names of the variables were `age_under_5_2017`, `age_over_65_2017`, `median_age_2017`, `white_2010`, `black_2017`, `no_move_in_one_plus_year_2010`, `speak_english_only_2017`, `hs_grad_2017`, `bachelors_2017`, `computer_2017`, `homeownership_2010`, `per_capita_income_2017`, `poverty_2016`, `employed_2015`, `uninsured_2017`, `fed_spending_2009`, and `density_2010`. A detailed description of each of these continuous variables can be found in the Appendix.

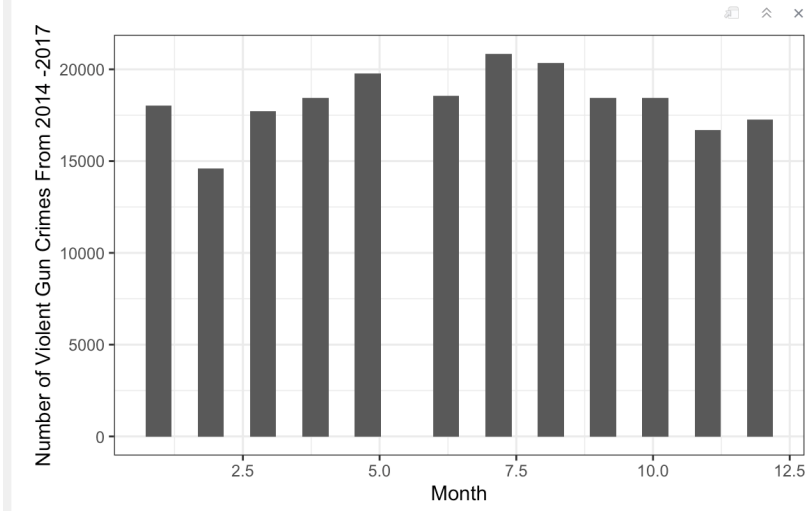
Allocation for Training and Testing

We used the standard 80/20 rule, allocating 80% of the data for training and 20% of the data for testing. To ensure that all models used the same training and test datasets so we could compare the models most accurately, we wrote the train and test datasets to their own files and imported the datasets in each of the scripts for each model.

Exploratory Data Analysis

Independent Analysis

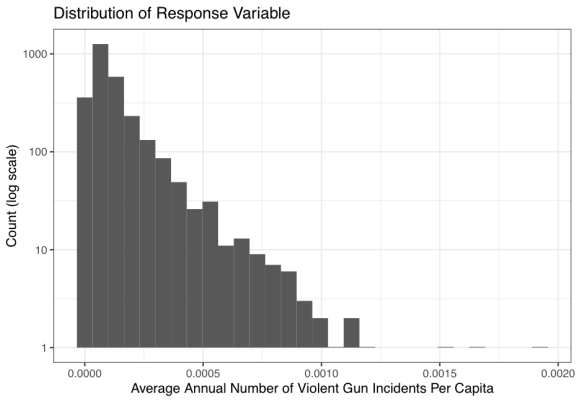
Before diving into the response variable and explanatory variables, we first investigated which counties had the most citizens die as a result of gun crime. Referring to the below tables, major cities like Cook County (Chicago), Los Angeles, Harris County (Houston), and Philadelphia face the largest gun violence problem in terms of total deaths.We also wanted to analyze seasonal patterns of gun crime and whether or not rates fluctuate based on the month of the year. We found that there wasn't much deviation between months, but there was marginally more gun violence in the summer than the winter months.

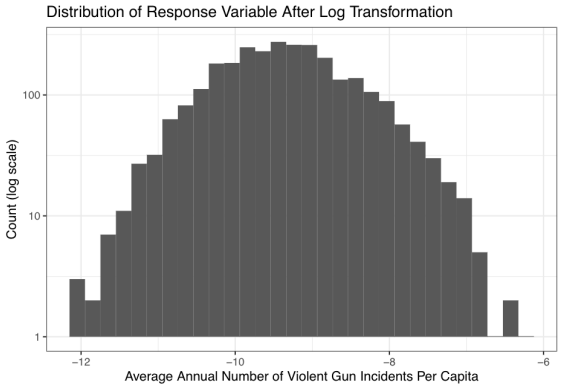


State	County	Total Gun Deaths From 2014-2017
Illinois	Cook County	2334
California	Los Angeles County	1724
Texas	Harris County	1331
Maryland	Baltimore city	996
Pennsylvania	Philadelphia County	854
Missouri	St. Louis city	769
Arizona	Maricopa County	721
Michigan	Wayne County	694
Louisiana	Orleans Parish	665
Nevada	Clark County	627

Response Variable

We then analyzed the distribution of our response variable and whether or not we should transform it so that our data had less outliers and was more normally distributed. The distribution of the response variable pre transformation (left) was highly skewed right, with a median of 0.0000828 and a variance of 0.0000000205. After computing a log transformation of the violent gun incident rate per capita, the distribution was more normal with a median of -9.398 and a variance of 0.7636.



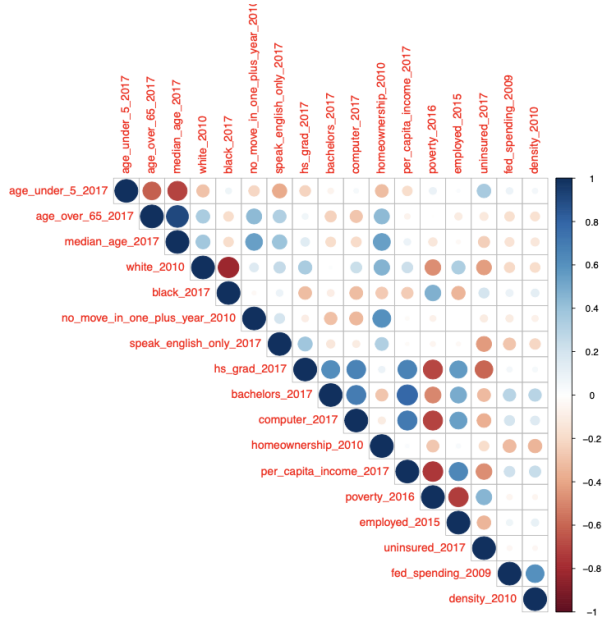


The counties with the top 10 average annual number of violent gun incidents per capita, as shown below, contained major cities like Baltimore and St. Louis. However, smaller counties like Lauderdale, Mississippi, Muscogee, Georgia, and Natrona, Wyoming, and even a county in Alaska also fall in the top 10. This confirms that rural areas also face a violent gun usage problem.

State	County	Incidents Per Capita Per Year
Louisiana	Orleans Parish	0.0019285600
Missouri	St. Louis city	0.0016794624
Maryland	Baltimore city	0.0015219944
Illinois	Peoria County	0.0012084568
Mississippi	Lauderdale County	0.0011516031
District of Columbia	District of Columbia	0.0011387098
Alaska	Lake and Peninsula Borough	0.0010829208
Georgia	Muscogee County	0.0010165059
Wyoming	Natrona County	0.0009937904
Virginia	Richmond city	0.0009545757

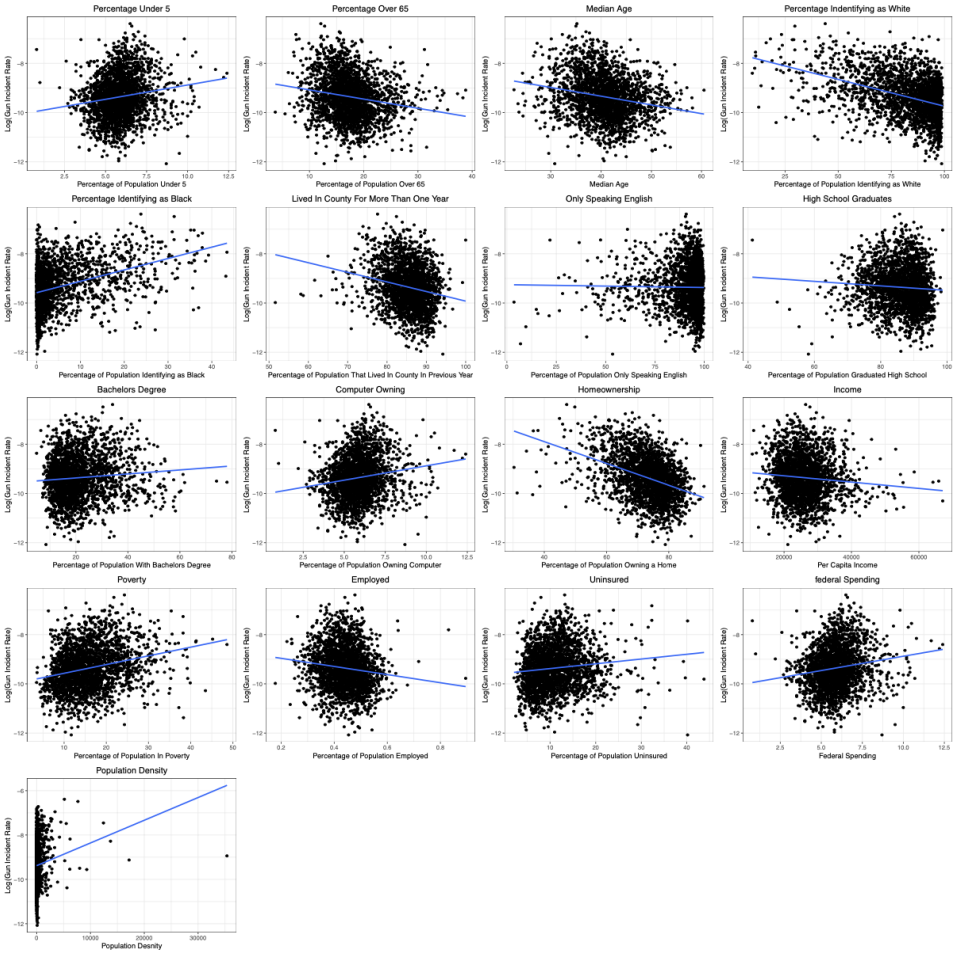
Explanatory Variables

Next, we analyzed the correlation between the explanatory variables, as well as their individual variances and relationships with the response variable.



From the correlation grid above, we observed that the percentage of the population under 5 and the population over 65 were unsurprisingly negatively correlated and had opposite

relationships with homeownership, insured percentages, and the proportion of the population that only speaks english. Similarly, the percentage of the population that was black versus the percentage of the population that was white was also negatively correlated. Higher rates of high school graduation and people with bachelor degrees were negatively correlated with poverty and the percentage of the population that was uninsured, and they were positively correlated with income, computer ownership, and employment rates. The percentage of the population that is black is negatively correlated with homeownership, education, and income, while the percentage of the population that is white has positive correlations with these same variables.



Using the above plot, we found that percentage of population under 5, percentage of black population, percentage uninsured, federal spending, and population density are all positively correlated with the response variable in isolation. On the other hand, percentage of population over 65, white population, percentage of population that has recently moved into the county, homeownership, and employment are all negatively correlated with the response variable. An initial insight is that variables relating to race and wealth are most strongly correlated with violent gun incident rates.

Modeling

Regression Methods

Ordinary Least Squares Regression

The first model we constructed to predict the gun incident rate per capita was ordinary least squares regression using all 17 explanatory variables. The summary of the model can be found below. Overall, the model was able to explain 27.34% of the variation in the response variable. The variables with statistically significant P values were `age_under_5_2017`, `white_2010`, `black_2017`, `speak_only_english_2017`, `bachelors_2017`, `computer_2017`, `homeownership_2010`, `per_capita_income_2017`, `poverty_2016`, and `fed_spending_2009`. The sign of each of these variables align with the direction of the relationship between the variable and the response variable in isolation. When the trained model predicted the test set, the resulting RMSE was 0.8134264, which is about 10% less than the RMSE of the intercept only model (0.907).

```
##
## Call:
## lm(formula = log_incident_rate ~ ., data = gun_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.62099 -0.47820  0.02348  0.49887  3.03927
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -1.111e+01  8.303e-01 -13.381  < 2e-16 ***
## age_under_5_2017                1.093e-01  2.181e-02   5.010 5.87e-07 ***
## age_over_65_2017               -1.119e-02  1.029e-02  -1.088 0.276884
## median_age_2017                 3.508e-02  1.036e-02   3.387 0.000718 ***
## white_2010                     -7.583e-03  2.379e-03  -3.187 0.001455 **
## black_2017                      2.087e-02  4.856e-03   4.298 1.80e-05 ***
## no_move_in_one_plus_year_2010 -3.232e-03  5.311e-03  -0.609 0.542912
## speak_english_only_2017         9.109e-03  2.417e-03   3.769 0.000168 ***
## hs_grad_2017                    2.249e-03  5.790e-03   0.388 0.697724
## bachelors_2017                  1.029e-02  3.760e-03   2.738 0.006229 **
## computer_2017                   1.545e-02  4.345e-03   3.555 0.000386 ***
## homeownership_2010             -2.834e-02  3.468e-03  -8.173 5.00e-16 ***
## per_capita_income_2017         -1.325e-05  6.240e-06  -2.124 0.033795 *
## poverty_2016                    2.392e-02  5.973e-03   4.004 6.42e-05 ***
## employed_2015                   4.665e-01  3.715e-01   1.256 0.209321
## uninsured_2017                 -5.073e-04  4.645e-03  -0.109 0.913050
## fed_spending_2009               1.856e-08  5.996e-09   3.095 0.001989 **
## density_2010                   -3.613e-05  1.943e-05  -1.859 0.063147 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7445 on 2232 degrees of freedom
## Multiple R-squared:  0.2734, Adjusted R-squared:  0.2679
## F-statistic: 49.41 on 17 and 2232 DF,  p-value: < 2.2e-16
```

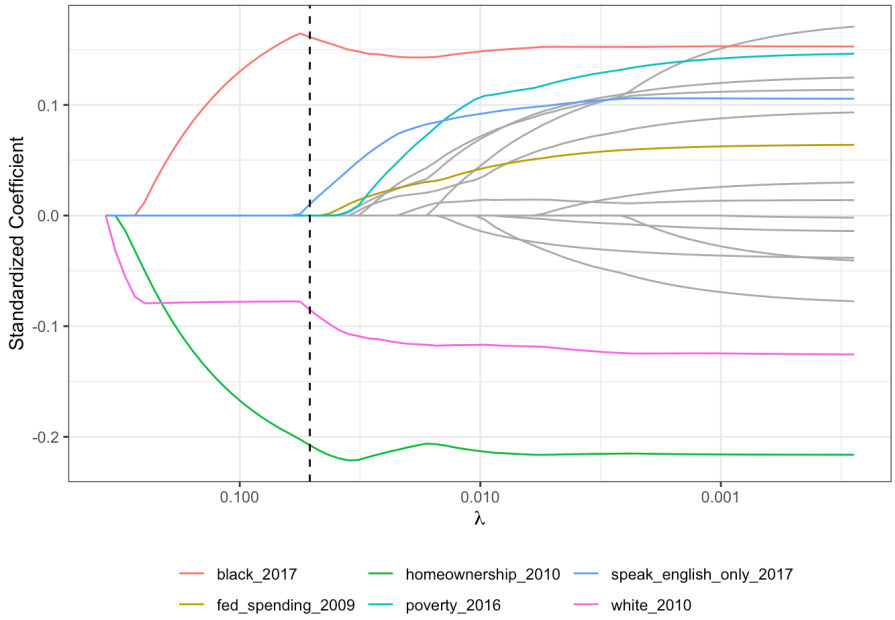
LASSO, Ridge, and Elastic Net Regression

Since the training dataset contains 17 features, we suspected that the ordinary least squares model, while having a decent RMSE, could have too high of a variance and thus imperfect predictions. Therefore, we also employed shrinkage methods to our regression models. We ran LASSO (Least Absolute Shrinkage and Selection Operator), Ridge, and Elastic Net Regression models, and for each model we used cross validation and chose the respective shrinkage factor based on the one standard error rule[7].

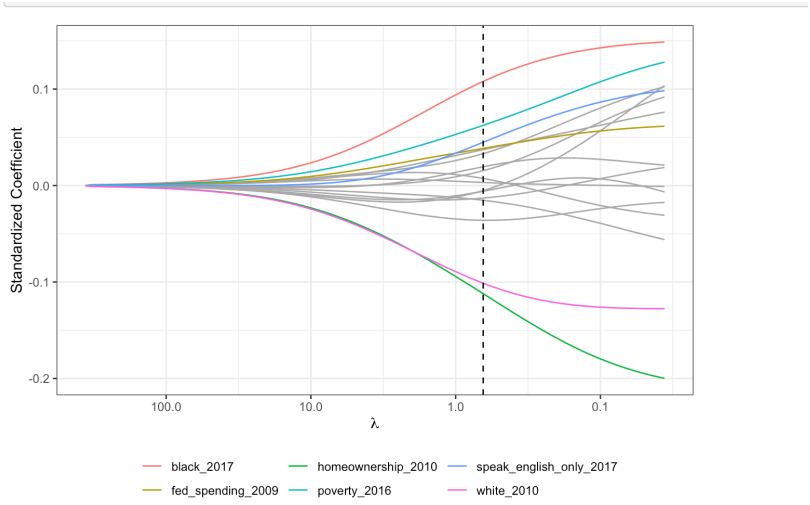
The LASSO regression model selected only 4 non-zero features to predict the response variable - `white_2010`, `black_2017`, `speak_english_only_2017`, and `homeownership_2010`. This simplified model greatly reduces the variance in the predictions

[7] See appendix for detailed CV plots for each model

on the test set; however, the RMSE fell to 0.831. Below is a plot relating the top six features and lambda, where the dotted line is the lambda value selected for our model.

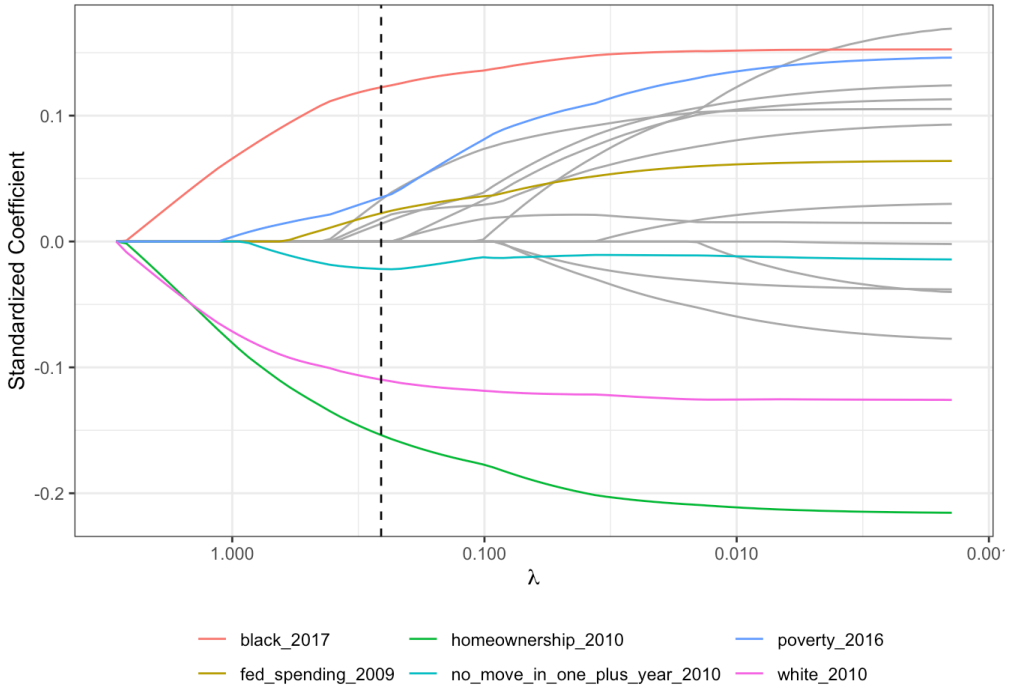


The tuned ridge regression model's top features were identical to the LASSO model. Similarly, it's RMSE was higher than the ordinary least squares model (0.825), but the variance in its predictions was less since the coefficients for the explanatory variables are closer to zero.



Finally, the optimized elastic net model performed worse than the ridge model but better than the LASSO model with an RMSE of 0.827547. It selected nine total non-zero coefficients, including the 4 involved in the LASSO model as well as `age_under_5_2017`, `no_move_in_one_plus_year_2010`, `bachelors_2017`, `poverty_2016`, and `fed_spending_2009`. The model had the same top 6 features in the feature importance plot

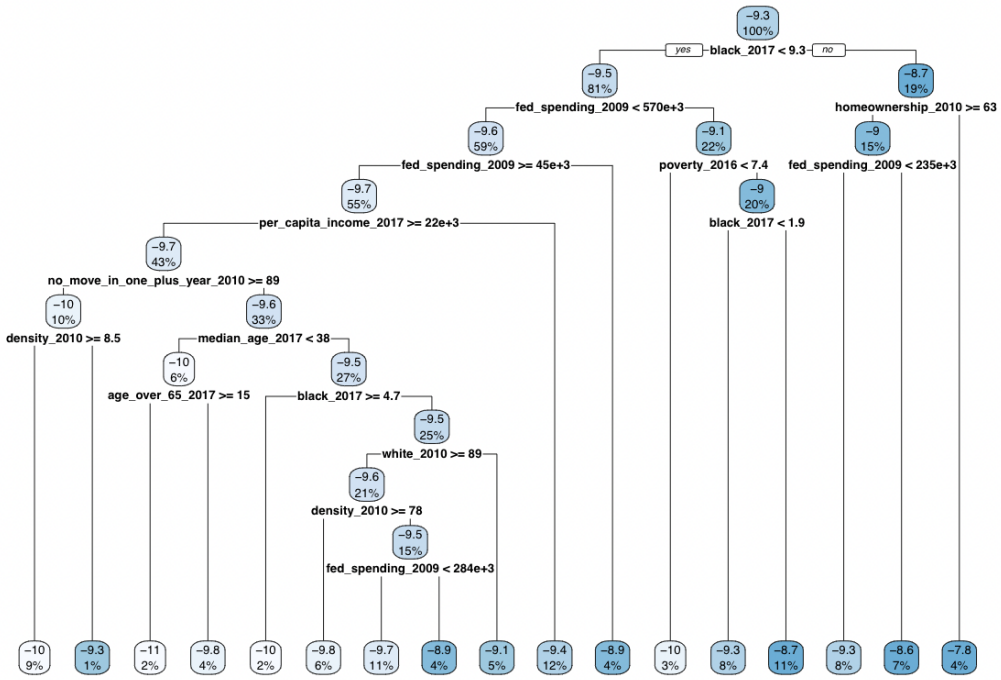
as the ridge and LASSO models, except it included `no_move_in_one_plus_year_2010` instead of `speak_english_only_2017`.



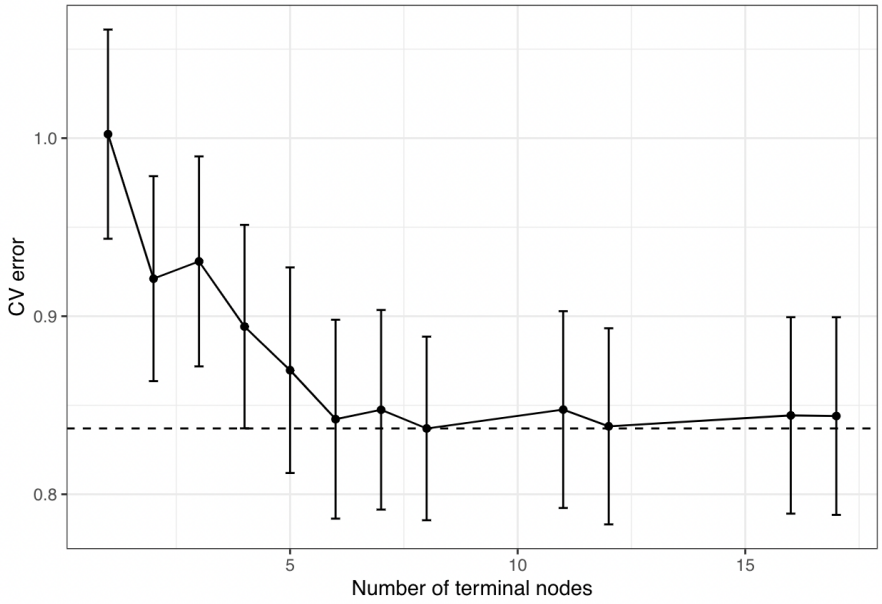
Tree-Based Methods

Regression Tree

We then proceeded to fit a regression tree to the data. We initially fit our regression tree to the data using the default parameters in the r.part function, where minsplit is equal to 20, minbucket is equal to minsplit/3, maxdepth is equal to 30, and cp is equal to 0.01. Minsplit is the minimum number of observations that must exist in a node in order for a split to be attempted, and minbucket is the minimum number of observations in any terminal node. The larger these numbers, the fewer nodes there will be in the tree and the more. The maxdepth parameter prevents the tree from growing past a certain depth. By setting the max depth of our regression tree equal to 30, we allow for our model to become very complex and thus possibly overfitting the data.

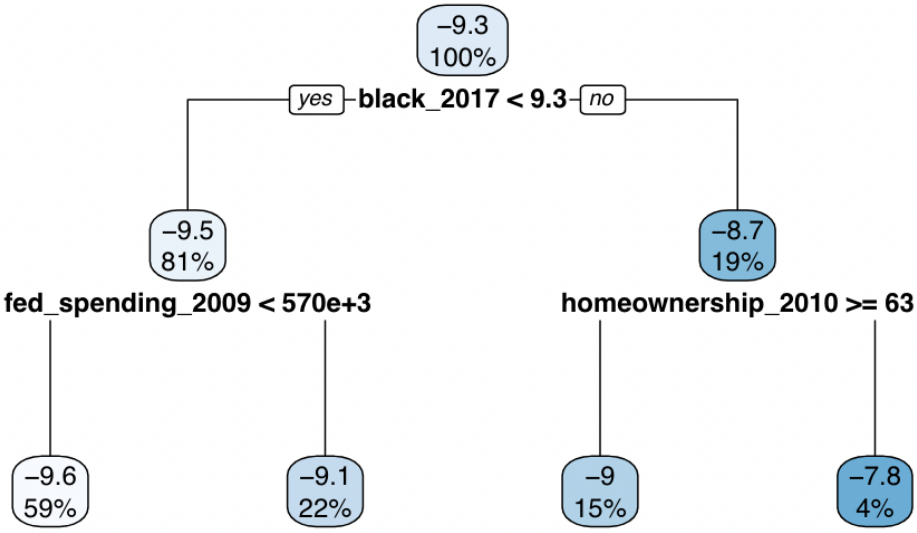


To combat a model with high variance that overfits our data, we pruned our original regression tree by plotting the cv error as a function of the number of terminal nodes and choosing the complexity parameter associated with the optimal model chosen based on the 1 standard error rule.



```
## # A tibble: 1 x 5
##       CP nsplit `rel error` xerror   xstd
##    <dbl>  <dbl>       <dbl>  <dbl>  <dbl>
## 1 0.0359      3       0.780  0.909 0.0589
```

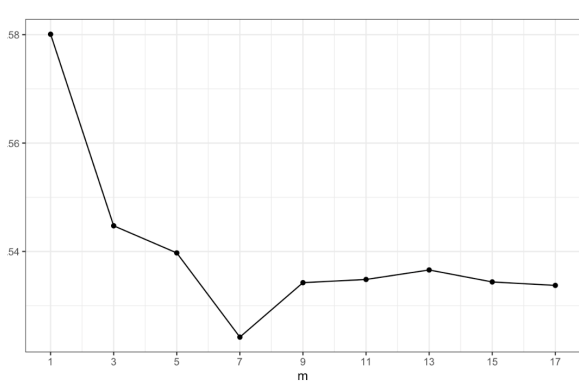
We found that the optimal regression tree makes only 3 splits and has 4 terminal nodes.

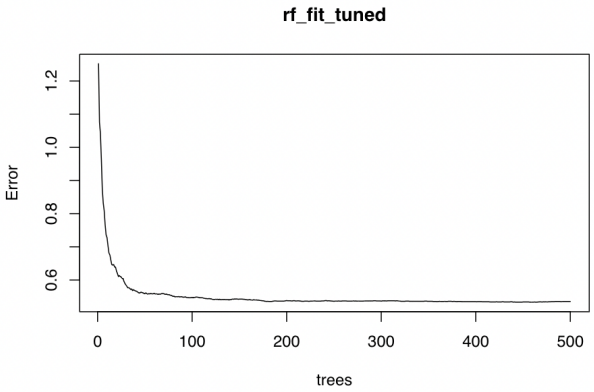


This is a much smaller tree than the original decision tree created using the default parameters. In the bias-variance tradeoff, we see that in our tree less complexity has greater predictive power. Making predictions on the test data, we see that the RMSE for our regression tree was 0.8009077.

Random Forest

It is straightforward to see that regression trees are an example of a model with low bias and high variance. The tree makes almost no assumptions about the target function but it is highly susceptible to variance in data. Therefore we fit a random forest to the data. While bagging is performed when all 17 explanatory variables are considered at each tree split, leading to a suboptimal predictive power and higher variance in our model, we tuned the random forest model for the optimal value of m, or the number of features to consider at each tree split, by training the model on different values of m, ranging from 1 to 17. A plot consisting of the out of bag error (OOB error) for each value of m can be seen below. We observe that the out of bag error is minimized when m = 7, after which the OOB error follows a roughly upward trend. We then tuned the B parameter which controls the number of bootstrapped samples. As shown in the plot on the right below, the cross validated training error initially decreases sharply as the number of trees increases and then eventually lateuas starting around B = 200.

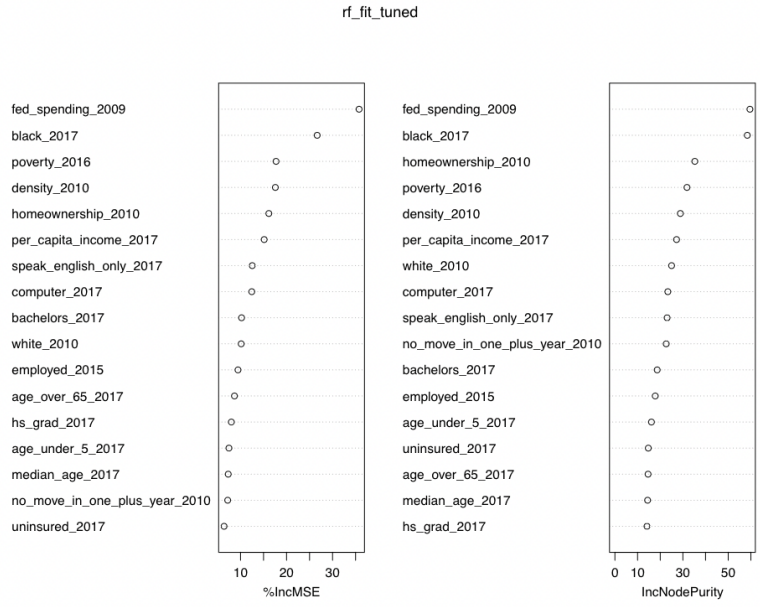




We then fit our tree using the m and B parameters specified above and assessed variable importance. When the trained random forest predicted the test set, the resulting RMSE was 0.720, which is about 10% less than the RMSE of the regression tree model (0.819).

In the context of random forests, there are two measures of importance given for each variable: purity based importance and OOB variable importance. Purity based importance is a measure of the degree of improvement in node purity that results from splitting on a given feature. OOB variable importance is a measure of the reduction in prediction accuracy that results from scrambling a given feature out of bag. The results of both of these variable importance measures in our random forest model are summarized below.

OOB and Purity Variable Importance:

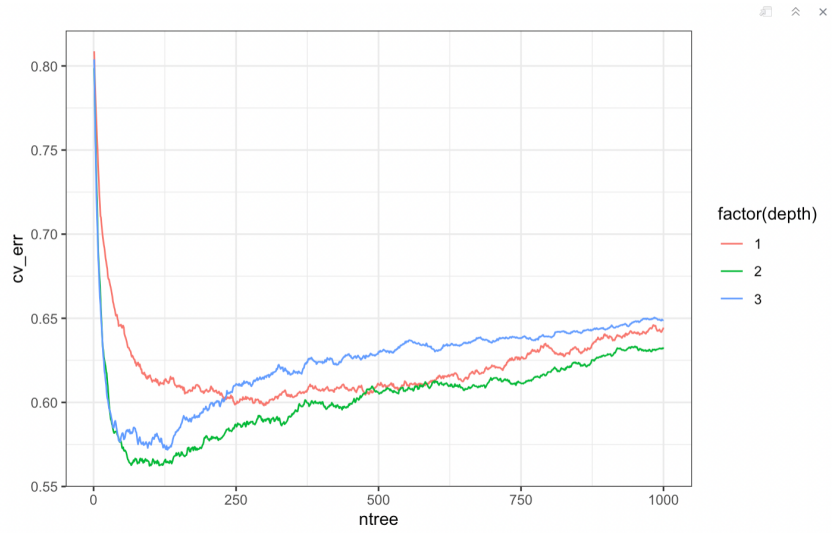


Looking at the purity based importance and OOB Error improvement plots above, we observe that `fed_spending_2009` and the `black_2017` variable measuring the percentage of the population in the county that is black have the highest importance as measured by both

metrics. The rest of the variables have much lower value for importance by both metrics suggesting that these two variables are most important in predicting the gun incident rate per capita across different counties in the US.

Boosting

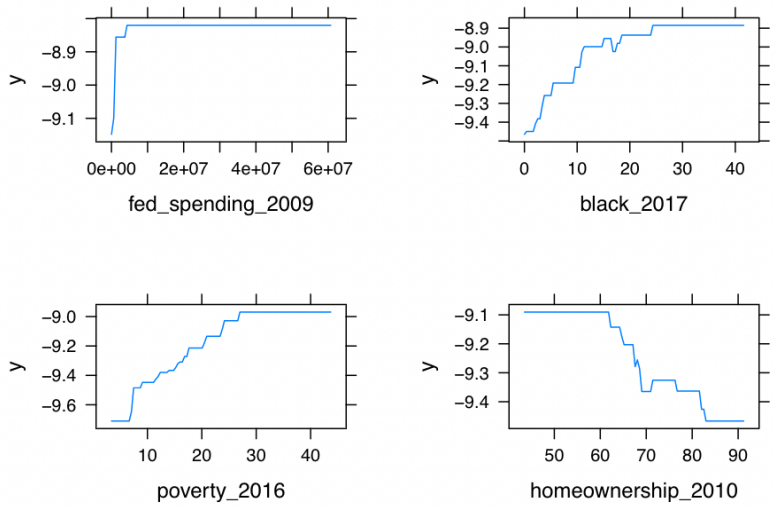
Unlike many ML models which focus on high quality prediction done by a single model, boosting algorithms seek to improve the prediction power by training a sequence of weak models, each compensating the weaknesses of its predecessors. As a result, we proceeded to fit our data using a boosted model. We began creating our boosting model using the default parameters: 100 trees, a shrinkage factor of 0.1, an interaction depth of 1, and a subsampling fraction π of 0.5. To optimally tune our boosted model, we proceeded to fit models with a differing number of trees as well as differing interaction depths. We found that the optimal number of trees is typically between 100 to 250 across a number of different model parameters. We then tuned the interaction depth of the model, testing interactions depths of 1, 2, and 3. As per the CV plot shown below, we can see that the model with an interaction depth of 2 attains the minimum cross validated error at 100 trees. When the optimally tuned boosting model predicted the test set, the resulting RMSE was 0.730, which is slightly greater than the RMSE of the random forest model model (0.720) and about 10% less than the RMSE of the regression tree model (0.819).



Using our tuned boosted model we derived above, we then assessed how important each feature was in the construction of the boosted decision trees within the model by two measures of variable importance: purity-based importance and partial dependence plots. We first found the results of purity-based importance and ranked the top 10 variables based on their contributions to improvements in node purity.

	var <chr>	rel.inf <dbl>
fed_spending_2009	fed_spending_2009	19.469213
black_2017	black_2017	13.611235
poverty_2016	poverty_2016	10.009729
homeownership_2010	homeownership_2010	8.339245
speak_english_only_2017	speak_english_only_2017	7.138818
per_capita_income_2017	per_capita_income_2017	5.627203
computer_2017	computer_2017	5.185453
density_2010	density_2010	5.157430
employed_2015	employed_2015	4.963160
age_over_65_2017	age_over_65_2017	3.757071

We then produced partial dependence plots for the most important four variables: `fed_spending_2009, black_2017, poverty_2016,`and `homeownership_2010.`



Looking at the top-left plot, as the absolute amount of fed spending in 2009 increases from 0, we see that the log_incident_rate drastically increases. This suggests that the amount the federal government spent in each county could be a negative factor that significantly increases a community's prevalence of gun violence. For the percentage of a county that is black as well as the percentage of a county that is in poverty, the percentage of gun violence in a county increases as those percentages increase respectively. Lastly the incident rate percentage decreases as the percentage who own their home in the county decreases. Taken together, these factors suggest that counties with impoverished inhabitants and higher percentages of African American individuals might be more dangerous and face more incidents of gun violence.

Conclusion

Comparison of Method Performance

The root mean squared errors (RMSEs) of each of our models are presented below:

Model	Test RMSE
Intercept-Only	0.907
Ordinary Least Squares (Log-Transformed Model)	0.8134264
Ridge Regression	0.825
LASSO Regression	0.831
Elastic Net Regression	0.827547
Regression Tree	0.8009077
Random Forest	0.720
Boosted Model	0.730

As shown above, the random forest and the boosted model have the lowest test errors. This is reasonable given these models' tendencies to have high predictive accuracy. Between the two, the random forest model has the lowest test error, with a root mean squared error of 0.720, but it is closely followed by boosting, which has a mean squared error of 0.730. However, the ridge, LASSO, and elastic net regressions do not perform as well, with test RMSEs of 0.831, 0.825, and 0.827547, respectively. The unpenalized OLS model interestingly has a slightly smaller testing error than the other regression models with an RMSE of 0.8134264, suggesting that the reduction in variance obtained through the penalized regression models was outweighed by the bias they introduced. Also as our data exhibited a nonlinear trend, it is clear why the tree-based methods performed better than their regression based counterparts.

Regardless of these differences in test RMSE, the methods overlap significantly in their identification of important variables from the larger set. For instance, the elastic net regression selects the following variables among others, which are also selected by LASSO, Ridge, and deemed significant in the OLS model: `white_2010`, `black_2017`, `speak_english_only_2017`, and `homeownership_2010`. The random forest and boosting models both include `fed_spending_2009`, `black_2017`, `poverty_2016`, and

`homeownership_2010` in the top 5 most important variables, as measured by their contributions to node purity.

Overall Conclusions, Recommendations, and Takeaways for Stakeholders

Our results point to a few key socioeconomic and racial features that, given their impact on violent gun incident rates, policymakers should consider when aiming to improve gun control laws overall. Our results also can help us potentially better understand the root social causes of gun violence and homicides and reduce these rates. The random forest model, which had the strongest predictive performance, suggests that federal spending in 2009 is the most important variable in predicting a county's violent gun incident rate. The percentage of the population that is African American, percentage of population in poverty, and percentage of population who own their home were also highly important in this model in comparison to the other explanatory variables included. These variables are identified across all tree-based models suggesting that these relationships are robust. The percentage of the population that is African American, in poverty, and housing cost burdened are demographic measurements and socioeconomic factors that affect an individual's likelihood to be a victim of gun violence. In addition, the amount of federal spending in 2009 can be attributed to a surge[8] in relief spending in poor communities as they experienced the effects of the Great Recession. Therefore, it is not a surprise that these facts would have a greater ability to predict differences in gun violence rates across counties in the United States. While some other variables are found to be significant, including the percentage of the population that is white in 2010, the percentage of the population that has a bachelors, as well as per capita income, the variable importance ranking from the random forest model provides an interpretable ranking of the influential factors from the total set of features.

Given that socioeconomic factors were the strongest predictors of violent gun incidents, it seems that in each county, gun violence rates are most associated with community poverty levels. More specifically, individuals living in impoverished communities with a higher African American population are much more susceptible to be impacted by a violent shooting. Our results suggest that the prevalence of these shootings can be captured by measures of socioeconomic inequality and poverty. The identification of the counties that exhibit these factors should serve as a warning to their respective local governments to reallocate tax dollars to further mitigate this problem and fight to decrease gun violence rates.

Limitations and Future Directions

Dataset Limitations

As described in the Data Description and Exploration section above, one main limitation we found using these datasets was the inconsistency of the year in which the features were

[8] https://www.pewresearch.org/fact-tank/2017/04/04/what-does-the-federal-government-spend-your-tax-dollars-on-social-insurance-programs-mostly/

calculated. The dataset that contained the number of violent gun incidents ranged from January 1, 2013 until March 3, 2018, while the census data included annual measurements primarily from the years 2010 until 2019. Therefore many of the features contained in the census data regarding the year 2019 had to be discarded. Additionally in the gun violence dataset, we had to extract the variables for each measurement that were calculated closest to our desired time window, which could have affected the credibility of our models. Another limitation of the dataset was some of its most important features being calculated in 2010 while the rest of the features were calculated in 2017, such as the percentage of the population that is white or owns their home. This left us with just 17 features. Furthermore the gun violence dataset contained many null values for the location of the shooting which greatly reduced the number of observations in our final dataset. Since each observation represents a different US county, many counties were thus left out of our analysis. Another limitation is that, as described in the Exploratory Data Analysis section, there is evidence of correlation amongst some of our explanatory variables. As a result, some variables could be confounding and mask the true relationship between the measured variables. Moreover, given how variable selection works, the variables selected in our LASSO regression, elastic net regression, and tree-based methods might be misleading.

Analysis Limitations

Although we provide different methods for the interpretation of the variables, our analysis only incorporates a fraction of the total socioeconomic and demographic measurements that describe the population. The results of our analysis may change if we included a number of different variables. For example, if we included the number of high-risk offenders in each county, such as previous domestic violence offenders or persons convicted of violent misdemeanor crimes, or the percentage of males in a community. Next, although splitting our dataset into training and testing data allows for a more unbiased test of the models and reduces the risk of overfitting our data, the random split in the data can skew our model and our conclusions. Therefore, splitting the data into our train and test sets using a different seed number may have yielded different selected variables in the shrinkage methods and different levels of variable importance in the tree-based methods.

Recommended Follow-Up Analyses

To compensate for the limitations described above, more concrete analysis can be done by acquiring complete datasets for gun violence incidents as well as yearly socioeconomic and demographic data. In addition, our response variable, gun violence incident rate, did not account for whether the event was a homicide or suicide. In 2019, 60% of of deaths from firearms in the U.S. are suicides[9] and in general, firearms are the means in approximately half of suicides nationwide. Future analysis on the impact of different socioeconomic and demographic measurements on gun violence incidents should take into account whether the incidents were suicides as that introduces a completely separate issue. Additionally, given that many observations needed to be omitted in our dataset as they contained NA fields, we recommend that our analyses be reconducted once the missing data is collected.

[9] https://health.ucdavis.edu/what-you-can-do/facts.html

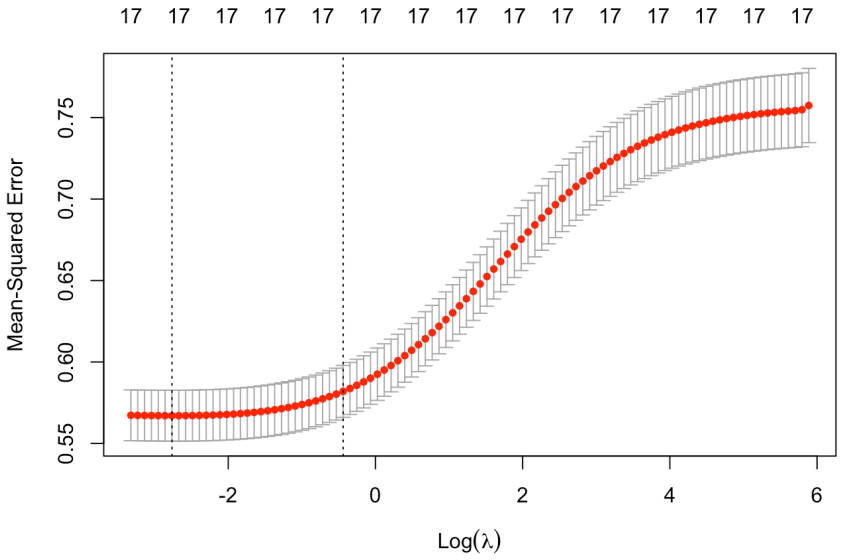
Appendix

Definition of features, copied from the source documentation:

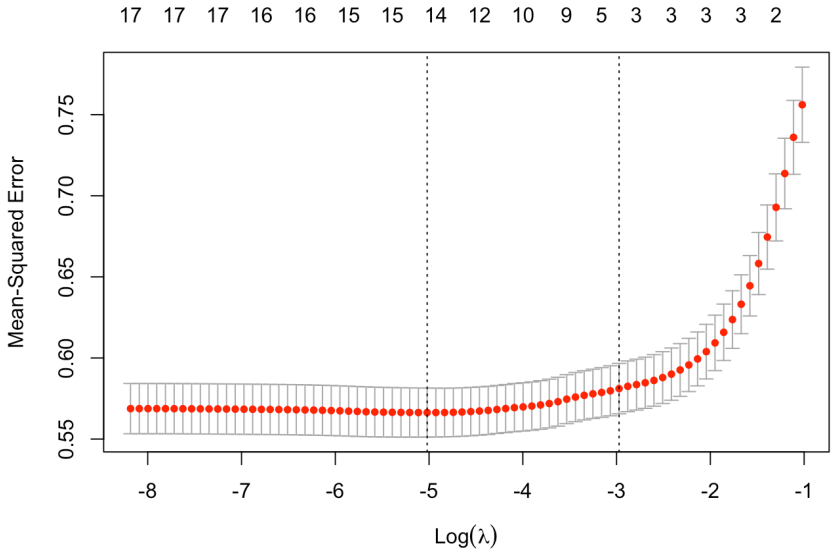
- Age_under_5_2017
    - Percent of population under 5 (2010).
- Age_over_65_2017
    - Percent of population over 65 (2017).
- Median_age_2017
    - Median age (2017).
- White_2010
    - Percent of population that is white (2010).
- Black_2017
    - Percent of population that is black (2017).
- No_move_in_one_plus_year_2010
    - Percent of population that has not moved in at least one year (2006-2010).
- Speak_english_only_2017
    - Percent of population that speaks English only (2017).
- Hs_grad_2017
    - Percent of population that is a high school graduate (2017).
- Bachelors_2017
    - Percent of population that earned a bachelor's degree (2017).
- Computer_2017
    - Percent of population who has access to a computer (2017).
- Homeownership_2010
    - Home ownership rate (2006-2010).
- Per_capita_income_2017
    - Per capita money income in past 12 months (2017 dollars, 2017)
- Poverty_2016
    - Percent of population below poverty level (2012-2016).
- Employed_2015
    - Number of civilians employed in 2015. (we divided this by our population estimate to get the employment rate).
- Uninsured_2017
    - Percent of population who are uninsured (2017).
- Fed_spending_2009
    - Federal spending, in thousands of dollars (2009)
- Density_2010
    - Persons per square mile (2010).

CV Plots:

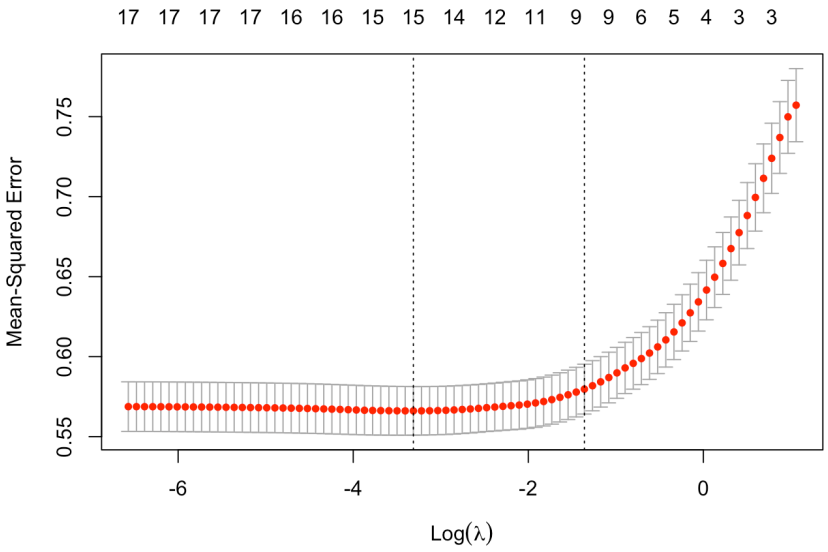
Ridge



LASSO



Elastic Net


17 17 17 17 16 16 15 15 14 12 11 9 9 6 5 4 3 3
Mean-Squared Error
0.55
0.60
0.65
0.70
0.75
-6
-4
-2
0
Log(λ)