Global Terrorism - Final Project

MATH 4685/6685 (Fall 2019)

**Group Members:**

James Willson (*Data Scientist*)

Sean Kugele (*Data Scientist/Team Lead*)

# Introduction

Our objective for this project was to create statistical models for the purposes of prediction and inference using the Global Terrorism Database[[1]](#footnote-1) as our data set. Each data scientist independently selected several questions of interest. The questions selected included:

(1) predicting if an attack will be successful based on a variety of different factors (**Section 3.1**)

(2) estimating the number of casualties resulting from a successful terrorist attack (**Section 3.2**)

(3) predicting the terrorist group responsible for perpetrating a terrorist attack (**Section 3.3**)

(4) estimating the risk of attack based on temporal and geospatial variables (**Section 3.4**)

The remainder of the paper is organized into the following sections: Section 2 provides additional detail on the Global Terrorism Database (our dataset). Section 3 describes the modeling efforts for each research question, including question-specific data cleansing, generated statistical models, model comparisons, and summaries of findings. Section 4 concludes the paper with a discussion of our findings, general observations about our data modeling efforts, and possible directions for future work.

# Background

The Global Terrorism Database (GTD) contains over 180,000 observations and 135 variables. Each observation corresponds to a terrorist event that occurred between 1970 and 2017 (excluding 1993). Terrorism is broadly defined as “the threatened or actual use of illegal force and violence by a non-state actor to attain a political, economic, religious, or social goal through fear, coercion, or intimidation." Data in the GTD is sourced from unclassified media articles (e.g., electronic news archives), and it is currently maintained by the National Consortium for the Study of Terrorism and Responses to Terrorism (START) group at the University of Maryland.

The type of information available for each event includes, but is not limited to, *temporal variables* (e.g., year, month, day, etc.), *geospatial variables* (e.g., latitude and longitude, region, country, city, province, etc.), *event descriptive variables* (e.g., attack type, duration of incident, weapons used, targets, etc.), *perpetrator descriptive variables* (e.g., terrorist group name, number of perpetrators, etc.), and *event outcome variables* (e.g., total number of fatalities, total number injured, extent of property damage, success or failure indicator, etc.).

## Geospatial and Temporal Distribution of Data

As expected, the spatial and temporal distribution of the terrorist events in the GTD are highly non-uniform.

Figure 2-1 shows the temporal distribution of the events from 1970 to 2017. The number of events appears to increase dramatically from 1970 to the 2010s; however, this is likely driven by differences in media coverage and data collection rather than a reflection of actual underlying trends.

A screenshot of a video game

Description automatically generated

Figure ‑ – Histogram showing frequency of terrorist events between 1970 and 2017.

Figure 2-2 shows the geospatial distribution of terrorist incidents. The highest density of events occurred in the Middle East & North Africa () followed by South Asia (). The lowest density of events occurred in North America () followed by Eastern Europe (). The countries with the greatest number of terrorist events were Iraq (), Pakistan (), Afghanistan (), India (), and Colombia ().

|  |  |
| --- | --- |
| Terrorist Events (1970 – 2017) | |
| A close up of a map  Description automatically generated | A close up of a map  Description automatically generated |
| A close up of a map  Description automatically generated | |
|  | |

Figure ‑ – Geospatial distribution of terrorist events.

Figure 2-3 shows changes in the geospatial distribution of terrorist events over time. Note that the areas of highest concentration have shifted considerably. Between 1970 and 1989, the greatest number of attacks occurred in South America () and Western Europe (). Between 1990 and 2009, the Middle East & North Africa emerged as the region with the highest density of attacks () and that trend continues into the 2010s where the Middle East & North Africa accounts for terrorist events. By comparison, South America accounted for only terrorist events in the same time period.

|  |  |  |
| --- | --- | --- |
| Changes in Terrorist Event Density Over Time | | |
| A close up of a map  Description automatically generated | | |
|  |  |  |
| *Time* | | |

Figure ‑ – Changes in terrorist incident density with time. Note that not only do high-density areas change with time, but the density’s variance decreases with time.

## Data Quality

While the GTD contains 135 variables, only a subset of these variables can be easily used for model creation. 41 variables contain redundant, free-form textual descriptions of the codes used in other variables and were, therefore, unusable. Among the 94 remaining variables, 44 have over 90% NA values. (Not surprisingly, running na.omit on the data set results in 0 rows.) Most of the “usable” variables are categorical; however, many numerical variables are also present (e.g., nwounded, nkill, nperps, etc.). Some of the categorical variables contain many distinct categories (e.g., country that contains 205 distinct values), making them challenging to use directly because of the large number of variables they will generate after encoding. Finally, the frequency with which each category occurs in most of these categorical variables is highly unbalanced, complicating our efforts.

Due to these complications, we tailored our data cleansing and feature selection to each problem rather than trying to cleanse the data up-front in a problem-agnostic fashion, which would have sacrificed too many observations. The specifics of these problem-specific data cleansing efforts are described in Section 3.

# Modeling

## Q1—Predicting Successful Attacks

## Q2—Estimating Number of Casualties

## Q3—Predicting Responsible Terrorist Groups

The goal of this research question is to develop statistical models that predict the terrorist group name (gname) based on a subset of the GTD’s available features. The GTD contains terrorist events attributed to over 3,500 different groups, however, many of these groups correspond to general classes of perpetrators (for example, “gunman” and “anarchists”) and not terrorist organizations, such as the Taliban or Al-Qaida. Furthermore, only 122 of these groups have over 100 attacks attributed to them, and, among these, only 13 operate in more than a single region. These 13 terrorist groups, that is, organizations having over 100 associated terrorist events and operating in at least 2 regions, will be the focus of this modeling effort. Table 3-1 lists these terrorist organizations and the number of attacks attributed to each between 1997-2017.

|  |  |  |
| --- | --- | --- |
| Terrorist Group | Identifier | # of Attacks (%) |
| Al-Qaida in Iraq | A | 606 (3.35%) |
| Al-Qaida in the Arabian Peninsula (AQAP) | B | 877 (4.84%) |
| Al-Qaida in the Islamic Maghreb (AQIM) | C | 225 (1.24%) |
| Hamas (Islamic Resistance Movement) | D | 297 (1.64%) |
| Hezbollah | E | 107 (0.59%) |
| Islamic State of Iraq and the Levant (ISIL) | F | 4,274 (23.61%) |
| Kurdistan Workers' Party (PKK) | G | 1,074 (5.93%) |
| Liberation Tigers of Tamil Eelam (LTTE) | H | 614 (3.39%) |
| New People's Army (NPA) | I | 1,573 (8.69%) |
| Revolutionary Armed Forces of Colombia (FARC) | J | 1,121 (6.19%) |
| Salafist Group for Preaching and Fighting (GSPC) | K | 182 (1.01%) |
| Taliban | L | **5,912 (32.66%)** |
| Tehrik-i-Taliban Pakistan (TTP) | M | 1,240 (6.85%) |

Table 3-‑ – Terrorist organizations with over 100 attributed attacks and operating in at least 2 regions. Note that the number of terrorist events per group is very unbalanced with the Taliban and ISIL accounting for over half of the total events. A character identifier has been assigned to each group to serve as an alias (used later in confusion matrices).

### Feature Selection

The features used for this modeling effort were limited to those with relatively low percentages of NA / unknown values to prevent the loss of too much data following data cleansing. Variables with over 30 distinct categorical values were also eliminated from consideration[[2]](#footnote-2). Table 3-2 summarizes the features selected as *candidates* for consideration—not all of these features were used in the final models.

|  |  |  |
| --- | --- | --- |
| Variable Name | Description | Data Type  (with cardinality for categorical variables) |
| attacktype1 | method of attack (e.g., assassination, hijacking, kidnapping) | categorical (9) |
| claimed | group claimed responsibility (“Yes” or “No”) | categorical (2) |
| extended | incident extended more than 24 hours (“Yes” or “No”) | categorical (2) |
| iday | day of the month | numerical |
| imonth | month | numerical |
| iyear | year | numerical |
| latitude | latitude | numerical |
| longitude | longitude | numerical |
| multiple | part of multiple incidents (“Yes” or “No”) | categorical (2) |
| nkill | number of fatalities (both perpetrator and victim) | numerical |
| nkillter | number of terrorists fatalities | numerical |
| nperps | number of perpetrators | numerical |
| nwound | number of non-fatal injuries (both perpetrator and victim) | numerical |
| property | property damage occurred during incident (“Yes” or “No”) | categorical (2) |
| region | geographic region (e.g., North America) | categorical (12) |
| success | attack successful (“Yes” or “No”), where definition depends on attack type | categorical (2) |
| suicide | suicide attack (“Yes” or “No”) | categorical (2) |
| targtype1 | type of target/victim (e.g., business, police, military) | categorical (22) |
| weaptype1 | general type of weapon used (e.g., biological, chemical, firearms) | categorical (12) |

Table 3-‑ – Features selected for use in initial modeling effort.

### Data Cleansing

In addition to limiting the observations to those corresponding to terrorist organizations operating in at least 2 regions and with over 100 attacks, we also removed any data for incidents prior to 1997. This decision was made for two reasons: (1) several variables of interest were not available prior to 1997 (2) data collection methodology was different prior to 1997.

The values for each selected feature were scrubbed to ensure that no “unknown” or NA values remained after data cleansing. Generally, this involved dropping observations with NA or “unknown” values; however, for several of the numerical variables (nkill, nkillter, nwound), NA values were transformed to 0. In addition, “unknown” or NA values for claimed were transformed to “no” rather than dropping those observations.

Following cleansing, the data set contained 18,102 observations over 19 features and 1 response variable (gname). The data set was then split into a training and test data set using an 80/20 split.

### Methods and Evaluation

Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Decision Tree, Random Forest, and SVMs were applied to the cleansed data set. Linear kernels produced the best SVM performance (when compared to “tuned” radial, polynomial, and sigmoid kernels); therefore, only linear SVMs are presented below.

The initial formulas used for all models (other than QDA) consisted of all variables listed in Table 3-2. QDA produced “rank deficiency” errors when using most of the categorical variables, so it was limited to only numerical variables.

During initial modeling, it was determined that balancing the observations over the response variable improved performance. Both *undersampling* and *oversampling* strategies were attempted, but oversampling was found to provide the best model performance. Note that oversampling increased the training data set to 62,010 observations (though only 14,481 of those represented distinct events).

A close up of a map

Description automatically generated

Figure ‑ – Plot of terrorist events per group by latitude / longitude. This plot shows a clear geographic clustering of attacks by group.

Initial results showed that all classifiers obtained high accuracies (up to 98.5% for Random Forest) on the test data when the models were trained using all features in Table 3-2. Analysis of the most significant variables showed that latitude, longitude, region, and iyear were highly predictive of the group name and that other variables were far less significant predictors of the response variable. Plotting the distribution of attacks for each terrorist organization by latitude/longitude (see Figure 3) makes it clear why this is the case. Each terrorist group has a relatively distinct area of operation, which the algorithms were able to use to their advantage. The plot of the decision tree in Figure 3-4 demonstrates this clearly, as the algorithm appears to be isolating year-sensitive “bounding boxes” for the terrorist events attributed to each groups and ignoring all other variables.

A close up of a map

Description automatically generated

Figure ‑ – Visualization of decision tree. Notice that the decision tree only uses latitude, longitude, region, and iyear.

Ideally, we would like to identify an “attack signature” for each of these groups that transcends a particular place and time. For example, if a terrorist organization expanded operations into a new region we would like to be able to continue to have a useful classifier. To this end, a second set of models was created that removed latitude, longitude, region, and iyear. All other variables were retained. The original models (that used all features) are referred to below as “Model Set 1”. The later models (with the reduced feature set) are referred to below as “Model Set 2”.

|  |  |
| --- | --- |
| *Model Set 1* | *Model Set 2* |
| A picture containing screenshot  Description automatically generated | A screenshot of a social media post  Description automatically generated |
| Accuracy | Accuracy |

Figure 3-5 – Accuracy on test data for each model set.

Figure 3-5 shows a comparison of the accuracy of each algorithm (LDA, QDA, Decision Trees, Random Forests, and SVM with linear kernel) on Model Set 1 and 2. Most algorithms performed well (greater than 80% accuracy) on Model Set 1, but considerably worse (less than 55% accuracy) on Model Set 2. Random Forests obtained the best performance for both model sets based on prediction accuracy on the test data—98.5% accuracy on Model Set 1 and 51.6% accuracy on Model Set 2.

In spite of this, one could argue that the sacrifice in performance resulting from Model Set 2 may be warranted in order to remove the models’ dependency on geospatial variables and the year of attack. Unfortunately, the confusion matrix for the Random Forest algorithm on Model Set 2 (see Table 3-3) shows that the false positive rate (FPR) varies greatly between predicted groups. In most cases, the classifier’s performance appears to be directly proportional to the number of terrorist events per group; that is, the classifier performs best on groups with the most events (in spite of data balancing). Given these high FPRs for most classes, it’s highly questionable whether this model would be useful for prediction in practice.

Figure 3-6 shows importance plots for the Random Forests from Model Sets 1 and 2. As previously mentioned, latitude, longitude, region, and iyear were the most important features on Model Set 1. Interestingly, iday and imonth were the most important features on Model Set 2 followed closely by targtype1. This would suggest that different groups prefer to attack on different days of the month and different months of the year. While this would be an interesting finding, it is not clear that this is the case. Removing iday and imonth does reduce the Random Forest’s accuracy to 40% (from 50%) suggesting some correlation, but additional analysis is needed to determine the cause of this effect.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Predicted Terrorist Group | | | | | | | | | | | | |
|  |  | **A** | **B** | **C** | **D** | **E** | **F** | **G** | **H** | **I** | **J** | **K** | **L** | **M** |
| ­­­Actual Terrorist Group | **A** | **77** | 5 | 1 | 2 | 0 | 7 | 3 | 6 | 1 | 2 | 1 | 12 | 6 |
| **B** | 3 | **58** | 3 | 2 | 5 | 20 | 14 | 16 | 14 | 14 | 6 | 23 | 9 |
| **C** | 2 | 3 | **4** | 1 | 1 | 0 | 3 | 2 | 4 | 3 | 0 | 8 | 5 |
| **D** | 0 | 3 | 1 | **26** | 3 | 9 | 1 | 1 | 1 | 1 | 1 | 9 | 3 |
| **E** | 1 | 2 | 0 | 1 | **9** | 2 | 2 | 1 | 0 | 1 | 0 | 2 | 2 |
| **F** | 34 | 31 | 8 | 27 | 12 | **513** | 16 | 21 | 16 | 25 | 8 | 115 | 37 |
| **G** | 2 | 9 | 1 | 0 | 6 | 2 | **85** | 14 | 20 | 27 | 12 | 16 | 8 |
| **H** | 2 | 8 | 4 | 1 | 2 | 9 | 21 | **36** | 9 | 19 | 8 | 21 | 1 |
| **I** | 6 | 19 | 5 | 2 | 9 | 3 | 30 | 10 | **195** | 18 | 8 | 28 | 5 |
| **J** | 1 | 12 | 1 | 1 | 2 | 7 | 28 | 15 | 22 | **100** | 11 | 20 | 5 |
| **K** | 1 | 2 | 0 | 0 | 0 | 1 | 5 | 4 | 8 | 4 | **7** | 9 | 2 |
| **L** | 17 | 27 | 13 | 29 | 24 | 105 | 71 | 63 | 41 | 36 | 15 | **632** | 69 |
| **M** | 8 | 13 | 2 | 6 | 3 | 17 | 3 | 13 | 19 | 11 | 2 | 61 | **81** |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| FPR |  | 50.0% | 69.8% | 90.7% | 73.5% | 88.2% | 26.2% | 69.9% | 82.2% | 44.3% | 61.7% | 91.1% | 33.9% | 65.2% |

*Table 3-3‑3 - Confusion matrix for Random Forest on test data for Model Set 2. Groups F (ISIL) and L (Taliban) have the greatest number of terrorist events and lowest false positive rates (FPRs). Groups E (Hezbollah) and K (GSPC) had the least number of associated events and had very high FPRs.*

|  |  |
| --- | --- |
| *Model Set 1* | *Model Set 2* |
| A screenshot of a social media post  Description automatically generated | A screenshot of a cell phone  Description automatically generated |

Figure 6 – The importance of features determined by the Random Forest algorithm on both model sets.

## Q4—Estimating Attack Risk Using Temporal and Geospatial Features

The goal of this research question is to develop statistical models that predict the “risk of attack” based on geospatial and temporal variables. The data set was filtered to only include terrorist incidents occurring between 2010 – 2017, as we are only focused on the “current” risk pattern and not historical risk patterns.

### Feature Selection

Our models were limited to three features: imonth, weekday, and cluster\_id. weekday (day of week) was derived from iday, imonth, and iyear. cluster\_id was derived by applying kmeans to the latitude and longitude of each event using 32 clusters (see Figure 3-7).

The response variable (risk\_level) was derived by (1) grouping terrorist events into “event groups” by imonth, weekday, and cluster\_id, (2) summing up the number of attacks in each event group, (3) applying a low/high risk-threshold to classify each event group as “low” or “high”. The low/high risk threshold value used was , where is the sample mean of the number of terrorist events over all event groups and the sample standard deviation. Figure 3-8 (left panel) shows a histogram with a vertical line depicting this threshold value.

### Data Cleansing

All NA values were removed from imonth, iday, iyear, latitude, and longitude. After grouping by imonth, weekday, and cluster\_id, the training set consisted of 2,594 observations. 91% of these observations were classified as “high risk”. Oversampling was applied to the dataset, which resulted in 3,813 observations with approximately equal percentages of low and high classifications.

A close up of a map

Description automatically generated

Figure 3-7 – World map with cluster\_ids (determined by kmeans from latitude and longitude of terrorist events).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| A screenshot of a cell phone  Description automatically generated | |  |  | | --- | --- | | Risk Level | # Event Groups | | low | 2,361 (91%) | | high | 233 (9%) | |

Figure ‑8 – Separation of event groups into high risk and low risk groups.

### Methods and Evaluation

Linear Discriminant Analysis (LDA), Decision Tree, Random Forest, and SVMs were applied to the cleansed data set. Radial kernels produced the best SVM performance (when compared to “tuned” linear, polynomial, and sigmoid kernels); therefore, only radial SVMs are presented below.

Figure 3-9 shows the ROC curves for all models. The AUC for Random Forest was 0.981, which was slightly higher than for other methods. The confusion matrices for Random Forest and SVM (radial) are shown in 3-10.

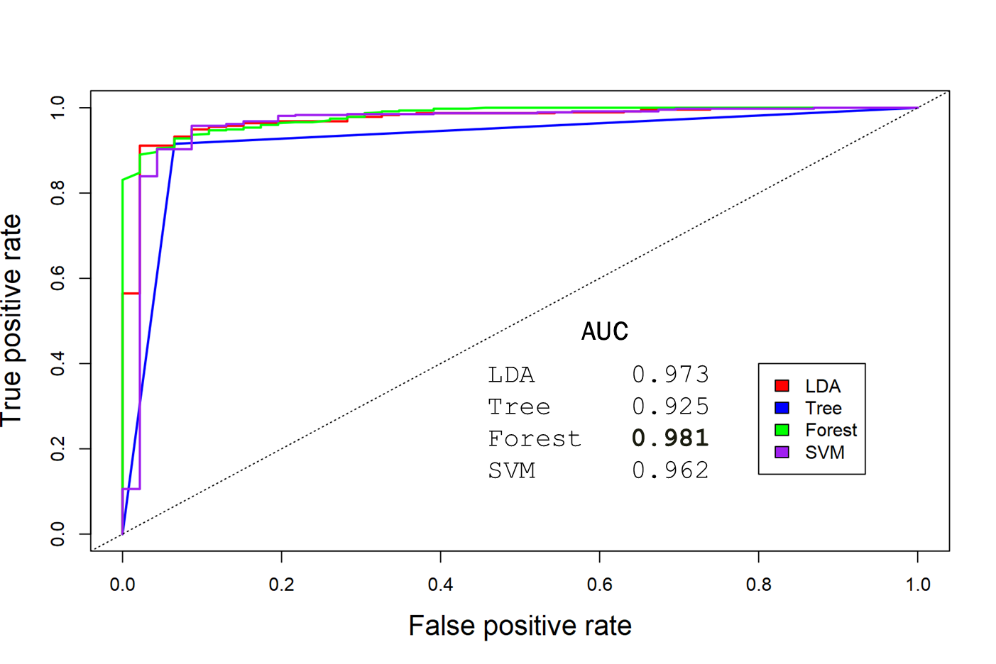


Figure ‑9 – ROC curves showing model performance for Q4. AUC values for each attempted model appear inside of the graph.

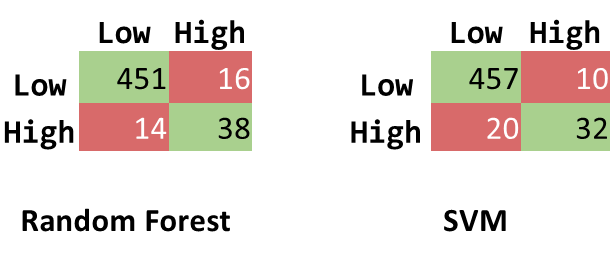


Figure ‑10 – Confusion matrices for Random Forest and SVM for Q4.

Examining the importance plot (Figure 3-11) for the Random Forest shows that cluster\_id considerably more important than imonth and weekday. This is not surprising given the density plots in Figures 2-2 and 2-3.

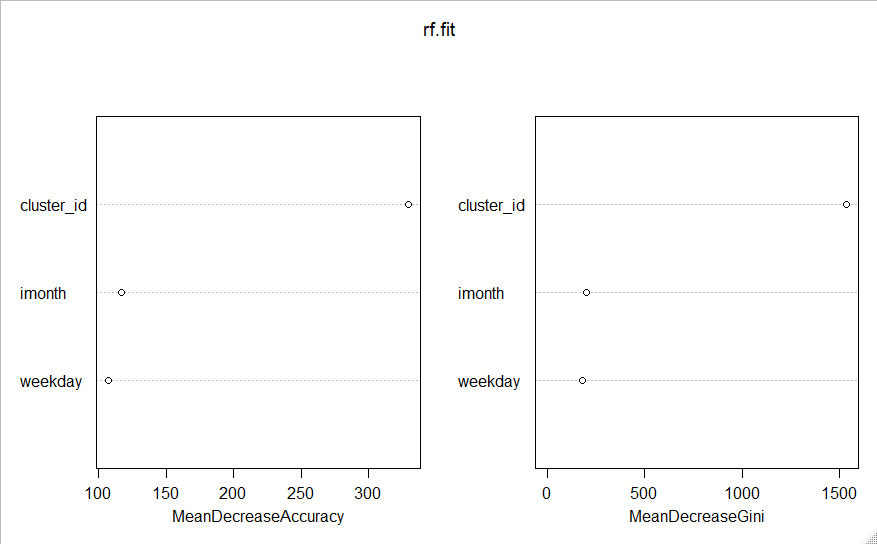


Figure 3‑11—The importance of features determined by the Random Forest algorithm for Q4.

The decision tree chose to only use cluster\_id, ignoring imonth and weekday (see Figure 3-12).

A screenshot of a cell phone

Description automatically generated

Table ‑12—Decision Tree for Q4 using formula with cluster\_id, imonth, and weekday. Note that only cluster\_id was considered significant.

In spite of the importance of cluster\_id, there does seem to be non-random variability in the number of attacks per weekday (see Table 3-4) and imonth (see Table 3-5). For example, assuming an equal probability of attack per weekday, the mean of a multinomial distribution over the 86,207 terrorist events from 2010-2017 would be approximately 12,315 and the standard deviation approximately 103. That places Tuesdays at almost 9 standard deviations above the population mean and Monday at over 13 standard deviations below the population mean. This is strong evidence that the probability of attack is not equal for each weekday. Based on this, a second set of models was created using the formula risk\_level~imonth + weekday. The Decision Tree for this model had an accuracy of nearly 60% (depicted in Figure

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | **1** | Monday | **10,954** | | **2** | Tuesday | **13,226** | | **3** | Wednesday | 11,925 | | **4** | Thursday | 12,672 | | **5** | Friday | 12,026 | | **6** | Saturday | 12,693 | | **7** | Sunday | 12,711 | |  |
|  | Terrorist Events per Weekday |

*Figure 3‑12—Number of attacks per day of week over terrorist events from 2010-2017. The blue line in the right panel represents the population mean for a multinomial distribution with and .*

A close up of a map

Description automatically generated

Figure ‑1—Decision Tree for Q4 using formula with only imonth and weekday as features.

# Discussion

1. https://www.kaggle.com/START-UMD/gtd [↑](#footnote-ref-1)
2. Most algorithms failed when presented with a formula containing a categorical variable with more than a few dozen values. The only exception was SVMs. LDA returned a “variable appears to be constant within groups” error, QDA returned a “rank deficiency” error , decision trees returned “factor predictors must have at most 32 levels”, and random forests returned “can not handle categorical predictors with more than 53 categories”. [↑](#footnote-ref-2)