Influencing Factors of Gun Violence:

Demographics, Policies, and Regulations

Team C-HOT

Chelsea P Daniels (cpd317)

Oussama Fatri (of282)

Hiroshi Horikawa (hh1504)

Tianze Zhou (tz504)

December 9, 2017

A. Problem Definition

Our project seeks to inform how specific policy interventions could be used to prevent gun violence. We are interested in examining how certain policies related to gun use and ownership, combined with environmental demographic characteristics impact gun-related deaths across the United States – specifically, can the level of gun-related deaths in an area be accurately predicted by these factors, and which factors are most significantly related to gun deaths. We are interested in parsing out whether gun-specific policy – things like open-carry, assault weapon prohibition, safety classes - can be used as a means of preventing gun-related deaths, or if policy attention is better placed toward social welfare initiatives - things like education and employment opportunities. Data mining allows us to combine data from multiple sources to derive a prediction model and then rank feature importance, and track which features drive prediction for a given area. This might allow policy makers to make inferences about how to best allocate resources. The intent of this project is to contribute to non-partisan, empirical understandings of how policy can be used to prevent gun violence across and within specific regional contexts.

B. Data Exploration

1. Data Understanding

Our data combines demographic characteristics at the census tract level with state-level gun ownership and use laws. Our target variable is gun-related deaths and gun-related incidents at the census tract level. Demographic data was aggregated from the 2010 Census of Population and Housing and the American Community Survey spanning 2011-20151. We include measures of an area's composition in terms of urbanicity, citizenship, gender, age, gender by age, race, educational attainment, income, and employment.

Our variable choices were informed by existing literature related to gun violence. Education is an important factor when measuring gun violence. Research has shown that children with better education (along with several other factors such as health, finance, and feeling of safety) decreases their risk of being involved with violent behaviors. Children who grew up in

[1] Steven Manson, Jonathan Schroeder, David Van Riper, and Steven Ruggles. IPUMS National Historical Geographic Information System: Version 12.0 [Database]. Minneapolis: University of Minnesota. 2017.

communities which lacks access to these resources will have exposure and have a high tendency to be involved with violent activities and this may lead to causing violence with guns. Education is not the sole factor which determines if a person is going to cause gun violence but it can be used as a measure to understand the environment the person gone through.

Tying this to the issue with race, Blacks and Latinos tend to have less opportunity to education which may lead to differences with income and which may ultimately lead to people be involved with crime (since they will somehow need to earn to be breadwinners). A large population of offenders of gun violence have not finished higher education such as college education (there is not data showing people who studied at vocational schools). Interestingly the percentage of the population that owns a gun with respect to the education level of high school or less, some college, and Bachelor degree or above is 31%, 34%, and 25%.

Additionally, gun violence rates differ by gender. First off, men are more likely to have guns compared to women as the percentage of men who said they own a gun is 39% where as the percentage of women who said they own a gun is 22%. Also interestingly enough, the percentage of men who said they don't own a gun but has someone else who owns a gun in the household is 5% whereas the percentage is 18% for women respondents.

Additional predictive features include indicator variables for specific gun-related laws including prohibition of open-carrying of weapons, restriction of the number of weapons that can purchased at a time, waiting period for weapons, requirements for concealed carry, and prohibition of assault weapons and machine.

In summary Our data comes from four different files:

- Demographic data at the census tract level;
- Gun violence data at the census tract level;
- Gun laws at the state level;
- Geographic information at the census tract level (latitude / longitude for visualization purposes);

A description of the different features can be found in Appendix 1.

2. Data Preparation

Preparing the data in the right format for data mining required a few manipulations:

- Converting numerical variables that were stored as strings to floating numbers (e.g. census tract population was stored as a string with the thousands comma serparator, demographic data was stored as strings with the % sign, …);
- Converting all strings to lower case, which will be important for merging data from different files (e.g. merging demographic data and state-level gun laws will be done by joining on the state column which was in lower case in one file and upper case in the other);
- Census tract data unique identifiers are different in different data sources (e.g. FIPS = 010010020100 on the gun violance file and GISJOIN = G0100010020100 for the demographic data file correspond to the same tract). Some effort was spent to make these unique codes uniform across files[2];
- Some variables were renamed for ease of presentation purposes (in particular, Appendix 1 shows our codebook for different laws);
- Duplicate entries were removed (e.g. duplicate information on census tracts with slightly different longitude / latitude coordinates).

3. Defining a Data Instance

After preparing and merging data from our different sources, we managed to define an instance of the data at the census tract level with its corresponding demographic characteristics, state-level gun ownership and use laws as well as the number of gun violence incidents (and the resulting number of deaths) observed. Our instances could potentially have spatial correlation but we can assume that there is no temporal correlation as we only focus on the 2015 data. An example of an instance is shown in Appendix 2. We have 8876 such instances.

[2] FIPS stands for Federal Information Processing Standards (https://en.wikipedia.org/wiki/Federal_Information_Processing_Standards)

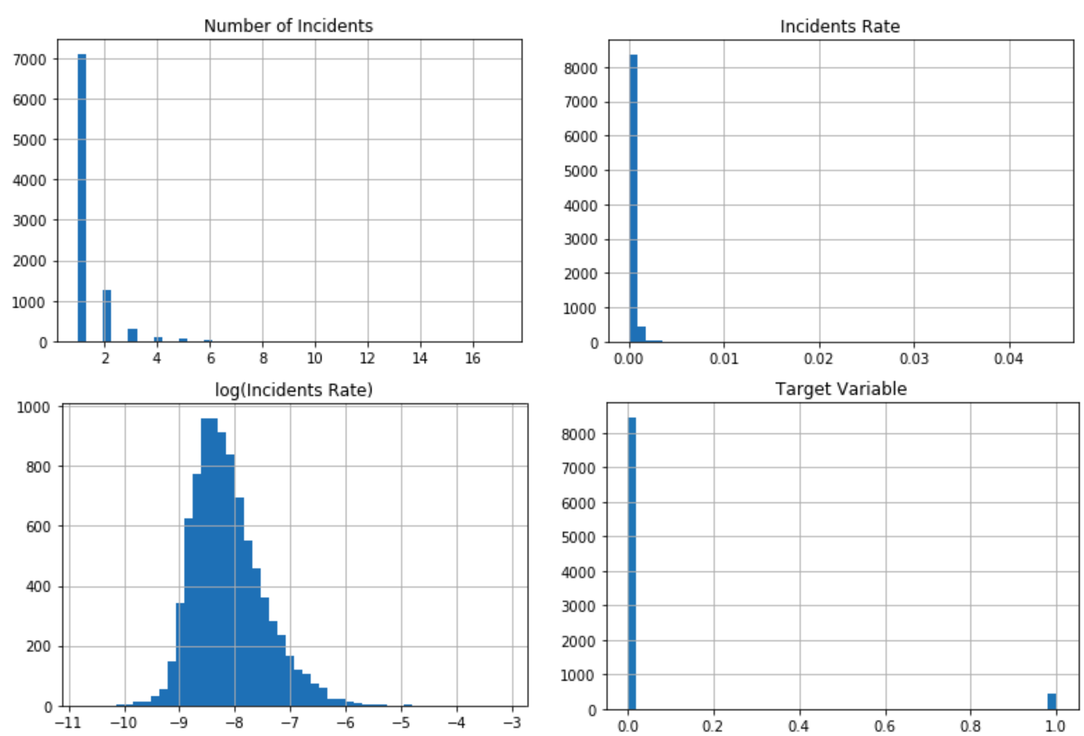
4. Defining the Target Variable

Only tracts with at least one gun violence incident are present in our final sample i.e. we only observe positive instances. Therefore, our analysis will not focus on predicting positive instances (presence of gun violence in a particular census tract) vs. negative instances (absence of incidents) but rather focus on two categories: high vs. low incident rate as defined below:

a. For each data instance, use the number of incidents (we could have used the number of deaths resulting from the incidents. We check that our results are not sensitive to this choice);
b. Normalize the number of incidents by dividing it by the tract population;
c. The resulting distribution is skewed, so we apply the log function;
d. Classify the data points that are above one standard deviation from the mean as positive instances.

Charts A to D from Exhibit 1 illustrate the steps followed in coming up with our binary target variable. Note that the resulting base rate is: 4.8%.

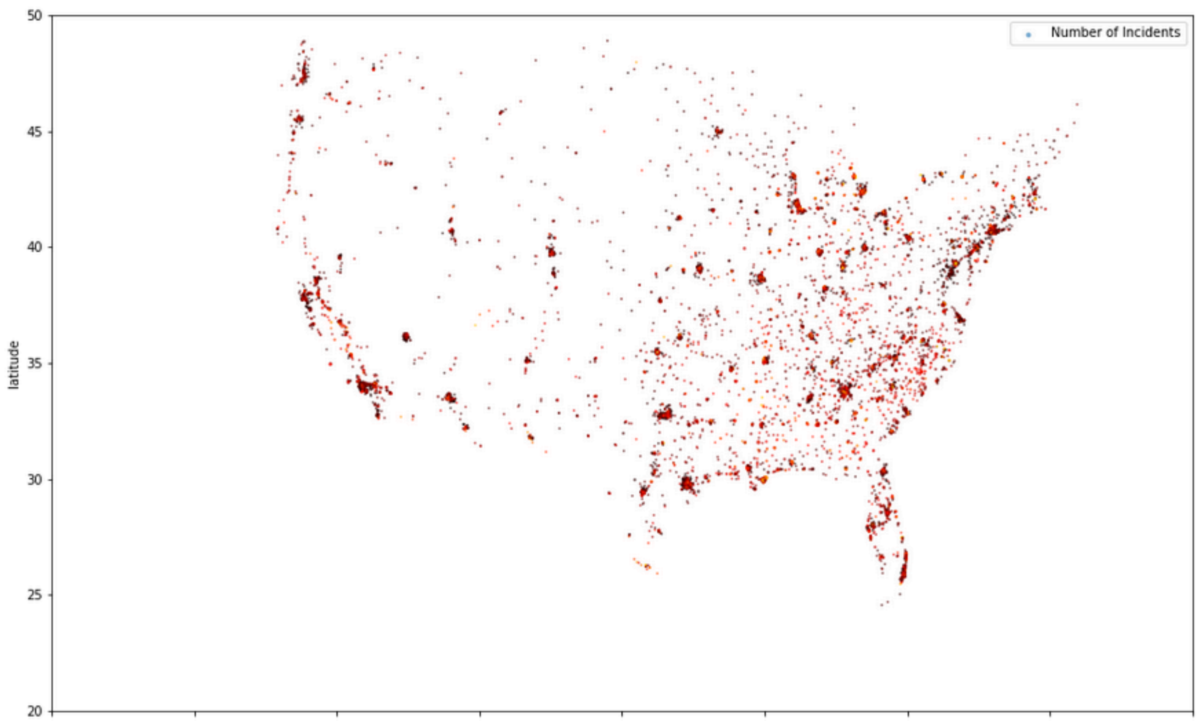
Exibit 1: Target Variable Construction



5. Preliminary Data Exploration

In 2015, there were 11606 incidents that resulted in 13050 deaths. As we can see on the map below (Exhibit 2), these incidents - even after controlling for the census tract population – are not uniformly distributed across the US territory.

Exibit 2: Distribution of the Number of Incidents across the US territory



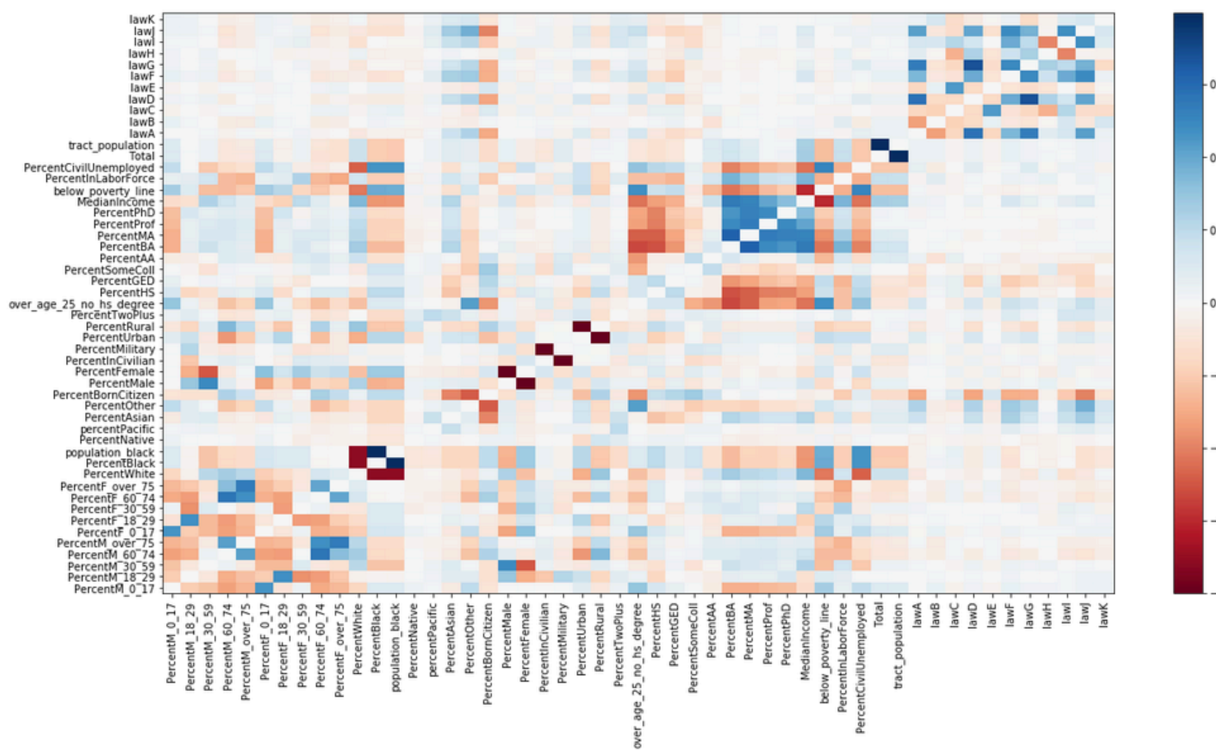
As a first step we will try to find which features are informative in explaining/predicting our target variable.

a) <u>**Predictor Correlation**</u>

After reordering the predictors so that different subcategories of the features are next to each other (urbanicity, citizenship, gender, age, race, educational attainment, income, and employment, …) we take a look at their pairwise correlations. As we can see on Exhibit 3, we can easily identify variables that are highly correlated, either positively and neatively:

- *Percent Urban* and *Percent Rural* are perfectly negatively correlated given that they add up to one; Similarly for *Percent Male* and *Percent Female*, and *Percent Military* and *Percent Civilian*. We drop the redundant variables;
- The age ranges for the female and male population are highly correlated (e.g. Percent Males between 0 and 17 and Percent Females between 0 and 17). We choose to create a new variable which represents the percentage of the population (across genders) below 29 as a representative of the age characteristics and drop all the other related variables;
- "Total" and "Tract Population" have a correlation of 1 (they represent the same figure from different sources). We drop the former;
- Eduation variables are very highly positively correlated, so we only keep a representative variable of this category: *Percent of people above 25 with no High School degree*.

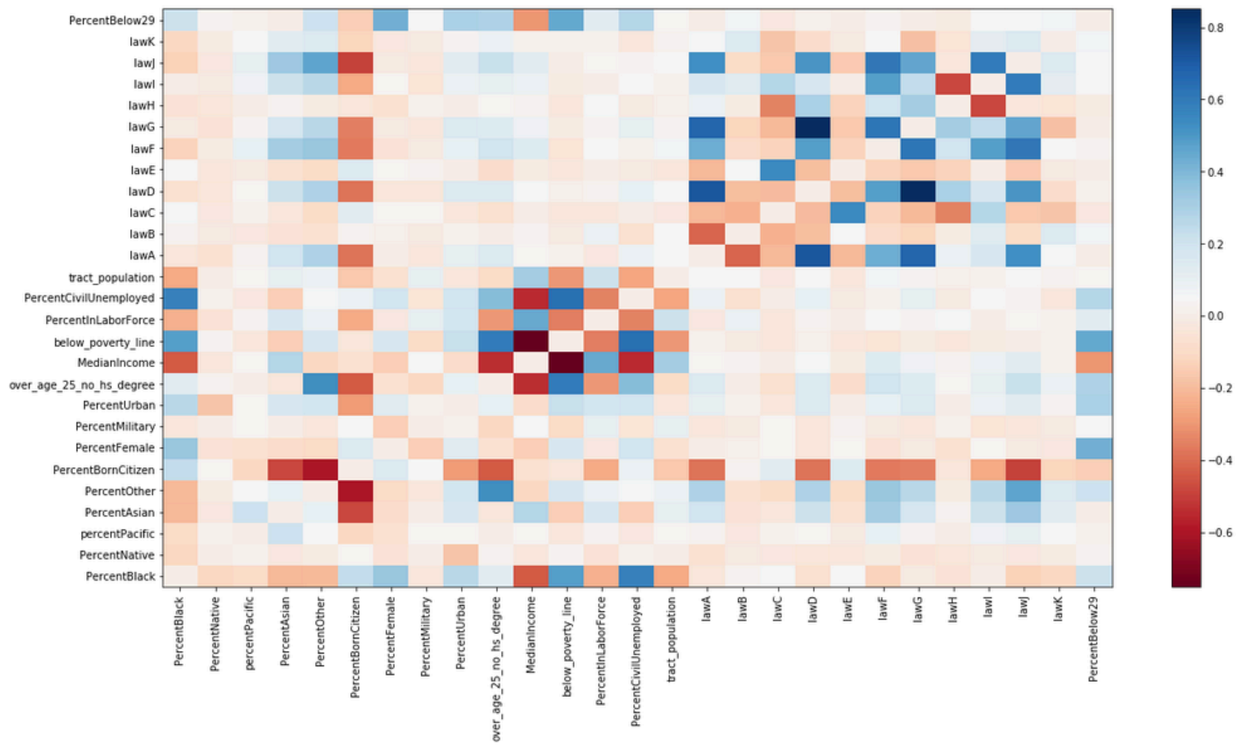
Exibit 3: Correlation Between Predictors



We do not want to constrain ourselves to any particular machine learning techinc at this point. Given that some algorithms are sensitive to multicolinearity (like logistic regression), we believe our choice of dropping redundant variables is necessary and parcimonious.

The chart below (Exhibit 4) shows the correlation matrix after dropping the variables discussed above.

Exibit 4: Correlation Between Predictors after Dropping Redundant Features



We can see that a few other variables are highly correlated although they represent different information, so we decided to keep them:

- *Percent Black popul*ation is positively correlated to *Percent Unemployment* and *Below Poverty Line* and negatively correlated to *Median Income*;
- *Percent Population over age of 25 with no High School Degree* is positively correlated to *Percent Unemployment* and *Below Povery Line* and negatively correlated to *Median Income*;

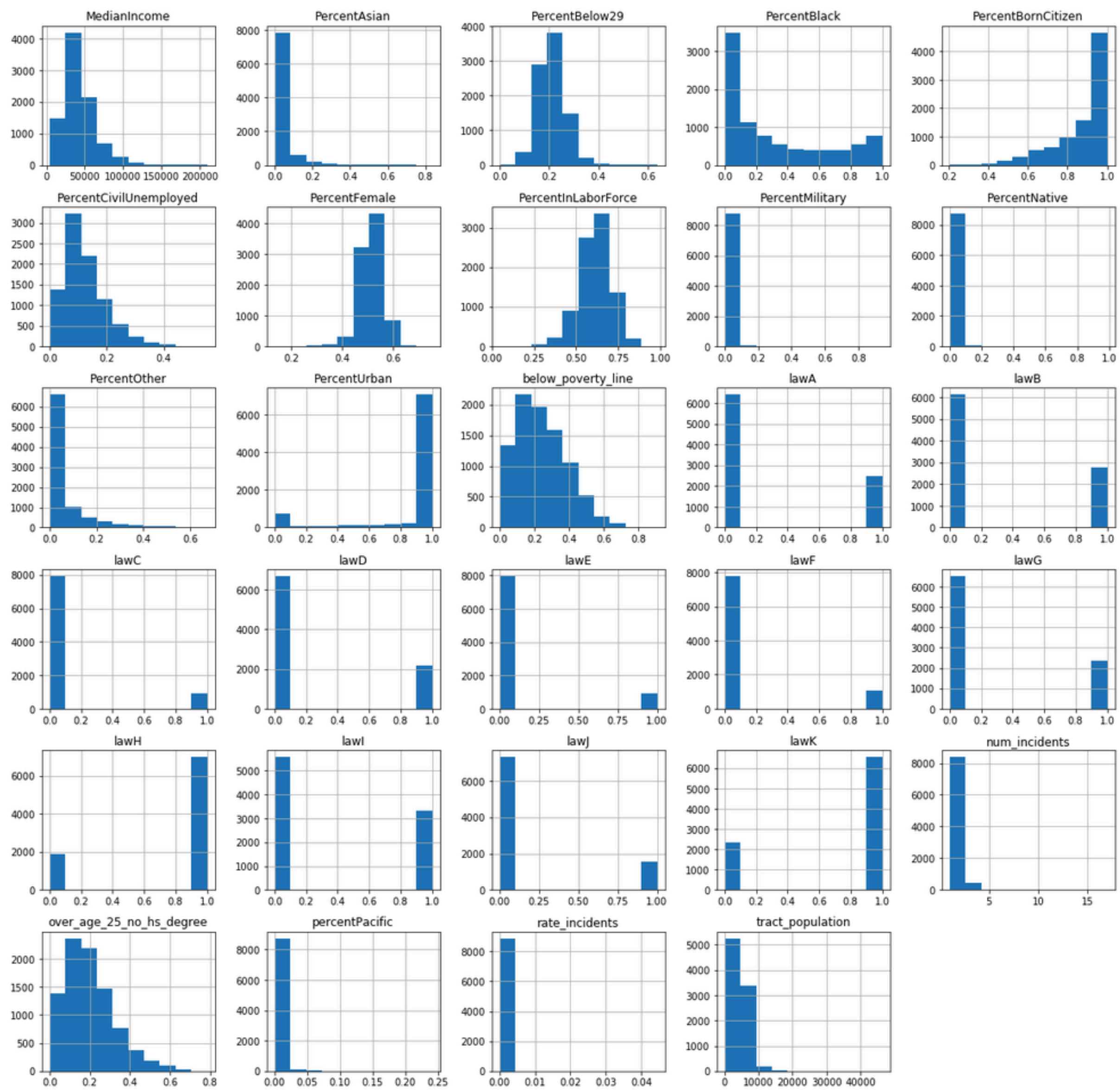
b) **Distribution of Predictors**

Next, we take a look at the distribution of the features (Exhibit 5). We can make the following observations:

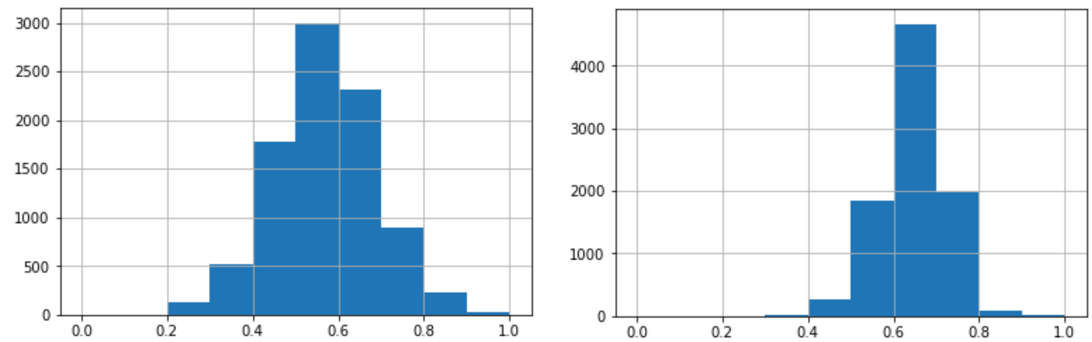
- Most variables are expressed as percentages and are therefore already on the same scale;

- The state-level gun related variables are binary;
- Median income and Tract population are highly skewed, so we apply the log function to both and then normalized using min-max scaling. Exhibit 6 shows these two variables after the transfomations.

Exibit 5: Features Distribution



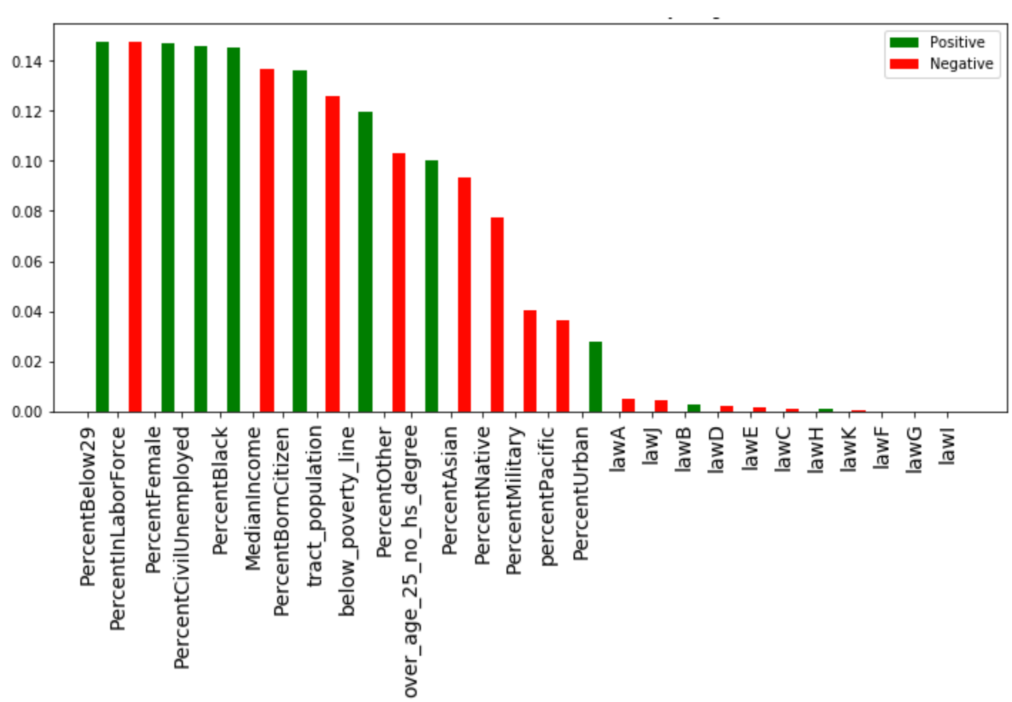
Exibit 6: Median Income and Tract population after applying log and min-max scaling Transforamtion



c) Relationship between the Target Variable and the Predictors

Since we are performing a supervised learning exercise, we would like to get some insights from looking at the relationship between our target variable and the predictors. The chart below (Exhibit 7) shows the mutual information between each feature and the target variable as well as their correlation sign.

Exibit 7: Mutual Information Between Predictors and the Binary Target Variable



We can observe that the target variable is positively correlated to:

- The percentage of the population below 29 years old;
- The percentage of unemployed civilians;
- The percentage of the black population;
- The percentage of the people born citizens;
- The percentage of the people below the poverty line;
- The percentage of the urban population;
- Percent population over age of 25 and no High School degree .

It is negatively correalted to:

- The percentage of the population in the labor force;
- The median income;
- The tract population.

Some other relationships are less intuitive like the positive correaltion to the percentage of the female population.

More interstingtly, the mutual infrmation between our target variable and state-level gun laws is the lowest amongst our predictors. We decided to drop these variables.

C. Modeling and Evaluation[3]

Now that we have a good understanding of our data and what features are informative, we turn to the modeling part of the problem and explore different supervised classification models for our binary target variable.

Since we have no particular constraints that dictate what type of algorithms we should be using – our data set has a manageable size both in the number of instances and number of features dimensions – we will run a horserace between different data mining algorithms.

[3] For this section we use and modify code from the GitHub repository supporting Intro to Data Science (DS-GA-1001) for the NYU Center for Data Science by Brian d'Allesandro https://github.com/briandalessandro/DataScienceCourse

1. Choice of algorithms

We will explore the following algorithms: Logistic Regressoin, Decision Tree and Random Forest. For each one of them, we will run a grid search over a appropriate range of hyper-parameters.

2. Data splits

Before turning into the modeling part, we split our dataset into training (80%) and testing (20%) subsets. For model selection we will run cross-validation within each family of algorithms, pick one specification using the "one-standard error rule" and then combine the selected using ensemble learning.

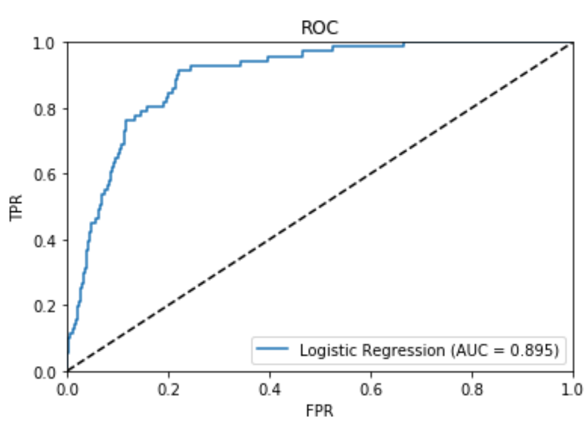
3. Evaluation Criteria

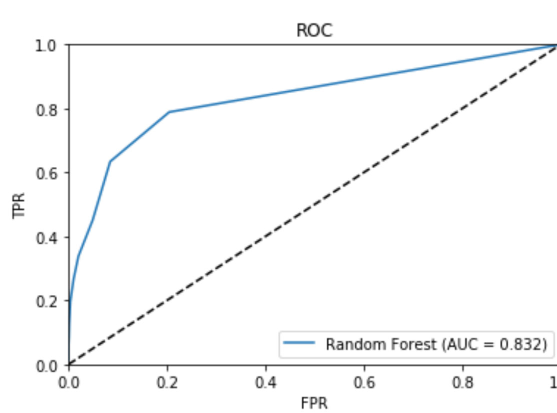
Our goal is to rank the census tracts according to the probability of having a high incident rate. The AUC seems to be an appropriate evaluation measure.

4. Baseline model

As a baseline model, we run a simple logitic regression without tuning any parameter (using the default parameters: L2 penalty and C = 1.0). The ROC chart below (Exhibit 8A) shows a decent performance of this simple model with an AUC of 0.895. Note that we split the 80% training data into training and validation to generate the charts below in order to preserve the test sample for our final evaluation. Can we improve on this?

Exibit 8: ROC Curves for a Logistic Regression Model and a Random Forest Model





We run a Random Forest and get an AUC of 0.832 (Exhibit 8B). The fact that the Random Forest – that typically has the tendency to overfit – has a lower AUC might suggest that there is no much improvement to be made on our base model. The next section explores what kind of enhancements we can hope for.

5. *Grid Search and Model Selection*

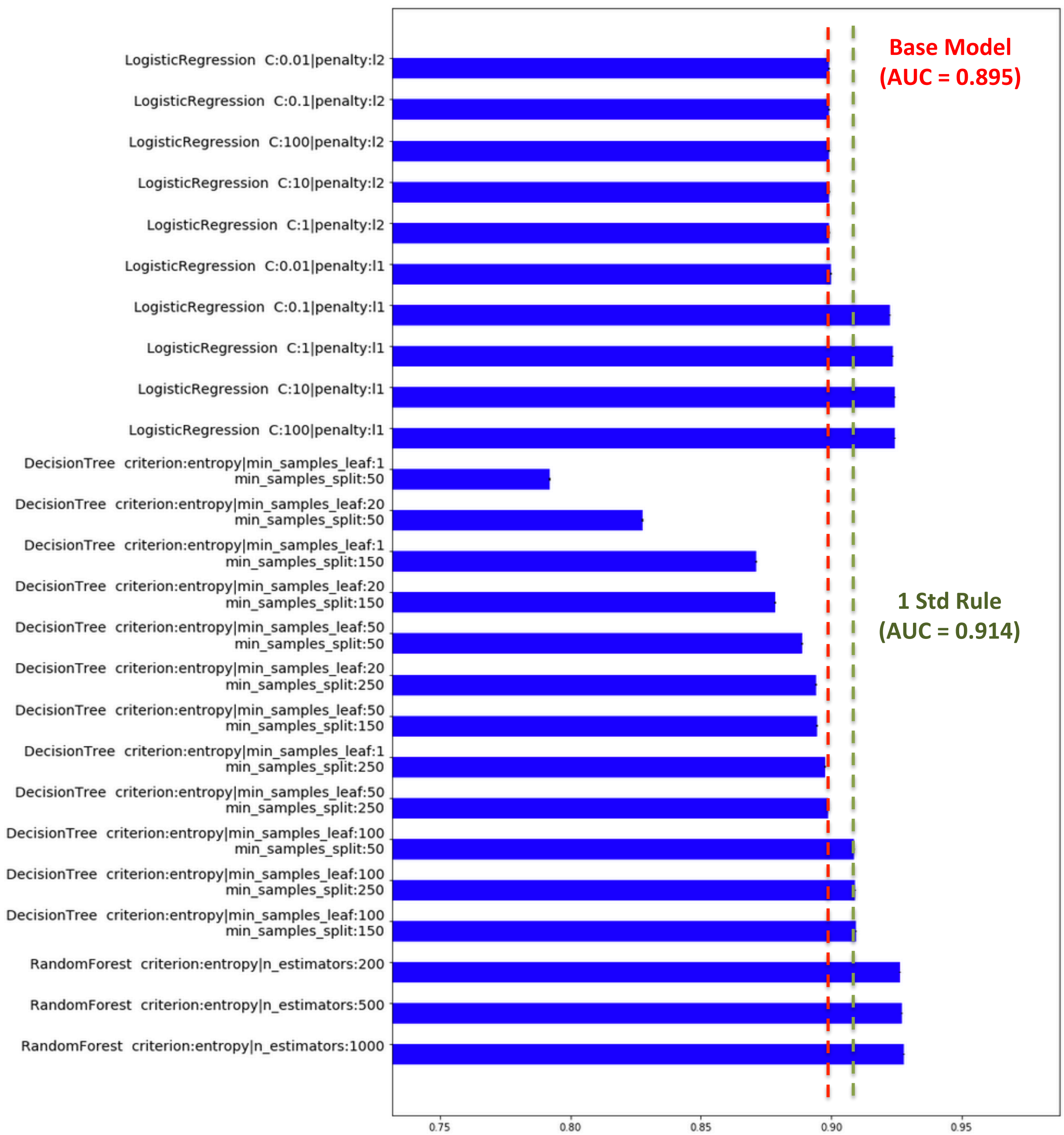
We perform a 10 k-folds cross-validation exercise on our training data to perform a grid search for the following algorithms and their corresponding hyper-parameters:

- LogisticRegression: {'C':[10**i for i in range(-2, 3)], 'penalty':['l1', 'l2']}
- DecisionTreeClassifier(): {'min_samples_leaf':[1, 20, 50, 100], 'min_samples_split':[50, 150, 250], 'criterion':['entropy']}
- RandomForestClassifier():{'n_estimators':[200, 500, 1000], 'criterion':['entropy']}

Exhibit 9 shows the correspong AUC numbers. We can make the following observations:

- Our base model ranks pretty high compared to other sepecifications we get from the cross-validation (red line on the chart);
- Changing the parameters for the Random Forest from their default values improves the AUC from 0.832 for the default parameters (Exhibit 8B) to 0.928 which is the highest we were able to achive
- Applying the one standard error rule to be more conservative a avoid overfitting, we can select a model with an AUC equal to 0.914 (gree line on the chart).

Exhibit 9: Grid Search and corresponding AUC

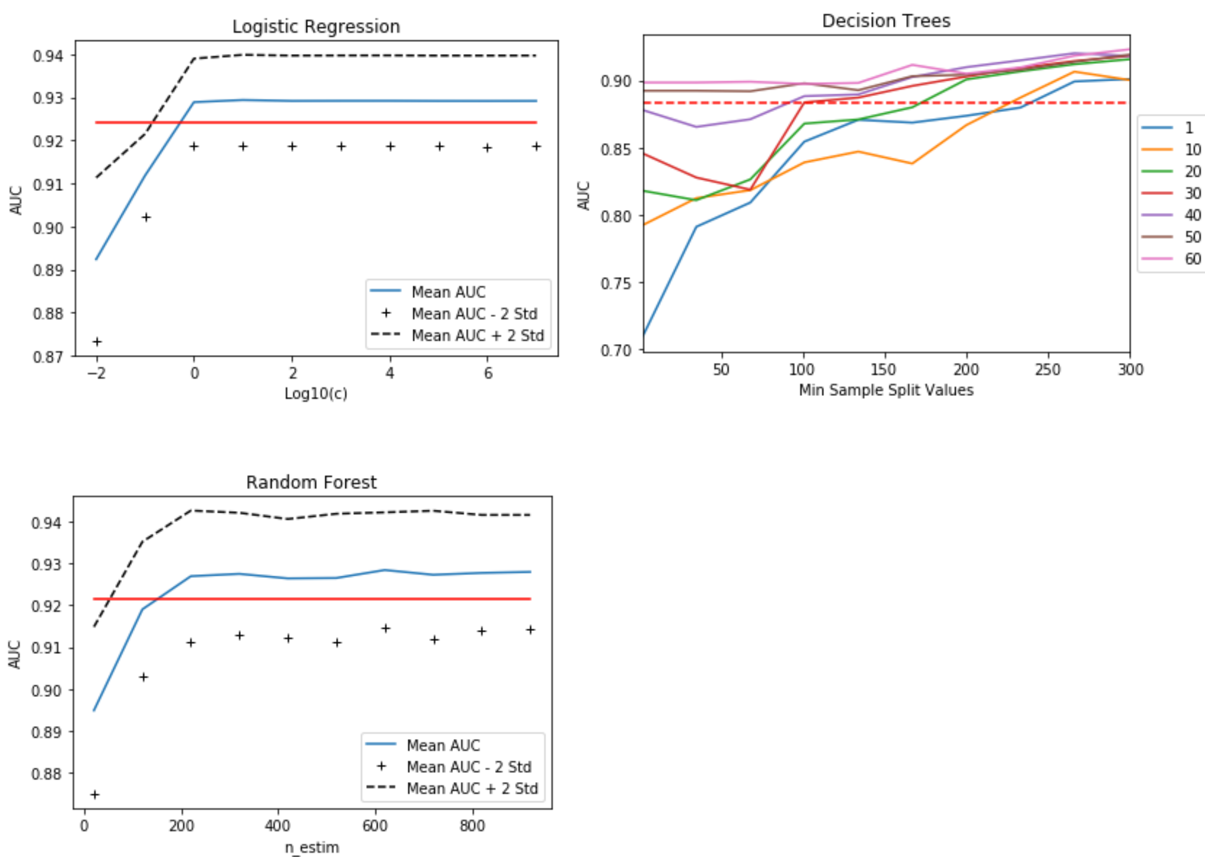


6. Ensemble Learning

We decide to go one step further and choose the best model within each family of learning algorithms and combine them using the Scikit-Learn ensemble module. As you can see on Exhibits 10, appplying the one standard deviation rule to each one of the different learning algorithms, end up with the following models:

- Logistic Regression: c = 1, penalty = ‘l1’;
- Decision Tree: min_samples_leaf = 10, min_sample_split = 200, criterion = 'entropy';
- RandomForest: n_estimators = 200.

Exhibit 10: Applying the One Standard Error Rule to different Algorithms



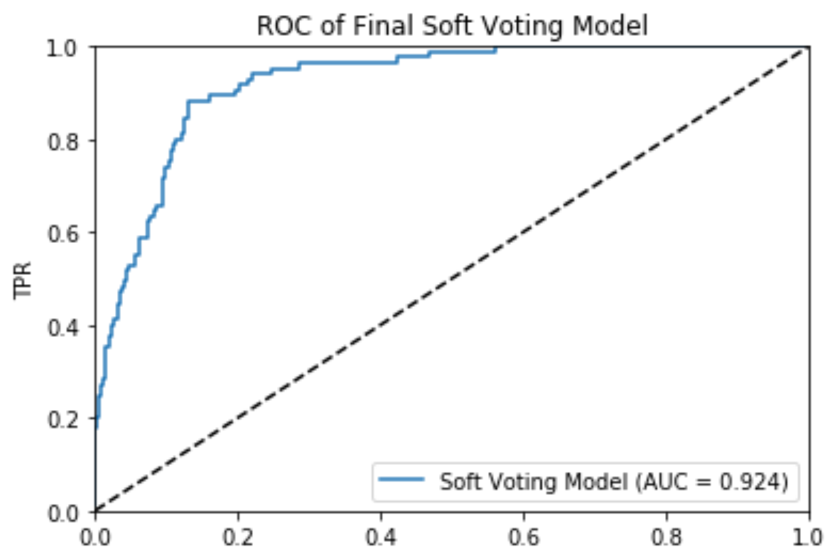
All our models can estimate class probabilities so we combine them using the “soft” voting parameter of the Scikit-Learn voting classifier.

7. Evaluation

We turn to the evaluation step on the testing set after fitting our final model on the entire training set. As we can see on Exhibit 11 A, the voting classifier improves on the individual models and has an AUC of 0.924, higher than each one of the individual models and higher than our base model. Exhibit 11B shows the ROC curve for our final model vs. our base model.

Exhibit 11: AUC of our selected models and the final Voting Classifier on the Test Set

LogisticRegression 0.922753678645
DecisionTree 0.907844296796
RandomForest 0.916947855428
Voting 0.923901624517



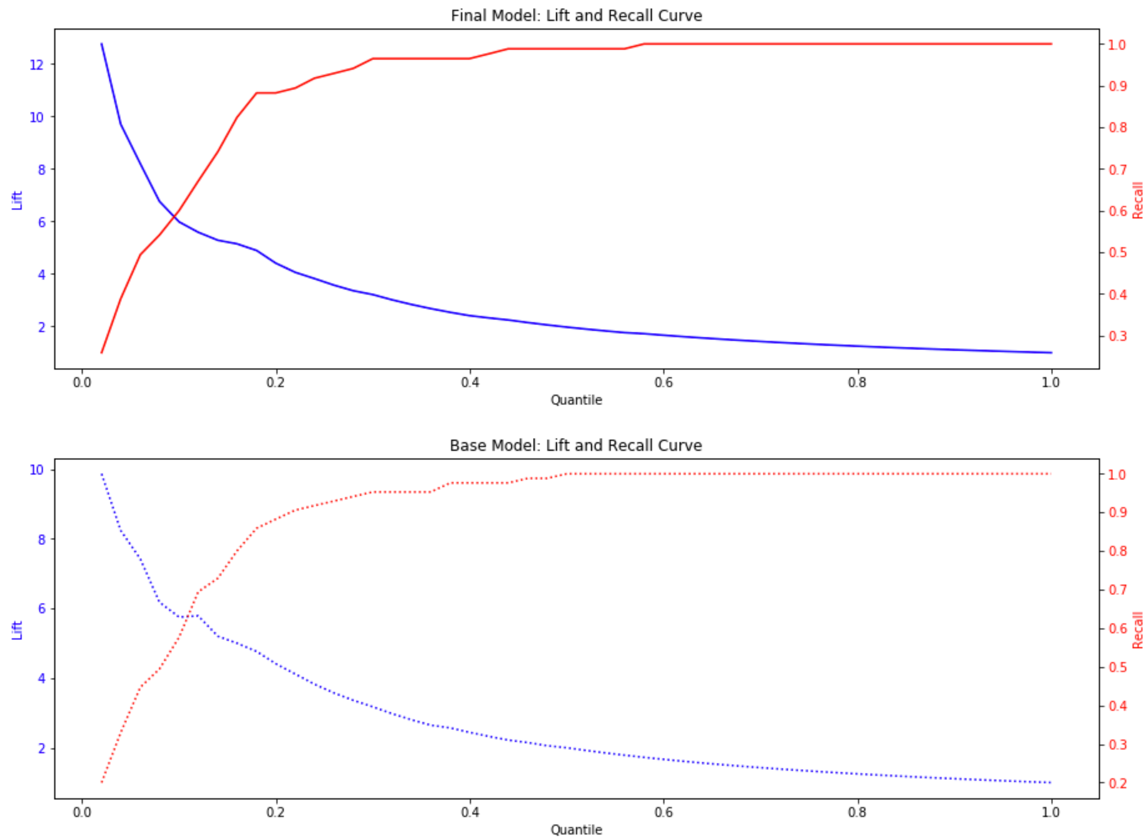
Now let's apply this model on the particular example of the state of New York. The table on Exhibit 12 shows the predicted probailities of high incident rate (sort from highest to lowest). While our goal was not to produce very accurate estimates of these probabilities, we can easily use them to rank the census tract and focus on the ones with the highest risk.

Exhibit 12: Predicted Probabilities og High Incident Rate in the State of New York

tract_fips	proba
36055001500	0.819457
36047035200	0.675350
36047089800	0.404139
36029016500	0.363572
36055006400	0.347725
36081018800	0.312918
36067005800	0.305659
36061011900	0.277901
36047036700	0.267795
36055005100	0.267536

Let's assume that the NY state is using this model and has a budget constraint: they can only focus on 10% of the census tracts. The lift curve on Exhibit 13A shows that using this model would be 6 times better than random targeting and the recall curve shows that 60% of those tracts would actually have high incident rates (by our definition of incident rate being one standard deviation higher than the mean across all the census tracts). This gives a marginal improvement over the base model where the lift is 5.5 and the recall is 50% (Exhibit 13B).

Exhibit 13: Lift and Recall Curves for the Final Model and the Base Model



D. Risks, Limitations and Ethical Considerations

Before we started our project, we thought it's highly likely to see a strong correlation between the incident rate and the public policies and regulations on gun control. However, our findings of the mutual information between incident rate and policies & regulations give us a complete contrast result – comaperd with any other features that we studied, the public policies and regulations on gun control somehow have the lowest correlation with the incident rate. One reasonable explanation for this surprising result would be that the illegal possession and carry of the guns has a most crucial influence on the gun violence. According to Sheley & Wright (1998), even high school students in some places reported that they have ‘little’ or ‘no’ trouble if they wanted a gun, which illustrates the fact that due to the existence of these underground gun markets, some regulations such as permission on gun purchase is to some extent nominal. Even though this research is questioned by Cook et al. (2007), who find out that the underground gun markets have become thinner, it does not eliminate the correlation between the underground gun markets and the gun violence. In the meantime, Cook et al. (2007) mention another feature that would support our explanation – gangs. According to Hoyert & Xu (2012), in both 2010 and 2011, 80% of gun homicides are gang-related. If we can assume that gang members usually don’t obey the public policies and regulations on gun control, then the fact that most of the gun violence incidents are conducted by the people who do not care about these related policies and regulations, effectively explains why the policies and regulations have low correlation with the incident rate in our project.

One way to deal with this fact is to make new policy to enhance the implementation of the policies and regulations on gun control. According to Sherman et al. (1995), directed patrol can in a way reduce the amount of gun violence in the high gun crime areas and is in average “three times cost-effective than normal uniformed police activity citywide.” Also, in the paper it mentions that the implementation of the directed patrol policy was successful in reducing the number of gun crimes in Kansas City at that time. Today, there are other areas, which have already implemented the directed patrol policy, such as New York City. In the future study, we could add a new feature “directed patrol” as a binary feature, and we could use the mutual information between this feature and incident rate to see if the directed patrol policy is universally effective in reducing gun violence national wide.

Another reason why the correlation between policies and regulations, and gun violence is negligible in our model is that the data we collected and analyzed on is based on a snapshot in time. The problem here is that unlike other features, such as percent of female, and median income, which can be regarded as internal elements, policies and regulations is like an external force or external element. Evaluating the influence of this external element in a parallel sense, as we did in our model, is logically irrational. For example, if a policy is implemented in two different places, it's to some extent impossible for us to consider these two places as an equivalent class in our model because even though we can try to control all the other features which are taken into consideration in our model, there are still many other features such as culture, religion, and geography, which might have huge impact on the implementation of this policy however we haven't paid attention to in our algorithm. Also, it is not possible for us to know how well this same policy is implemented in these two places only by the binary data we have. If this policy is not implement at the same level at two places, it's intuitively not correct for these two places to have the same value for this feature. However, on the other side, it is theoretically impossible for us to collect such data, which could describe the "situation", or "behavior" of a policy or a regulation in a census level, which means, if we compare these data in a parallel sense, the error would be essential.

One way to alter this situation is to compare the data based on time. According to Webster et al. (2014), researchers from Johns Hopkins University, repealing the background check requirement in Missouri has resulted in a 25 percent increase in gun homicide rate each year. Consequently, in the future study of the correlation between policies and regulations and gun violence, we shall collect the data from a long time period including some special time points such as when a policy or regulation starts to take effect, and when it is repealed.

According to our model, generally, there is an inverse correlation between the economy and gun violence. This correlation is deduced by a. The negative correlation between median income and incident rate, and b. The positive correlation between the unemployment rate and the incident rate. The contrast of the correlations mentioned above implies the fact that the weak economy of an area would result in more gun violence. Turn the other way around, the increase in the gun violence would lead to the fallback of the economy. The first implication is reasonable: a weak economy might result in an insufficient-funded public security system, which further result in a higher amount of gun crimes. The second implication is supported by the research report presented by Irvin-Erickson et al. (2016). Based on the findings from three specific cities, their studies conclude that the increase of the gun crimes in neighborhoods would jeopardize the local economy by lowering "the retail and service business establishments, the number of new jobs created, and the volume of sales in business establishments." As we can see from these implications, there is a bad reciprocity between the gun violence and the economy, and this interaction gives us a hint that, devoted in reducing the gun violence, the government shall not only pay more attention to the implementation of policies and regulations, but it shall also put more effort into economic development.

Based on the result of our model, one thing we have to mention is the huge positive correlation between the incident rate and the percent of black. From the ethical side, it is definitely morally unhealthy and irresponsible, and logically problematic to claim that black people are the reason of the gun violence but it doesn't mean then that our model is not trustworthy. From an analytical perspective, there is a strong relationship between the gun violence and the composition of races, and a higher percentage of black might leads to a higher incident rate. There are probably a lot of reasons behind this correlation, and these reasons could be intricate. One possible reason would be that compared with other races, on gun violence, black people are sometimes more vulnerable. This point coincides with one of the results from the research made by the research team from Evertown (2017). The result shows that making up of only 14 percent of the total population in the United States, Black Americans, represent more than half of the victims of gun homicides, based on the data given by the U.S. Census Bureau (2010), and the Centers for Disease Control and Prevention (2015). Consequently, a higher percentage of black implies higher percentage of potential victims of gun violence which further implies a potentially higher incident rate.

References

- Braga, A.A., Cook, P.J., Ludwig, J., MARKETS, U.G., & Venkatesh, S.A. (2007). Underground Gun Markets.

- “Fatal Injury Reports," Injury Prevention & Control: Data & Statistics (WISQARS), accessed January 25, 2015, http://1.usa.gov/1plXBux.

- GitHub repository supporting Intro to Data Science (DS-GA-1001) for the NYU Center for Data Science https://github.com/briandalessandro/DataScienceCourse by Brian d’Allesandro

- Gun Violence by the Numbers. (2017, June 22). Retrieved December 08, 2017, from https://everytownresearch.org/gun-violence-by-the-numbers/#foot_note_9

- Hands-On Machine Learning with Scikit-Learn and TensorFlow by Aurélien Géron (O’Reilly). Copyright 2017 Aurélien Géron, 978-1-491-96229-9

- Hoyert DL, Xu JQ. (2012) Deaths: Preliminary data for 2011. National vital statistics reports; vol 61 no 6. Hyattsville, MD: National Center for Health Statistics.

- Irvin-Erickson, Y., Bai, B., Gurvis, A., & Mohr, E. (2016). The Effect of Gun Violence on Local Economies. Urban Institute.

- Sheley, J.F. and Wright, J.D. (1998). High School Youths, Weapons, and Violence: A National Survey, Washington, DC: National Institute of Justice. NCJ 172857.

- Sherman, L.W., Shaw, J.W., & Rogan, D.P. (1995). The Kansas City Gun Experiment. National Institute of Justice.

- U.S. Census Bureau. 2010. http://www.census.gov/prod/cen2010/briefs/c2010br-06.pdf

- Webster, D., Crifasi, C.K., & Vernick, J.S. (2014) Erratum To: Effects of the Repeal of Missouri’s Handgun Purchaser Licensing Law on Homicides. Journal of Urban Health 91, no. 3

Appendix 1: Feature Codebook

Variable	Desciption
Demographic Data	
MedianIncome	Median Income of total population
PercentAA	% of total population that has an associates degree
PercentBA	% of total population that has a bachelors degree
PercentBlack	% of total population is black
PercentBornCitizen	% of total population who are US citizens by birth
PercentCivilUnemployed	% of people in the labor force who are unemployed civilians
PercentF_0_17 – PercentF_over_75	% of total population that is female between the ages of 0-17, etc.
PercentFemale	% of total population that is female
PercentGED	% of total population that has a GED
PercentHS	% of total population that has a high school diploma
PercentInCivilian	% of people in the labor force who are in the civilian labor force (non military)
PercentInLaborForce	% of the total population in the labor force
PercentM_0_17 – PercentM_over_75	% of total population that is male between the ages of 0-17, etc.
PercentMA	% of total population that has a masters degree
PercentMale	% of total population that is male
PercentMilitary	% of people in the labor force who are in the military
PercentNative	% of total population that is American Indian or Native Alaskan
PercentOther	% of total population is another race
PercentPacific	% of total population that is pacific islander or native Hawaiian
PercentPhD	% of total population that has a doctorate
PercentProf	% of total population that has a professional degree
PercentRural	% of total population living in urban areas
PercentSomeColl	% of total population that has completed some college
PercentTwoPlus	% of total population that is two or more races
PercentWhite	% of total population that is white
Total	Total population
state	State
GISJOIN	Tract code
Gun Violence Data	
tract_fips	Federal Information Processing Standards (FIPS)
num_killed	Number of people killed
num_incidents	Number of incidents
city_or_county	City or county
square_mileage	Square mileage of the census tract
tract_population	Tract population
population_black	% of total population is black
over_age_25_with_no_hs_degree	% of total population 25 or older with no High School degree
below_poverty_line	% of total population below the poverty line
State-Level Gun Laws	
lawA	States that Prohibit Open Carrying of Handguns
lawB	States that Require a Permit or License to Openly Carry Handguns
lawC	States that Otherwise Restrict Open Carrying of Handguns in Public Places
lawD	States that Generally Prohibit Open Carrying of Long Guns
lawE	States that Restrict, But Do Not Prohibit, the Open Carrying of Long Guns
lawF	Restriction on # of weapons
lawG	Waiting period
lawH	Safety Course for Concealed Carry
lawI	State or local Dealer License
lawJ	Assualt weapons Prohibited
lawK	Machine Guns phobitied
State-Level Gun Laws	
tract_fips	Federal Information Processing Standards (FIPS)
longitude	Tract longitude
latitude	Tract latitude

Appendix 2: 10 Randomly Selected Data Instances

tract_fips	PercentM_0_17	PercentM_18_29	PercentM_30_59	PercentM_60_74	PercentM_over_75	PercentF_0_17	PercentF_18_29	PercentF_30_59	PercentF_60_74
12083000402	8%	9%	25%	9%	2%	7%	6%	18%	12%
48113008701	16%	4%	11%	5%	2%	20%	12%	20%	6%
45011970300	11%	7%	15%	5%	1%	23%	7%	19%	8%
13135050610	18%	5%	23%	4%	1%	14%	5%	25%	4%
45051030103	8%	3%	14%	18%	6%	6%	6%	17%	16%
29159480600	22%	6%	19%	4%	1%	13%	7%	20%	5%
13115002100	11%	11%	24%	6%	1%	12%	7%	18%	4%
48309002700	14%	8%	22%	6%	1%	17%	6%	20%	5%
39049009373	18%	5%	18%	3%	0%	19%	11%	21%	4%
17031250200	18%	12%	18%	4%	0%	14%	9%	18%	7%

PercentF_over_75	PercentWhite	PercentBlack	population_black	PercentNative	percentPacific	PercentAsian	PercentOther	PercentBornCitizen	PercentMale
4%	98%	0%	0%	0%	0%	1%	1%	96%	54%
3%	4%	94%	94%	0%	0%	0%	1%	96%	38%
4%	38%	60%	58%	1%	0%	0%	0%	99%	39%
1%	73%	15%	15%	1%	0%	7%	0%	87%	51%
5%	70%	28%	32%	0%	0%	2%	0%	97%	50%
4%	66%	12%	12%	0%	0%	0%	10%	90%	52%
4%	49%	42%	41%	0%	0%	0%	7%	93%	53%
2%	60%	22%	20%	0%	0%	1%	12%	86%	51%
1%	26%	65%	70%	0%	0%	3%	0%	91%	44%
1%	19%	57%	56%	0%	0%	1%	22%	87%	51%

PercentFemale	PercentInCivilian	PercentMilitary	PercentUrban	PercentRural	PercentTwoPlus	r_age_25_no_hs_deg	PercentHS	PercentGED	PercentSomeColl
46%	100%	0%	0%	100%	0%	20%	34%	15%	13%
62%	100%	0%	100%	0%	3%	35%	38%	5%	23%
61%	100%	0%	67%	33%	1%	15%	30%	2%	25%
49%	100%	0%	97%	3%	9%	5%	15%	4%	21%
50%	100%	0%	0%	100%	1%	10%	37%	1%	22%
48%	100%	0%	92%	8%	24%	34%	28%	4%	21%
47%	100%	0%	82%	18%	3%	40%	24%	13%	19%
49%	100%	0%	100%	0%	9%	21%	28%	10%	24%
56%	100%	0%	100%	0%	12%	12%	33%	3%	33%
49%	100%	0%	100%	0%	1%	27%	30%	5%	20%

PercentAA	PercentBA	PercentMA	PercentProf	PercentPhD	MedianIncome	below_poverty_line	PercentInLaborForce	ercentCivilUnemploye	Total
6%	3%	4%	0%	0%	32,639	22%	46%	13%	4,876
2%	3%	0%	0%	0%	17,509	52%	52%	10%	4,221
13%	10%	3%	0%	1%	28,962	38%	58%	15%	4,170
9%	29%	12%	4%	1%	100,708	4%	67%	8%	16,825
12%	10%	4%	2%	0%	46,350	19%	48%	7%	2,982
10%	3%	0%	0%	0%	20,446	45%	52%	23%	2,125
2%	5%	1%	1%	0%	26,071	32%	49%	18%	4,553
4%	8%	3%	0%	1%	31,677	21%	67%	15%	4,109
8%	8%	2%	1%	0%	43,764	34%	67%	14%	6,627
9%	5%	3%	0%	0%	38,250	28%	65%	23%	2,860

tract_population	lawA	lawB	lawC	lawD	lawE	lawF	lawG	lawH	lawI
5,046	1	0	0	1	0	0	1	1	0
4,016	0	1	0	0	0	0	0	1	0
4,127	1	0	0	0	0	0	0	1	0
16,678	0	1	0	0	0	0	0	0	1
3,015	1	0	0	0	0	0	0	1	0
2,201	0	1	0	0	0	0	0	1	0
4,457	0	1	0	0	0	0	0	0	1
4,171	0	1	0	0	0	0	0	1	0
6,327	0	0	0	0	0	0	0	1	0
2,968	1	0	0	1	0	0	1	1	0

lawJ	lawK	square_mileage	city_or_county	latitude	longitude	state	num_killed	num_incidents
0	0	87.77	Fort Mc Coy	29.3906	-81.9337	florida	1	1
0	1	1.54	Dallas	32.6956	-96.7693	texas	1	1
0	1	26.7	Blackville	33.3022	-81.3282	south carolina	1	1
0	1	12.63	Dacula	34.0471	-83.9174	georgia	1	1
0	1	31.7	Longs	33.9747	-78.7425	south carolina	3	3
0	1	3.35	Sedalia	38.7138	-93.2323	missouri	1	1
0	1	10.27	Rome	34.2452	-85.1866	georgia	1	1
0	1	1.06	Waco	31.5447	-97.1797	texas	1	1
0	0	1.21	Columbus	39.9129	-82.8456	ohio	3	3
0	1	0.19	Chicago	41.9098	-87.7462	illinois	2	1

Appendix 3: Contributions of each team member

Chelsea P Daniels: Business Understanding & Data Understanding

Oussama Fartri: Data Preparation & Modeling & Evaluation

Hiroshi Horikawa: Literature Review & Deployment

Tianze Zhou: Literature Review & Deployment