Global Terrorism

Capstone Project by ANN

HDSC Winter 2022

# 

# Abstract

“*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*.” - GTD.

As a generation, we may not have faced World War I and II. But the events of the past and the present year are testaments to the fact that humankind is still far from “Live and Let Live”. The constant parlay of power grabs, and political, trade, and religious conflicts still suffocate all of us.

Defense against terrorism is difficult because of terrorism's surprise advantage, since we do not know the type, extent, timing or precise location of the next attack. Preventive measures are still the best defense against terrorism.

Hence, this project studies such factors that make an attack successful. Knowing the target region and extent of an attack can be highly useful for saving countless lives. The past and current scenarios can be examined with the aid of data visualization. Further, we aim to predict the target region of an attack and whether or not an attack will be successful with the help of various machine learning and deep learning models.

# Source

<https://www.kaggle.com/datasets/START-UMD/gtd>

The Global Terrorism Database (GTD) is an open-source database including information on terrorist attacks around the world from 1970 through 2017. The GTD includes systematic data on domestic as well as international terrorist incidents that have occurred during this period and now includes more than 180,000 attacks. The database is maintained by researchers at the National Consortium for the Study of Terrorism and Responses to Terrorism (START), headquartered at the University of Maryland.

# Dataset Description

This dataset consists of 181691 rows and 135 columns containing information about various terrorist attacks across the world over a while of 1970 to 2017, except 1993. The large number of columns includes information about location, tactics, perpetrators, targets, and outcomes.

# Exploratory Data Analysis

Due to the sizable number of features present in the dataset, the first step was to check for missing values and redundant features to prepare the data for visualization and modelling.

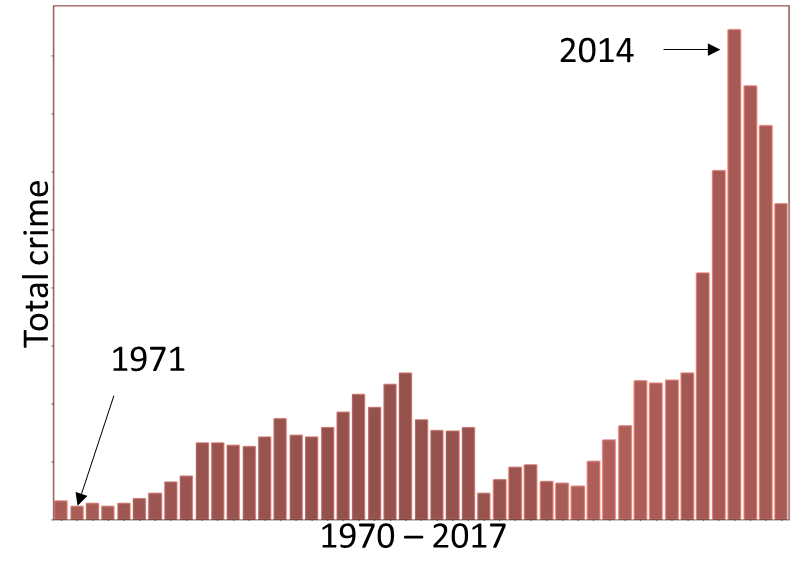
## Data Wrangling

* The entire dataset was divided into 7 sections (column-wise) and each section was separately checked for missing values and redundant features, which were then dropped. After the initial data cleaning, 27 columns were chosen out of 135 which were then further scaled down to fit the specific purposes of the problem definitions.
* The missing values for the numerical features were replaced with the median of that respective column, and the same for string features was replaced with ‘Unknown’.
* Duplicate rows were dropped.

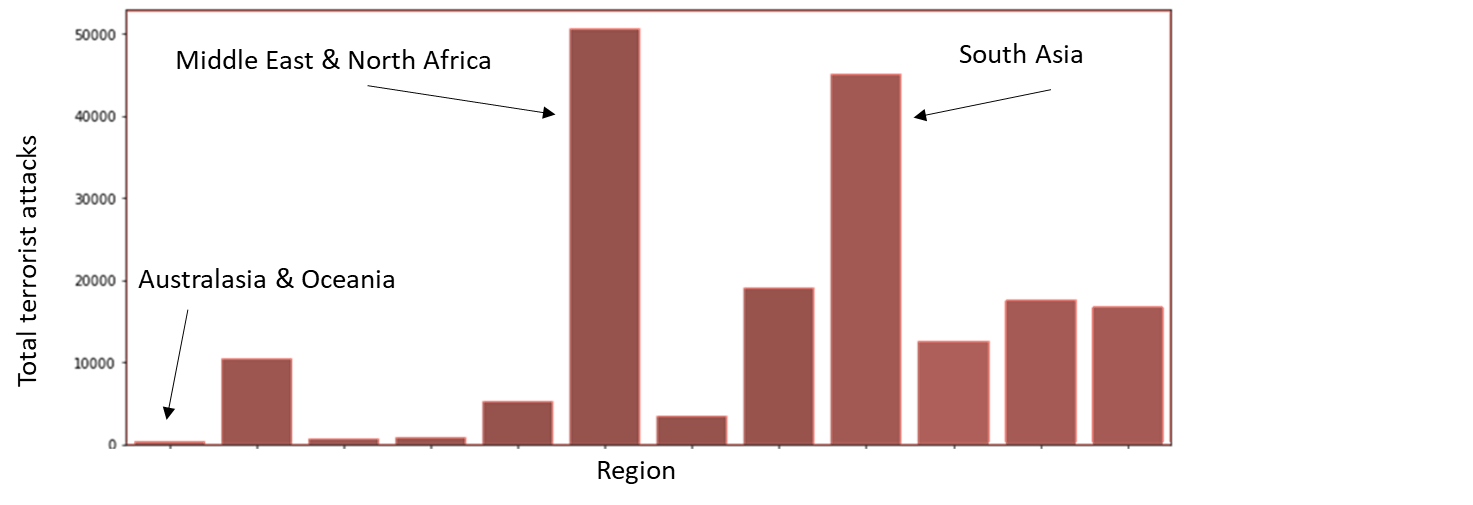
## Data Visualization

* Global terrorism rate over the years:

**Insight:** The rate of global terrorism increased over time with 2014 recording the highest rate of global terrorism while 1971 with the least recorded terrorist attacks.

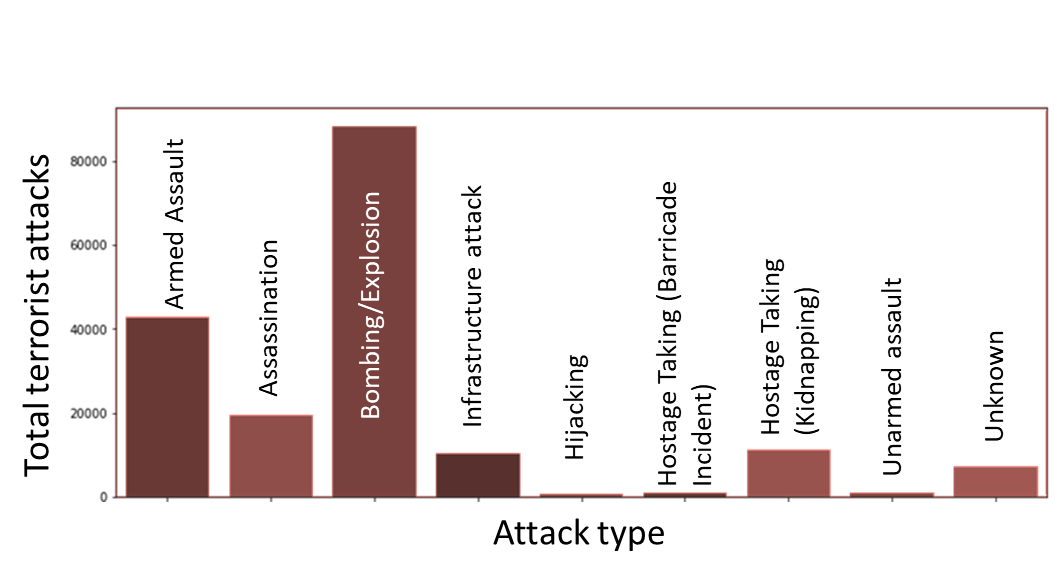
****

* Terrorist attacks vs. region



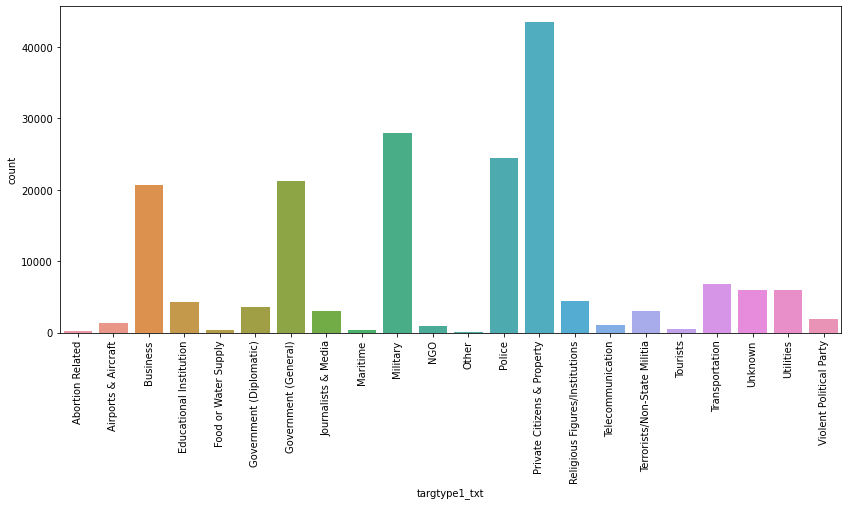
**Insight:** The regions with the most occurrences of global terrorism are the Middle East & North Africa, and South Asia while Australasia and Oceania is the region with the least number of attacks.

* Terrorist attack types and their significance:



**Insight:** The predominant form of attack in global terrorism has been carried out by the use of Bombing/Explosion followed by Armed Assault, while Hijacking has been the least form used.

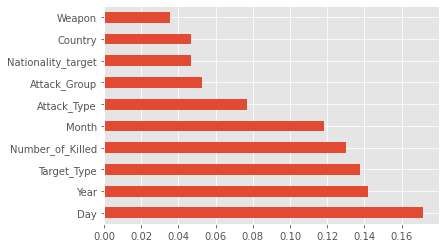
* Terrorist Attacks and their targets:



**Insight:** The most targeted area of global terrorism has been Private Citizens & Property followed by the Military and the Police.

# Feature Engineering

* Date features (such as year, month, day) were merged into a date column with a uniform format.
* Label Encoder was used on categorical features.
* During this phase, it was observed that there existed an imbalance in the dataset nearly in the ratio of 1:8 favouring successful attacks. Hence, oversampling was performed on the minority category to tackle this problem.
* ‘Extratrees’ classifier was used to discern the important features in the dataset.

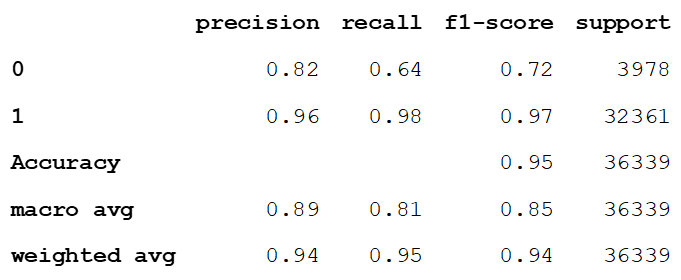


# Model development

This project tackles two different problem statements for this dataset. For each problem statement, different machine learning and deep learning models have been developed.

## Aim 1 - Prediction of a successful attack

In order to predict whether a terrorist attack is going to be successful or not, a deep learning model - Sequential, was developed. This model was fine-tuned with the ‘adam’ optimizer and ‘Binary cross-entropy’ loss function, resulting in the following classification report.



From the above classification report, it can be observed that the deep learning model performed fairly well on the dataset by achieving an 95% accuracy. The slight discrepancies in the precision for both the classes (successful and unsuccessful attacks) arose because of oversampling of the minority class i.e., unsuccessful attacks.

## Aim 2 - Prediction of the target region of a terrorist attack

For predicting the target region of attack, several ensemble machine learning models have been used. Aided by the performance metrics results, Random Forest emerged as the most accurate model among the others with an accuracy and recall of 93%.



Figure 1- Table for comparing ensemble models

The confusion matrix generated by the chosen model, Random Forest, gives a more detailed insight into the cases.



Figure 2- Confusion matrix for Random Forest classifier

# Conclusion

This project dives deep into the Global terrorism dataset providing important visualizations that help in extracting useful information like the rate of global terrorism over the years, regions targeted and weapons used. Moreover, using deep learning and machine learning models, predictions of successful attacks and target regions were generated. Sequential deep learning model for predicting the success or failure of an attack provided results with 95% accuracy, while Random Forest Classifier model for predicting the target region achieved an accuracy of 93%. Some of the main challenges faced during this project were tackling a huge dataset which made feature extraction tricky, and the class imbalance which, even though was tackled by oversampling, resulted in slightly biased results. An obvious improvement on the problem statements would be to collect more raw data on unsuccessful attacks so as to reduce the class imbalance problems. Further, the implementation of more deep learning models can produce better and more accurate results.