Capstone Project 1

Josh You

Problem

Terrorism is a major social problem because of the direct damage caused by terrorist attacks and because of its broader negative side-effects to society, such as widespread fear and political instability. This project seeks to analyze what factors predict the outcome of a terrorist attack; namely, whether a given terrorist attack or attempted terrorist attack will be successful, and whether it will lead to any casualties.

Client

The client for this project would be government agencies and security officials that are seeking to prevent acts of terrorism. The results of this analysis may help them make better decisions about which sources of potential terrorist attacks to focus on (for example, whether to focus on individuals or organizations), and which types and means of attack are most likely to succeed. The analysis may be relevant for both broad regulatory responses to terrorism (such as weapons control), and how governments should respond to specific threats of suspected terrorist plots.

Data

The data for this project will come from the Global Terrorism Dataset (GTD) compiled by the National Consortium for the Study of Terrorism and Responses to Terrorism (START), which is a research center based at the University of Maryland. The dataset contains information on 170,000 terrorist attacks between 1970 and 2016. The dataset is publicly available here, and a Codebook providing a guide to the dataset and its variables can be found here.

START compiles the dataset by combing through approximately one million media articles published daily around the world, using natural language processing to identify potential articles describing terrorist attacks, and then manually reviewing these candidate articles. The definition of terrorism can be ambiguous. For the purposes of this dataset, the researchers define terrorist attacks as having all three of the following conditions: 1) the incident is intentional, 2) the incident entails violence of the immediate threat of violence, and 3) the incident is carried out by sub-national actors (the researchers chose not to include state terrorism in the dataset).

Additionally, an incident must meet at least *two* of the following three criteria in order to be included (quoting from the Codebook):

> "Criterion 1: The act must be aimed at attaining a political, economic, religious, or social goal. In terms of economic goals, the exclusive pursuit of profit does not satisfy this criterion. It must involve the pursuit of more profound, systemic economic change.

Criterion 2: There must be evidence of an intention to coerce, intimidate, or convey some other message to a larger audience (or audiences) than the immediate victims.

…

Criterion 3: The action must be outside the context of legitimate warfare activities."

In addition to defining terrorism, START also needed to define what counts as an "attack". From the Codebook:

"The GTD does not include plots or conspiracies that are not enacted, or at least attempted. For an event to be included in the GTD, the attackers must be "out the door," en route to execute the attack."

Thus, the dataset contains both attacks that were successfully executed and those that were attempted, but only those attempts involving a tangible action. Plans that were not acted upon are not included.

The dataset contains over 100 variables describing various characteristics of each attack. The main outcome of interest is success, which is a binary categorical variable, indicating whether the attack achieved a tangible outcome. For example, a bombing attack is successful if a bomb detonates, though the bombing may not necessarily lead to fatalities or achieve the perpetrators' intended long-run goals. The dataset also contains information on the number of people killed and injured in the attack, and the extent of property damage. Other variables of interest include the region where the attack occurred, attack type (which includes categories such as assassination, armed assault, bombing, hijacking, and others), weapon type, and characteristics of the target.

There are several other organizations who compile data on terrorist attacks, and these datasets would be potential alternatives to the dataset I use. The Chicago Project on Security and Terrorism (CPOST) has compiled a database of all suicide attacks from 1974 through June 2016. While this data is restricted to suicide attacks, it contains additional biographical information on the individuals who carried out the attacks. The RAND Corporation maintains a dataset of terrorist attacks from 1968 through 2009.

Data Cleaning and Wrangling

The first step I took in cleaning my dataset was identifying which columns were relevant to my project. The original dataset had over 100 columns. I narrowed down the variables I am interested in to year, month, region, attack type, target type, weapon type, success (whether the attack was successful), number killed, latitude, and longitude.

Next, I dealt with missing values - The dataset did not contain values for every column I was interested in. Many values were missing in that the fields were blank:

The dataset also encoded missing values using specific codes indicating that the value for a column is unknown. For example, attack type is encoded by a number between 1 and 9, where 1 through 8 represent types of attacks, while 9 indicates that the type is unknown. I removed data points with missing values, though I had to be careful in which columns I used to filter. In this dataset, attack type, target type, and weapon type are represented by several columns because some events fall under multiple types of attack, target, and weapon. However, most rows have only one type for each of these sets of columns, so I only filtered based on whether the first column had a missing value (if the first column has a missing value, the second and third will have missing values too).

Cleaning the dataset to remove missing values brought the number of data points down to 140928, compared to the original 170,350.

The next data wrangling step was pivoting the categorical features. This meant separating variables with multiple possible values into a binary (dummy) variable for each of these values. Thus, each of the nine attack types was given its own column, with a value of 1 or 0 depending on whether a given attack had that type.

There are two motivations for this. First, weapon type, attack type, and target type were represented by multiple columns, because a terrorist incident may involve multiple weapons and targets, and may be described using more than one attack category. However, the second, third and fourth columns for these categories had many missing values, so turning these columns into dummy variables allowed me to retain all of that information while removing missing values. Second, the machine learning libraries I used required that categorical variables be converted to dummy variables in order to be interpreted correctly. Without converting them to dummy variables, these variables would be interpreted as quantities. For example, an attack type of 8 would be treated as 8 of *something*, and as more similar to an attack type of 9 than an attack type of 2.

Exploratory Data Analysis

I conducted exploratory data analysis on the cleaned dataset, and found a few interesting patterns. Here is a stacked area chart showing the number of attacks per year in the seven regions with the highest number of attacks.

Attacks by Region and Year
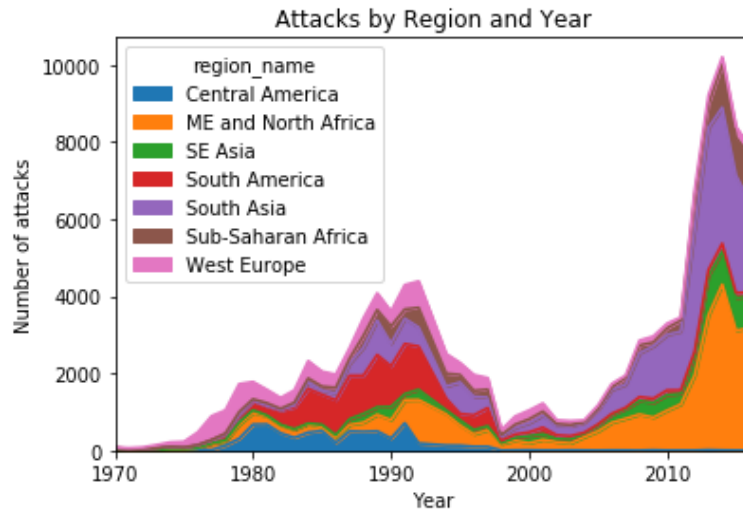
As this plot shows, there is a spike in terrorist attacks in the late 1980s and early 1990s in this dataset, and a larger spike occurring after 2010. There are clear differences in the distribution of attacks by region over time. South Asia and the Middle East and North Africa appear to be the main contributors to the recent spike in attacks. South America and Western Europe have seen large declines in both the absolute number of attacks, and their share in the world total, over time.

In addition, I found that terrorist attacks carried out by individuals had a lower success rate in this dataset, as shown in this graph.


Success Rate by Individual or Group
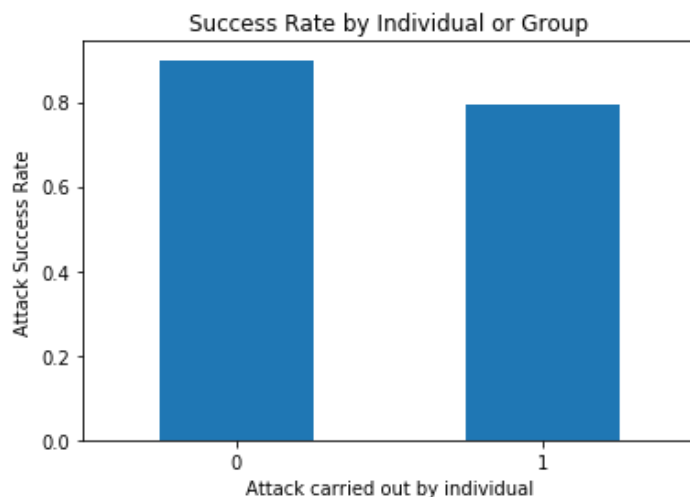
There were also large differences in success rates based on the type of weapon used. Radiological attacks rarely succeed, while attacks with sabotage equipment, incendiary weapons, and firearms usually succeed.
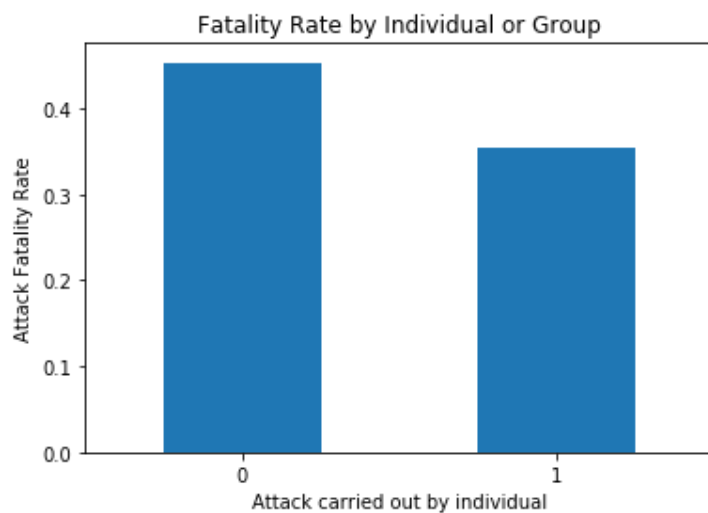
Success Rate by Weapon Type
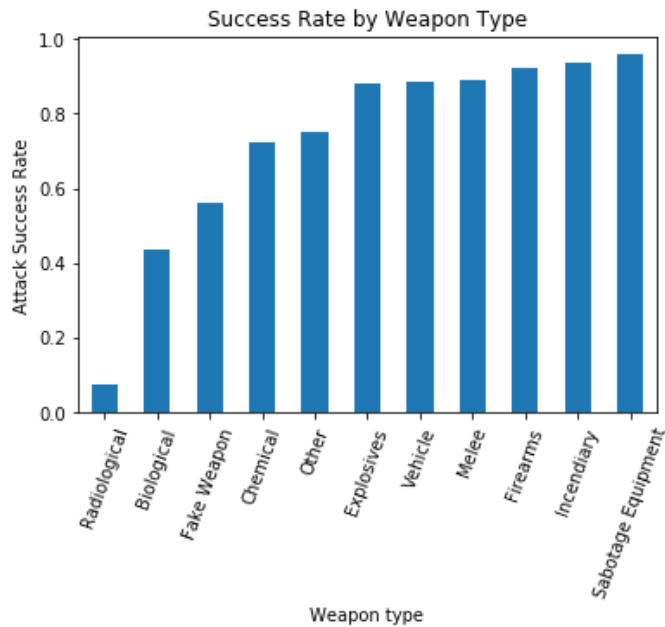
Success rates do not differ as much between attack types. However, assassinations are less successful than other types of attacks, while kidnappings and hostage attacks are usually successful.


Success Rate by Attack Type
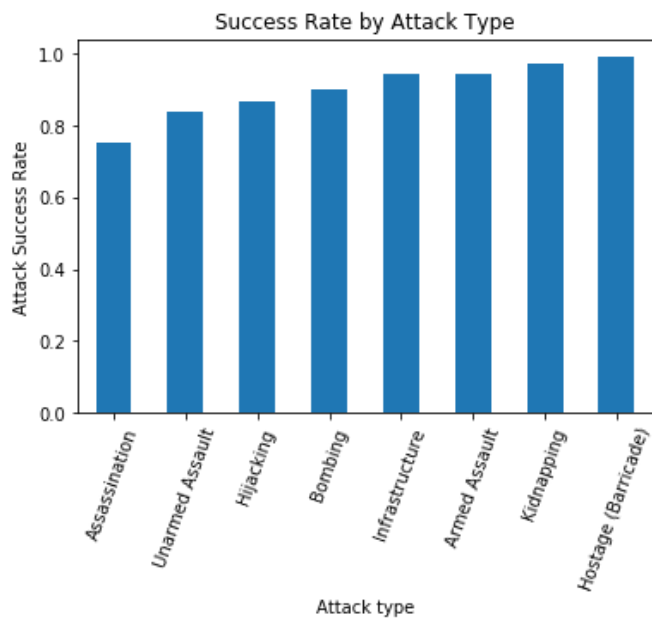
Next, I explored which factors affected the fatality rate of terrorist attacks (the proportion of attacks that result in more than one death). First, I show fatality rates by region:

Fatality Rate by Region
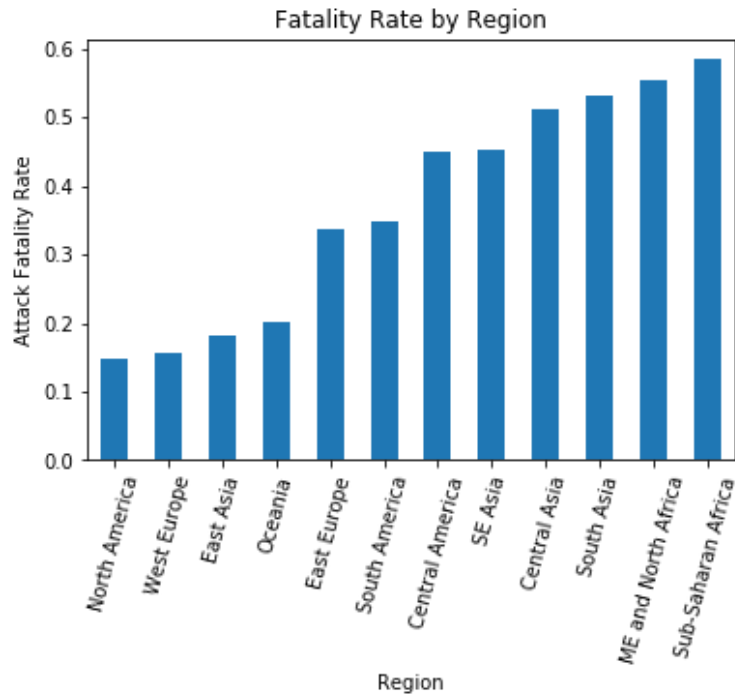
Most attacks lead to fatalities in Africa and the Middle East, while most attacks do not lead to fatalities in North America, Western Europe, and East Asia. There appears to be a relationship between the economic success and political stability of a region and a low fatality rate of attacks.

Fatality Rate by Individual or Group

Attacks carried out by individuals have a lower fatality rate, in addition to having a lower success rate.

Fatality Rate by Weapon Type
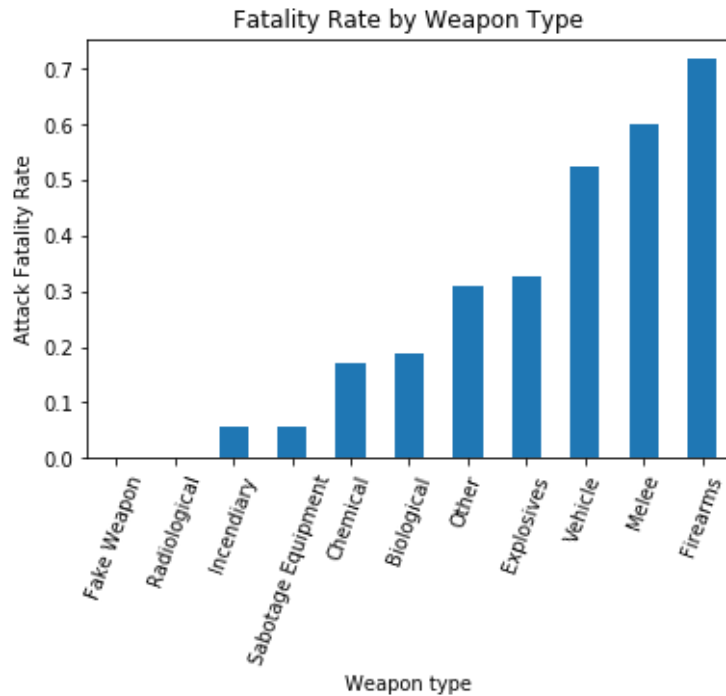
This plot shows large differences in fatality rates between weapon types. Unsurprisingly, firearms have a relatively high fatality rate. Attacks using fake weapons and radiological weapons do not lead to any fatalities in this data set.

Fatality Rate by Attack Type
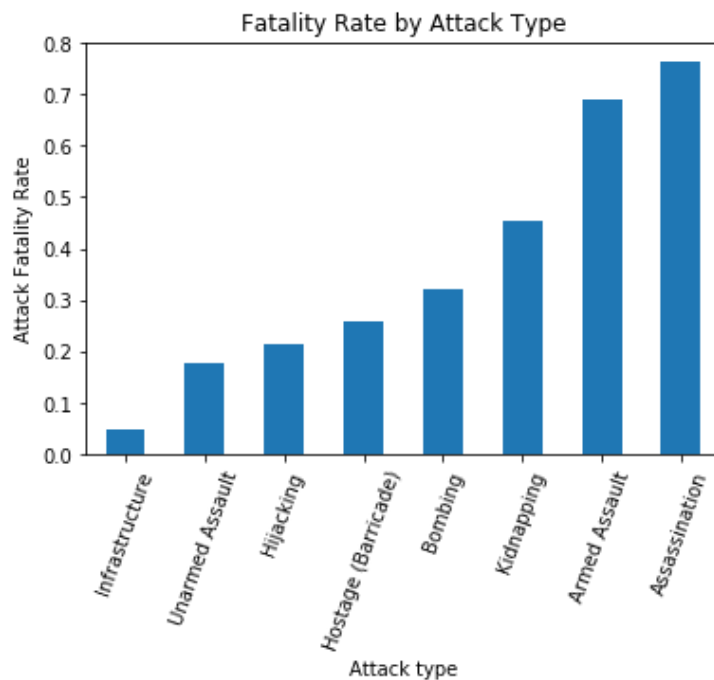
Finally, are large differences in fatality rate based on attack type. There differences are largely as one would expect: most assassinations and armed assaults lead to fatalities. Most infrastructure attacks and unarmed assaults do not lead to any deaths.

Inferential Statistics

I conducted several hypothesis tests based on the patterns found in the exploratory data analysis.

First, I tested the null hypothesis that the success rate of attacks attempted by unaffiliated individuals is the same as the success rate of attacks attempted by groups, against the alternative hypothesis is that the success rate for attacks attempt by unaffiliated individuals is smaller than the success rate for group attacks. I used a z-test for the difference in proportions, resulting in a p-value of $2.004*10^{-12}$. Thus, I reject the null hypothesis and conclude that group attacks have a higher success rate. In addition, I found that the 95% confidence interval for the difference in success rates was [0.067, 0.148], i.e. 6.7 percentage points to 14.8 percentage points.

Next, I used a chi-square test for independence to test the null hypothesis that the success rate does not differ by region, compared with the alternative hypothesis that success rates do differ by region. This resulted in a p-value of $4.96*10^{-277}$, so I reject the null hypothesis and conclude that attack success rates do vary by region.

I also used a chi-square test for independence to test the null hypothesis that success rate does not change based on attack type against the alternative hypothesis that success rate does vary by attack type. The p-value was small enough to be reported as 0.0 by the software package I used, so I reject the null hypothesis and conclude that success rates do vary by attack type.

Classification analysis

I used several machine learning algorithms to build a classifier that will predict the outcome of a terrorist attack based on its characteristics. The two types of classifiers I used were random forests and logistic regression.

First, I tried using logistic regression and random forests to predict the whether or not a terrorist attack will be successful. Because the data set is unbalanced (around 90% of attacks in this dataset are successful), every model I trained had very high overall accuracy, but poor recall for unsuccessful attacks. Recall is defined as the percentage of members of a label or class that the model correctly classifies. The logistic regression model had 91% accuracy overall after five-fold cross-validation, and 12% recall for unsuccessful attacks, meaning that out of the attacks that were in fact unsuccessful, the model correctly classified 12% of those attacks. Meanwhile, random forests performed somewhat better on this metric with 29% recall for unsuccessful attacks, and 89% accuracy overall after cross-validation.

I decided that I would change focus from trying to predict the success of terrorist attacks. From a practical standpoint, there is no real benefit to trying to distinguish the small percentage of observable terrorist attacks that are unsuccessful - a 90% risk of a terrorist attack succeeding is

not much different from a 100% risk. If I had a different dataset containing all the potential attacks in earlier planning stages that were not known to the public, then predicting the success of those plots would have greater practical value.

Thus, I turned to classifying whether a terrorist attack will lead to any fatalities. I first created a binary variable encoding whether the number of individuals that were killed was greater than zero, which would be the outcome variable.

Using logistic regression, I was able to achieve 75.68% accuracy after five-fold cross-validation on the training data. Using random forests, I achieved 79.37% accuracy after cross-validation on the training data. Due to the random forest classifier having the best cross-validated performance, I evaluated this classifier on the test data, resulting in 79.86% accuracy. This classifier had similar recall for both classes – 78% for attacks that do not lead to fatalities, and 82% recall for attacks that do lead to fatalities.

Model Intepretation

I will now turn to discussing the implications of these models. The software package I used computes the feature importance of each explanatory variable in a random forest classifier. This is a measure of how much a given variable reduces the error rate of a classifier, compared to if that feature did not exist. Here is a list of the ten most important features of the random forest classifier that I trained:

        1: ('longitude', 0.16768033050604564)
        2: ('latitude', 0.15080243328875254)
        3: ('Firearms (weapon type)', 0.10370863118224923)
        4: ('Armed Assault (attack type)', 0.10218115340790215)
        5: ('suicide', 0.07362376434450732)
        6: ('Assassination', 0.048022902258343728)
        7: ('Incendiary', 0.042557764982068966)
        8: ('Private (target type)', 0.03682039133140385)
        9: ('Bombing (attack type)', 0.036545864653674667)
        10: ('Explosives (weapon type)', 0.031064310050692701)

The logistic regression coefficients also provide a useful way to interpret the model. A coefficient in a logistic regression model determines how much the value of that variable affects the log odds of the outcome that is being predicted.

        1: ('suicide', 3.0213458205203239)
        2: ('Assassination', 2.016834843375956)
        3: ('Utilities', -1.3726656725329718)
        4: ('Bombing', 1.1736991486671537)
        5: ('Armed Assault', 1.0935978196358054)
        6: ('West Europe', -1.0888653258315837)
        7: ('Firearms', 1.0610168016587411)
        8: ('Incendiary', -0.99936114010553978)

9: ('Sub-Saharan Africa', 0.86498810035013007)
10: ('Private (target type)', 0.82521010612328349)
11: ('Explosives', -0.78830915800003554)
12: ('Terrorists (target type)', 0.76783482688930815)
13: ('North America', -0.73435826417999772)
14: ('East Europe', -0.66294058038591441)
15: ('Diplomatic (target type)', -0.64522009593268215)
16: ('ME and North Africa', 0.43658414221862041)
17: ('Police', 0.4029393779987997)
18: ('Media', -0.37174302673366888)
19: ('Melee (weapon type)', 0.36950300532127023)
20: ('crit2', -0.32414370636593465)
21: ('Military', 0.32233559215248081)
22: ('crit3', -0.27710300889933615)
23: ('Education', -0.23607871410845474)
24: ('South Asia', 0.20794125029178565)
25: ('Infrastructure', -0.19430931760737868)
26: ('Government', -0.16305932093938039)
27: ('Religious', 0.11704170738379295)
28: ('South America', 0.088595545824458027)
29: ('latitude', 0.013550001838052219)
30: ('Transportation', 0.011274177657975525)
31: ('longitude', 0.001104685038329375)

Above is the list of coefficients for the regularized logistic regression model. Every predictive variable except for latitude and longitude is a binary variable, so we can interpret the relative importance of these variables by how high the absolute value of the coefficient is. When an attack is a suicide attack or an assassination attempt, the attack is more likely to be deadly. Attacks on utilities targets, and those using incendiary weapons, are less likely to be deadly. Geographically, attacks in Europe and North America are less likely to be deadly, and attacks in Sub-Saharan Africa and the Middle East and North Africa are more likely to be deadly.

Discussion and Recommendations

The high importance of latitude and longitude indicate that the country or geographic location is a major factor in whether an attack will prove deadly. For officials interested in global terrorism, rather than just terrorism in their country or province, this suggests that resources and attention should be focused on certain areas. Based on the logistic regression coefficients, Africa and the Middle East should be prioritized, though more analysis should be done on which specific areas within those regions are most predictive.

The high importance of the Firearms and Bombing variables are also of interest. Both of these variables have positive coefficients in the logistic regression model. Curiously, however, the Explosives weapon type has a negative coefficient in logistic regression. Security officials should

consider policies that restrict the availability of firearms and bombs to those who may wish to use them for an attack.

These recommendations are largely consistent with existing counter-terrorism policy. Nonetheless, it may be useful to know that existing policy does have empirical support in terms of targeting the attacks that are most likely to have deadly consequences.

Next Steps

This analysis looked at attacks from around the world, while some may be interested in an analysis specific to one country or region.

In addition, I did not utilize every explanatory variable in the dataset. There are other variables that can potentially be added, after any necessary data cleaning steps.

There are also other outcome variables that can be investigated. For example, one could do user linear regression to predict the number of deaths that result from an attack, not just predicting whether an attack will lead to any deaths.