Terrorism: Data Analysis & Predictive Modelling

Shivprasad Kathane | Minor in AI & Data Science | C-MInDS IIT Bombay

Introduction

- Terrorism has become a huge threat to society in recent times as terrorist attacks lead to loss of lives, damages to property & mental trauma for many.
- Counter-terrorism measures are employed by security forces to prevent occurrence of such attacks. An important basis for such measures is data.
- Global Terrorism Database (GTD) by START is a comprehensive database on terrorist attacks and popular in data analytics studies on terrorism.
- Data analysis & visualization reveals key insights from historical trends.
- Theoretical studies help find factors or causes for terrorism.
- But can both be integrated and utilized for predictive applications?

Literature Review

- Alaa S. Alsaedi et al (2019) [1] used 3 machine learning (regression) algorithms: K-nearest neighbor (KNN), Naïve Bayes (NB) and Random Forest (RF) for prediction of attacks and attackers based on GTD. RF performed the best.
- M. Irfan Uddin et al (2020) [2] used 5 different deep neural network (classification) models to predict whether attack is successful or not, suicidal or not, the type of weapon used, the mode of attack and the region of attack. One DNN model demonstrated more than 95% accuracy compared to other techniques in ML.
- Terrorism has been attributed to many potential factors, including economic factors (Piazza 2006; Ljujic et al. 2017), political factors (Bakker 2006; Criado 2017) and social factors (Alcalá et al. 2017; Fu et al. 2012) under theoretical frameworks like exploitation theory, game theory, utility maximization theory etc.

Building upon prior work

- Shuo She et al (2020) [3] used a negative binomial regression model based on data of 49 countries under China's proposed Belt & Road program from multiple sources from 1999 to 2014. They explored the impact of political, economic and social factors on the number of terrorist attacks and resulting casualties.
- Building upon this idea of national indicators as factors correlating with terrorism, all countries were included and popular ML regression models were utilised to predict same outputs with varying input data for time period from 1970 to 2017.
- This analysis expands the time space domain of input data and builds models for predictive applications rather than just determining significant causes for terrorism.

Datasets

- Global Terrorism Database (GTD)
  Documents more than 190,000 terrorist attacks that occurred worldwide from 1970 to 2018.
- Polity5 Project Database (PPD)
  Records political regime characteristics and transitions from 1800 to 2018.
  It has assigned scores for some parameters of a political system.
- Database on Political Institutions (DPI)
  It has absolute data on many political variables but not included in primary dataframe.
- Penn World Tables (PWT)
  Has a dataset on economic and social indicators from 1950 to 2017.
- ourworldindata.org (OWD)
  Annual data on population in urban and rural areas for each country from 1960 to 2017.

Variables (Primary Dataframe)

Variables from PWT

- HCI: Human Capital Index, based on years of schooling & returns to education
- %Emp: Calculated as proportion of population engaged in workforce
- GDP: Real GDP at constant 2011 national prices (in mil. 2011US$)
- Rconna: Real consumption at constant 2011 national prices (in mil. 2011US$)
- XR: Exchange Rate, national currency/USD (market+estimated)
- CP: Price level of household consumption relative to that of USA in 2011

Other Variables:

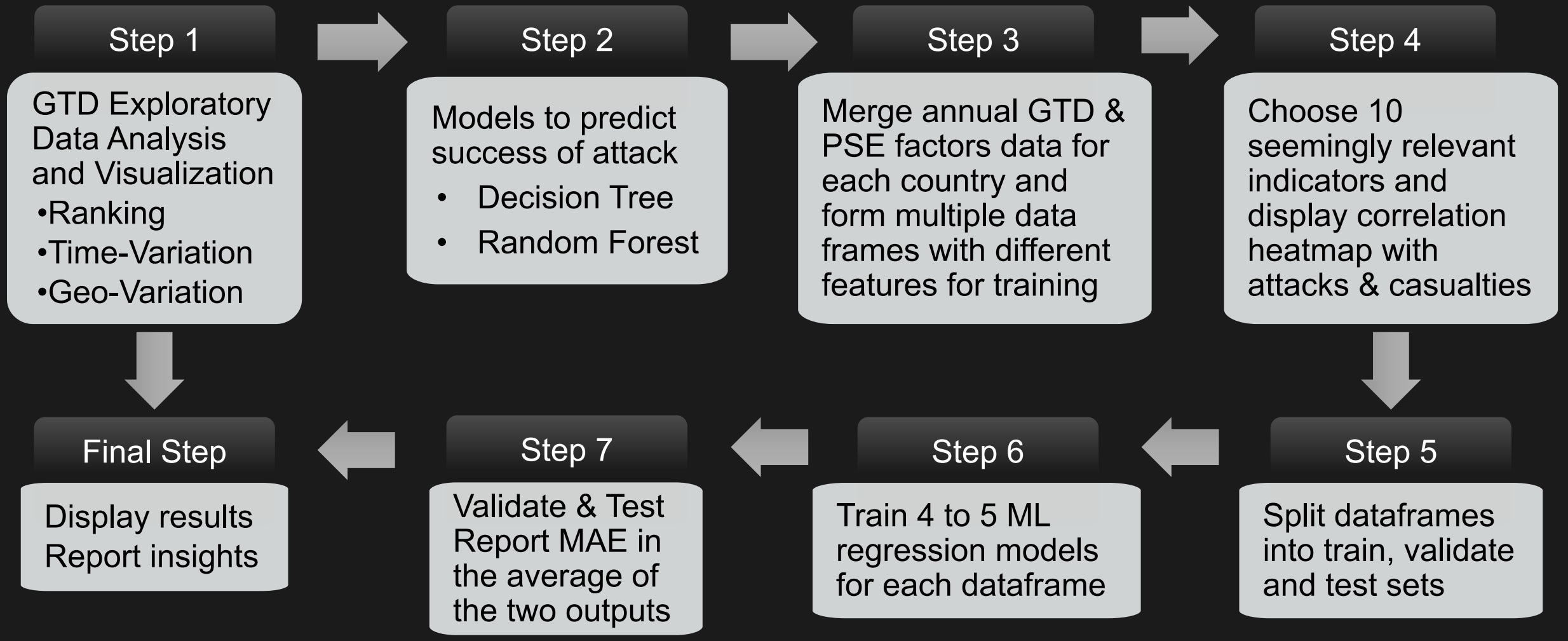
- DS: Level of democracy (score 0-10 by PPD)
- AS: Level of autocracy (score 0-10 by PPD)
- PSS: Political Stability Score by PPD (length of years for which same party rules a country)
- %Urb: Taken as percentage of population living in urban areas (OWD)
- Nattacks: No of terrorist attacks (GTD)
- Ncasual: No. of casualties=killed+injured (GTD)

The last 2 variables were predicted (outputs) based on the rest 10 variables as inputs to ML models.

Research

- Exploratory Data Analysis on GTD for world in general and India in specific
  - Variable values were ranked according to their popularity in attacks
  - Time-variation in values of these variables was studied
- Predicting the success of an attack based on GTD variables
- Merging GTD data with data on political & socio-economic indicators of a country
- Noting correlations between PSE factors and terrorism
- Applying Machine Learning Algorithms on aggregated data:
  Objective: to be able to predict the number of terrorist attacks and casualties in a country in a year given data on political, social & economic fronts reported for that country in that year
- Sample insights and their influence on counter-terrorism planning

Analysis Pipeline



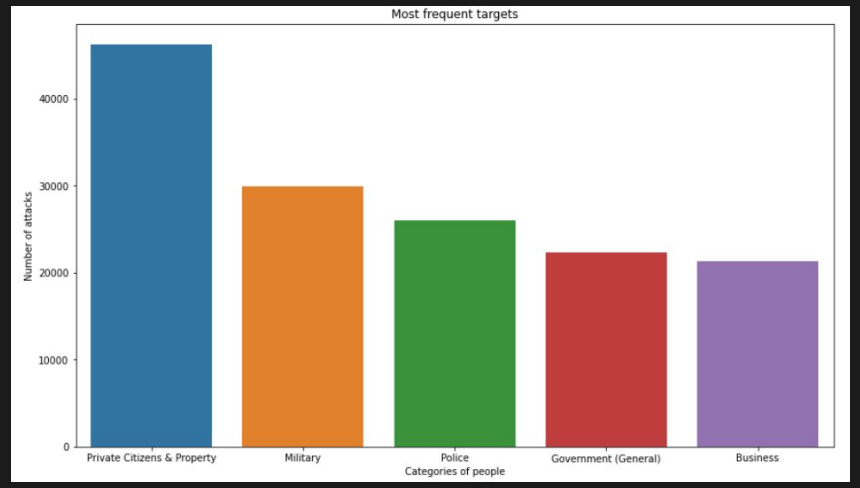
GTD Analysis Observations

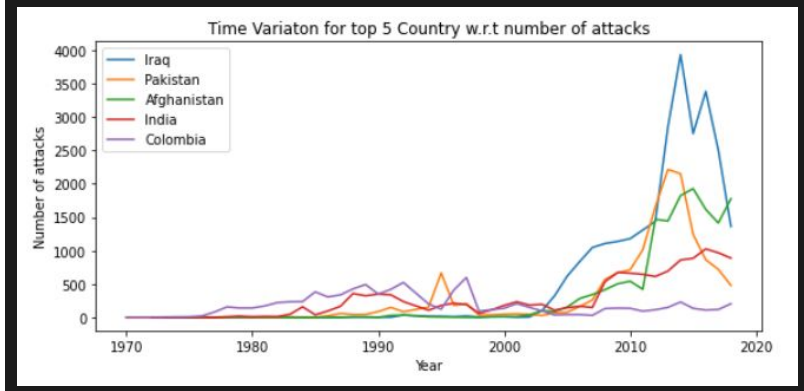
World:

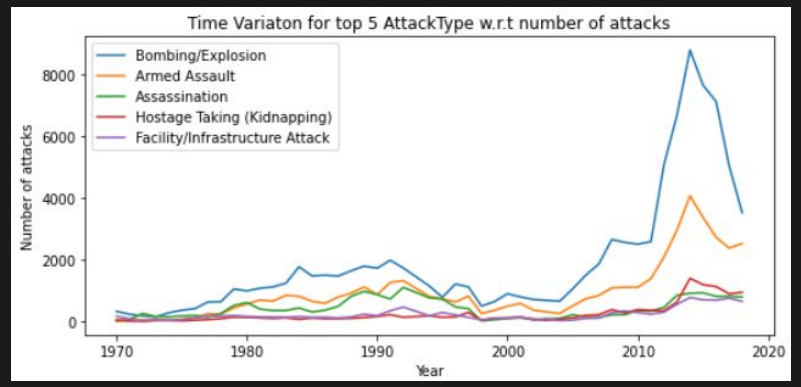
- Steep rise in attacks post 2010 in South-Asia, Middle East & Africa.
- Most active terrorist groups: Taliban, ISIL & Al-Shabaab
- Iraq, Afghanistan, Pakistan & India have suffered the most
- Popular targets: Private citizens
- Most dominant mode of attack: Bombing using explosives

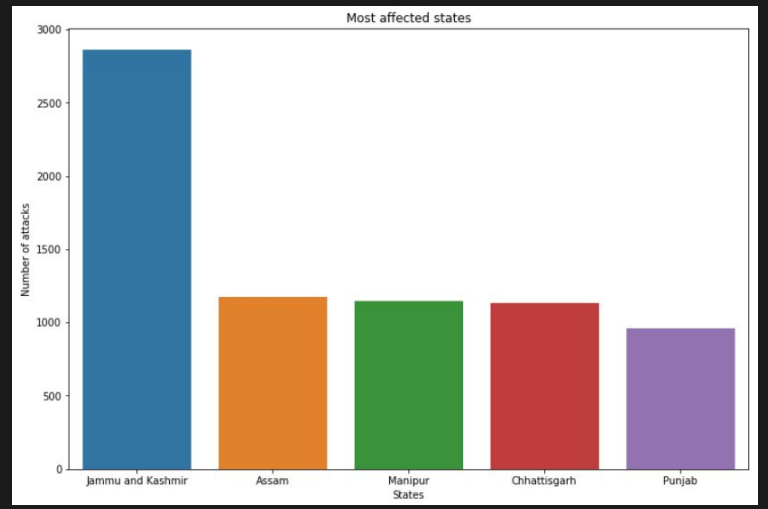
India:

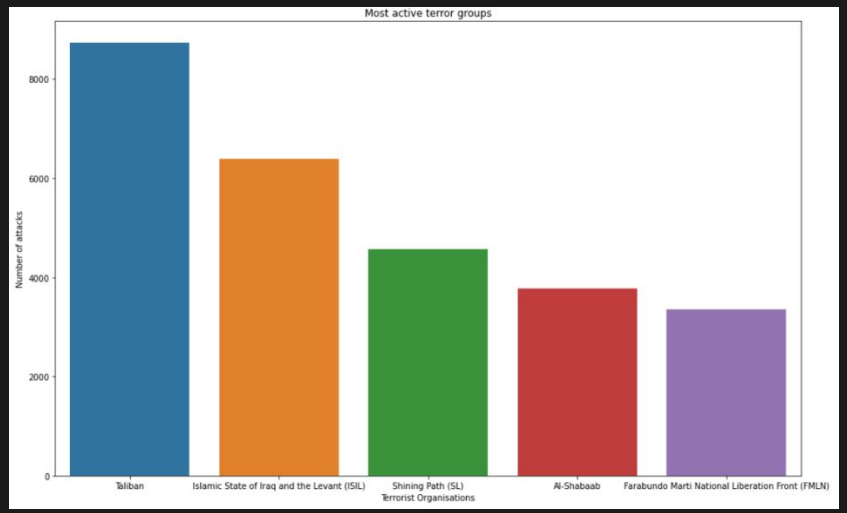
- Rising attacks on Police recently
- Most no. of attacks by Maoists
- Kashmir: Worst affected state

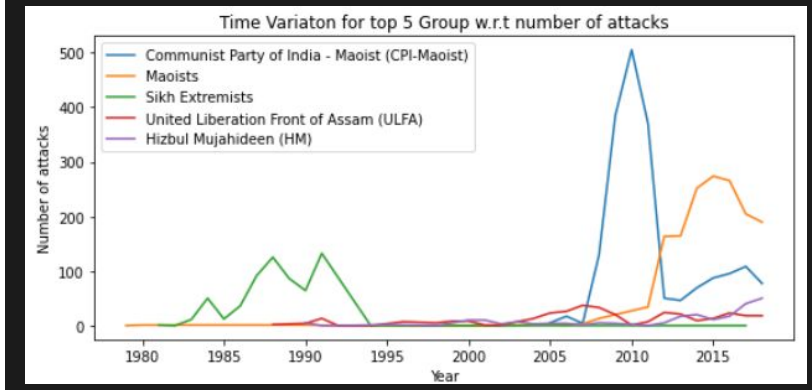












Machine Learning Results

Machine Learning Model	10 features from 3 datasets	All PWT variables	All DPI variables
	Best Validation MAE		
Linear Regression	236.58	297.5	212.02
Lasso Regression	210.26	238.64	169.3
Random Forest Regression	97	140.46	92.81
Neural Network (MLP) Regression	198.13	-	-
KNN Regression	88.13	114.84	75.39
Test MAE	121.77 (KNN)	111.83 (KNN)	86.52 (KNN)

PWT+DPI data was trained using KNN and it yielded MAE=99.54 on the test set

Conclusions

- Review of existing literature utilising GTD primarily reveals machine learning models for prediction of attacks and attackers [1], deep learning models for prediction of future terrorist activities [2] and study of correlation factors influencing terrorist attacks [3].
- Decision tree classifier used to predict the success of a terrorist attack yielded 85% weighted average accuracy which suggests that GTD is good enough for obtaining data-driven insights.
- In general, as per the correlation heatmap terrorism is found to be low where there is more education, employment, urbanisation, political stability, economic strength though the correlations weren't strong enough to determine significant factors and this is expected for a diverse database.
- KNN performed the best in all the 3 cases (input dataframes) for ML. The reason could possibly be that it maps terrorism to its geography as the inherent nature of this algorithm suggests.
- Least test set MAE=86.52 was obtained on DPI data. This seems to indicate that political factors influence terrorism more dominantly than others inline with findings of [3].
- Extension of work: Developing problem specific algorithms and choosing relevant indicators.

Impact & Relevance

- India witnessed its single most deadly terror attack as an explosion in Mumbai local trains in 2006 and this was followed by a series of attacks in the same city in 2008. Hence, the issue of terrorism has been key since then and taking ‘data-driven’ decisions has become an important part of security also in the current information-heavy world.
- Practical review of existing data analysis and findings in the domain of terrorism via elegant visualisation:
  - Enhanced security and combat requires knowledge on the popular weapons used in the attacks, most active terrorist organisations, most targeted sections of society, popular modes of attack etc.
  - Which are the cities, countries and regions prone to terrorism governs deployment of security forces.
  - Further, addressing how these have varied with time in the past provides better temporal insights.
- The ultimate desirable goal for counter-terrorism can be to develop real-time location based threat perception and formulate anti-terrorism policies for the nation based on the current situation in the country. This work aims to aid the same via an effective data mining practice for analysing terrorism as a function of national PSE factors.
- The results for the machine learning part are significant. Although the error (MAE) is large compared to the mean number of attacks (50) it still is much better compared to the maximum value for the number of casualties (30769).

Acknowledgement & References

Done as part of DS203M Programming for Data Science Course Project at IITB

Thankful to course profs & TAs, project partner and VRSS organisers & attendees

[1] A. S. Alsaedi, A. S. Almobarak and S. T. Alharbi, ”Mining the Global Terrorism Dataset using Machine Learning Algorithms,” 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA), Abu Dhabi, United Arab Emirates, 2019, pp. 1-7.

[2] M. Irfan Uddin, Nazir Zada, Furqan Aziz, Yousaf Saeed, Asim Zeb, Syed Atif Ali Shah, Mahmoud Ahmad Al-Khasawneh and Marwan Mahmoud, ”Prediction of future terrorist activities using Deep Neural Networks,” Wiley Hindawi Complexity, vol. 2020.

[3] She, S., Wang, Q. Weimann-Saks, D.,”Correlation factors influencing terrorist attacks: political, social or economic? A study of terrorist events in 49 “Belt and Road” countries,” Qual Quant 54, 125–146, 2020.

[4] Global Terrorism Database (GTD), START, University of Maryland

[5] Feenstra, Robert C., Robert Inklaar and Marcel P. Timmer (2015), ”The Next Generation of the Penn World Table” American Economic Review, 105(10), 3150-3182.

[6] Polity5 dataset version 2018, Polity Project, Center for Systemic Peace

[7] Hannah Ritchie (2018) - ”Urbanization”. Retrieved from: ”https://ourworldindata.org/urbanization” [Online Resource]

[8] Database of Political Institutions (DPI)