GLOBAL TERRORISM 1970-2017

By NISHIT PARMAR

EM 623
FINAL
PROJECT

Tutor:
Dr. CARLO
LIPZZI

TABLE OF CONTENTS

I) BUSNIESS UNDESTANDING

Terrorism has been a major concern over the years and analysis needs to be done. We will be applying CRISP-DM on the Terrorism Database. The geography included is worldwide. The goal is to analyze the dataset and bring out conclusions and try to recommend measures which can decrease/diminish terrorist activities.

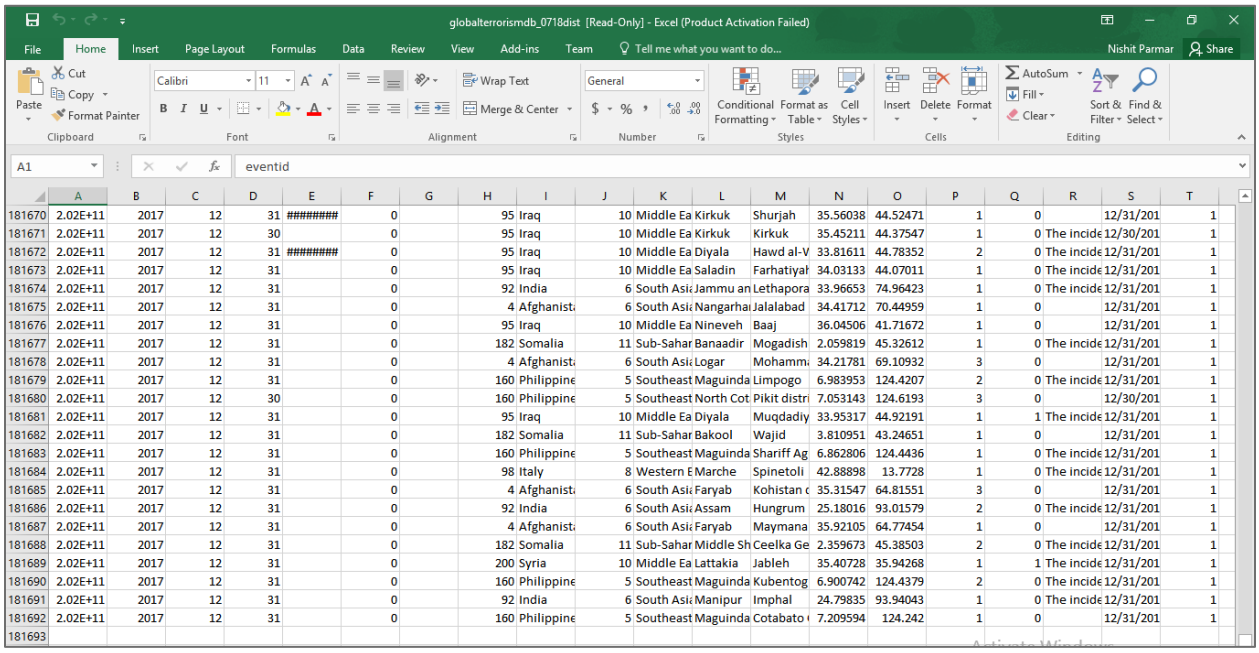
II) DATA UNDERSTANDING

The dataset is a compilation of terrorist activities over the globe. The timeline of the dataset is from the year 1970 till 2017. It includes systematic data on domestic as well as international terrorist happenings that have occurred over the years. There are around 180,000 observations in the dataset. The dataset is rich in variables, around 100 variables.

SOURCE: *National Consortium for the Study of Terrorism and Responses to Terrorism (START), headquartered at the University of Maryland.*

VARIABLES: *>100 variables on location, tactics, perpetrators, targets, and outcomes*

III) DATA PREPARATION



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
181670	2.02E+11	2017	12	31	########	0		95	Iraq	10	Middle Ea	Kirkuk	Shurjah	35.56038	44.52471	1	0		12/31/201	1
181671	2.02E+11	2017	12	30		0		95	Iraq	10	Middle Ea	Kirkuk	Kirkuk	35.45211	44.37547	1	0	The incide	12/30/201	1
181672	2.02E+11	2017	12	31	########	0		95	Iraq	10	Middle Ea	Diyala	Hawd al-V	33.81611	44.78352	2	0	The incide	12/31/201	1
181673	2.02E+11	2017	12	31		0		95	Iraq	10	Middle Ea	Saladin	Farhatiyah	34.03133	44.07011	1	0	The incide	12/31/201	1
181674	2.02E+11	2017	12	31		0		92	India	6	South Asi	Jammu an	Lethapora	33.96653	74.96423	1	0	The incide	12/31/201	1
181675	2.02E+11	2017	12	31		0		4	Afghanist	6	South Asi	Nangarhar	Jalalabad	34.41712	70.44959	1	0		12/31/201	1
181676	2.02E+11	2017	12	31		0		95	Iraq	10	Middle Ea	Nineveh	Baaj	36.04506	41.71672	1	0		12/31/201	1
181677	2.02E+11	2017	12	31		0		182	Somalia	11	Sub-Sahar	Banaadir	Mogadish	2.059819	45.32612	1	0	The incide	12/31/201	1
181678	2.02E+11	2017	12	31		0		4	Afghanist	6	South Asi	Logar	Mohamma	34.21781	69.10932	3	0		12/31/201	1
181679	2.02E+11	2017	12	31		0		160	Philippine	5	Southeast	Maguinda	Limpogo	6.983953	124.4207	2	0	The incide	12/31/201	1
181680	2.02E+11	2017	12	30		0		160	Philippine	5	Southeast	North Cot	Pikit distri	7.053143	124.6193	3	0		12/30/201	1
181681	2.02E+11	2017	12	31		0		95	Iraq	10	Middle Ea	Diyala	Muqdadiy	33.95317	44.92191	1	1	The incide	12/31/201	1
181682	2.02E+11	2017	12	31		0		182	Somalia	11	Sub-Sahar	Bakool	Wajid	3.810951	43.24651	1	0		12/31/201	1
181683	2.02E+11	2017	12	31		0		160	Philippine	5	Southeast	Maguinda	Shariff Ag	6.862806	124.4436	1	0	The incide	12/31/201	1
181684	2.02E+11	2017	12	31		0		98	Italy	8	Western E	Marche	Spinetoli	42.88898	13.7728	1	0	The incide	12/31/201	1
181685	2.02E+11	2017	12	31		0		4	Afghanist	6	South Asi	Faryab	Kohistan d	35.31547	64.81551	3	0		12/31/201	1
181686	2.02E+11	2017	12	31		0		92	India	6	South Asi	Assam	Hungrum	25.18016	93.01579	2	0	The incide	12/31/201	1
181687	2.02E+11	2017	12	31		0		4	Afghanist	6	South Asi	Faryab	Maymana	35.92105	64.77454	1	0		12/31/201	1
181688	2.02E+11	2017	12	31		0		182	Somalia	11	Sub-Sahar	Middle Sh	Ceelka Ge	2.359673	45.38503	2	0	The incide	12/31/201	1
181689	2.02E+11	2017	12	31		0		200	Syria	10	Middle Ea	Lattakia	Jableh	35.40728	35.94268	1	1	The incide	12/31/201	1
181690	2.02E+11	2017	12	31		0		160	Philippine	5	Southeast	Maguinda	Kubentog	6.900742	124.4379	2	0	The incide	12/31/201	1
181691	2.02E+11	2017	12	31		0		92	India	6	South Asi	Manipur	Imphal	24.79835	93.94043	1	0	The incide	12/31/201	1
181692	2.02E+11	2017	12	31		0		160	Philippine	5	Southeast	Maguinda	Cotabato (	7.209594	124.242	1	0		12/31/201	1
181693																				

Figure1: Original Dataset

The dataset contains 181,692 rows and there are more than 100 variables.

DATE

The dataset had three columns for Year, month and date. With the help of =DATE function a new column was created which had all three in one.

MISSING VALUES

Working on a data with so many missing values was not efficient and therefore it required a lot of cleaning. The means adopted to clean the data was to filter out all the rows and columns which had missing values: blank and "NA" with a proportion of more than 50%. Excel was used to do this,

using the formulas: COUNTIF and COUNTBLANK. This resulted in a comparatively cleaner dataset with a lot of columns deleted.

CORRELATION MATRIX

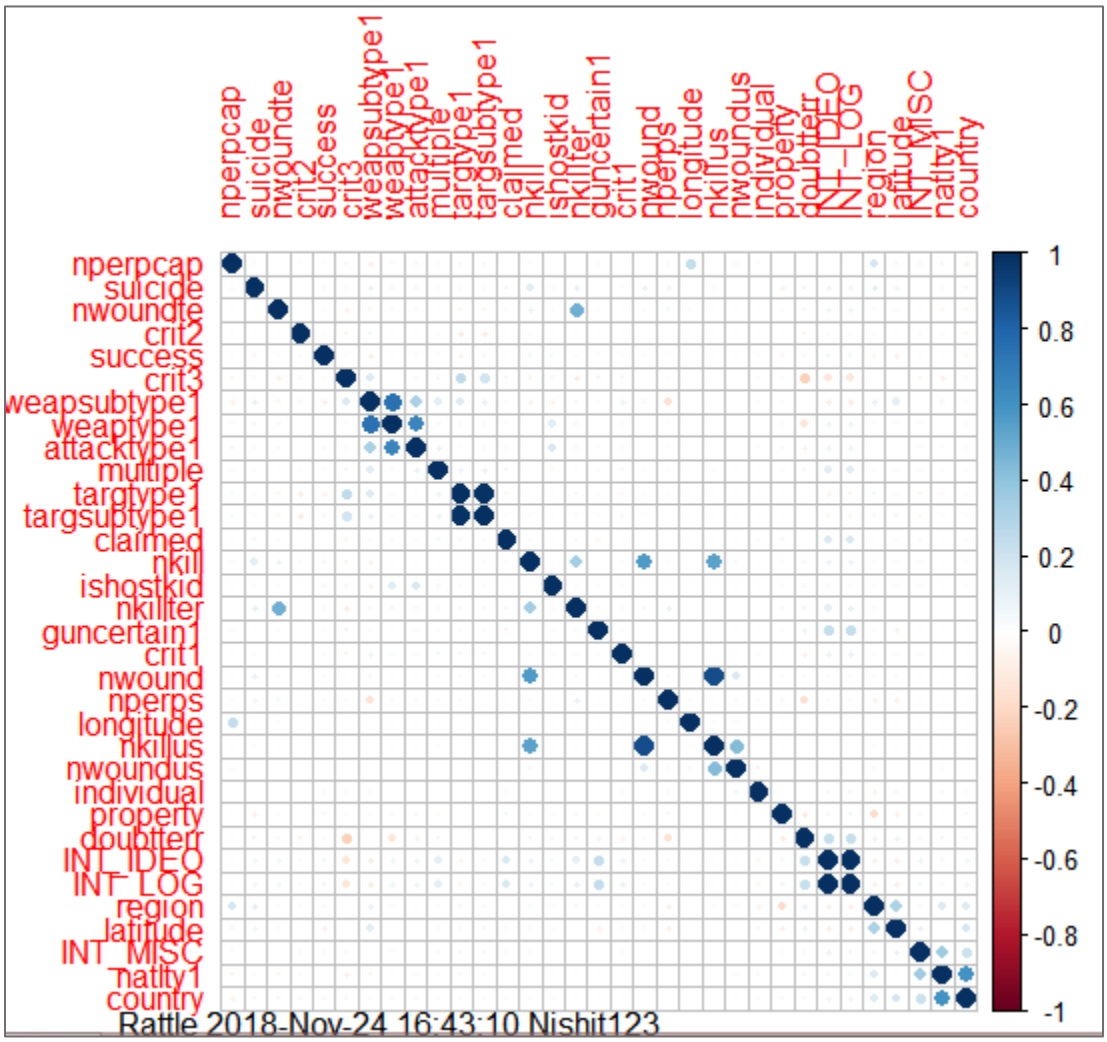


Figure2: Corelation Matrix

IV) MODELING

The dataset was split into two parts for training and testing.

i) 70% of the data for training
ii) 30% of the data for testing

A decision tree was plotted on the basis of that:

The parameters adopted to form the tree were

Min. split = 20

Min bucket = 7

Max depth = 5

Complexity = 0.01

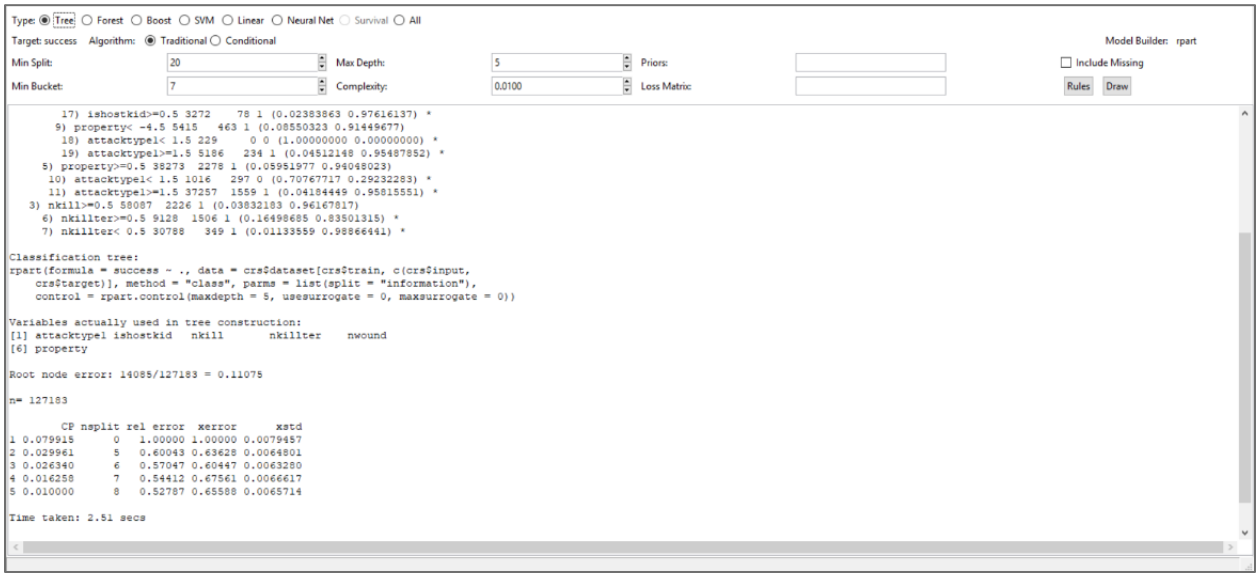


Figure3: Decision Tree Parameters

The overall error is 0.1107%

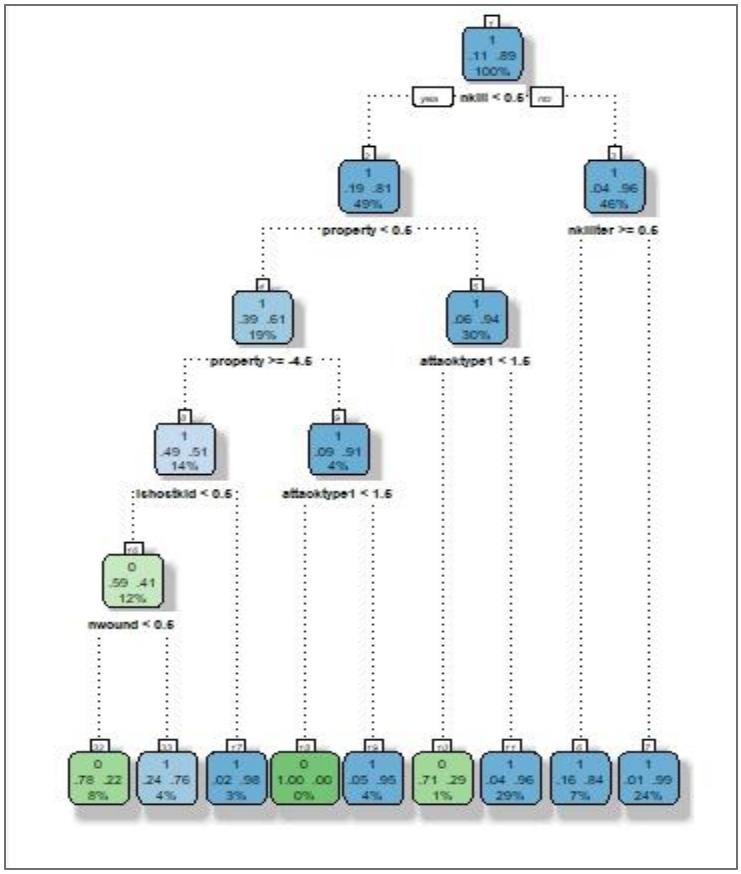


Figure4: Decision Tree

V) EVALUATION

ERROR MATRIX

An **error matrix** shows the true outcomes against the predicted outcomes. There are two tables as the output of the error matrix. First is the count of the observations and the second one is the count of proportions from the top left going clockwise, as the True Negatives,

False Positives, True Positives, and False Negatives.

Evaluation was done on the testing data.

Overall error is 6.5% which is efficient and averaged class error is 21.15%

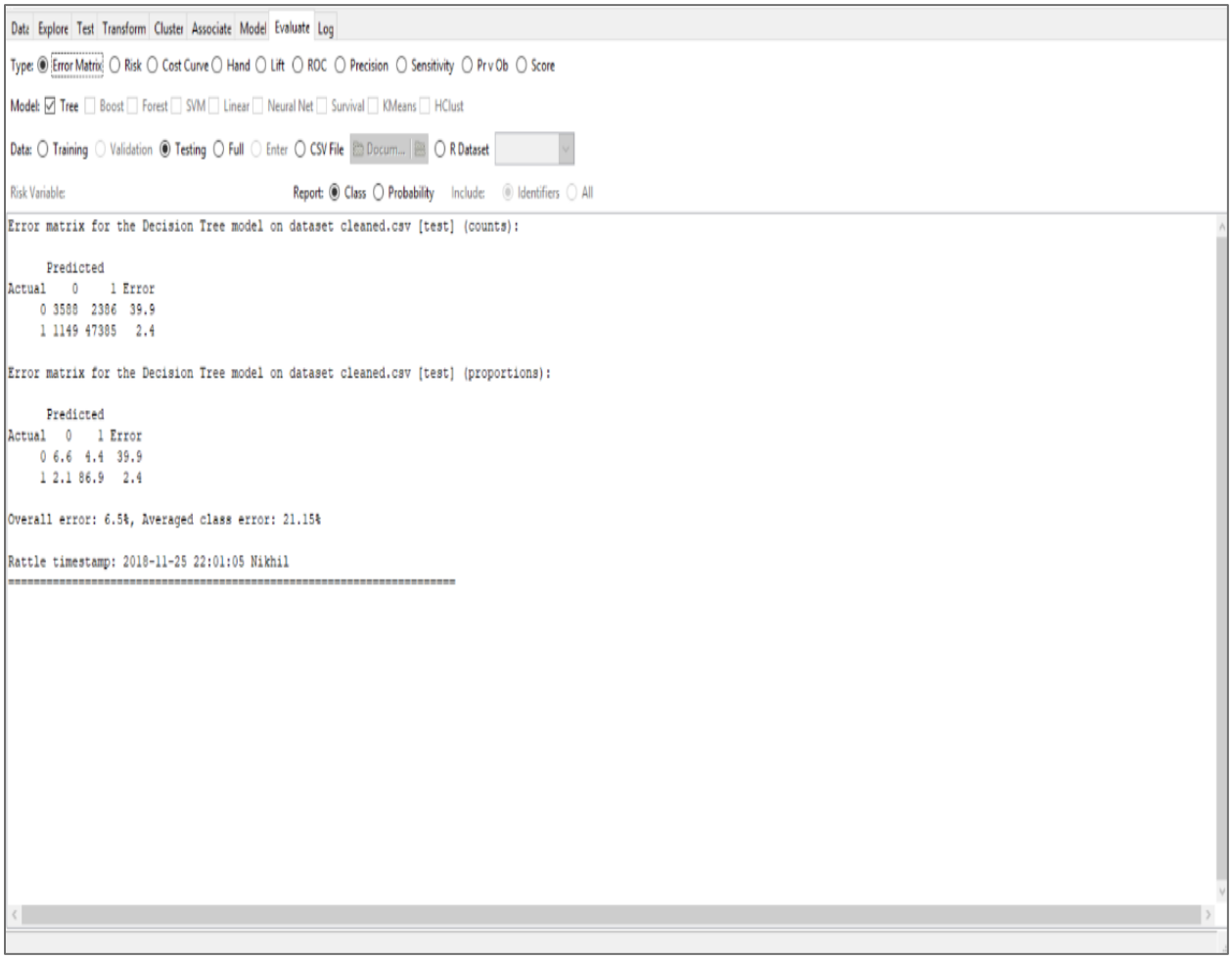


Figure5: Error Matrix

RECIEVER OPERATING CHARACTERISTICS (ROC)

The ROC curve plots the true positives against the false positives

Greater the area under the curve means a better model performance. AUC- 0.89

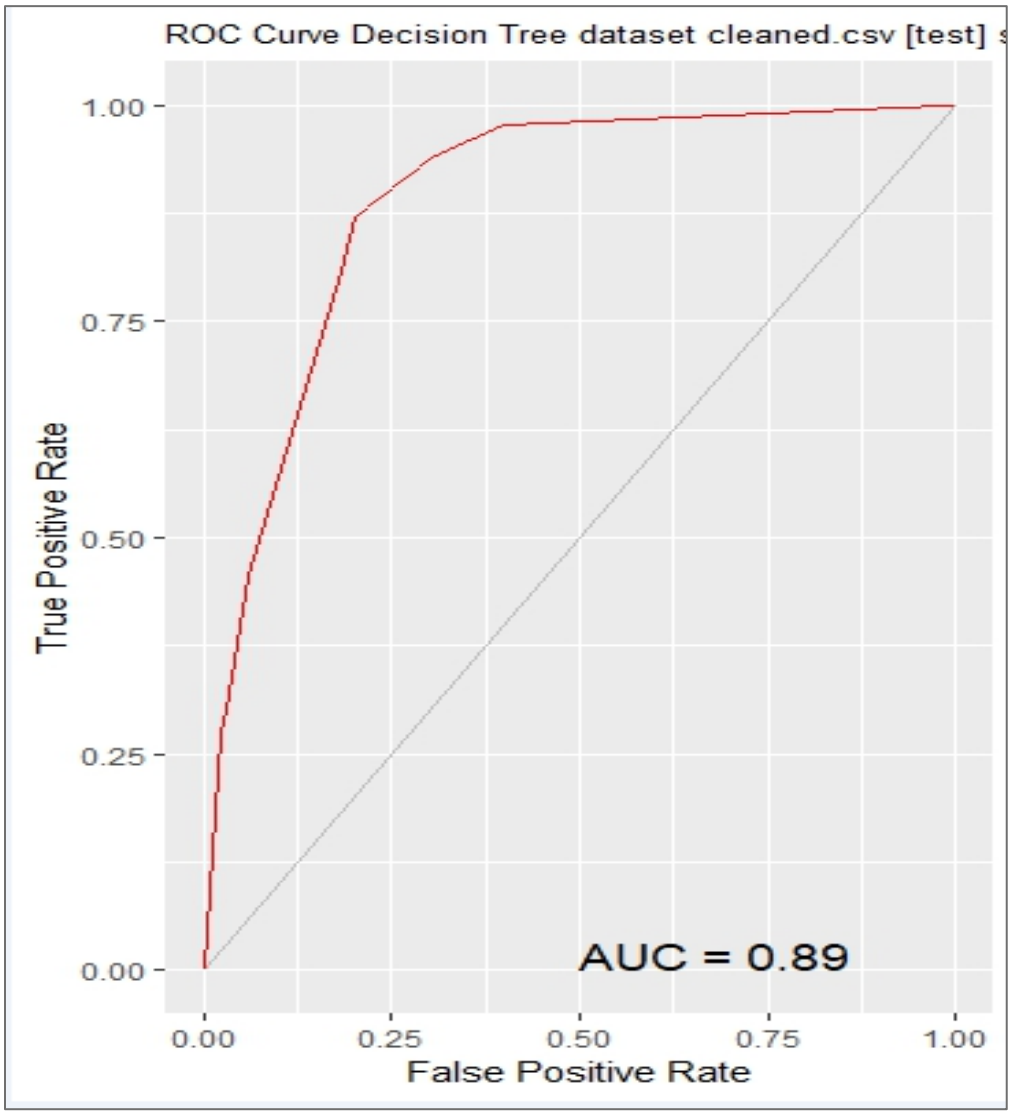


Figure6: ROC Curve

VI) CONCLUSION

To draw conclusions on several parameters, visualizations were created.

1) Year-wise

It is seen that over the last decade terrorism has drastically increased. However it is noticed that since 2014, the numbers have decreased.

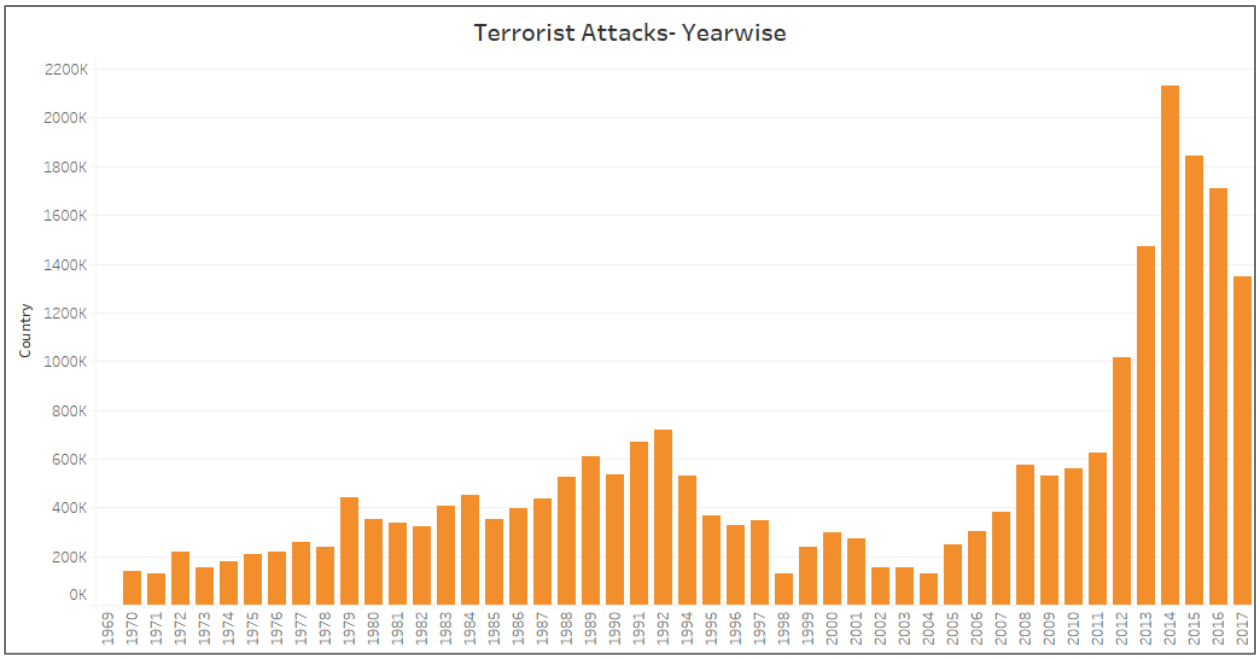


Figure7: Year wise terrorist attacks

2) Country-wise

It can be seen that Iraq is the most prone to terrorism followed by Afghanistan and Pakistan.

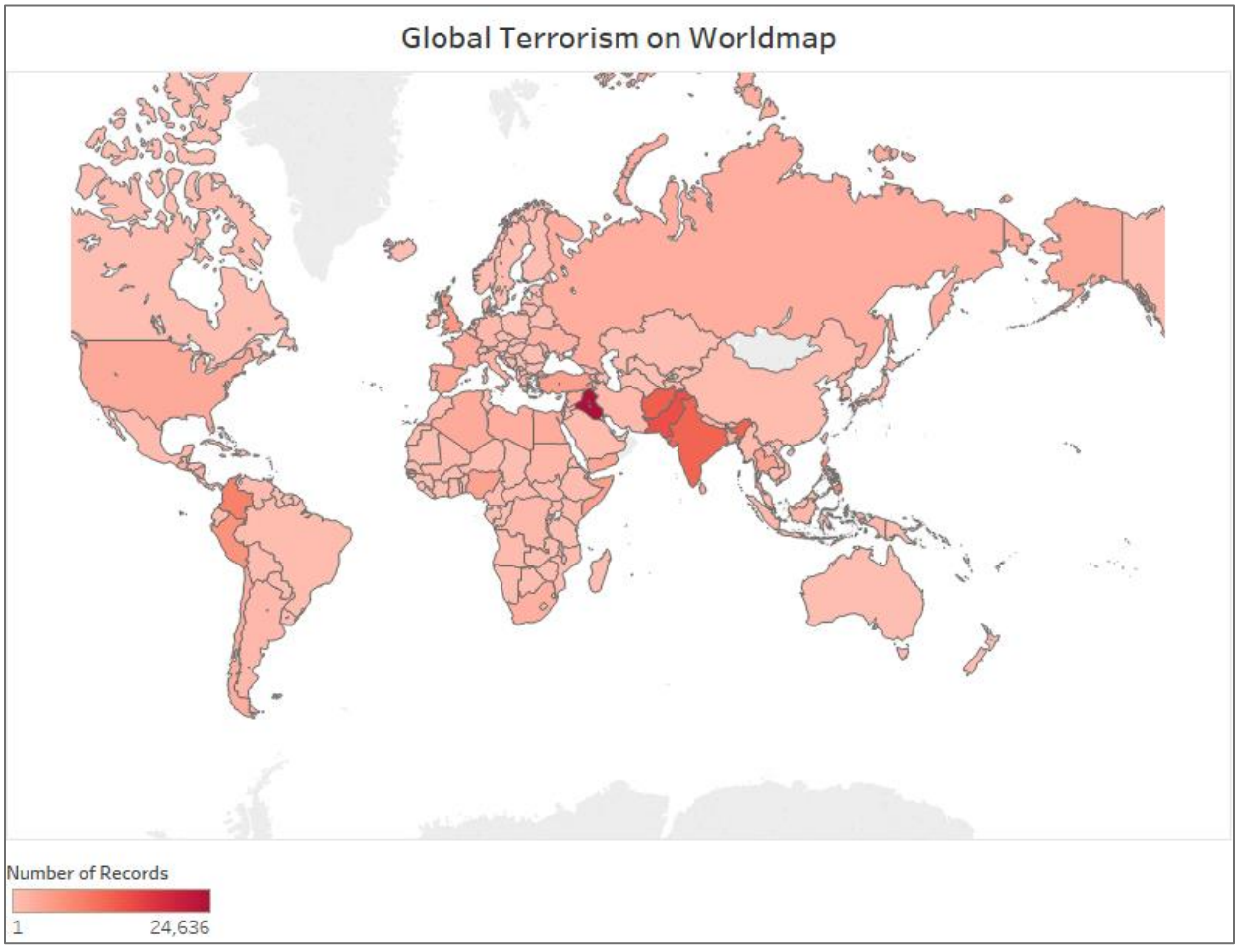


Figure8: Terrorism worldwide.

3) Region-wise terrorism trend

It is seen that Middle East and North Africa have been affected by terrorist activities more than any region and measures need to be taken in these regions in order to diminish it.

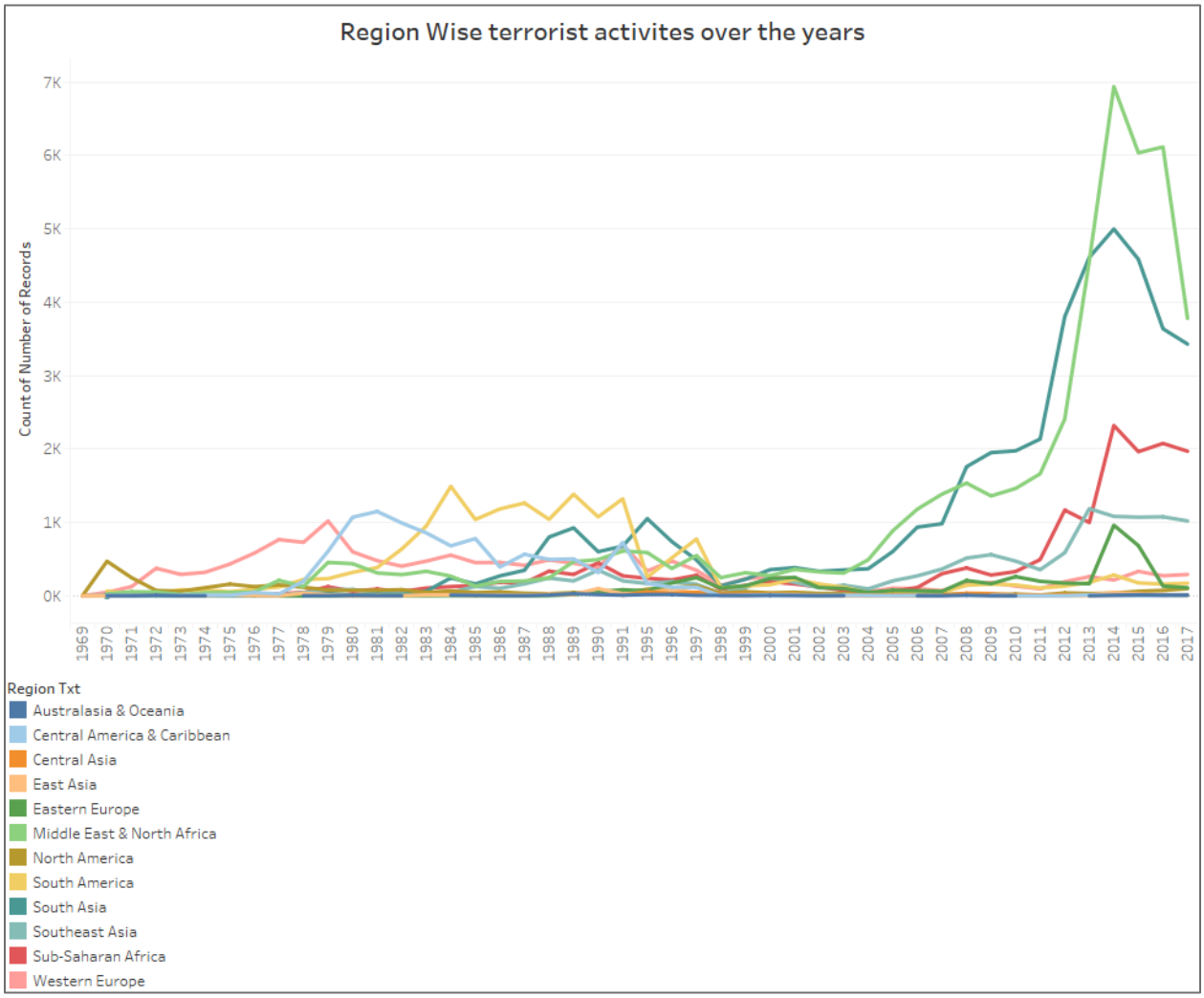


Figure9: Region wise trend in terrorism

4) Most frequent types of attacks adopted by terrorists

The category of bombing surpasses every other category of attacking means. There have been nearly 90,000 cases of bombing. Following bombing is, Armed assaults which is approx. 45,000.

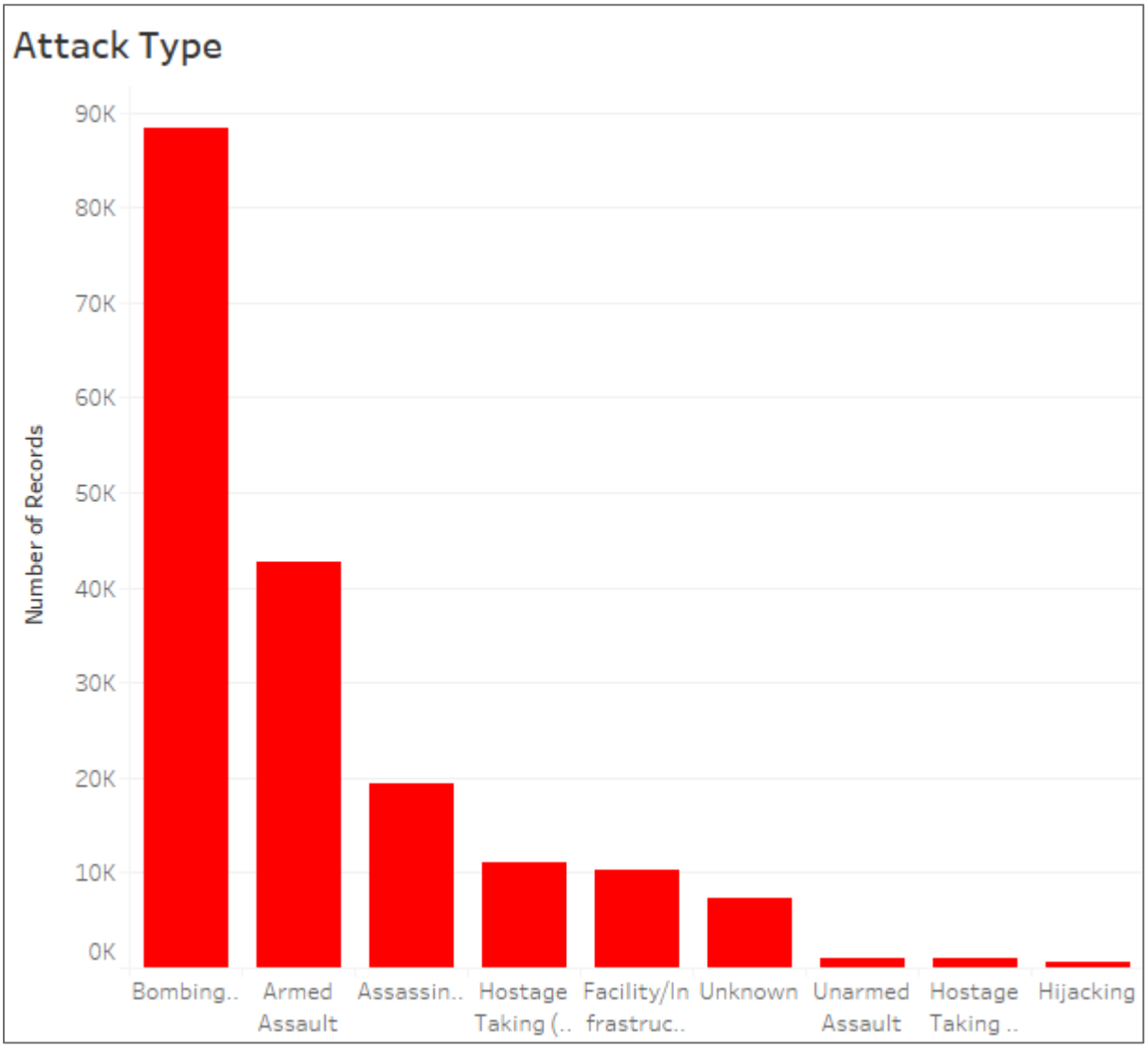


Figure10: Most frequent means of attacks`

5) Frequently used weapons for attacks

After plotting the graph it was seen that, the primary types of weapons and explosives used is still unknown. It needs to be identified. The frisking and checking needs to be done strictly. Adoptive measures need to be taken in order to fight terrorism

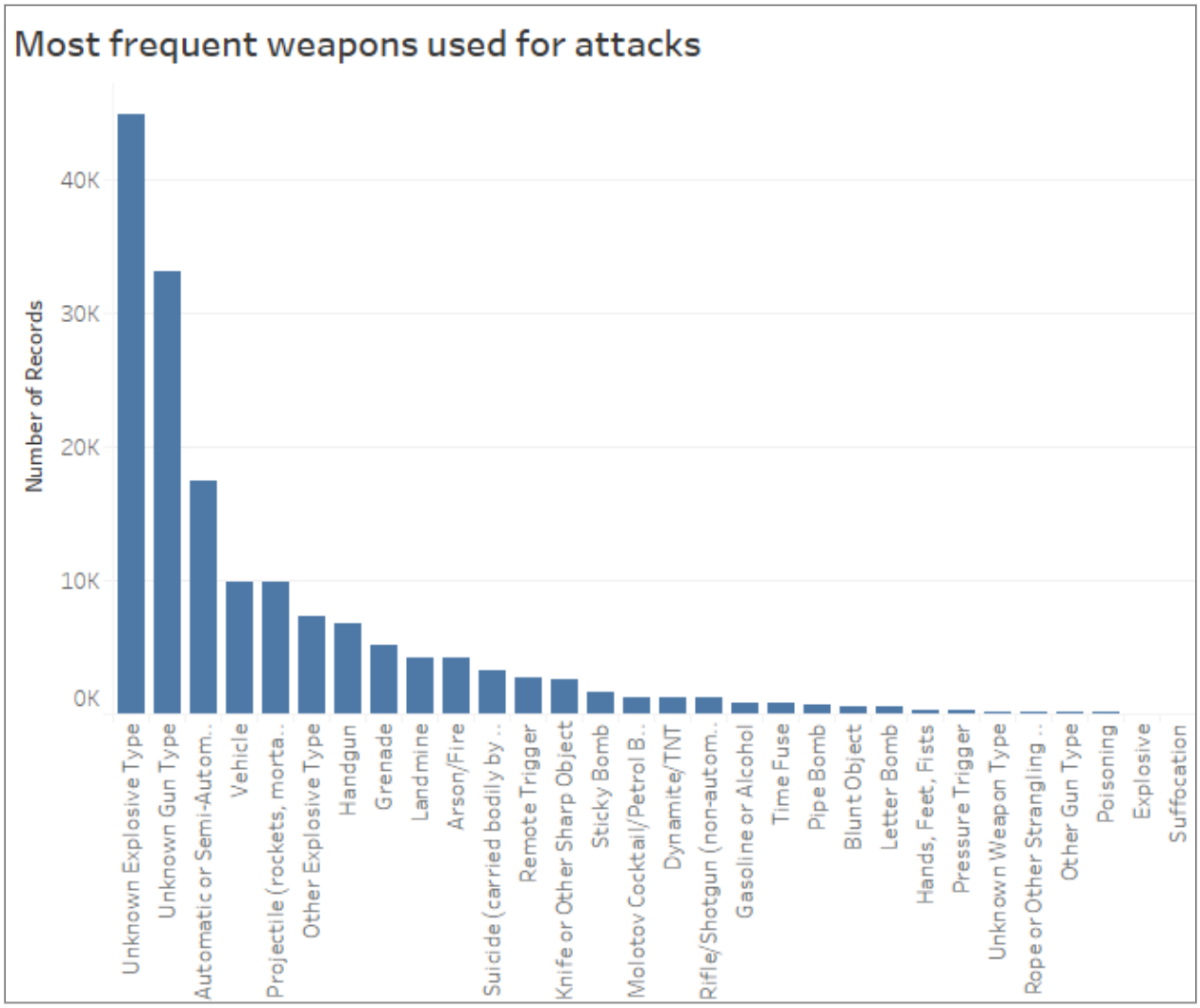


Figure11: Weapon types

VII) DEPLOYMENT

Now that we have performed analysis on this dataset, and come up with a model. This model can be used on similar datasets to find out interesting information and conclusions. An extended version of this analysis could be used by organizations which try to fight terrorism globally.

VIII) OVERALL

Handling this dataset was quite a task, in terms of cleaning it, figuring out what to focus on, from a huge list of variables but then it was very interesting and challenging. Hope this model adds value or at least help in having a better understanding.