Analysing and Classifying Global Terrorism Using Machine Learning and Statistical Techniques

***Abstract:***

According to GTD, *‘Terrorism’* is 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."* It has become a massive global concern for the last couple of decades. Terrorist attacks and events can cause grave consequences for the society and can also adversely affect the safety, security, confidence, economic and social conditions of the citizens. Hence, this project is a wholehearted attempt for purpose of analyzing, modelling and classifying the terror attacks or events based on a set of historic data by using some appropriate machine learning and statistical algorithms.

We have a few objectives in the study, which mainly aims to distinguish and identify different terrorist groups according to certain concerned parameters in an incident scene and give an idea about the ‘near repeats’ of a particular terrorist attack, given it has been occurred at least once. Various classification models have been used for the following purpose followed by feature engineering by using Boruta Wrapper method. The classifiers are Decision tree, Random Forest, Bagging. At the end a spatio-temporal ‘Near-Repeat’ analysis has been done for certain selected terrorist group. Hence, this statistical analysis is concerned mainly in arriving to suitable measures to effectively decrease the terrorism rates in the future and to predict the different responsible organizations for future events.

This project has been a great experience for me. There were various information which I have gathered and has given me a broader picture to this field. This experience and exposure have helped my personal development. This experience has shown me a glimpse of how life is and will be in days to come.

***Keywords:*** *GTD, Classification models, Decision Tree, Random Forest, Bagging, Near-Repeat.*

***Introduction:***

*Origin of the report:*

This report, based on a thorough analysis of GTD data, helped me to gather practical information, which is necessary for my future life. I would like to express my deep respect to my institute guides of University of Calcutta for giving me their valuable time and all the necessary guidance, which helped me to prepare this report.

*Aims:*

1. To look upon few data visuals to get an overview of the current scenario of the attack frequencies done by different terrorist groups. Also, to get an idea about the spatial regions and intensity of terrorist attacks around the globe.
2. Classification modelling to classify a ‘terrorist attack’ based on certain covariates and to predict the most probable group for the attack.
3. To conduct a “Near-Repeat Analysis” for predicting the future of probability of an attack at a specific geographical area, given that an attack has already took place there.

*Collection of Data:*

All the datasets are collected from the excellent digital initiative of *University of Maryland* and *START (National Consortium for the Study of Terrorism and Responses to Terrorism),* [***Global Terrorism Database (GTD)***](https://www.start.umd.edu/gtd/) which maintains more than 200k records for the years 1970 to 2020.

*Organization:*

The project has been divided into three sections and is organized as follows. Section 1 is an exploratory data analysis which includes a few data visuals and chloropleth maps of the terrorist attacks and the place of attacks. Section 2 is an attempt to fit five different classification models to predict the responsible terrorist group of an occurred incident and to compare for the best model. Section 3 elaborates about the methodologies and procedures about the Spatio-Temporal Near-Repeat analysis for analyzing the chance of re-hit (reattack) at a place within specified temporal and spatial bands.

***Section 1***

In the first section, we will look into the frequency distribution of the terrorist groups according to the number of attacks. We have considered data from 1970 till 2020.

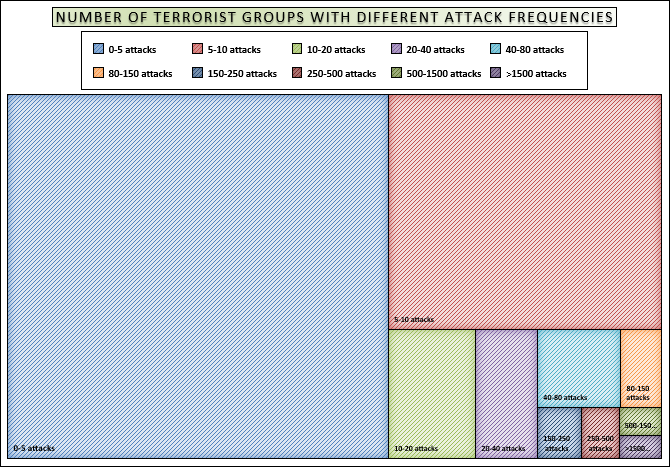
**Figure 1: Bar Chart of Terrorist Group Attacks with Attack Frequency >500**

From the above bar chart, it can be observed that there are 29 groups who have attacked for at least five hundred times with respect to the above-mentioned definition of ‘terrorism’. The top 3 groups are Taliban, Shining Path (SL) & Farabundo Marti National Liberation Front (FMLN) with attack frequencies of 5486, 4300 and 2952 respectively.

Now, from some frequency distribution of different classes with predefined attack frequencies, we can visualize the distribution of number of terrorist attacks and terrorist groups.

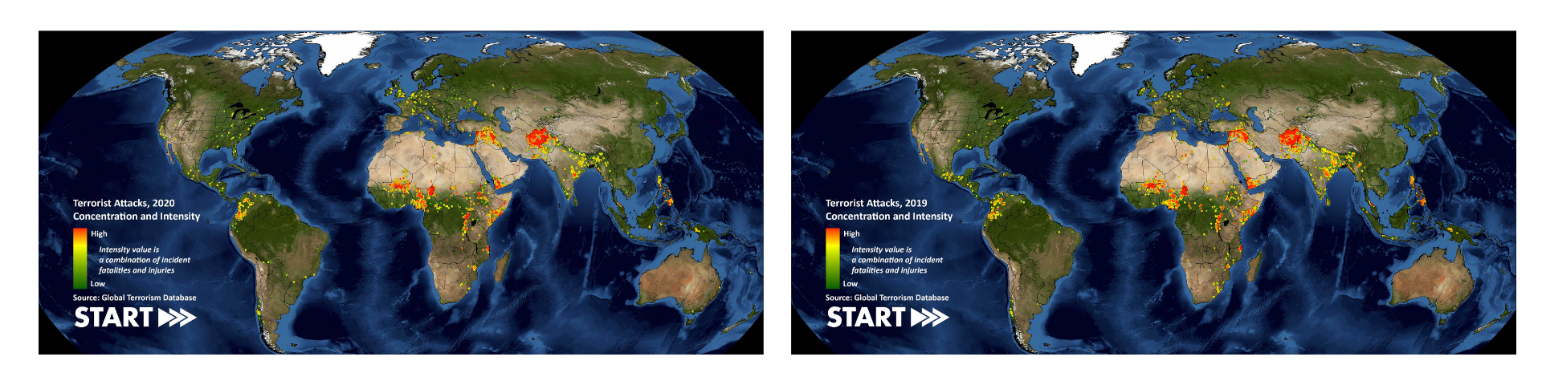
**Table: Frequency Distribution of Number of terrorist attacks and terrorist groups**

|  |  |
| --- | --- |
| **Groups with No. of Attacks** | **Frequency** |
| 5-500 | 1713 |
| 500-1000 | 12 |
| 1000-2000 | 10 |
| 2000-5000 | 6 |
| >5000 | 1 |

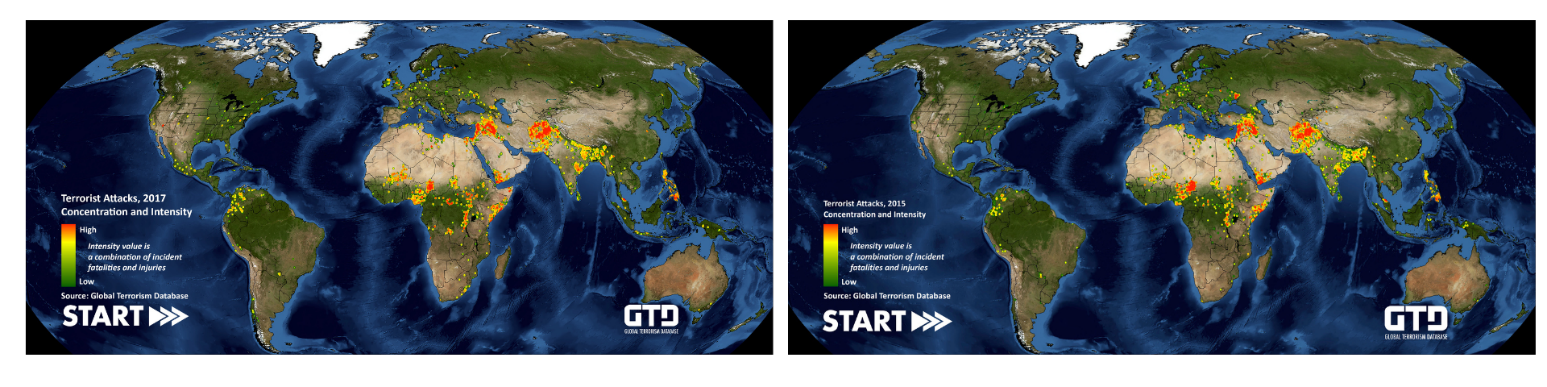


**Figure 2: Tree Map for the No. of Terrorist Organizations according to number of attacks**

Now, we will investigate some chloropleth maps [R1] from the GTD for the Key Regions of Terrorist attacks of past few years. Note that in these maps the intensities of the terror attacks are categorized from low to high based on the number of casualties and degree of property losses. According to the diagrams below [figure 3], we can see that some of the main terrorist attack prone areas are Central Asia, Middle East and some parts of Africa.



**Figure 3(a): Heatmap of Global Attacks, 2020 Figure 3(b): Heatmap of Global Attacks, 2019**

****

**Figure 3(c): Heatmap of Global Attacks, 2017 Figure 3(d): Heatmap of Global Attacks, 2015**

From the above figures, it is observed that there is surge in the attack numbers in the Central Asian Zone from the year 2015-2020. Mainly, the south-eastern region should be given special care as maximum clusters can be noticed at that part. Elsewhere, Middle East and Central & North-eastern parts of Africa also shows significant increase in the number of terrorist attacks in the given five years. According to the shown patterns, one can expect more growth in the numbers at the specific regions in the near future.

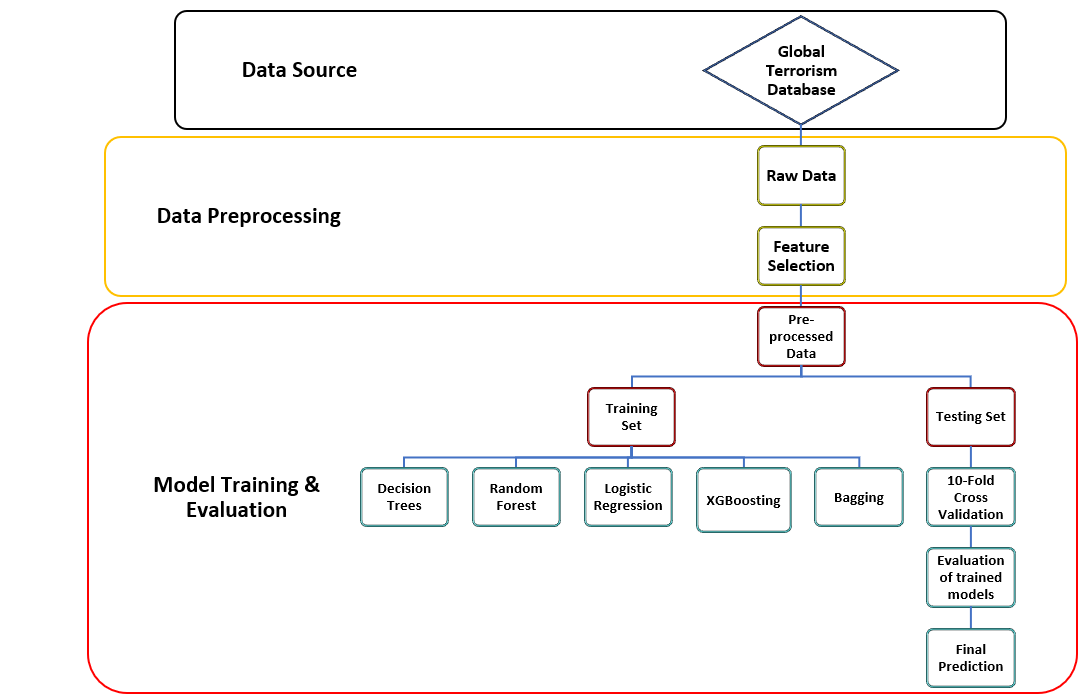
According to some previous research works [R2], it is shown that for the year 2015-2017, the top five targets of terror attacks were Citizens (29.2%), Military (20.1%), Police (13.9%), Governments (8.3%) and Businesses (7.3%). Hence, we can comment that at the present time, the protection on these 5 sections, especially on Citizens and Militaries should be strengthened more.

**Figure 4: Statistics on Top Terrorist Targets for the year 2015-2017**

***Section 2***

***Objective:***In this section, our objective is to fit appropriate classification models to our data so that, the terrorist organization responsible for a particular terrorist attack can be predicted. We have divided this section into two parts, namely: Feature selection and Model Training & Validation.

***Plan:***The plan for the execution of the second section has been divided into 3 stages, namely: data collection, data pre-processing and model training & evaluation. The raw data has been collected from GTD and then by using certain appropriate techniques of feature selection, the preprocessing of the data has been done. The data has been partitioned into ‘training set’ and ‘testing set’ and five classifier models (Decision Tree, Random Forest, Bagging) were constructed on the training set and then is used to classify the testing data, hence validating the models. The below mentioned diagram gives a better understanding of the same.



**Figure 5: Overview of the plan for the Classification of Terrorist Organizations**

***Understanding the Dataset:*** In the dataset, there are 17 covariates in total, based on which we are trying to classify the terrorist groups. The variables are region, lat, lon, multiple, success, suicide, attacktype, nter, claimed, weapontype, extended, nkilled, nkilledter, nwounded, nwoundedter, property, propertyextent based on these we want to predict ‘gname’. According to the GTD codebook [R3], the definition of the above-mentioned variables is as follow:

1. *region (categorical):* 1 = North America, 2 = Central America & Caribbean, 3 = South America, 4 = East Asia, 5 = Southeast Asia, 6 = South Asia, 7 = Central Asia, 8 = Western Europe, 9 = Eastern Europe, 10 = Middle East & North Africa, 11 = Sub-Saharan Africa, 12 = Australasia & Oceania.
2. *lat (numerical):* Latitude of the place of the incident.
3. *lon (numerical):* Longitude of the place of the incident.
4. *multiple (categorical):* 1 = "Yes" (The attack is part of a multiple incident), 0 = "No" (The attack is not part of a multiple incident).
5. *success (categorical):* 1 = "Yes" (The incident was successful). 0 = "No" (The incident was not successful).
6. *suicide (categorical):* 1 = "Yes" (The incident was a suicide attack), 0 = "No" (There is no indication that the incident was a suicide attack).
7. *attacktype (categorical):* 1 = ASSASSINATION, 2 = ARMED ASSAULT, 3 = BOMBING/EXPLOSION, 4 = HIJACKING, 5 = HOSTAGE TAKING (BARRICADE INCIDENT), 6 = HOSTAGE TAKING (KIDNAPPING), 7 = FACILITY / INFRASTRUCTURE ATTACK, 8 = UNARMED ASSAULT, 9 = UNKNOWN.
8. *nter (numerical):* Number of terrorists participated in the event (if unknown then assign 0).
9. *nkilledter (numerical):* Number of perpetrator fatalities.
10. *nwounded (numerical):* Number of total confirmed non-fatal injuries.
11. *nwoundedter (numerical):* Number of perpetrator non-fatal injuries.
12. *property (categorical):* 1 = “The incident resulted in property damage”, 0 = “The incident did not resulted in property damage”, -9 =” Unknown”.
13. *propertyextent (categorical):* 1 = Catastrophic (likely ≥ $1 billion), 2 = Major (likely ≥ $1 million but < $1 billion), 3 = Minor (likely < $1 million), 4 = Unknown.
14. *extended (categorical):* 1 = "Yes" (The duration of an incident extended more than 24 hours), 0 = "No" (The duration of an incident extended less than 24 hours).
15. *multiple (categorical):* 1 = "Yes" (The attack is part of a multiple incident), 0 = "No" (The attack is not part of a multiple incident).
16. *claimed (categorical):* 1 = "Yes" (A group or person claimed responsibility for the attack), 0 = "No" (No claim of responsibility was made).
17. *weapontype (categorical):* 1 = Biological, 2 = Chemical, 3 = Radiological, 4 = Nuclear, 5 = Firearms, 6 = Explosives, 7 = Fake Weapons, 8 = Incendiary, 9 = Melee, 10 = Vehicle, 11 = Sabotage Equipment, 12 = Other, 13 = Unknown The weapon type cannot be determined from the available information.

Based on the above mentioned 17 variables, we want to classify the terrorist organizations denoted by *‘gname’ (the name of the terrorist organization)* when an incident had taken place*.*

***Model Training & Validation:*** In the dataset, 73% of the terrorist organizations attacked for less than 5 times. And, in total they’re responsible for only 4% of the identified terror attacks. Now, as maximum number of the organizations have attacked for only time, there will be too many categories with very a smaller number of attack frequency, which will result to a 1-inflated model which can cause unusual errors training the model. Hence, we have excluded all the terrorist groups with less than 5 attacks.

Now for the purpose of classification we are considering five different classes and have coded them as 1,2,3,4 and 5 respectively. The classes are as follows:

**Table: Encoded groups with different Attack Frequencies**

|  |  |
| --- | --- |
| **Groups with Frequency of Attacks:** | **Coding** |
| 5-500 | 1 |
| 500-1000 | 2 |
| 1000-2000 | 3 |
| 2000-5000 | 4 |
| >5000 | 5 |

**Classifier Models:** In this research work, five classifier models are used for the purpose of classification of terrorist attack groups for certain occurred incident. The models which have been used are:

1. **Decision Tree:** Decision Trees or Classification Trees are binary tree structured classifiers, which are constructed by repeated splits of the subset of *X* into two descendant subsets, beginning with X itself. For uniform assumption, we say that the proportion of classes are equal, i.e.,

For any node ‘t’, suppose there is a split ‘S’ of the node which divides it into tL and tR, such that a proportion pL of cases in t go into tL and a proportion pR go into tR, and .

The goodness of split is defined as the decrease in impurity,

For a k class problem, let Nj be the no. of cases in class j . The prior probabilities are estimated by .

In a node t, let N(t) be the total no. of cases with Xn belongs to t and Nj(t), the no. of class j cases. The proportion of class j cases falling into t is . Hence, the probability that a case will both be in class j and fall into node t is,

So, for initial tree growing we need four elements:

1. A set Q of binary questions {for eg: is X belongs to A}.
2. A goodness of split criteria.
3. A stop-splitting rule.
4. A rule for arranging every terminal node to a class.

Initial Stop Splitting Rule: Let a threshold β>0 (however small) and declare a node t terminal if,

Class Assignment Rule: A class assignment rule assigns a class j∈{1,….,k} ro every terminal node t∈T with certain probability. The class assigned to node t is denoted by j(t). j(t) equals to that class for which P(j|t) is the largest.

Algorithm: At first, the tree begins with the root node ‘S’, containing the whole dataset → Then, the best attribute in the dataset is selected using ‘attribute selection measure’ → S is divided into subsets containing possible values of the best attribute → Recursively, this process is continued to make new subsets from the third step until a node reaches to a terminal node.

1. **Random Forest:** A random forest is an ensemble machine learning technique, that means it merges a number of classifiers for solving classification problems. Random Forest is a supervised method which consists of many decision trees. The main difference between a RF and DT classifier is that, in random forest getting root nodes and splitting the nodes is done at random by using Bagging method.

Algorithm: In Random Forest, n number of random records are taken from the data set having k number of records → Individual decision trees are constructed for each sample → Each decision tree will generate an output → Final output is considered based on Majority Voting or Averaging for Classification and regression respectively.

1. **Bagging:** Bagging is also known as Bootstrap aggregation. Bagging is an ensemble method which is used to reduce the variance of a highly varied dataset.

Algorithm: At first in Bagging method, **Bootstrapping** is done. If we have a fixed dataset, we can pick infinite samples (subsets) from the dataset with replacement. Hence, it results to resampling i.e., we have a large sample from a typically small dataset. It is to be noted that same value of the dataset can be present in the multiple samples for multiple number of times as the sampling is done with replacement → Once we get the bootstrapped samples, these samples are trained parallel with each other using base learners, this is known as **Parallel Training.** → At the end, a major number of estimates from the different samples are taken and pooled to compute an estimate with better accuracy. For classification problems, the class with most votes is accepted. This is known as hard voting or majority voting.

1. **XGBoost:** The XGBoost or extreme gradient boost is a classification algorithm which is made by integrating multiple decision trees. By using second order derivatives it improves the mis learned samples of previous round. In gradient boosting, a predictor corrects its predecessor’s error. XG Boosting is nothing, but implementation of gradient boosted decision trees.

**Data Partition:** In machine learning, we use to partition data into training and testing datasets as it is often the case that the variables of importance are unevenly distributed. Hence, according to the proportion of the target variables in dataset, the dataset is partitioned such that the proportion of the data in each category of the training set and the testing set is consistent with the proportion of the sample dataset. We will use *K-Fold Cross Validation* for this purpose.

**K-Fold Cross Validation:** Cross validation is a procedure for resampling, which is used for evaluation of ML/Statistical models, i.e. it resample the data from a single sample based on a single parameter ‘**k**’ which defines how many number of groups the sample should be split into.

In the next step, we use **‘k-1’** groups of the **‘k’** groups to train the model and with the remaining **1** group we validate the prediction based on the model built. In the next step, we repeat the same steps for the k set of combinations of the groups of the sample.

Chart

Description automatically generated with medium confidence

**Figure 7: K-Fold Cross Validation Structure**

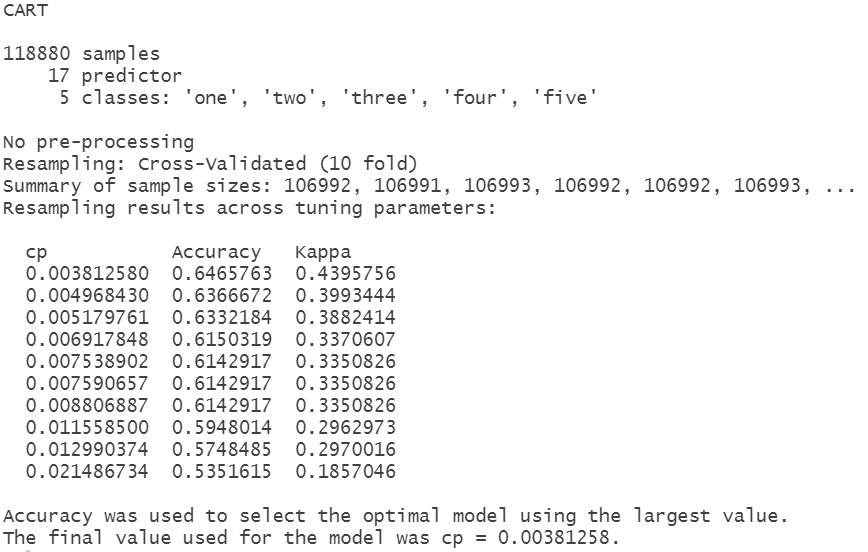
We have used 10-Fold Cross Validation for our case of validation of the different classifier models, i.e., value of parameter k=10.

**Results & Interpretation:**

a. **Decision Tree:**

We have 5 levels for our response variable “encoded”, namely, ‘one’,’two’.’three’,’four’,’five’. We are using 10-fold cross validation method for training our classifier. This is because, there are some variations in the data, i.e., some groups attacked a greater number of times while some other groups attacked a lesser number of times. Hence, training the model for 10 times and coming to a final model is an adequate plan to pursue.

Running the Decision Tree based on the 17 covariates, we get the following output:



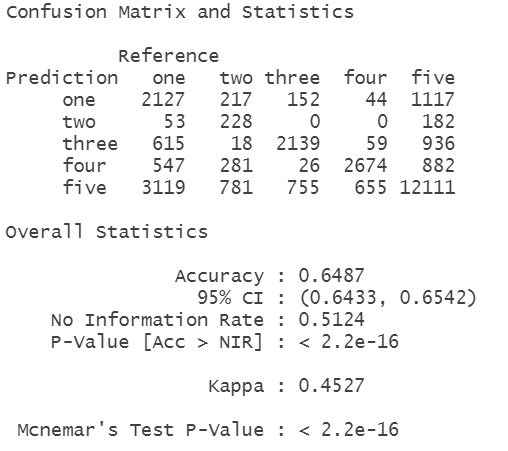
We can see from the output that there are different sample sizes due to the use cross validation. It is observed that the final value of the complexity parameter of the model ‘cp’ = 0.00381258 which gives the highest accuracy and kappa value for the model. The kappa metric = 0.4395 means that the model is moderate fit.

Diagram

Description automatically generated

**Figure 8: Decision Tree Model**

Now, based on the trained model we will validate our model by applying it to our test data (which is a 20% split of the complete dataset). We get the following confusion matrix by predicting on the testing data:



From the output, it is seen that the model is 64.87% accurate, i.e., we can tell that Decision tree is an average classifier for our purpose. [Accuracy=Sum of diagonal of the confusion matrix/Sum of its all elements].

Given, non-information rate (NIR) = 0.5124

We want to test, H0: Accuracy of the model is equal to NIR

vs. H1: Accuracy of the model is greater than NIR

i.e., if the model’s accuracy is better than the proportion of data with the majority class.

P-value of the test = 2.2\*10^-16

Hence, at 5% level of significance, P-value<α, hence. we may reject Ho.

And by empirical rule, Kappa=0.4527 signifies a moderate level of agreement of the model and also we can say that the data is 35-63% reliable.

The variables of importance are given in the table below,

Table

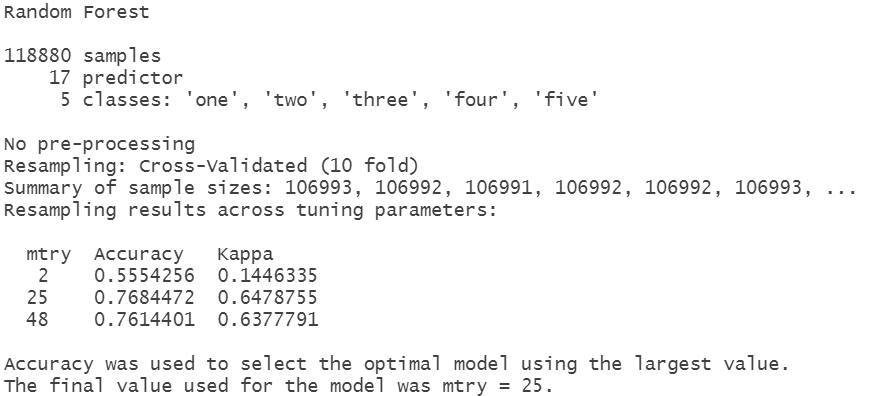
Description automatically generated

**Table: Variables of Importance according to DT**

b. **Random Forest**

**Out Of Bag (OOB) Error:** Out of bag error is method of measuring the prediction error of random forests and other ML models where Bagging is used. OOB error is defined as the mean prediction error on each training sample xj, using only those trees which do not have xj in their bootstrapped samples.

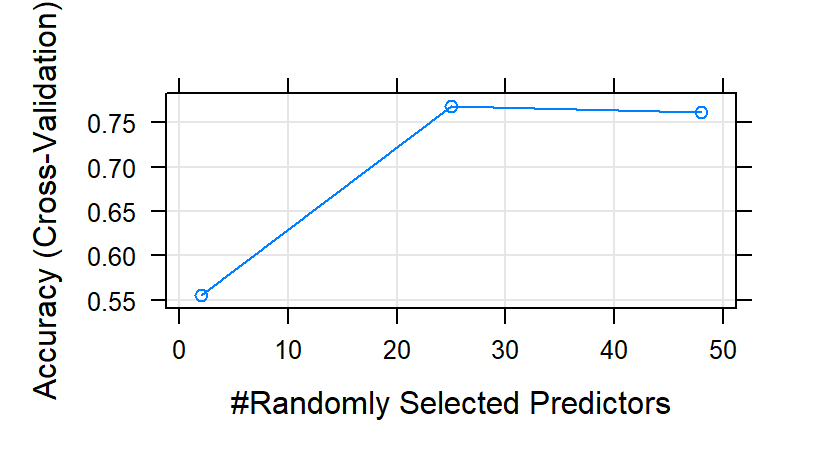
Now fitting the RF Classifier to our data, we get the following output:



From the output, the final mtry value used for the model, for which the accuracy is maximized = 25. This means that the number of variables randomly sampled as candidates at each split = 25 for which the model is giving the maximum accuracy.

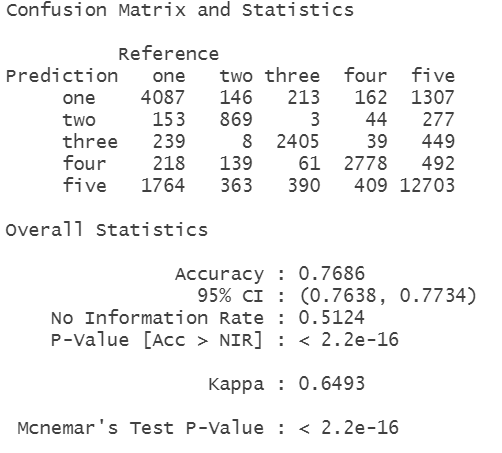
The OOB error of the model = 23.004%

Now OOB error is an estimate of the error rate of the model, i.e., 1-Accuracy. Hence, we can say that the estimate of error is 23% approximately for the model.



**Figure 9: Randomly Selected Predictors vs Accuracy (Cross Validation)**

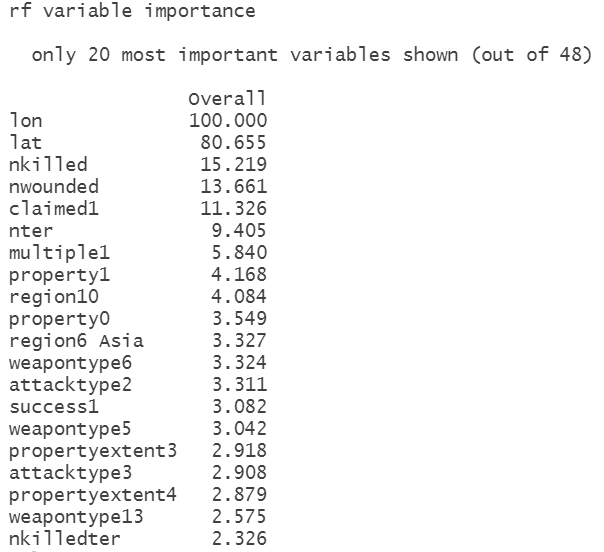
Now we validate the model by checking the accuracy from the confusion matrix, based on the testing data.



The accuracy of the model is equal to 76.86%, i.e., it is much higher than a Decision Tree model which was quite expected, as RF is a group of Decision trees (ensemble).

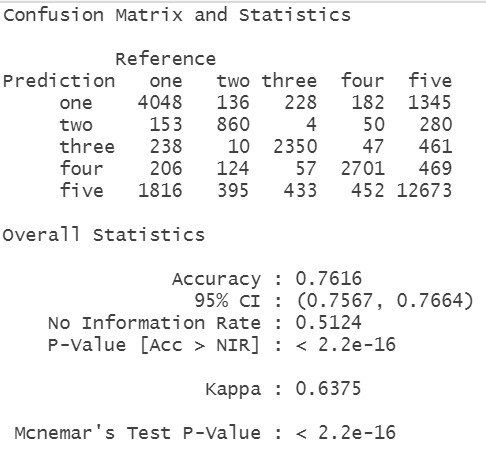
Kappa score is quite high; hence the data is >64% reliable, and the level of agreement of the model is moderate.

Variables of Importance according to RF classifier is given below:



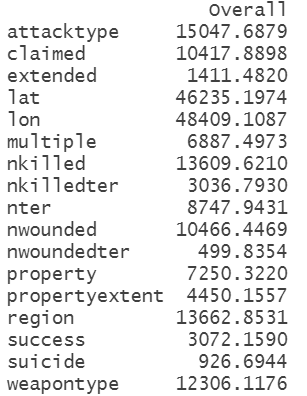
c. **Bagging:**

Bagging or bootstrap aggregating is an effective classification method. Applying it to our data, we get,



We can see that the model of Bagging is also performing quite well in the testing data. Accuracy of the model = 76.16% which is moderately good. Also, kappa is specifying a moderately reliable value.

Variables of importance according to Bagging:



***Section 3***

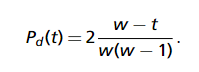
***Objective:*** In this section, we want to perform a Spatio-Temporal Analysis called as *Near-Repeat Analysis.*

A near repeat incident was first published in 2003 by Townsley, Homel, & Chaseling. Near-Repeat is defined as the occurrence of a crime event at a particular location, given it is at target of that crime events, and at least one event has occurred there. Near-Repeat Analysis is a branch of Criminology.

In our study, we have taken attacks by a group called ‘Talibans’ and have considered the the latitude, longitude and Date of each of the occurred event.

**Methodology:**

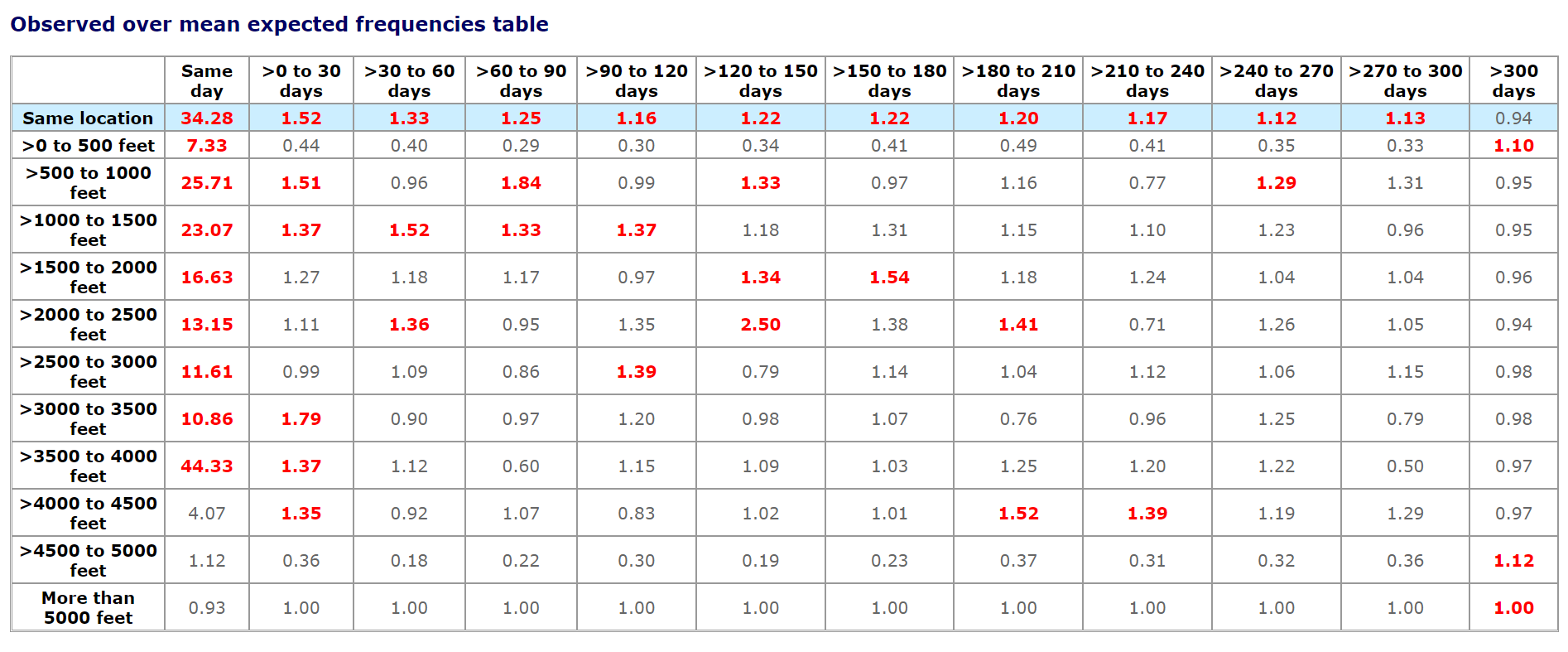
This is done by partitioning the given timespan into several windows of times ‘w’; within each window the time difference is 0<t<w and the spatial distance between every pair of events is calculated. The time distribution for events occurring within a maximum distance ‘d’ is subsequently compiled from all windows. This data-derived distribution is then compared to the random-event hypothesis:



**Results and Interpretation:**

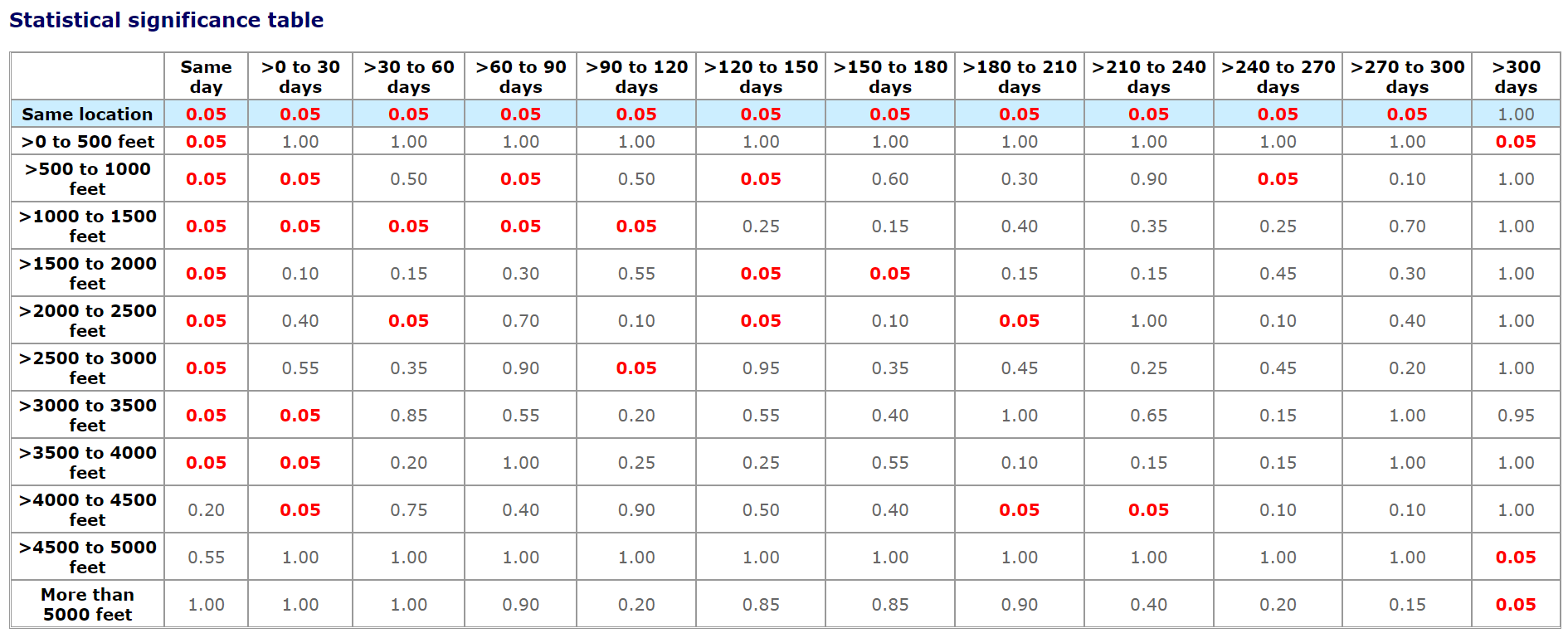
We have performed, Near-Repeat analysis on activities of Taliban. The latitude, longitude and Date of an event has been taken into consideration. We have done the Analysis in Near Repeat Calculator (Ver 2.0).

We have converted the latitude and longitudes distance from meridian (in feet), as the calculator does not support for degrees.



The observed over mean expected frequency table show us the expected number of attacks for different temporal as well as spatial distances.

The Knox’s ratio is defined as the test statistic for testing the the above spatio-temporal frequency distribution. We get the result as,



At 5% level of significance, the red values defines that the test are being rejected. [here, H0: There will no Near-Repeats at the discussed place, vs H1: H0 is not true]

Hence, we can observe that, at certain time points and distance points, there are chances of near-repeat of an event.

***Conclusion:***

From the 3 sections of the project, it was a wholehearted try for using Statistical/ML tools as an equipment of counter terrorism.

World security bodies can apply such best fitted classification models to classify for the terrorist organization responsible for a particular terrorist attack, based on the incident parameters. It would be a great deal for the several Investigation Bureaus for spotting the responsible groups of a concerned terrorist event. We have seen that Random Forest has given the best accuracy in the prediction of the testing set, hence we would choose for Random Forest among the 3 applied classification models. Once, the Terrorist organizations can be classified according to the encoded class (as discussed previously), we can further apply the same classification technique within each class to predict the exact name of the responsible group for an event.

In the third section, Near-Repeat analysis is also quite useful if we want to predict the probabilities and expected frequencies of ‘Near-Repeat’ attacks at a certain place and at a certain time, given we have some historic records of events occurred at different places over different times. The process is still under research and development phase, hence there are ample future scope to work on this topic to carry on this analysis.

***Bibliography:***

1. Data visualization – START - Global Terrorism Database

2. Quantitative Analysis of Global Terrorist Attacks Based on the Global Terrorism Database Zhongbei Li, Xiangchun Li, Chen Dong, Fanfan Guo, Fan Zhang and Qi Zhang

3. GTD - CODEBOOK: METHODOLOGY, INCLUSION CRITERIA, AND VARIABLES

4. Near Repeat Calculator - <https://figshare.com/articles/code/Near_Repeat_Calculator/11409996>

5. Blog – Near Repeat Analysis - <https://www.jratcliffe.net/near-repeat-analysis>

6. Xiaohui Pan - Quantitative Analysis and Prediction of Global Terrorist Attacks Based on Machine Learning

7. Olusola A. Olabanjo, Et. Al. - An ensemble machine learning model for the prediction of danger zones: Towards a global counter-terrorism   
8. Kanika Singh, Et. Al. - A Machine Learning Approach for Enhancing Defence Against Global Terrorism.

9. "Pattern Recognition and Machine Learning" by Christopher M. Bishop

10. "Classification and Regression Trees" by Leo Breiman, Jerome Friedman, Charles J. Stone, and R.A. Olshen

11. "Random Forests" by Leo Breiman

12. "Introduction to Statistical Learning" by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani

13. "Machine Learning: A Probabilistic Perspective" by Kevin P. Murphy

14. "Applied Predictive Modeling" by Max Kuhn and Kjell Johnson

15. "An Introduction to Machine Learning" by Miroslav Kubat