Predictive Modeling

I. Problem Structure and Label

In order to use the data in a predictive model to give insight into gun violence trends, we need to define the label that we want to predict. I decided that we should set the problem up as a classification problem rather than a regression problem because predicting the exact number of deaths would be infeasible.

I set the problem up as a binary classification with the following structure:

- Each row in our data represents an observation: features for a state in a given time period
- For each row, a 0/1 label is defined as whether **the number of gun homicide will increase from this period by more than 20% in the next time period**.
- The rows are sorted by date, and predictions will be made for each time period using **all data available** prior to that time period.

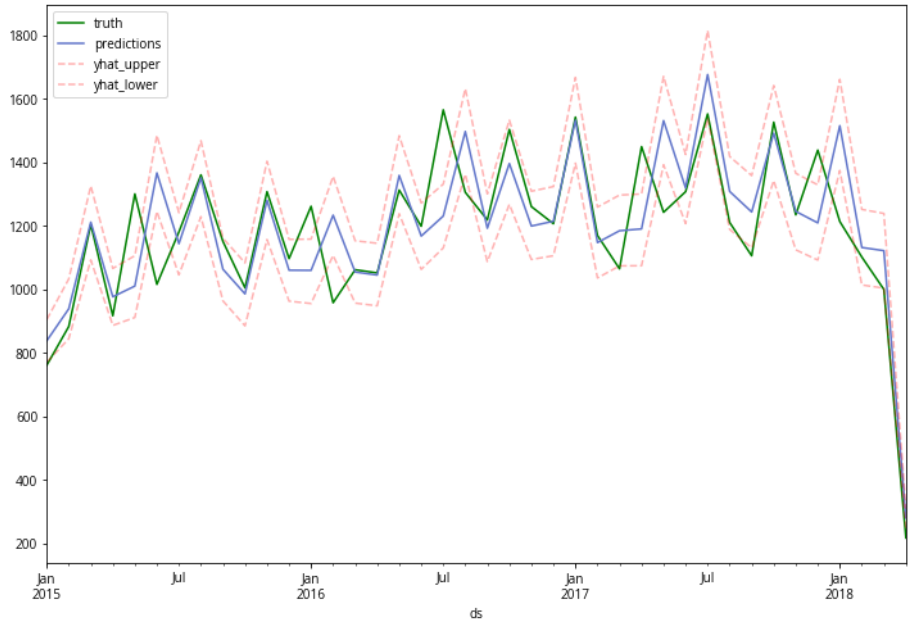
Example structure for monthly time period:

state	**time_period**	**features...**	**label**
Alabama	2014-02	...	0
Alaska	2014-02	...	0
...	...	...	...
Alabama	2016-12	...	1

II. Monthly Modeling

Pipeline and Modeling

In this section, we try to predict gun homicide increases for a state in each month. We have data on gun violence incidents for 2014-2017. This means we'll be training on 2014-2016 and predicting for each of the 12 months in 2017. From the visualization notebook, we saw that there is seasonality and trends in gun homicide rates. In order to use this in a model along with other features, I decided to use a time series analysis algorithm, Facebook's Prophet, to make numeric predictions on gun violence. Here's a visualization of the performance of the time series analysis (more information can be found in the time-series notebook):



I use the output of the time series analysis along with the other features as input into classification models to make predictions on the label. This process is repeated for each state separately. The reason for this decision is that each state has different gun violence patterns and feature effects. Here's a simple diagram of the pipeline:

I excluded states with less than 2 average monthly gun homicides. This was because gun homicides is not as significant of an issue in those states, and the label has high fluctuation.

Results and Challenges

I found that XGBoost had the best performance among the models trained. When training on years 2014-2015 and making predictions for 2016, the accuracy was 69%. Training on years 2014-2016 and making predictions for 2017, the accuracy was 75.5%.

Adding just one more year of training increased accuracy by 6.5%. It seems that the model can still learn, but suffers from lack of data.

Furthermore, the features we are using are annual features. This means we won't be able to capture the monthly changes for each feature, and the most useful feature was the prediction for the number of gun homicides from the time series analysis model.

III. Annual Modeling

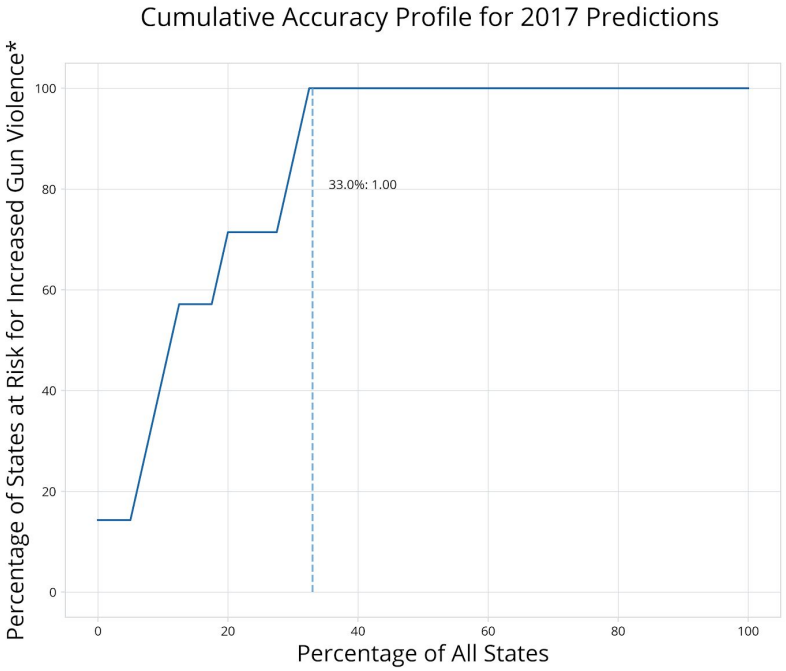
Pipeline and Modeling

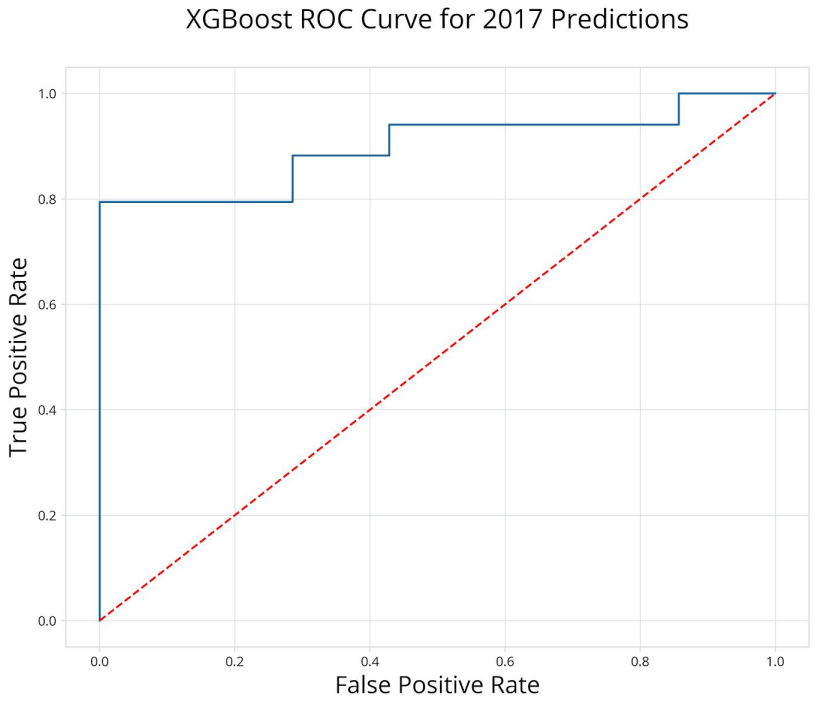
Here, we predict gun violence increases for each state at the annual level. We have gun violence information from 2000-2017, and we will be able to make better use of our annual features here. We have insufficient data points (16 data points from 2000-2016 for training) to be using time series analysis here. We also won't be training a separate model for each state because there are too many parameters to learn from the limited number of data points. Furthermore, we only used the features that were available from 2000 to 2017. This includes features on gun violence rates, gun control laws, and party support for presidential candidates.

Just as with the monthly modeling, I excluded states with less than 20 average annual gun homicides.

Results and Challenges

Once again, the best performing model was XGBoost. We ended up with 85.4% accuracy with 42.9% recall rate. By tweaking the sensitivity, we got 80.5% accuracy and 86% recall. All of the states that will see a 20% increase in gun homicides were within the top 33% states classified most at risk model. Here are the cumulative accuracy profile and the ROC curve:





The total number of laws and laws that restrict access to guns for children had high feature importance in the model. More details can be found in the annual_modeling notebook.

Due to the limited availability of some of the features, we were not able to use all of them. This was unfortunate because for many of these features, we saw correlation with gun violence during the visualization part.