# Global Terrorism Data Analysis

Project: AIT 614
May 13, 2018
George Mason University

*By Snigdha Cheekoty, Sai Hemanth Nirujogi*
*G01025599, G01065588*

## Introduction

The whole world was shocked and was traumatized on September 11, 2001, when al Qaeda hijacked four airplanes and attacked the World Trade Center, the Pentagon and tried to strike the U.S. Capitol.

Terrorist attacks around the world have seen a high average increase in the past decade. Terrorism has no accurate definition and defining it is getting harder, because the terrorism has become an urban term for different acts of violence which weren't considered as terrorism before.

"2017 is the second year in a row, the total number of deaths caused by terrorist attacks has been decreased. The reduction in deaths is encouraging, but 2016 was also the third deadliest year since 2000" [1]

9/11 attack was masterminded by a Pakistani, Shaykh Muhammad and Bin Laden, but this attack had eventually lead to the fall of the Al Qaeda Empire which had its base in Afghanistan, killing Osama Bin laden, which was indeed a remarkable achievement by the US Army. The terrorists were on the run and the Taliban army had gradually lost its control over Afghanistan and Pakistan.

This was one of the most famous attacks on terror itself, and has given hope to many countries and corp bodies to tackle such terrorists and their plots against innocent people.

In our project, we have studied the various terror attacks that happened all over the world through the years from 1970 to 2017. The type of attack, weapon type used, targeted agencies and groups of people, number of people injured and killed, property loss, success rate of the attacks, etc., are few of the variables that are intensively studied.

## Problem Statement

Could we identify the factors that lead to the success of the terror attacks that are happening all over the world?

In this project, we have analyzed the global terrorism database to generate summary statistics and visualizations. The variables with most effect on the incident will be determined from the analysis. We will determine the weights of attributes to understand the importance on the attack, which will in turn help us to gain insight into what determines the success of the attacks (successful or unsuccessful). Based on the study, there is a scope for future work that can strategize in general, as to what factors must be considered in defense of such hate and terror attacks. Defense study groups can get a generalized report of what factors are really dominant in these attacks, and what possible measures can be taken to avoid them from happening in the near future.

## **Data Source**

The dataset is originally collected from the Kaggle website. But the data is owned and maintained by the Global terrorism database research team. The global terrorism database is an open-source database curated with information from different open media sources with systematic data on all the terrorist incidents that have happened from 1970 through 2016. This database consists information on the terrorist attacks occurred all over the world including domestic as well as international terrorist attacks that have happened during the time period. This database is created and maintained by the National Consortium for the Study of Terrorism and Responses to Terrorism (START), located at University of Maryland.

- The dataset includes more than 170,000 cases of terrorist attacks.

- For each incident, details about the incident like date, time, location (with latitude and longitude), type of weapons used, target, number of deaths and is the incident a group or individual agenda.

- Information about more than 83,000 bombings, 11,000 kidnappings and 18,000 assassinations is included in the dataset.

- Each terrorist incident at least has information on 45 variables. [2]

## **Data Preprocessing**

Out of the 170350 records and 89 columns, there was a significant amount of missing data. (Figure 1) The attributes that had a major portion of the missing data are that related to weapon type such as weapsubtype4, weaptype4; claim- related such as claimmode3, claim3; attack form used like attcktype3, and various other variables like ransompaidus, targettype3, nhours, etc. More or less, these missing data and NA values wreak havoc on many models and must be removed, changed or imputed. We have performed methodological but limited imputation on these variables, the reason being a concern that imputing so many values might skew the data significantly and in turn skew the result, so we wanted to leave the blank values unchanged and do only required preprocessing for every model individually, whichever is best suited for them.

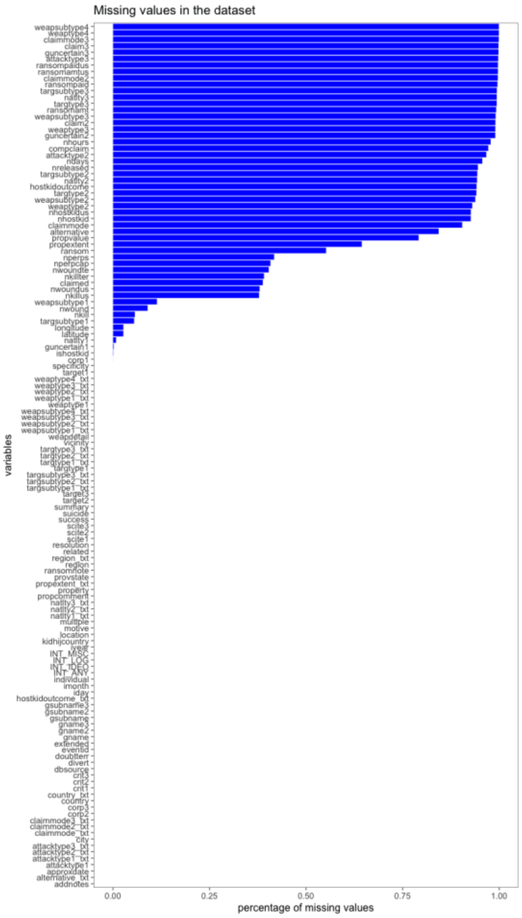


Figure 1: The variables having missing values shown in blue.

Firstly, we examined the pattern of missing values, in order to understand what method to employ for the data imputation. The plot for the pattern (Figure 2) shows the attributes individually with their missing data percentages.

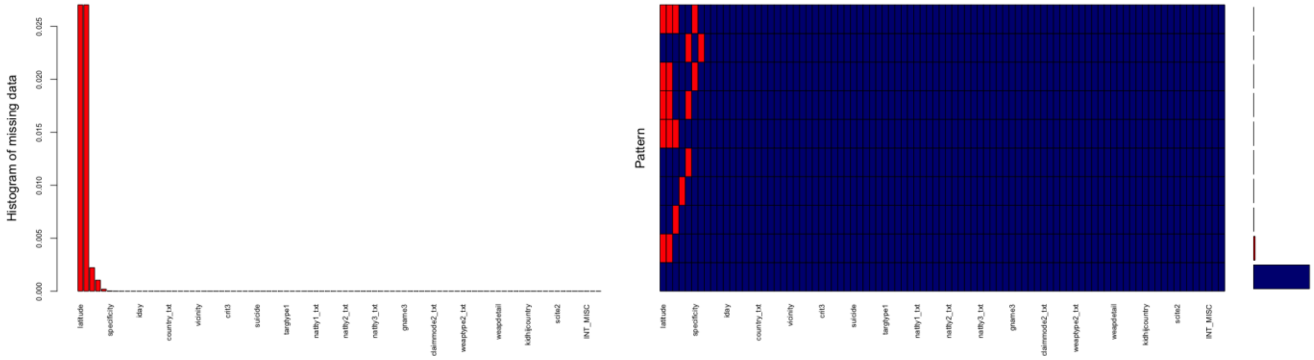
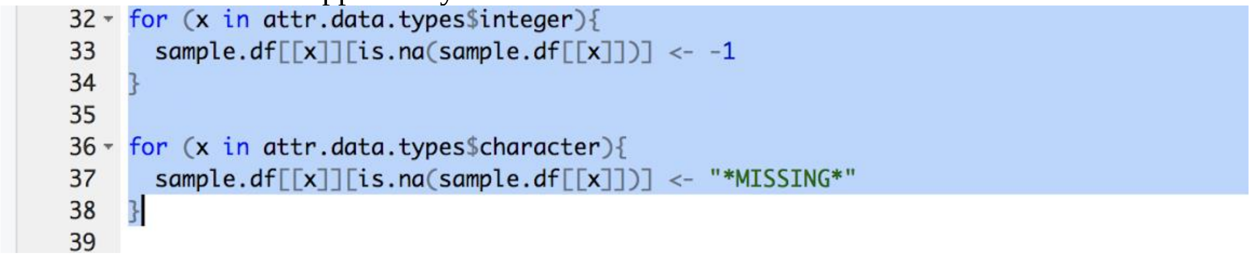


Figure 2: Pattern of missing data

The data imputation technique employed was quite simple. The original dataset was initially divided into numerical data and character data. The numeric data variables were imputed with 0 and the character data variables were imputed with the string “MISSING”, ensuring there are no null values left in the data. The variables with a very high percentage of missing data were removed, as the data can produce accurate results when fitted with different predictive and descriptive models. The threshold was set to 60% of missing data in a variable, for it to be removed from the dataset.

Once these columns were removed, a separate section of code was used to remove any rows which had blank values. The threshold to control how many values triggered a column for removal allowed us the opportunity to balance the ratio of rows:columns removal.



```
for (x in attr.data.types$integer){
  sample.df[[x]][is.na(sample.df[[x]])] <- -1
}

for (x in attr.data.types$character){
  sample.df[[x]][is.na(sample.df[[x]])] <- "*MISSING*"
}

```

Figure 3: Looping over the attributes for imputing the data for numeric and character data.

Next, we checked for near-zero variance predictors and stored them in a separate dataframe. Because removing them isn't a good idea as some of them are categorical predictors having '0' values. This dataframe of near zero variance data will be used in future for preprocessing the models that we are using.

## Correlations plot

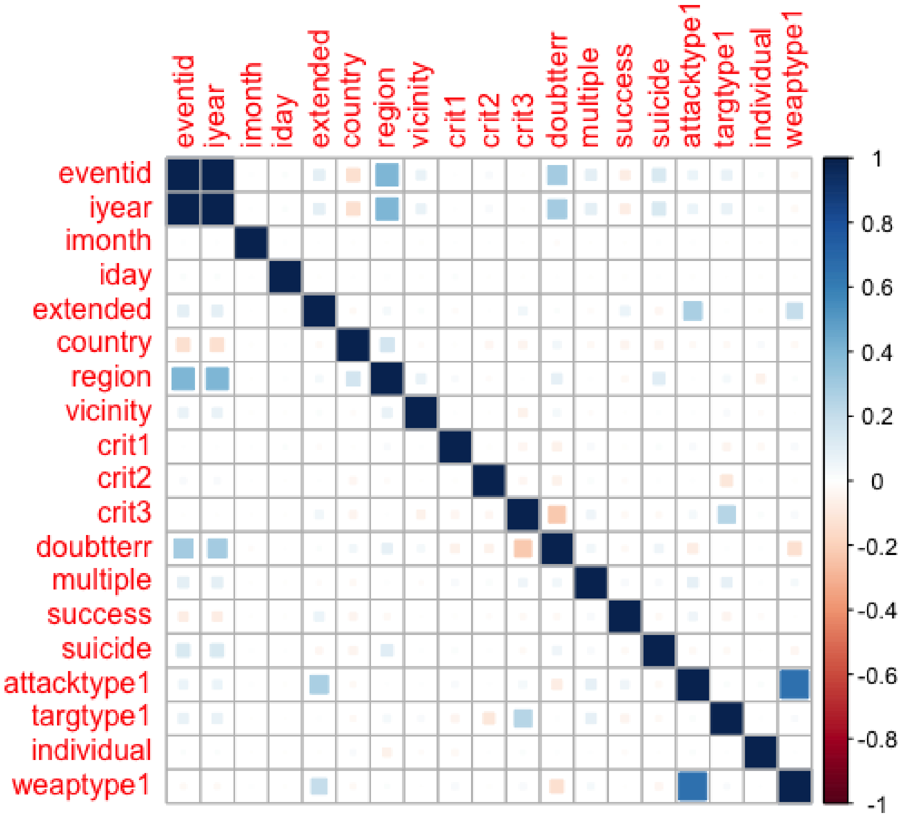


Figure 4: Correlation plot for the different predictors

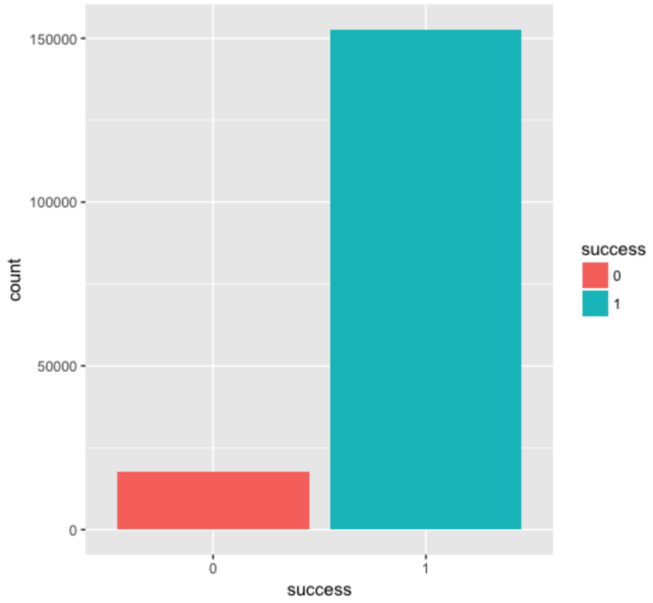
Correlations matrix is used to show the correlations between the attributes in the dataset. Correlations show the direction of the linear relationship between attributes. They indicate how one or more predictors vary with an increase or decrease in the value of other predictor(s). If there is a very high correlation between two variables in the dataset, then they must be considered for removal. The correlations affect the model performance.

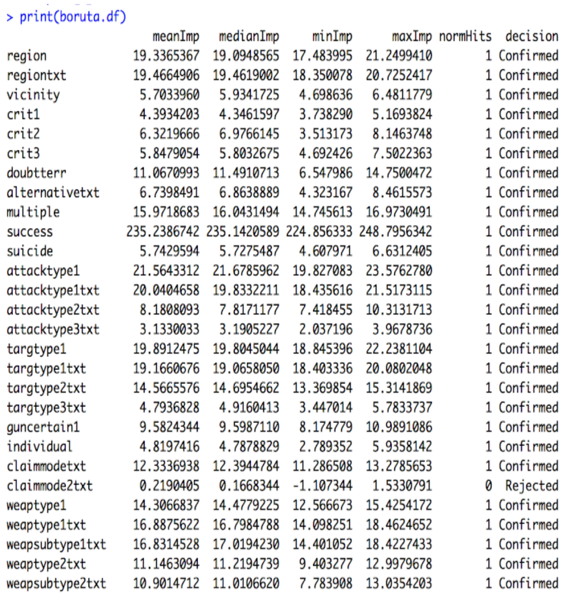
Our correlation plot (Figure 4) contains the correlations for different numeric attributes in the dataset. Almost all the attributes have very less correlation. Only weaptype1 and attacktype1 have relatively high correlation and they have bene taken care of, in the later sections.

## Feature Selection

Feature selection has been done using *Boruta analysis*. It gives the variable importance and shows whether the variable is accepted or rejected for the feature set. weaptype4txt, weapsubtype4txt, claimmode2txt are three variables that are rejected using this model. We have ignored these variables in our predictive modeling, and have just used them in data exploration. A code snippet of a small part of variable importance is shown below.

The target variable that we are considering in this dataset for all our models is the *success* variable. Figure 5 shows the count of *successful and unsuccessful terror attacks* globally. Successful terror attacks are denoted by 0 and unsuccessful by 1.





```
> print(boruta.df)
                    meanImp   medianImp     minImp      maxImp normHits  decision
region           19.3365367  19.0948565  17.483995  21.2499410        1 Confirmed
regiontxt        19.4664906  19.4619002  18.350078  20.7252417        1 Confirmed
vicinity          5.7033960   5.9341725   4.698636   6.4811779        1 Confirmed
crit1             4.3934203   4.3461597   3.738290   5.1693824        1 Confirmed
crit2             6.3219666   6.9766145   3.513173   8.1463748        1 Confirmed
crit3             5.8479054   5.8032675   4.692426   7.5022363        1 Confirmed
doubtterr        11.0670993  11.4910713   6.547986  14.7500472        1 Confirmed
alternativetxt    6.7398491   6.8638889   4.323167   8.4615573        1 Confirmed
multiple         15.9718683  16.0431494  14.745613  16.9730491        1 Confirmed
success         235.2386742 235.1420589 224.856333 248.7956342        1 Confirmed
suicide           5.7429594   5.7275487   4.607971   6.6312405        1 Confirmed
attacktype1      21.5643312  21.6785962  19.827083  23.5762780        1 Confirmed
attacktype1txt   20.0404658  19.8332211  18.435616  21.5173115        1 Confirmed
attacktype2txt    8.1808093   7.8171177   7.418455  10.3131713        1 Confirmed
attacktype3txt    3.1330033   3.1905227   2.037196   3.9678736        1 Confirmed
targtype1        19.8912475  19.8045044  18.845396  22.2381104        1 Confirmed
targtype1txt     19.1660676  19.0658050  18.403336  20.0802048        1 Confirmed
targtype2txt     14.5665576  14.6954662  13.369854  15.3141869        1 Confirmed
targtype3txt      4.7936828   4.9160413   3.447014   5.7833737        1 Confirmed
guncertain1       9.5824344   9.5987110   8.174779  10.9891086        1 Confirmed
individual        4.8197416   4.7878829   2.789352   5.9358142        1 Confirmed
claimmodetxt     12.3336938  12.3944784  11.286508  13.2785653        1 Confirmed
claimmode2txt     0.2190405   0.1668344  -1.107344   1.5330791        0  Rejected
weaptype1        14.3066837  14.4779225  12.566673  15.4254172        1 Confirmed
weaptype1txt     16.8875622  16.7984788  14.098251  18.4624652        1 Confirmed
weapsubtype1txt  16.8314528  17.0194230  14.401052  18.4227433        1 Confirmed
weaptype2txt     11.1463094  11.2194739   9.403277  12.9979678        1 Confirmed
weapsubtype2txt  10.9014712  11.0106620   7.783908  13.0354203        1 Confirmed
```

Figure 5: Count of the Successful and Unsuccessful attacks/ Boruta Analysis- variable importance

## Exploratory data analysis

Data exploration helped us to gain our initial insight into the data and focus our model fitting to the attributes that are important. We have made few observations through the exploratory analysis, which are discussed in this section.

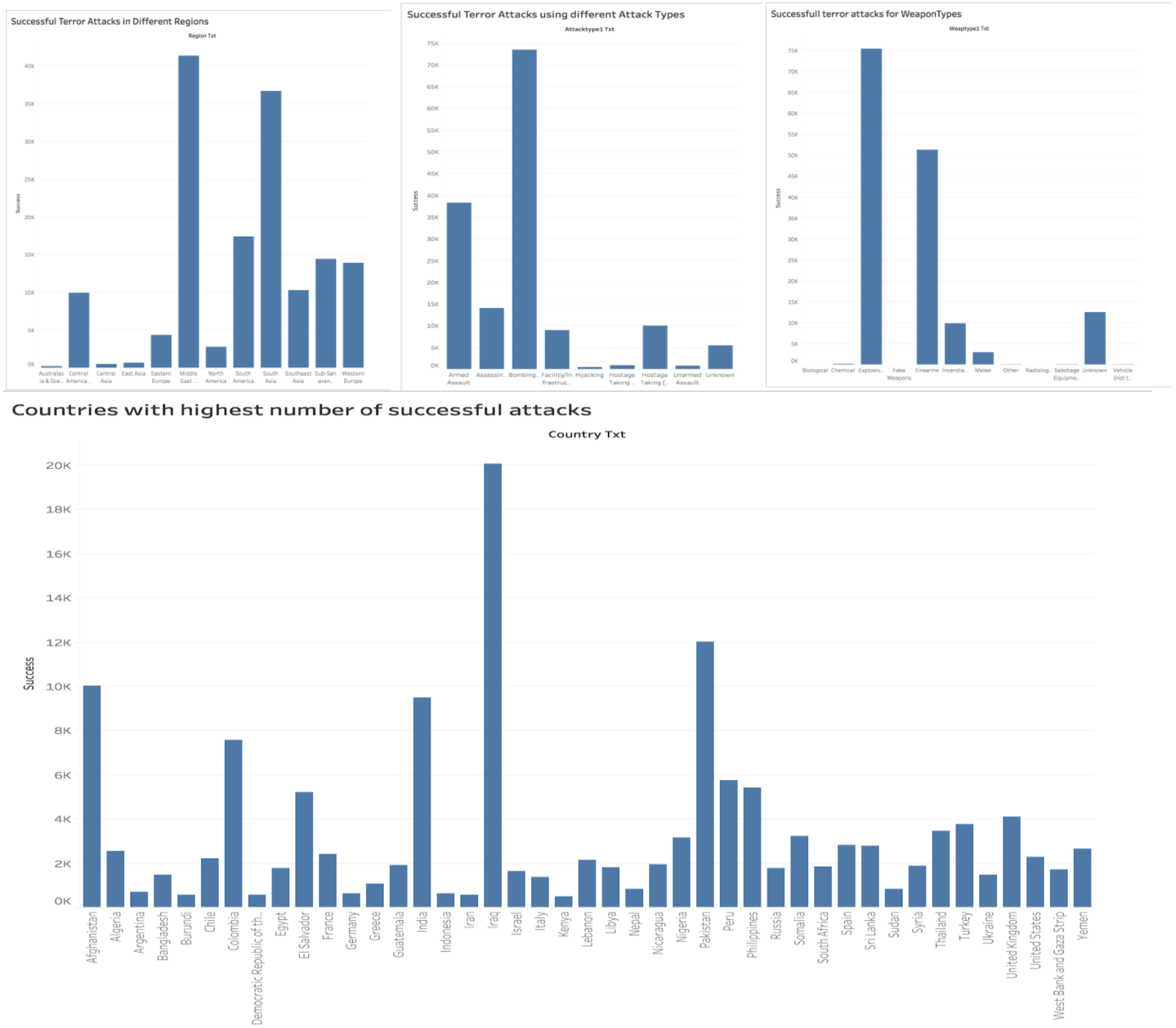


Figure 6: Shows graphs for terror attack count based on regions, attack types, weapon types, countries (left –right)

Starting from the left, the first bar graph shows the frequency of terror attacks region-wise. The highest terror attacks happened in the middle-eastern region between 1970 and 2016. The second highest terror attacks rate is seen in the South Asian region.

The second graph shows the frequency of the terror attacks using different attack-types. The highest attacks are made through bombing explosions and the second highest being armed assaults.

The third graph shows the frequency of the terror attacks made using different weapons. The highest being explosive bombs and the second highest using firearms.

The fourth graph shows the terror attacks country-wise. Iraq has the highest number of successful attacks, next highest being Pakistan.

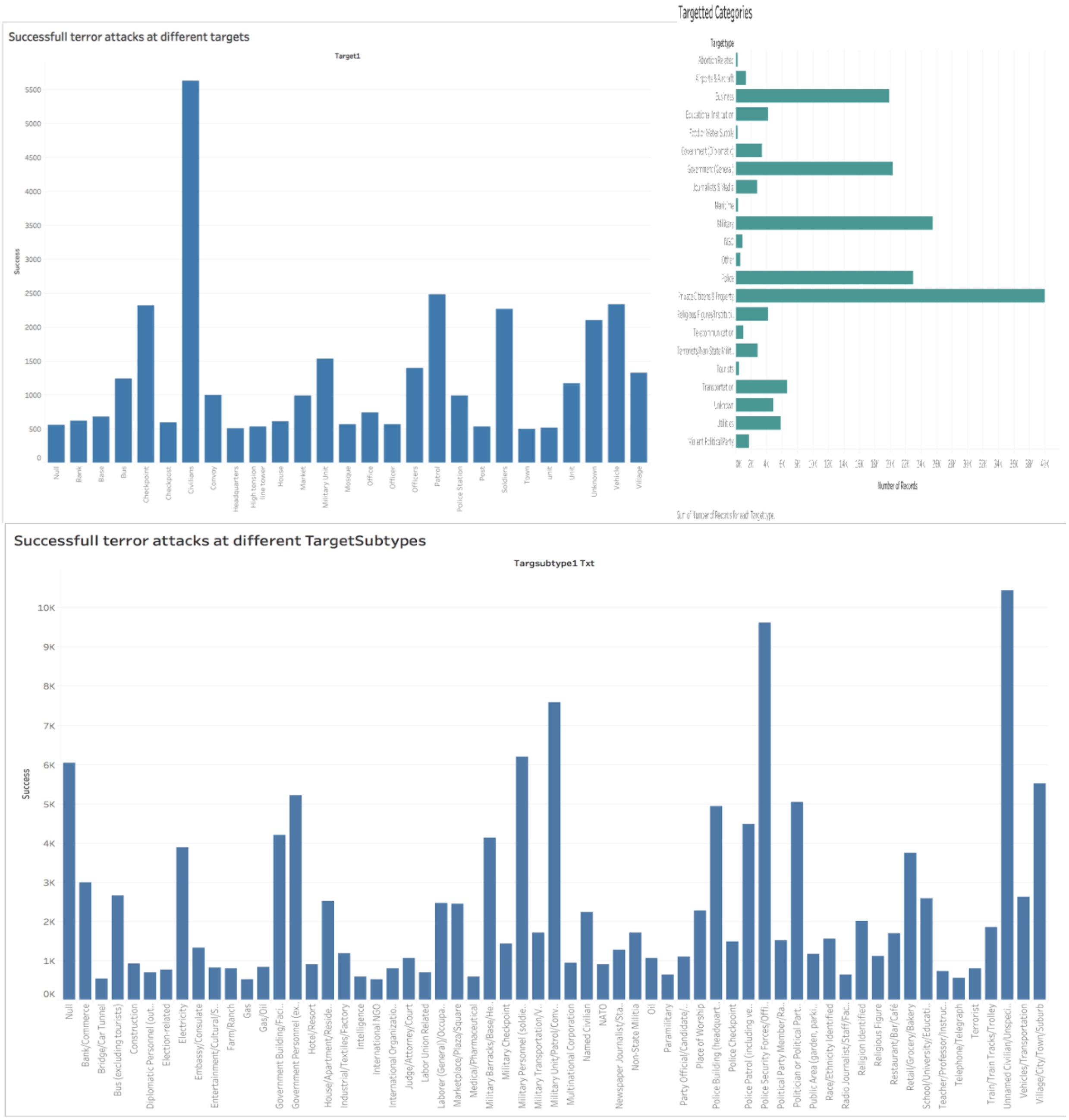


Figure 7: Success rate at different targets- targettype2 and targettype3, and sub target types

Starting from the left, the first plot shows the terror attack success rate at different groups of people like police officers, petrol officers, military, bankers, etc. The most targeted groups are the civilians. The second plot shows the number of terror attacks on another category of targeted groups. The highest in this case are the private citizens' properties. The third plot shows the terror rate count for different target sub-types- government buildings, intelligence, police departments, NATO, etc. The highest is seen for civilian property, police security forces, military units, etc.

At each target type, the granularity of the types is increasing, with target sub-type being the lowest and target type1 being the highest levels.

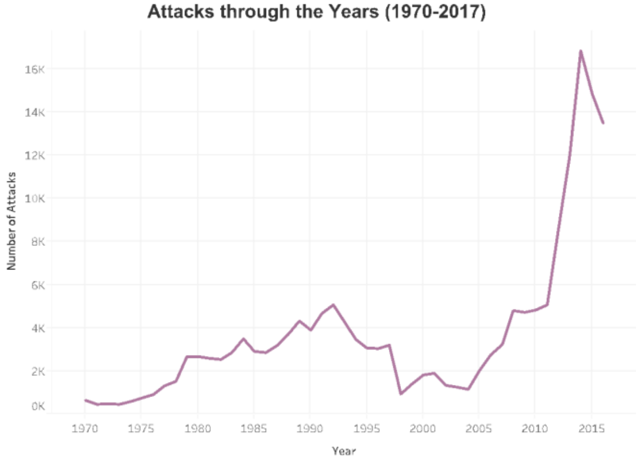
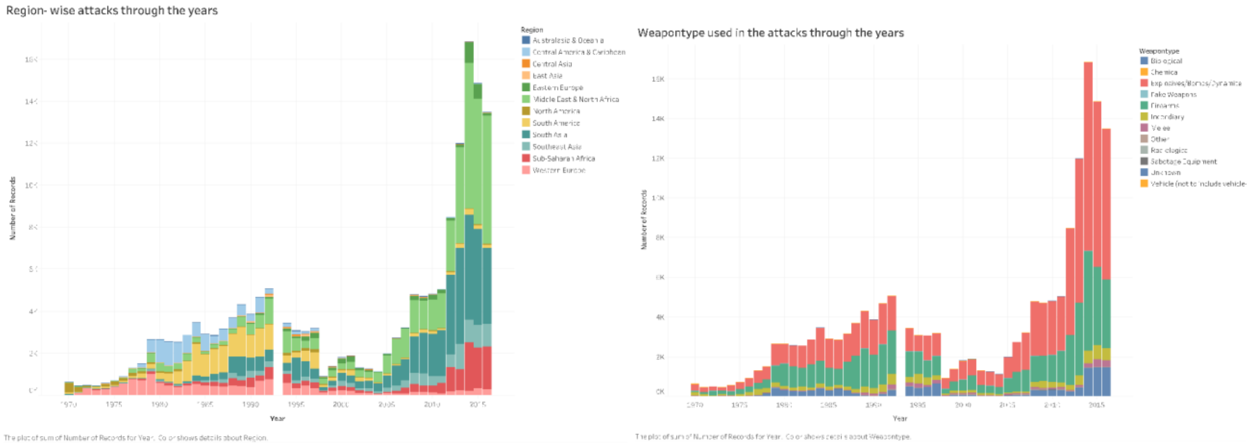


Figure 8: Attacks through the years from 1970- 2017

We can see three major phases of terror attacks through the years between 1970 to 2017. The attacks have increased since 1970 till 1993 and then there is a sudden fall in the number of attacks through the years 1994 to 1996 and the period from then on shows a sudden a rise.



The plot of sum of Number of Records for Year. Color shows details about Region.

The plot of sum of Number of Records for Year. Color shows details about Weapontype.

Figure 9: The trends throughout the years for different regions and weapon type used

The above figure shows the trends throughout the years between 1970 and 2017 for different regions(left) and weapon-type used(right). These trends are shown in form of *stacked bar plots* with each color representing different category of the variable (region/weapon type)

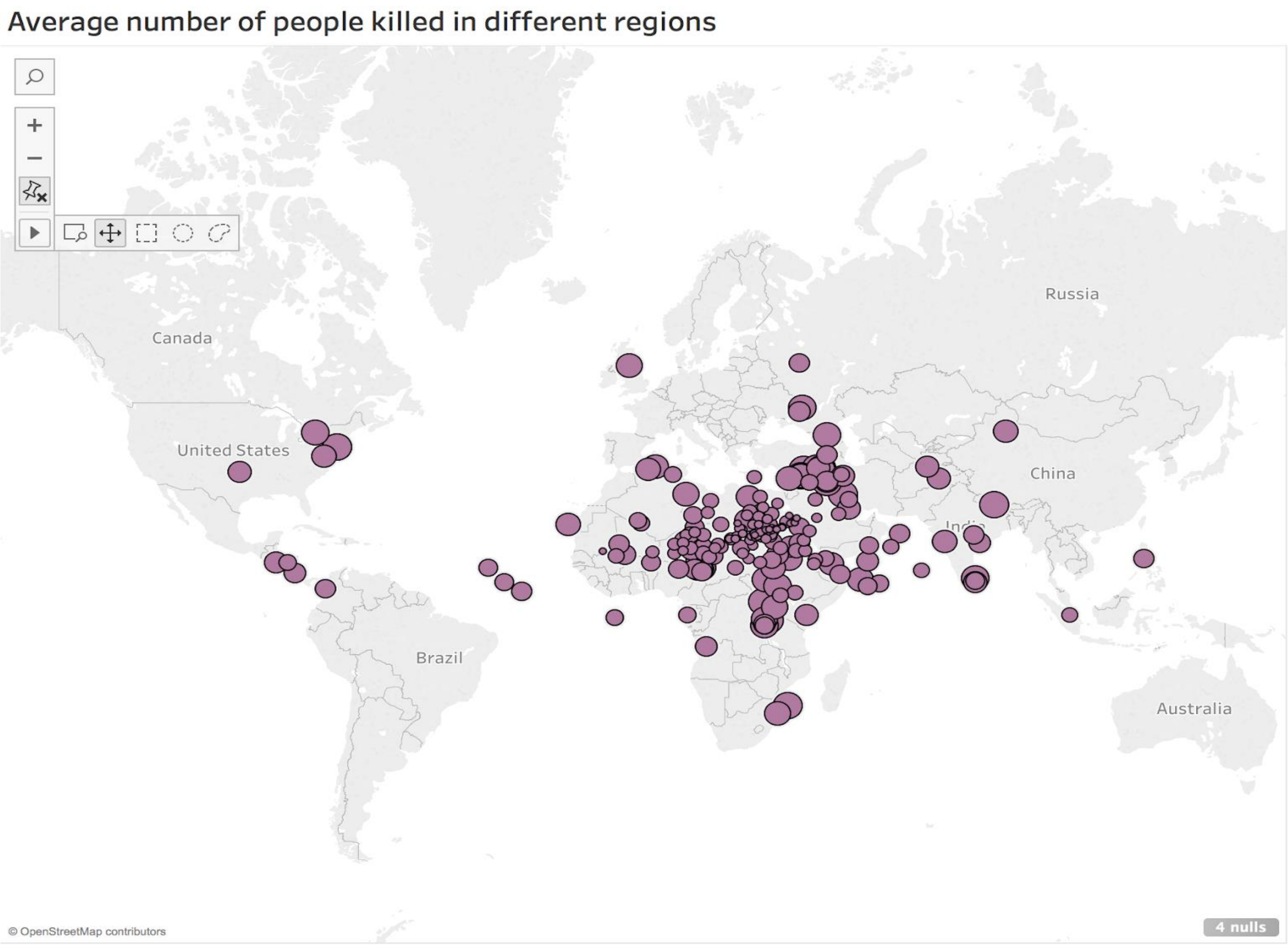

Average number of people killed in different regions
Russia
Canada
United States
China
India
Brazil
Australia
© OpenStreetMap contributors
4 nulls

Figure 10: The distribution of attacks throughout the world.

The average distribution of attacks throughout the world for all the years between 1970 and 2017. The data points are shown increasing in size with an increase in the number of attacks at every region. Figure 11 shows the distribution of the highest killings globally. The color scheme used is orange for the least and brown for the highest in that range.

The stretch along the east coast of Mexico, Arab Countries, Nepal border, Central Africa, Southern Europe show a huge number of attacks. The collection of data points on the Atlantic Ocean show the hijacks of airplanes in those regions.

Figure 12, on the next page, shows the terror attacks that have taken place in the different parts of Iraq, Pakistan and Afghanistan. These three countries show an *overall highest terror rate* in the entire world. There are reports of continuous violence towards targeted groups in these regions even today. World’s most terrifying militant groups, the Al Qaeda and the Taliban had once made these regions as their base for carrying out their operations to attack different countries throughout the world.

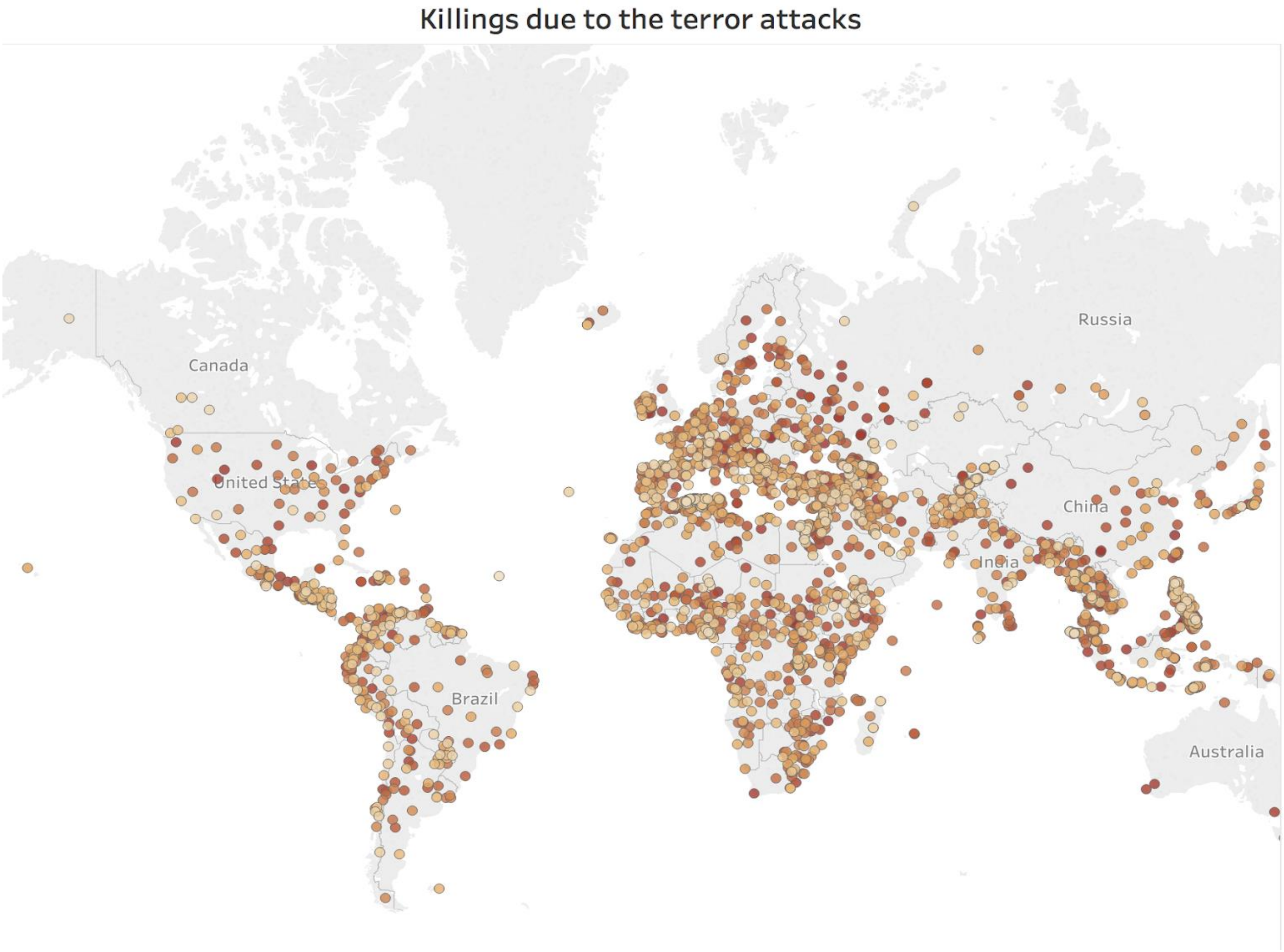


Figure 11: Killings due to terror attacks

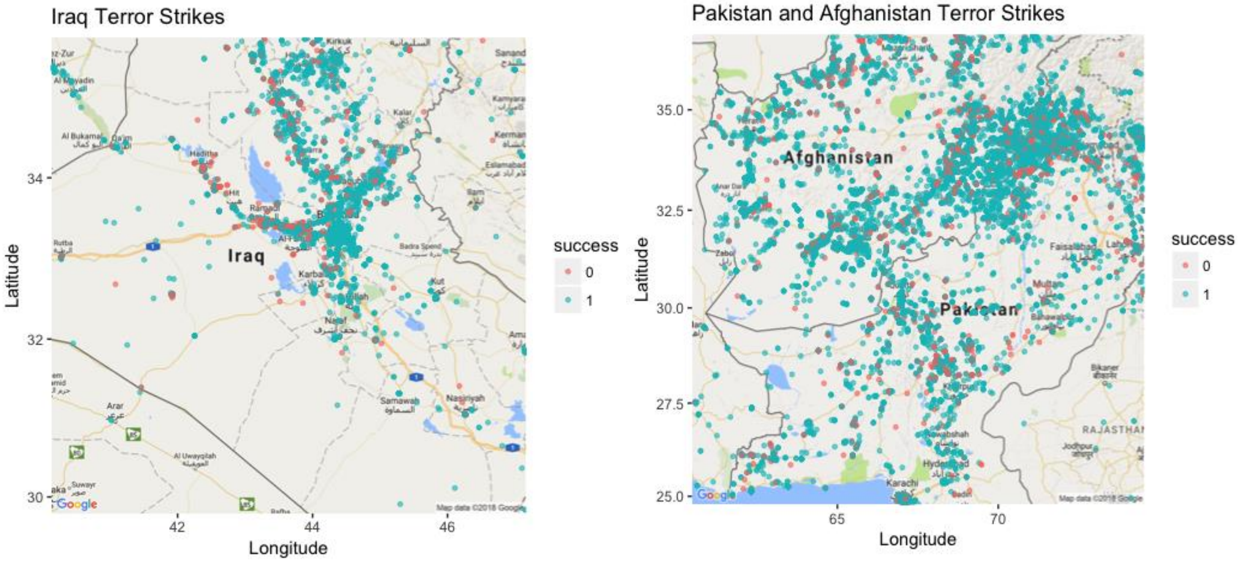


Figure 12: Terror attacks location-wise in Iraq, Pakistan and Afghanistan

## Model Building

We studied our data by fitting six different models and compared their performance using ROC curve. We have taken the *success* variable as the target variable for our analysis.
The models that we have used for our analytics are the Random Forests, Classification Trees, Logistic Regression, Support Vector Machine, Multi-Layer Perceptron(Neural Network), AdaBoost.

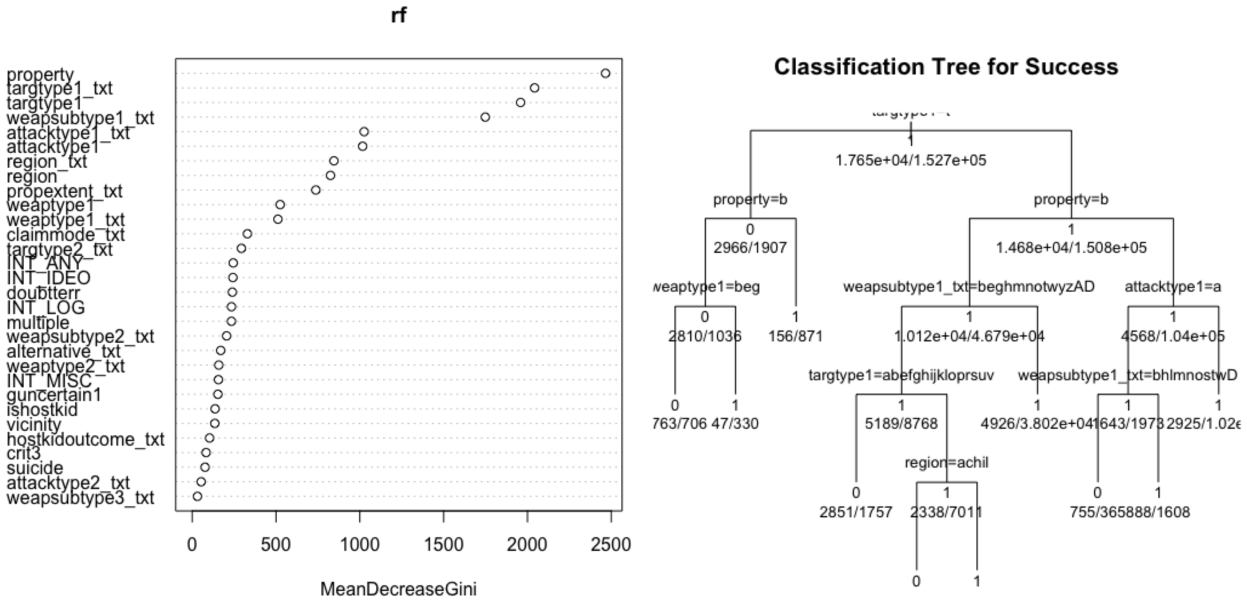
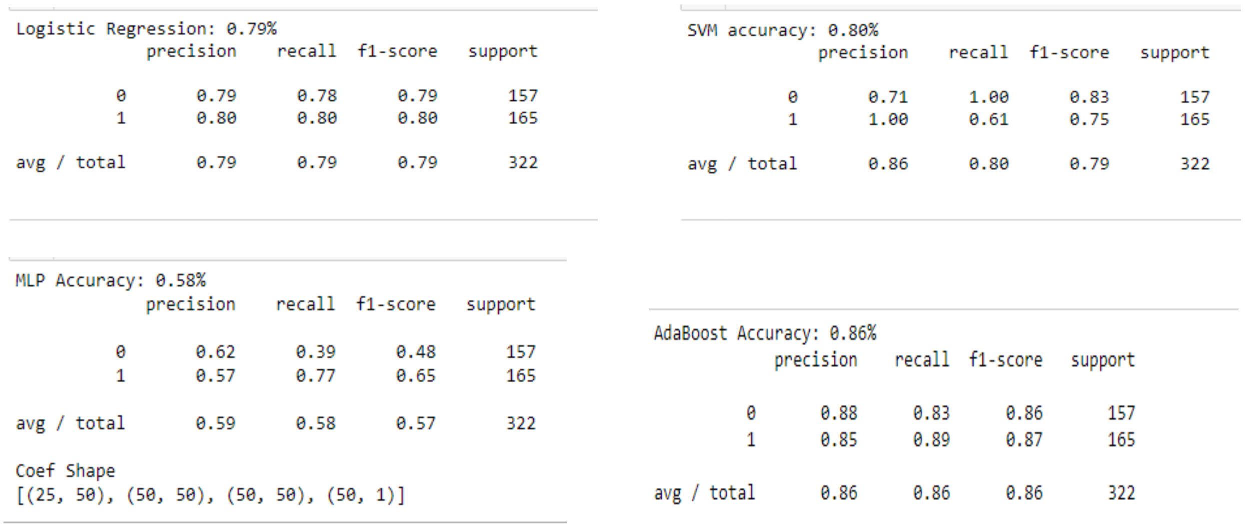


Figure 13: Variable Importance plot and Classification Decision Tree



```
Logistic Regression: 0.79%
                 precision    recall  f1-score   support

          0         0.79      0.78      0.79       157
          1         0.80      0.80      0.80       165

avg / total         0.79      0.79      0.79       322
```

```
SVM accuracy: 0.80%
                 precision    recall  f1-score   support

          0         0.71      1.00      0.83       157
          1         1.00      0.61      0.75       165

avg / total         0.86      0.80      0.79       322
```

```
MLP Accuracy: 0.58%
                 precision    recall  f1-score   support

          0         0.62      0.39      0.48       157
          1         0.57      0.77      0.65       165

avg / total         0.59      0.58      0.57       322

Coef Shape
[(25, 50), (50, 50), (50, 50), (50, 1)]
```

```
AdaBoost Accuracy: 0.86%
             precision    recall  f1-score   support

          0       0.88      0.83      0.86       157
          1       0.85      0.89      0.87       165

avg / total       0.86      0.86      0.86       322
```

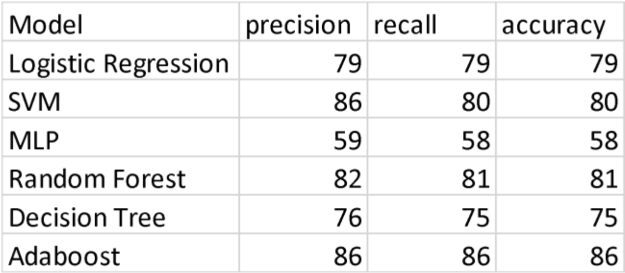
Figure 14: Results of Predictive Modeling for the data

The results shown in the Figure 13 and Figure 14, are all based on the predictive models used. Using the various models, we have created the variable importance plots, but since the random forests was amongst the ones that showed the highest accuracies, we have only displayed the variable importance plot created using Random Forests. The variables that have the highest weights are property, targettype1, weapon_subtype, attacktype, region, weaptype, targtype2,etc in the decreasing order of importance. We have used the purity metric to calculate the weights of the variables; specifically the *gini index*.

The metrics used in evaluating the performance of the models are precision, recall, f1-score and support.

## Model Performance Comparative Study/Results:

We have shown the precision, recall and accuracy values for the six models in the Figure 14. Also, we have employed the method of ROC- AUC metrics to measure the model performance. The best performing models are Random Forests and AdaBoost with AUC of 83 and 86 percentages respectively. Both of these models have been optimized for the overfitting-underfitting problem, popularly known as the variance-bias tradeoff.



| Model | precision | recall | accuracy |
|---|---|---|---|
| Logistic Regression | 79 | 79 | 79 |
| SVM | 86 | 80 | 80 |
| MLP | 59 | 58 | 58 |
| Random Forest | 82 | 81 | 81 |
| Decision Tree | 76 | 75 | 75 |
| Adaboost | 86 | 86 | 86 |

Figure 14: Comparison of the different models used

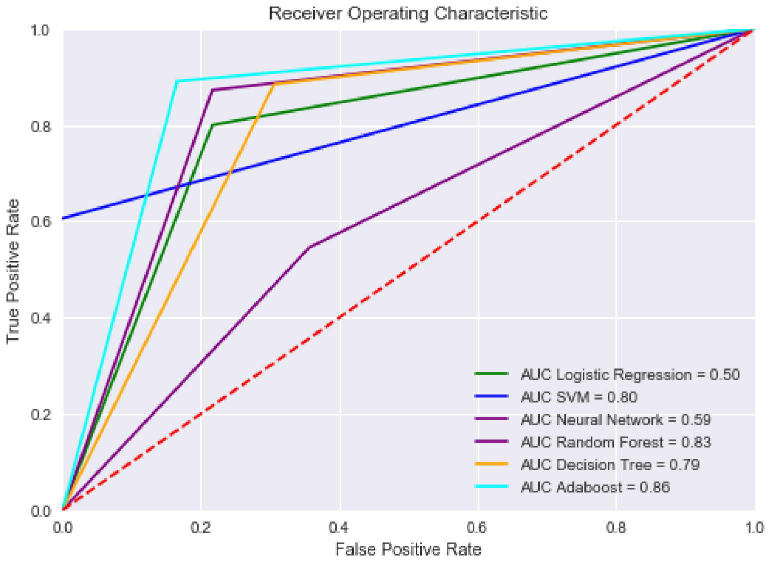


Figure 15: ROC curve for all the six models

Figure 15 shows the ROC (Receiver Operating characteristic) curve for all the models, created by plotting the true positive rates and false positive rates for each model at different thresholds. The comparison can be made using the AUC (Area under the curve) for the different models. The best performing model with the highest AUC value of 86% is the AdaBoost model.

## Conclusion

Our analysis shows that factors like property, target type, weapon type are the most important variables to be considered in determining the success of the incident. Region factor that includes the details of the region where the incident occurred also plays a significant role. Whether the attack was by an international group, attacker or perpetuator, and whether the attack took place at multiple places at the same time is also considered one of the main factors for determining the success of the incident. The variable for the incident occurring in the vicinity of the city also has a high weight of important along with the variable of whether the attack is a suicide attack or not.

Based on these important variables, one can determine to a certain extent of whether the attack is successful or not. This analysis can be put to future use by research studies to go in-depth work about what measures can be taken in order to stop or reduce these types of attacks.

## References

[1] Global Terrorism Index 2017", *ReliefWeb*, 2018. [Online]. Available: https://reliefweb.int/report/world/global-terrorism-index-2017. [Accessed: 13- May- 2018].

[2] "Global Terrorism Database | Kaggle", *Kaggle.com*, 2018. [Online]. Available: https://www.kaggle.com/START-UMD/gtd. [Accessed: 13- May- 2018].

## Appendix

## Few important variables description:

1. iyear – Year of the incident
2. imonth – Month of the incident
3. Iday – Day of the incident
4. approxdate - Approximate Date
5. country – Country number of occurrence
6. country_txt- Country label
7. region – Region number of occurrence
8. Region_txt – Region label
9. provstate – Specific Province
10. vicinity - Vicinity
11. specificity - Geocoding Specificity
12. attacktype - Attack Type
13. success -Successful Attack
14. weapsubtype1- Weapon Sub-type
15. weaptype1 - Weapon Type
16. weapdetail- Weapon Details
17. targtype1 - Target/Victim Type
18. targsubtype1 - Target Sub-Type
19. killnumb - No. of. Deaths
20. summary - Incident Summary