**Real world Protest Violence Predictor using supervised learning model**

INTRODUCTION

Since the dawn of democracy, protests have played an essential role in maintaining the power of the people. Unfortunately, the powers that be have also endeavored to maintain that power, and often have done so violently. A report by the Center for Strategic and International Studies Risk and Foresight Group points to an annual increase of mass protests worldwide of an average of 11.5 percent from 2009 to 2019 ([report](https://csis-website-prod.s3.amazonaws.com/s3fs-public/publication/200303_MassProtests_V2.pdfchttps://csis-website-prod.s3.amazonaws.com/s3fs-public/publication/200303_MassProtests_V2.pdf)). The results of protest can range from accommodation to violence and casualties. But what leads to these violent reactions?

There are plenty of factors that are believed to increase the rate of protests, though the available data already suggest the trend will continue. Knowing this, it is important to understand the effect of demonstrations, however, this is highly dependent on the root motivation for a protest. One of the most prominent effects is the economic impact, though the scale can vary. This can be seen in many occasions where, for example, riots have had long term economic repercussions to the surrounding areas, with a potential to shape the quality of life of residents. Economic ramifications tend to be larger when tied to political instability. Some countries actively suppress protest movements and others have yet to develop a systemic response to mass mobilization. New ingredients are being added to the mix with an increase in technology use by citizens and governments alike.

Problem Statement

The aim of this project is to identify potential avenues to influence and drive change through mass mobilization, while sparing both protesters and state alike the resources associated with demonstrations, namely time and money.

THE DATA

In the data set explored (“[Mass mobilization Data Project Dataverse](https://dataverse.harvard.edu/dataverse/MMdata)”), David H. Clark (Binghamton University) and Patrick M. Regan (University of Notre Dame), recorded protest instances against governments across 162 countries worldwide noting protester demands and government responses. More than 17k observations were made across 1990-2020. For each event, the project notes protestor demands, government responses, data surrounding the location of the demonstration (country, region) as well as the identity and number of protesters.

**Data Wrangling and Cleaning**

The unit of observation for this data set is the protest-country-year, where each demonstration is recorded individually within country and year. There were 26 variables recorded for 17,145 events.

['id', 'country', 'ccode', 'year', 'region', 'protest', 'protestnumber','startday', 'startmonth', 'startyear', 'endday', 'endmonth', 'endyear','protesterviolence', 'location', 'participants\_category', ‘participants', 'protesteridentity', ‘protesterdemand1’,'protesterdemand2','protesterdemand3','protesterdemand4',

'stateresponse1', 'stateresponse2',’stateresponse3', 'stateresponse4','stateresponse5', 'stateresponse6', 'stateresponse7', 'sources','notes']

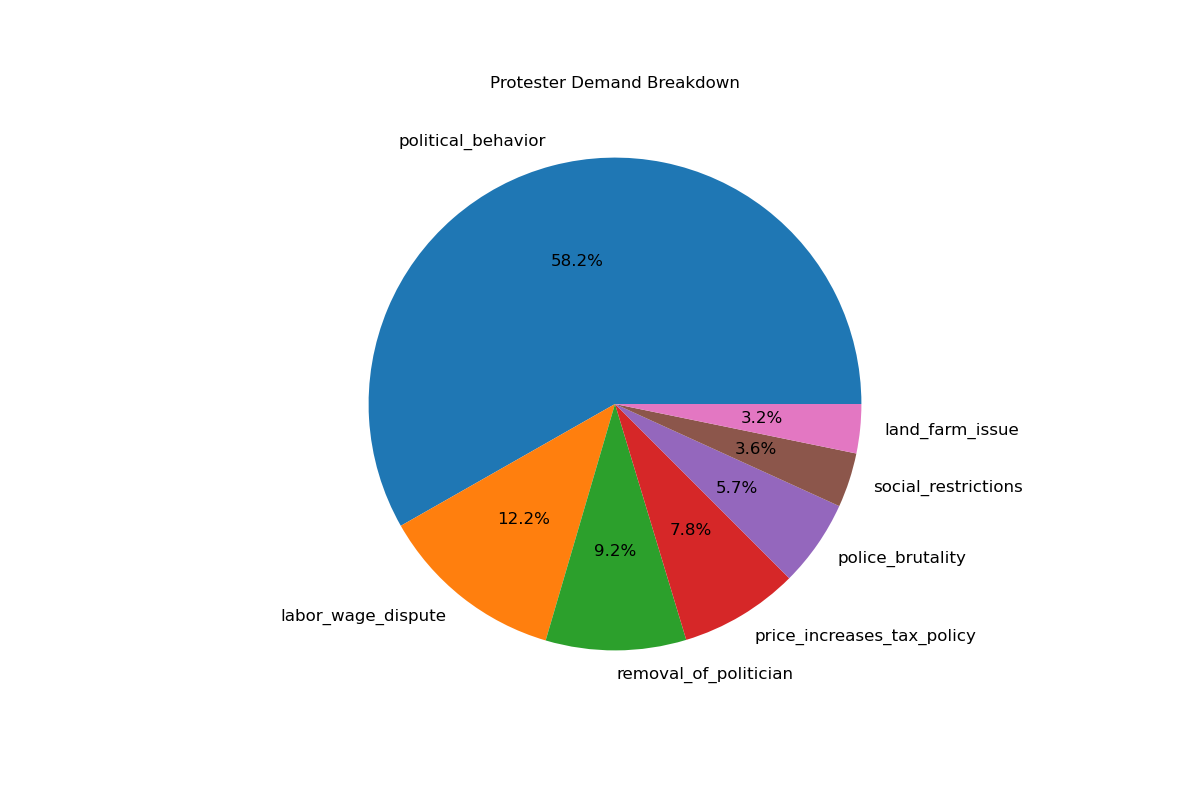
In addition to cleaning each column for data type, whitespace and placeholder characters, new columns were created to denote the presence or absence (0/1) of each protester demand (n=4) and state response (n=7). After this was done, helper columns like combo demands/responses, a count of each and first and last response were created so that they may be explored for significance.

Given that the number of participants and inherently the participant category could provide valuable insight into a state’s violent response there was an effort to fill in as many observations in the participants field and to complete the participants category column since it was added part way through the data collection endeavor. I first standardized the participant number column which contained a mix of data types. I considered participants category when available and chose the mid range of the category if available i.e if participants was noted as 100s, and participant category is 100-999, new participants\_int number is 450 . Otherwise the lowest notated number i.e if participants was noted as 100s, and participant category was not available, participant\_int is 100.

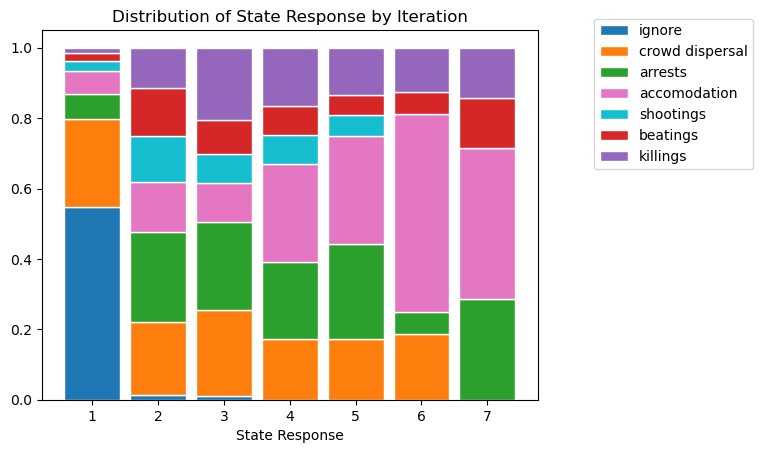
Any observations that were missing a state response were dropped since it is the target variable. The new data set became 53 columns and 14,482.

**Exploration Data Analysis**

All protest motivations were categorized as 7 types of issues that would inspire protest behavior, noting up to 4 for each observation.

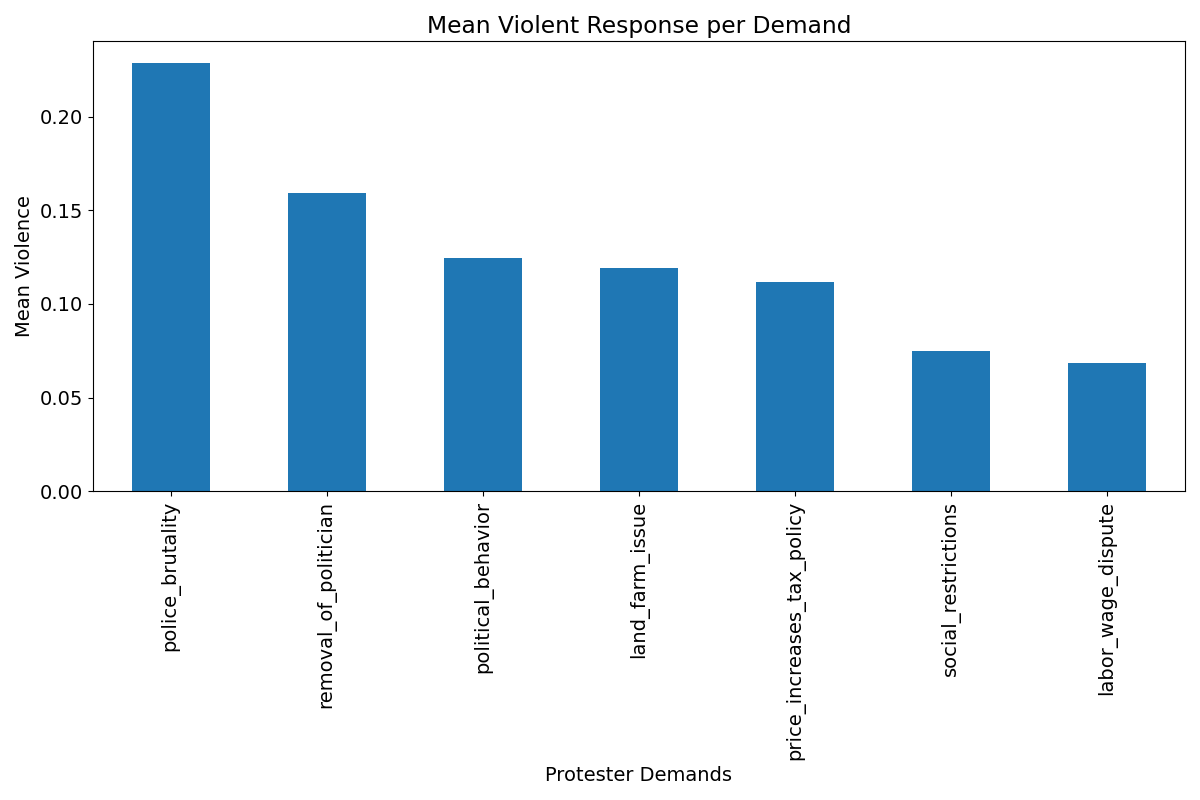
58.2% of protests were due to political behavior. This is consistent with the researcher’s expectations since the scope of this category is vast (includes who rules and how, who can participate in elections, decisions, as well as choices made by leaders that have both domestic and international impacts.)

Every state response was categorized into 7 classes, where up to 7 responses were noted per protest.



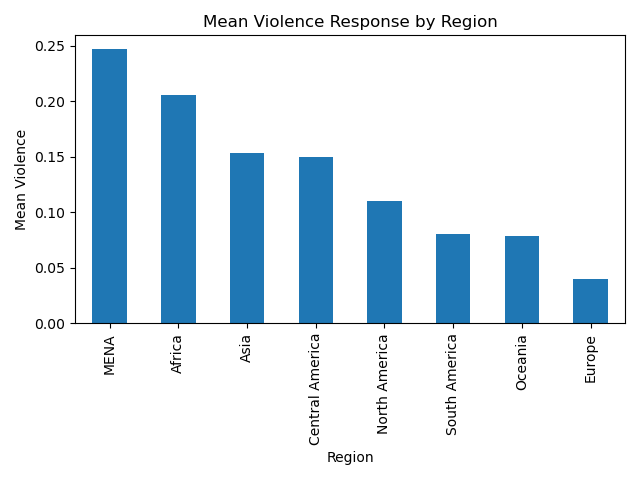
The state response number stacked graph shows that a trend by the states to ignore or disperse gatherings while there is also an increase in the more violent responses (shootings, beatings, killings) as the states continued to respond. It’s important to note that while the first response has a very small percentage of violent responses (shootings, beatings, killings) this increases to almost 40% of the second responses.

A summary of the distribution of protests, shows that most recorded observations are in Europe, followed by Africa and Asia. A majority of demonstrations were 1 day long and they ranged from 50 to over ten thousand participants, with most having roughly 1000 protestors.

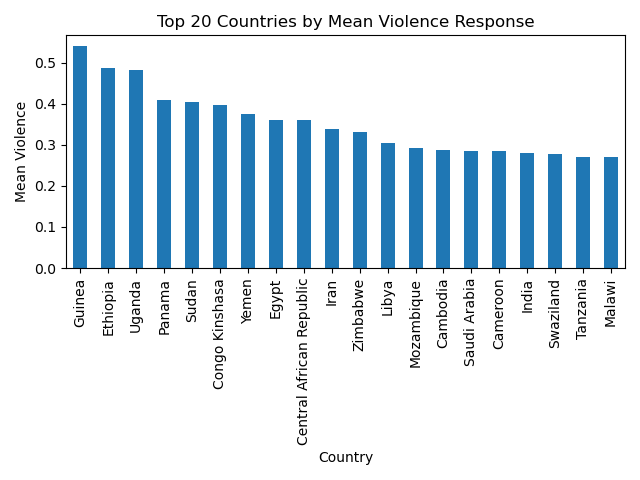
The pictures provided by the data are very different when seen through the lens of a violent state response. 

While political behavior was the most common reason to demonstrate, it was mass mobilizations against police brutality that elicited the most violence from the states. Sadly, this makes sense and is consistent with what one would expect. The demand that resulted in the second highest violence mean was removal of a politician. Again, unfortunately, it makes sense that the reaction of a leader whose power is being threatened is to suppress anyone who is against him/her. Lastly, I will note that all demands experienced violent responses, which means that although they might be indicators of this behavior, there might be other features driving this.

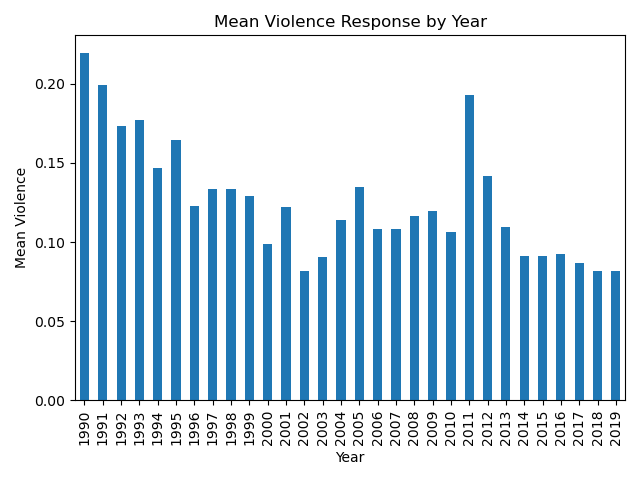
Police brutality, removal of politician, labor wage dispute and social restrictions were all statistically significantly different from the mean violent response overall.



Even though we see that most demonstrations occur in Europe (n=4808), Africa (n=3037) and Asia (n=2977), it is the Middle East (MENA, n=1093) that on average reacted most with violence to mass mobilizations. This could be due to the overwhelming presence of capitalist, industrial, wealthy, and developed countries within Europe, compared to the rest of the noted regions - though this is speculation. Another potential element to be considered is the type of government in a specific country. Looking at the breakdown of mean violence per country below we can more confidently claim these influences.

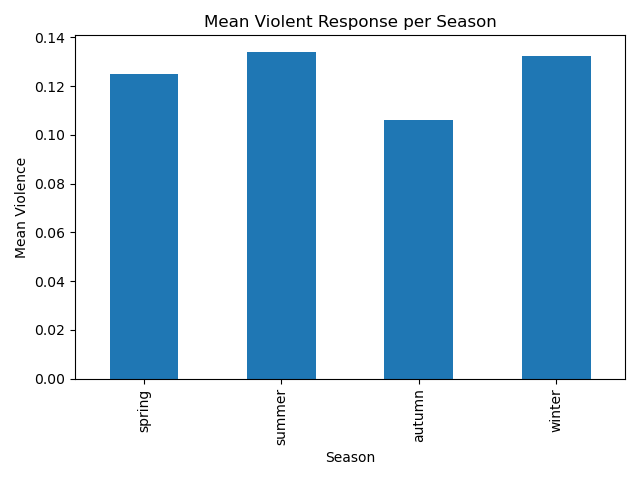
Continuing to explore the context and the setting of each protest. The year and date of each observation were noted and explored for potential trends or indications that they influence a violent state response.

A view of mean violence across the years shows that there are some years that recorded more violence than others. Additionally, slight trends of decreasing and increasing violence can be seen. It would be interesting to overlay world events to see if the behaviors observed coincide with our data.

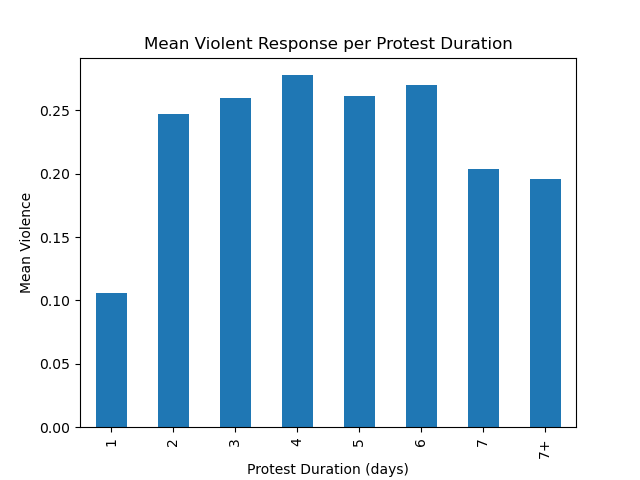


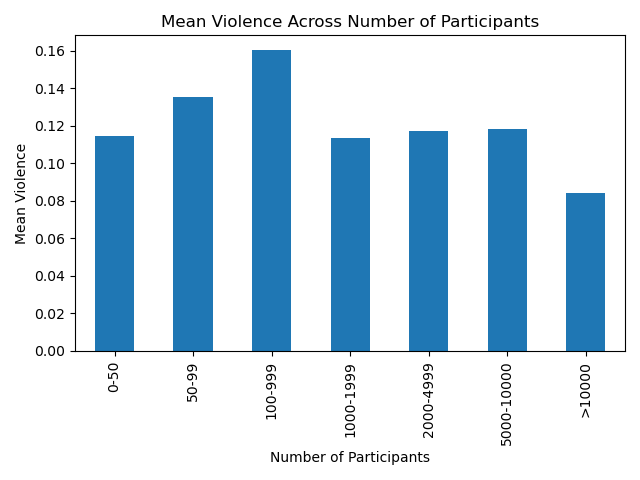
Another aspect that was explored by the researchers, was the time of year that the demonstration took place. Depending on the hemisphere where a certain country is located it meant different weather and overall conditions for the participants and the states alike. Multiple potential behaviors could be expected, such as, maybe a decrease in demonstrations in the colder months, or protests during election months. The data doesn’t show a particular trend for violence across the year as shown with the graph below.

In order to be able to see how violence was distributed across different seasons, I filtered each country by its location in the southern or northern hemisphere. After that, I mapped the start date to a season. This shows a slight increase in violence in both the summer and winter, presumably the harshest seasons in terms of weather for both hemispheres.

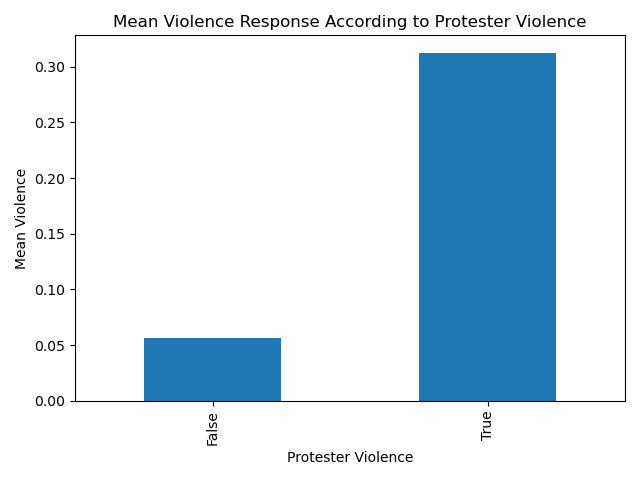


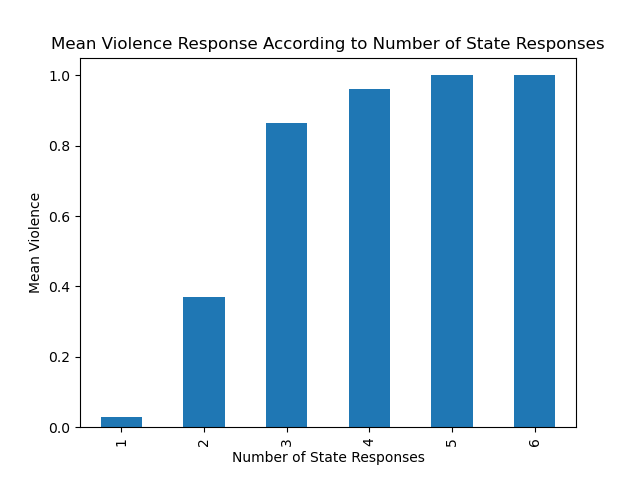
The length of a demonstration was investigated in this data set. An overwhelming number of protests were only one day. However, if a protest ran across multiple days, the chance of violence increased.

  
Another characteristic explored was the number of participants and the incidence of violence. As previously mentioned the majority of protests had between 1000 and 2000 participants. Curiously the largest protests are the ones that average the least violent responses, this could be because a larger mass of people might be more intimidating to the state, it could also mean more news coverage.



An additional, pertinent aspect of protests was whether there was protestor violence or not. The graph below does show that violence begets violence in the observed cases.



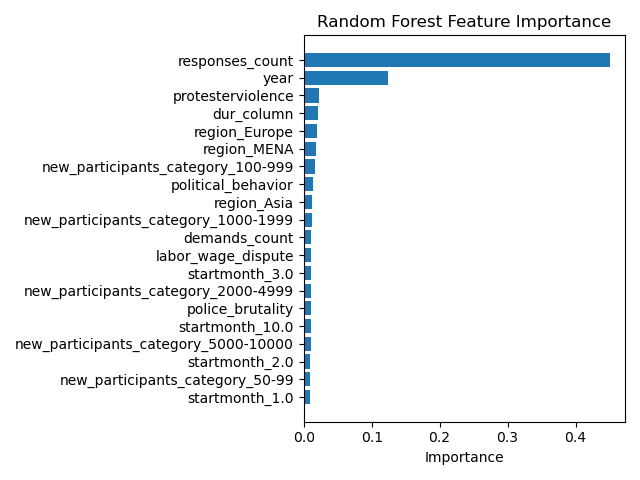
When observing whether a demonstration was brought upon by multiple demands, it seemed that if a group had more than 3 grievances (of note, up to 4 were noted in this data set), this would result in an increased chance of violence. 

Lastly, the data shows that an increased number of responses almost assuredly resulted in a violent final response. we can start to see this in the distribution of responses. There are no *ignore* responses after the third iterative response.

Out of the features explored above, the following displayed statistical significance:

labor\_wage\_dispute[demand], social\_restrictions[demand], removal\_of\_politician[demand], police\_brutality[demand], country, region, year, season, startmonth, duration, participant, response count.

**Machine Learning**

In order to further substantiate the feature selection for this project, the data set was modified to exclude country, participant number, any column related to the state responses (each noted response as well as first and last response) and the noted protester demand columns. Placeholder columns were created for each of the remaining variables. A random forest feature identification analysis was subsequently run. The findings coincide with some of the findings from previous data exploration analysis. 

**Preprocessing**

In order to create a data set that could model real life and be useful, the duration column, response count, and protester violence field were also dropped from the data set since these were not features that can be known at the start of a demonstration.

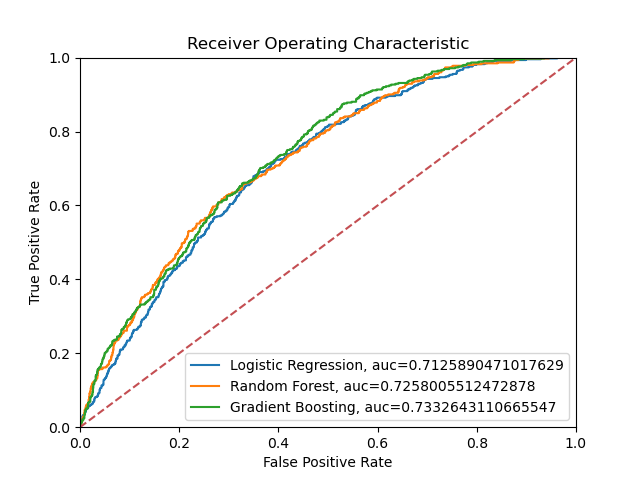
RESULTS

Three different models were created in order to try to predict state violence with the first day knowledge. GridSearchCV was used to find the best parameters and to test their ability to discern a violent response.

**Model Selection**

|  |  |  |
| --- | --- | --- |
| Classifier Performance | | |
| **Classifier** | **ROC-AUC Score** | **Best Hyperparameter Values** |
| **Logistic Regression** | 0.713106042631401 | {'C': 0.0001} |
| **Random Forest** | 0.741263156769126 | {'max\_depth': 7, 'n\_estimators': 300} |
| **Gradient Boosting** | 0.741930638152778 | {'n\_estimators': 200, 'max\_depth': 4, 'learning\_rate': 0.05} |

**Model Performance**

All three models performed similarly and there is no obvious reason to choose one over the other. 

CONCLUSIONS

In summary, there are features that might be able to indicate the potential for a violent response from the state at the beginning of a demonstration, but no true predictor. Given the findings, citizens can shift their approach when possible in order to avoid the losses associated with state violence. Some variables that increase the likelihood of a state’s violent response, namely, the year and country/region cannot be modified and remain an opportunity for change. Conversely, protesters can make an effort towards peaceful, short demonstrations. With the advent of new technologies, the visibility of a controlled demonstration can increase, bypassing the inherit costs of a protest.

The major findings of this project include the knowledge that the longer, and more times a state gets to respond to a demonstration the more likely it is for it to display violence.

Of note, this data set does not include any demonstrations in the USA, where we have experienced an increase in demonstrations and protests, especially in the last 4 years [Make note of concentrations political/race etc]

In an ideal world states would endeavor to learn about motivations and more importantly how to approach mass mobilizations moving forward so as to not compromise people’s safety whilst protecting the country of the monetary loss and \_of protests.

REFERENCES

Citation: Mass Mobillization Data project protests against governments, all countries, 1990-2020. Visit the project page at http://www.binghamton.edu/massmobilization/ (2019-02-07)