Gabrielle Agrocostea
Nandini Shah
Vidyavisal Mangipudi

Predictive Analytics of the Global Terrorism Database

Introduction:

Over the last decade, the world has witnessed an increase in violence and terrorism. The Global Terrorism Database (GTD) was created in 2006, and is maintained by the National Consortium for the Study of Terrorism And Response to Terrorism (START). The current GTD is the product of several phases of data collection efforts, each relying on publicly available, unclassified source materials.These include media articles and electronic news archives, and to a lesser extent, existing data sets, secondary source materials such as books and journals, and legal documents. Thus, it is no surprise that there will be a lot of sparse and missing data that may or maynot be relevant. In the remainder of this paper we will discuss the data we are working are working with, the development and implementation of different models and the analysis of the results on the test set.

What are the Questions we are trying to answer?

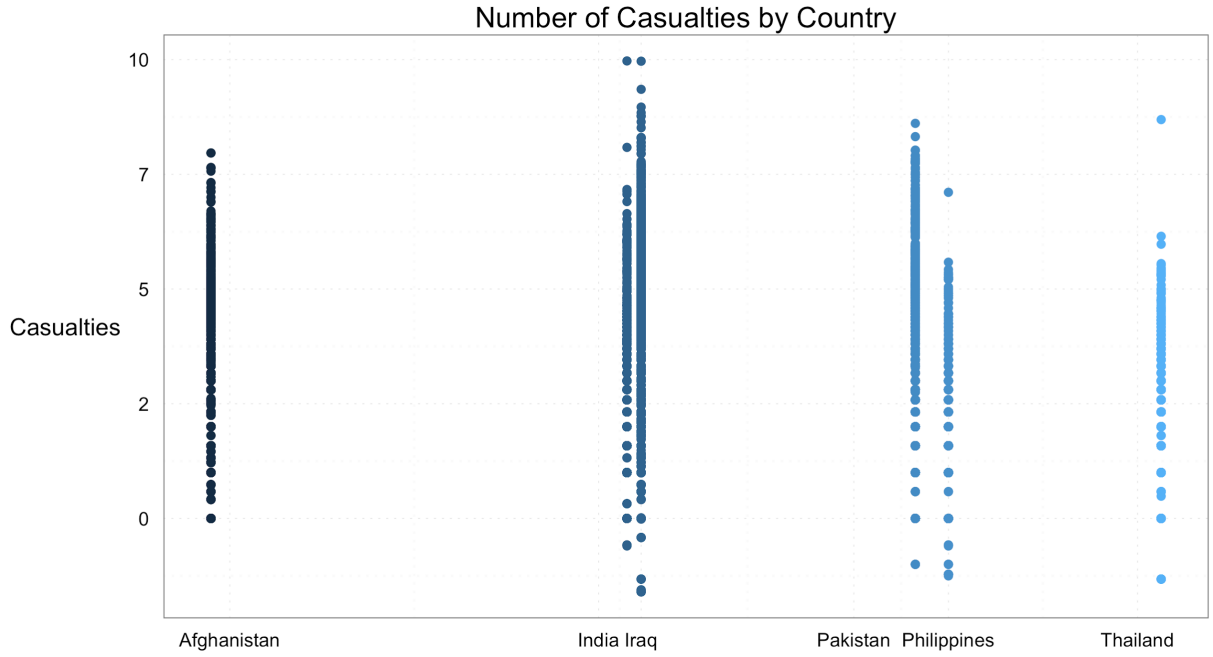
1. Predict whether a terrorist event will be successful: Success is defined in the GTD as an attempt by terrorists going ahead as planned, for example, a bomb explosion is considered a success if the bomb exploded, regardless of whether there were any casualties or not or whether the goal of the perpetrators was met.
2. Predict the number of casualties: The number of casualties is the sum of number of people killed and the number of wounded.

Exploratory Data Analysis:

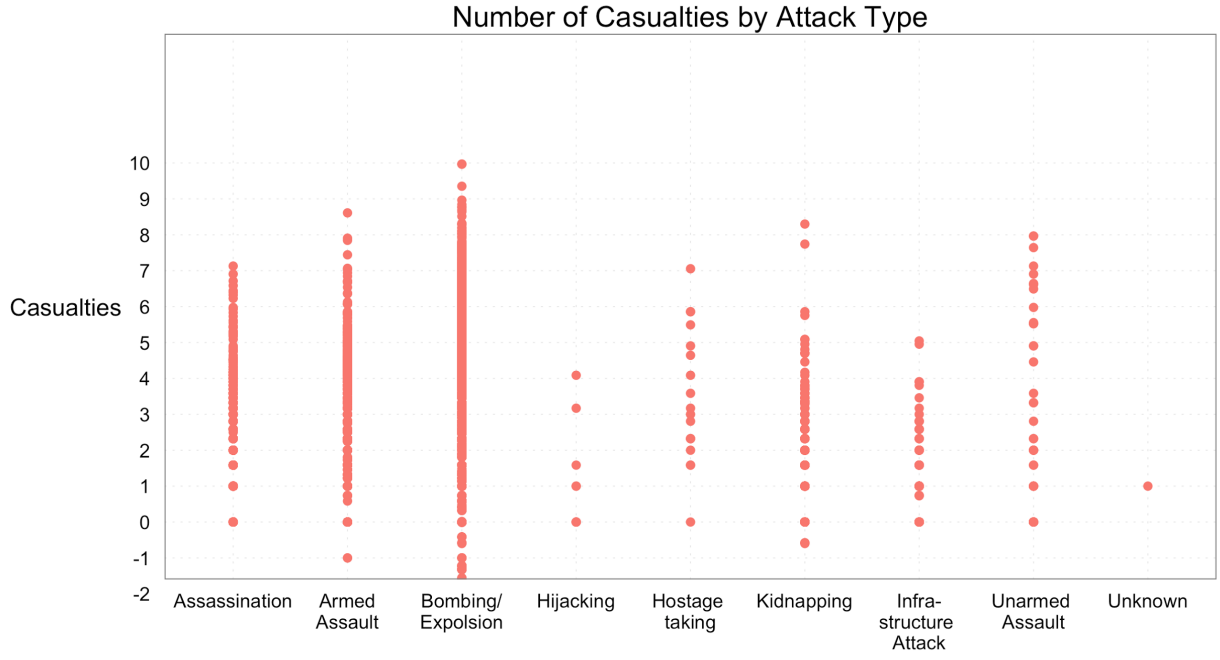
Before manipulating or cleaning any data from the data set, we did exploratory data analysis to identify patterns within features including any collinearity and/or correlations between variables. The original database contained approximately 45,000 instances and 136 features. One of the first things that was noticed was that approximately 70% of the instances in the database were associated with only six countries - Afghanistan, India, Iraq, Pakistan, Philippines and Thailand. Hence, we decided to filter the dataset such that we could build a stronger model within the limited scope of just these six countries.

(**Please note that all the values for casualties on the y-axis are in a log scale.)

Gabrielle Agrocostea
Nandini Shah
Vidyavisal Mangipudi

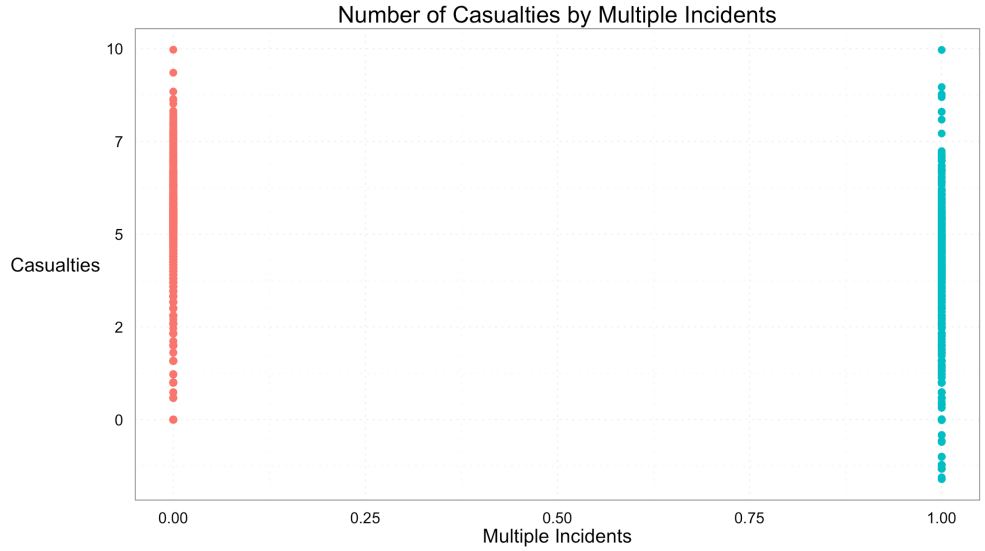


Another association we discovered was that different attack types did not have similar effects on the casualties. A larger portion of the data was limited to Assassinations, Armed assaults and Bombings.

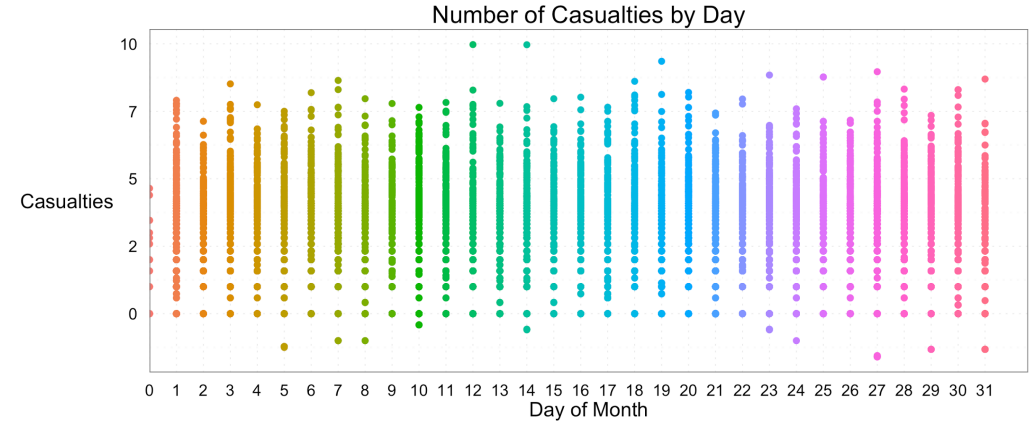


Gabrielle Agrocostea
Nandini Shah
Vidyavisal Mangipudi

The plot below shows that both single event terrorist attacks and multiple event attacks occur within a relatively stable frequency. A multiple event attack (1) is defined in the GTD documentation as a series of terrorist attacks linked to one another. A single event (0) conversely is an isolated attack.



Lastly, we decided to see if there was any pattern in regards to which days in a month had the increased frequency of attacks or increased casualties. Our hypothesis was that, terrorist groups may have favored certain days over the other or that we may have seen a spike in the frequency every 7 days. However, there appears to be no such correlation between the days of the month and the frequency or number of casualties.



Gabrielle Agrocostea
Nandini Shah
Vidyavisal Mangipudi

The Data:

The GTD dataset is one of the most comprehensive existing databases in regards to terrorism, containing data from 1970 to 2013. There are a lot of constraints and definitions for an event to be qualified as a terrorist attack. Additionally the 136 features include continuous, discrete and categorical variables along with some text variables. Using domain knowledge we identified that almost all of the text-based features were either news snippets or an attack of the summary that was compiled post occurrence of the event. Given that most of the data in the database was compiled using news reports, legal papers and unclassified documents, it was no surprise that there would be a lot of missing data in the dataset that would need to be accounted for. The process and methods used for this project will be discussed in the following section, Cleaning the data.

Cleaning the Data:

As mentioned above, all text features were eliminated from the database as they could not qualify as material that could be used to predict either the success/failure of an event or the number of casualties. Additionally, as mentioned in the Exploratory Data Analysis section, the dataset was filtered to include only the 6 countries that accounted for approximately 70% of the instances in the dataset. Any features that were not relevant for our prediction of the success of an event or the number of casualties were also eliminated from the dataset (example: event_id or Latitude and Longitude). All of the relevant categorical variables (example: Terrorist Group Names) was vectorized and indexed so as to simplify the process of manipulating and modeling the data.

The presence of missing data in the features was an anticipated problem due to the nature of the database. There were three forms of missing data that we came across: NA values, Null data points and features that contained a lot of (-9) or unknown values. Any feature that was more than 50% missing data was deleted from the dataset. Using the knowledge from the documentation, we were able to further delete some features that had very weak relationships (example: nationality). Lastly, we deleted the rows that had NA values in the important features. We considered this an acceptable loss as the deleted instances were under 5% of the existing data. The final dataset was filtered down to approximately 25,000 instances and 25 relevant features that were not collinear.

Gabrielle Agrocostea
Nandini Shah
Vidyavisal Mangipudi

Modeling the Data:

Question 1: Prediction of "success" of an attack.

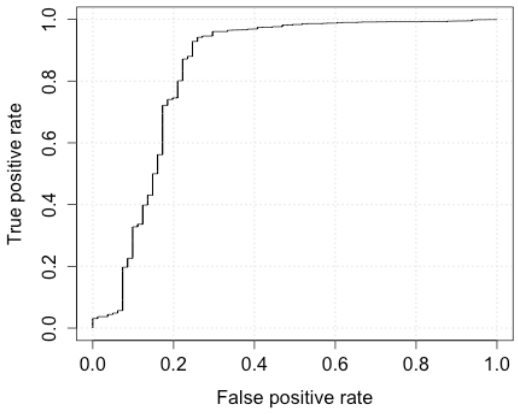
This is a binary classification problem. We used Logistic Regression as our choice of model. We got good results after implementation as below:
Area under the curve (AUC): 0.838
Accuracy: 96%
Confusion matrix:

Class	0	1
FALSE	13	15
TRUE	68	2131

(Actual Values are 0 and 1. Predicted values are FALSE and TRUE)

It is worth noting that the number of events with '0' success made up only ~8% of the dataset. One natural reasoning is a reporting bias due to the fact that more successful events are known, confirmed and recorded and hence included in the GTD dataset. Additionally, the way GTD defines success also contributes to this skew.
Thus, after this additional research we decided the "success" column is not necessarily the most useful data point.

ROC Curve:



Gabrielle Agrocostea
Nandini Shah
Vidyavisal Mangipudi

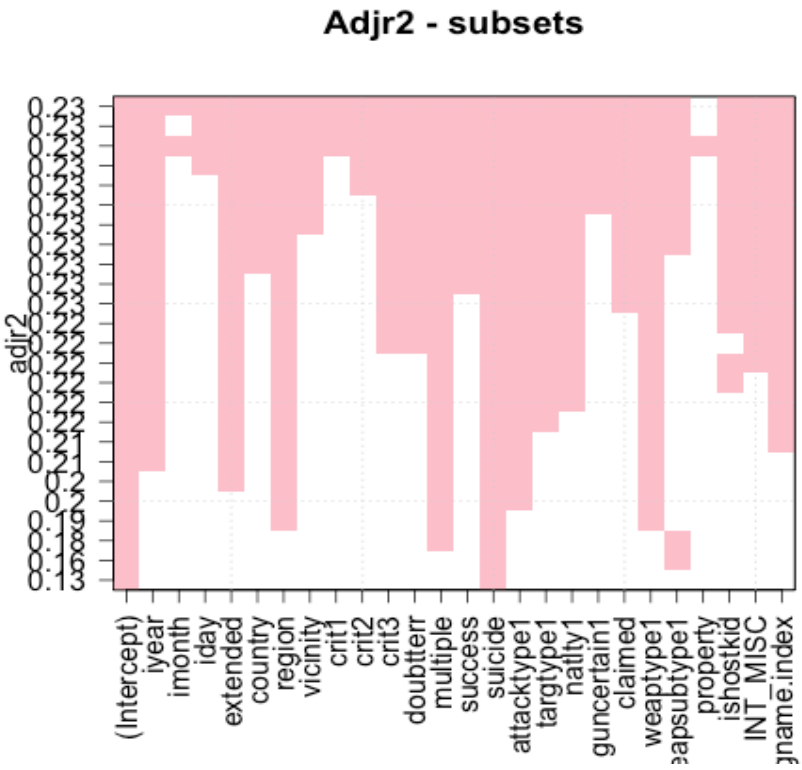
Question 1: Prediction of number of casualties.

Given the large number of features, during the exploratory data analysis phase of the project, we used subset selection to understand what features were most relevant to answer this question, and predict using these features.

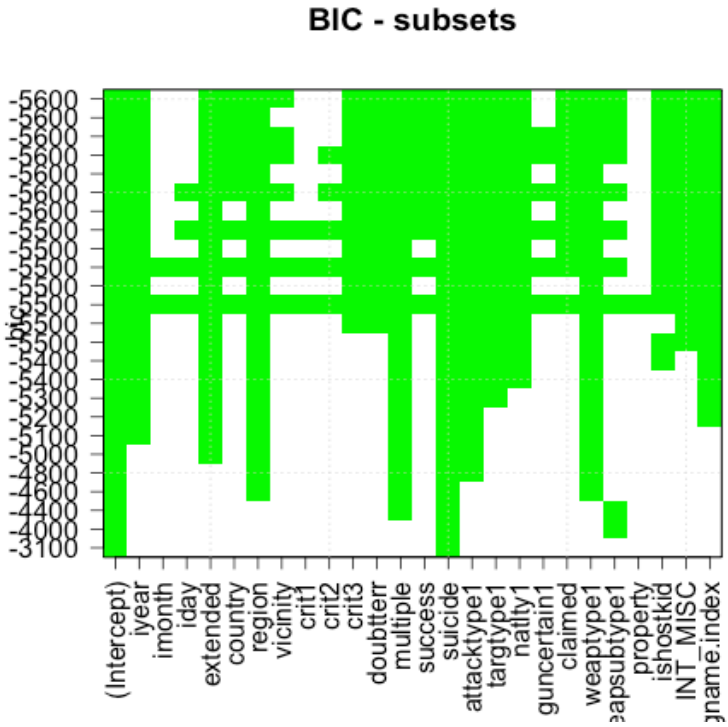
We also implemented Lasso Regression to generate a sparse model and avoid overfitting.

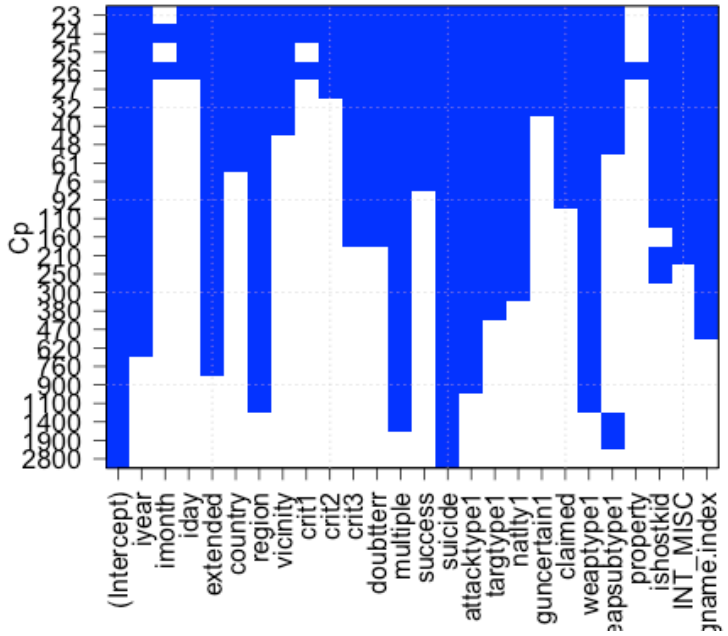
Technique 1: Subset selection

We used the leaps package in R to implement subset selection. The algorithm does an exhaustive search over the entire feature space for the features with minimum error for upto a specified 'k' value. In our case, we used k=25 driven by our data cleaning process. Below are some charts that help graphically understand scores obtained by different models based on the number of features k used.



Gabrielle Agrocostea
Nandini Shah
Vidyavisal Mangipudi





To do a valid comparison between the two techniques employed, we chose the best model for subset selection by using RSS as our evaluation metric.

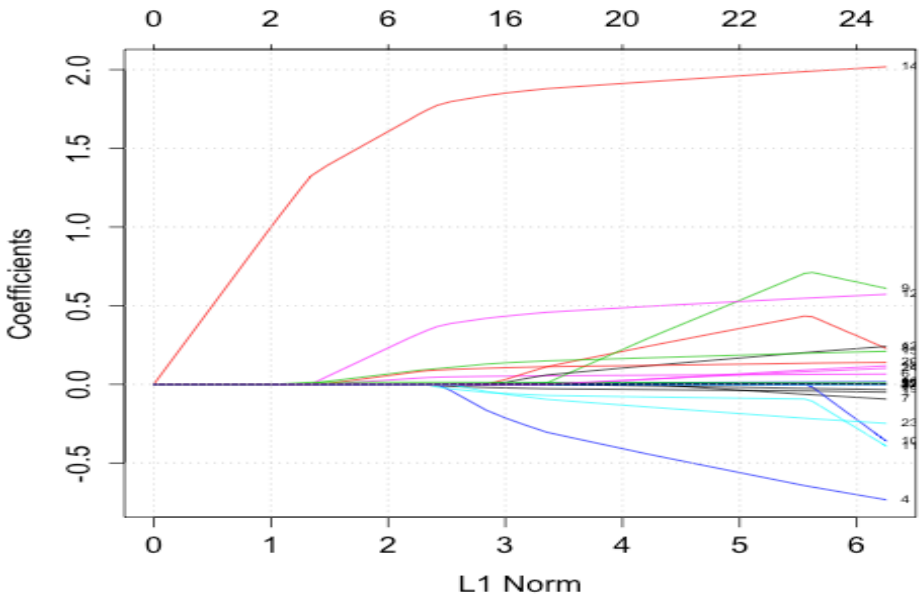
Gabrielle Agrocostea
Nandini Shah
Vidyavisal Mangipudi

Technique 2: Lasso Regression

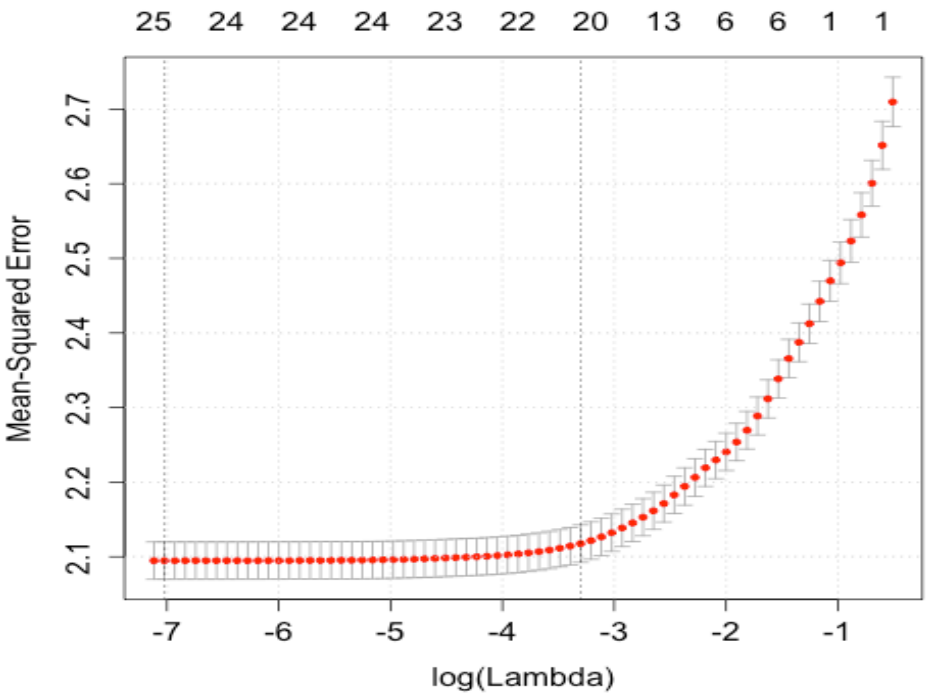
We used glmnet with cross-validation to implement lasso regression. Our implementation returned coefficients for 20 out of the 25 features.

Note: the only post-fact feature used by our model is “success”. Based on the way GTD defines success, we noticed this column has great impact on the number of casualties. The motivation for including this feature was for our model to be able to predict the number of casualties accurately whether or not the terrorist event was eventually successful, which one can obtain by running the model twice using both ‘0’ and ‘1’ flags.

See charts below reflecting the choice of lambda:



Gabrielle Agrocostea
Nandini Shah
Vidyavisal Mangipudi



Analysis of the Results

Comparison of the two techniques based on RMSE:

Characteristics	RMSE train	RMSE test	Model Complexity # features
Subset Selection	1.455	1.455	20/25
Lasso Regression	1.444	1.458	25/25

Feature Rankings:

The priority of features is very similar in both the techniques. This reinforces the validity of both the models.

Gabrielle Agrocostea
Nandini Shah
Vidyavisal Mangipudi

Feature	Lasso Ranking	Subset Selection Ranking
Intercept	1	1
suicide	2	2
intercept	3	3
crit2	4	4
crit1	5	6
attack_type	6	7
weapon_type	7	8
success	8	5
region	9	11
int_misc	10	10
gun_certain	11	9
weapon_sub_type	12	13
target_type	13	12
group_name	14	14
country	15	17
nationality	16	18
claimed	17	20
year	18	21
doubt_terror_event	19	25
is_hostage_kid	20	23
extended	21	16
month	22	19

Gabrielle Agrocostea
Nandini Shah
Vidyavisal Mangipudi

day	23	15
vicinity	24	22
property	25	24

In conclusion, since the size of data set and number of features is small currently, it appears to be method agnostic. However, as the data set increases and we want to re-train the model, subset selection will be the less efficient choice. Also, if the NAs reduce with time in the future, the number of features could be increased and lasso regression will still be an optimal method.

Gabrielle Agrocostea
Nandini Shah
Vidyavisal Mangipudi

References:

The GTD dataset can be found at:
http://www.start.umd.edu/gtd/contact/
To download the dataset from website, you need to provide name and email.

Else you can download the original dataset from the this github repository:
https://github.com/nandinishah/MSDFinalProject
The filename is: gtd_06_to_13.csv