Final Report: Terrorism Prediction and Classification using CRISP-DM

Submitted to:

Prof. Andrew McCaren
School of Computing
Dublin City University

Report Prepared By:

Bharathvaj Devarajan
16212388 bharathvaj.devarajan2@mail.dcu.ie
Juhi Shrivastava
16212548 juhi.shrivastava2@mail.dcu.ie
Manikandan Swaminathan
16212801 manikandan.swaminathan2@mail.dcu.ie
Nitin Yadav
16212304 nitin.yadav2@mail.dcu.ie
Subject: Data Mining, CA683
April 19, 2017

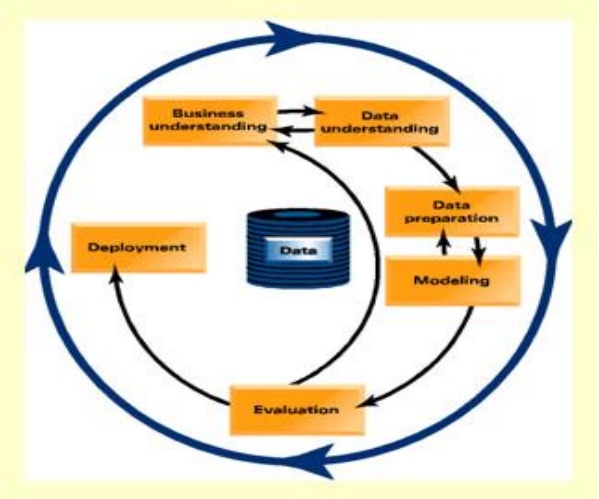
Declaration:

In submitting this project, we declare that the project material, which we now submit, is our own work. We make this declaration in the knowledge that a breach of the rules pertaining to project submission may carry serious consequences. We also regard that both of us have contributed equally to the work presented.

Introduction

This report aims to present an analysis done on the Global Terrorism Dataset to understand the trends and use them to avert future attacks. The analysis is done using the CRISP-DM lifecycle for ensuring a systematic approach in dealing this problem. The report is divided into 6 steps:

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

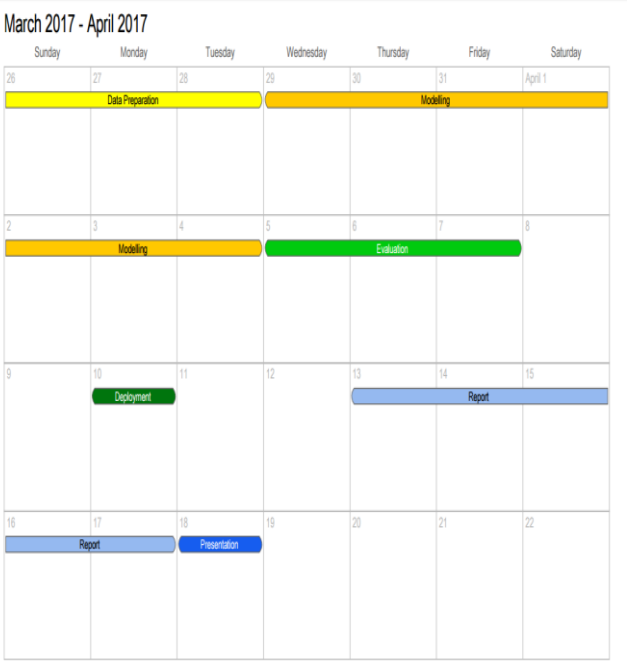


Project Plan:

A Gantt chart was created to fix the timeline of various activities involved in building the model.

Technologies Used: R and Python

Code Source:
A github repository was set up with a branch each for every team member.
https://github.com/Bharath25191/Data_Mining_DCU
Each team member pushed their code to their respective branch.



1. Business Understanding

1.1 Objective

The object of this project is to mine the terrorism dataset to explore patterns and predict terrorist activities to minimize loss of life in wake of such barbaric events.

1.2 Description

The dataset contains the record of all terrorist attacks which have occurred around the world from 1970 to 2015. The dataset includes more than 150,000 instances of both domestic and international incidents. It has more than 100 features on location, perpetrators, group responsible, targets, kills, outcome and regions. This dataset is maintained by National Consortium for the Study of Terrorism and Responses to Terrorism (START).

Dataset Characteristics:

Number of Records: 150667

Number of Columns: 137

1.3 Prediction Tasks

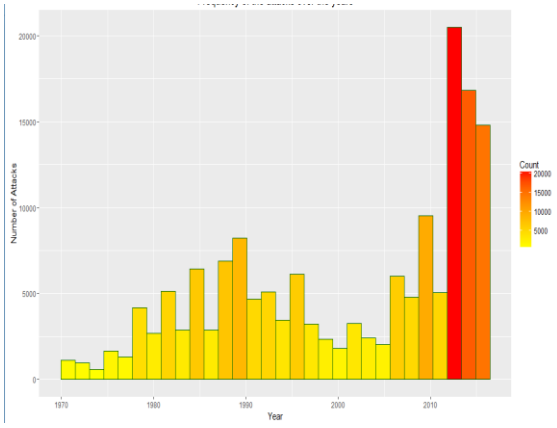
1. Predict the number of kills/scale of a terror attack based on the attack characteristics.
2. Predicting when a terrorist attack is successful (leading to structural damage or loss of life).
3. Predicting the region of the attack.
4. Identify the possible terrorist group responsible for the attack.
5. Can we predict future terrorist activities?

2 Data Understanding

2 Data Understanding

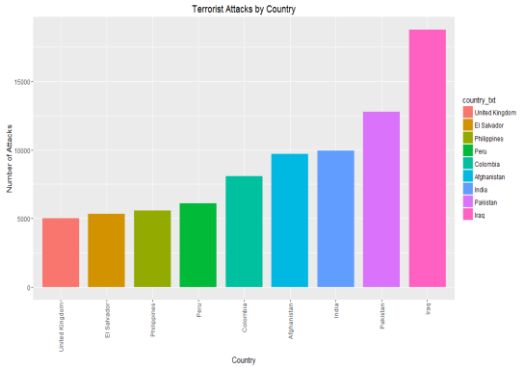
To understand the data, the dataset was uploaded as a dataframe in R studio. As there were lot many features to deal with, the exploratory analysis was done through Data Visualization.
The following 4 graphs explore the dataset in terms of questions posed and create the basis for the question – Predicting the kills?

2.1 Frequency of the attacks over the years?



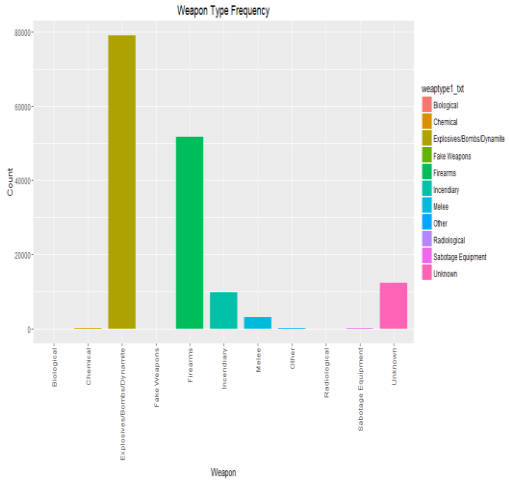
The above graph indicates that the frequency of the attacks has increased in the recent years, starting from 2010.

2.2 Country of attack?



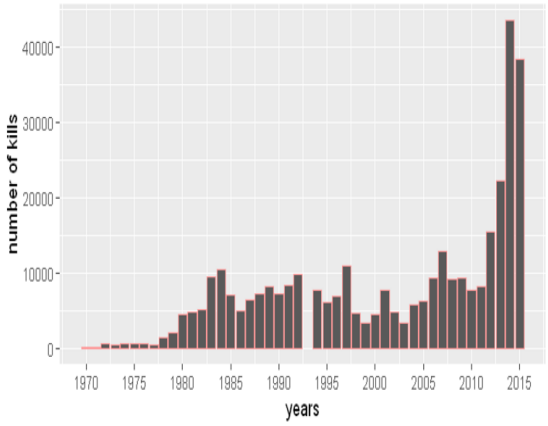
The countries in which the terrorists activities are prevalent are shown in the plot above and Iraq is the Main country where most of the terrorists activities have taken place.

2.3 Weapon preference?



The above bar plot reveals that the most common choice of weapons of the perpetrators is Explosives\Bombs dynamites.

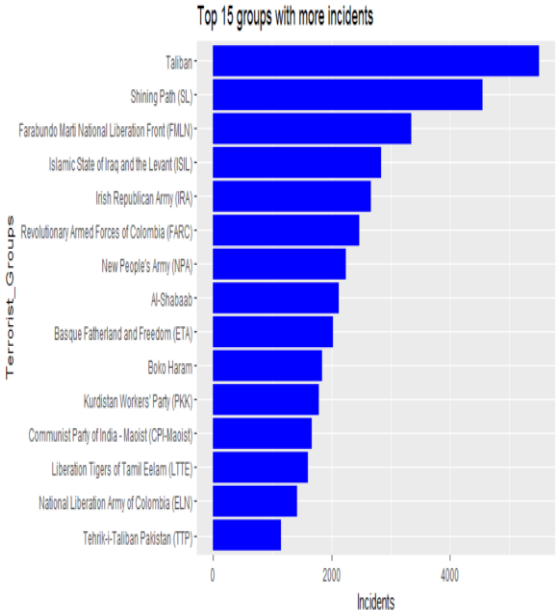
2.4 Number of kills over the year?



From the above plot it is clear that the no. of killing have increased tremendously from 2013.

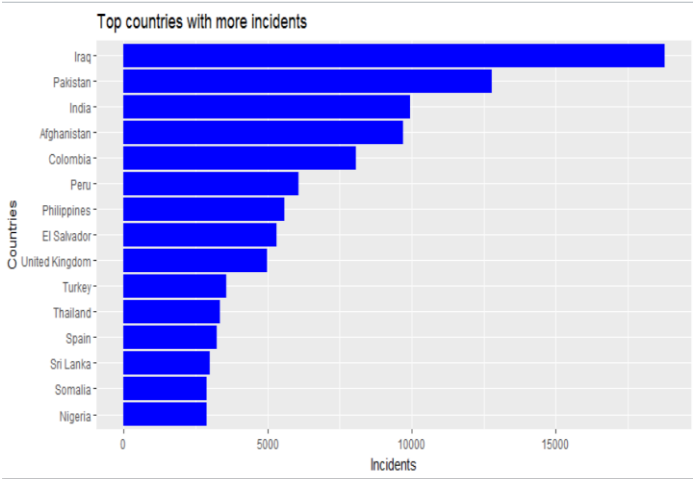
Now to describe and better understand the terrorist perpetrators the following questions were explored with the help of data visualizations:

2.5 Who are they?



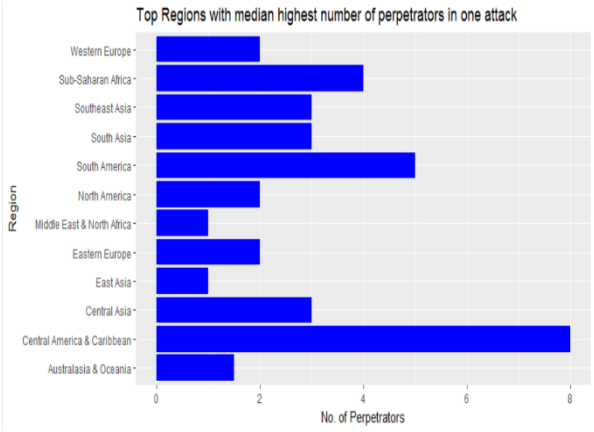
Taliban and Shining Path (Communist Party of Peru) are the most noticeable groups, both with more than 4000 incidents confirmed.

2.6 Where are they from?



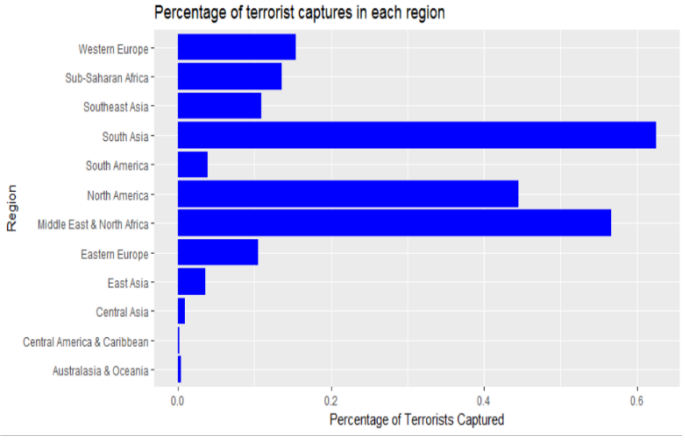
The above graph explicitly depicts that Iraq and Pakistan are the countries with most terrorists activities.

2.7 Do they work in groups? Are they working alone?



Central America and Caribbean dominates when it comes to perpetrators working in groups.

2.8 Are they captured? Does that depend on the region?



In terms of capture rates, South Asia and Middle East & North America have the higher values (> 0.7). On the other hand, East Asia, Central Asia, have the lowest values (< 0.2)

Clustering was performed based on different parameters to better understand the if the data can be sub grouped into clusters giving some insightful information. K Means clustering algorithm was used to check if the data can be segregated into the clusters.

The raw data was processed using the 1-hot-encoding of the categorical variables viz:

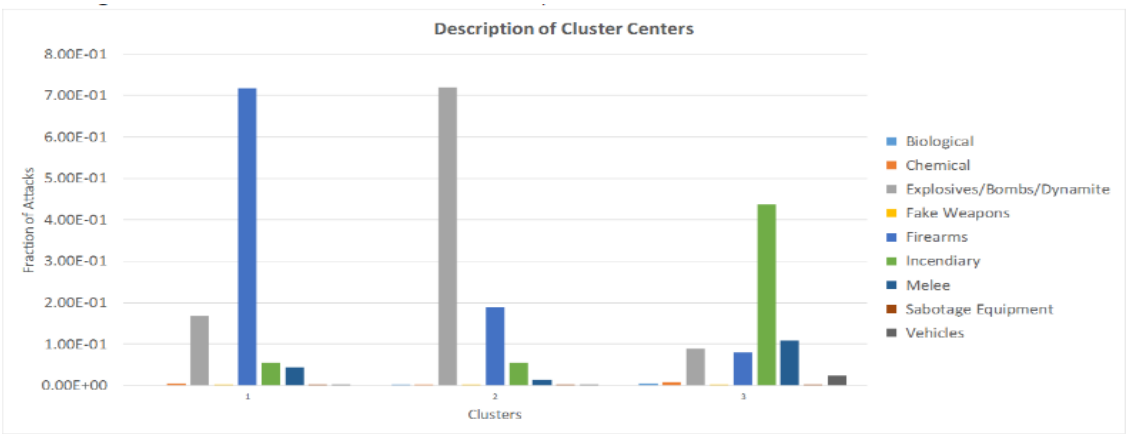
Weapon type

Attack type

And a profile was created for each group based on the frequency of each type.

The clustering algorithm was applied to the group profiles with different values of k, to obtain meaningful cluster centres. The results are presented below.

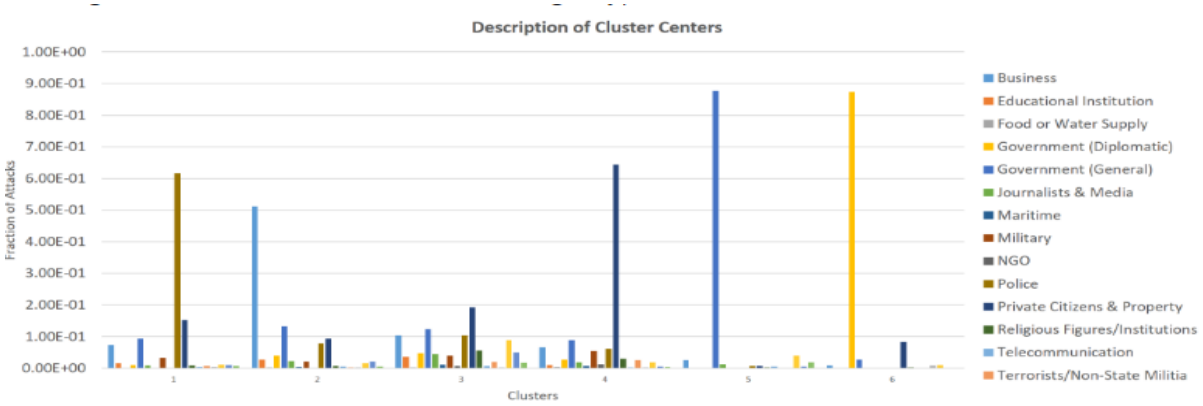
Category 1: Based on the Weapons Type:



Three clusters were formed as is evident from the above figure.

1. Firearm users
2. Explosives users
3. Incendiary users

Category 2: Based on the Attack type



The clusters formed are:

1. Bombers
2. Ars assaulters
3. Mixed
4. Infrastructure/Facility Attackers

3 Data Preparation

3.1 Converting Categorical values to Numerical attributes

As the models we would be using works better with numerical data, we needed to convert categorical attributes to numerical. Also, Principal Component Analysis can only be done with numerical attributes. The process is rather straight forward in R by using the function as.numeric().

3.2 Handling Missing Values

This dataset is rather sparse with a lot of missing values in most columns. It was not possible to impute every missing column as some of these columns were not missing at random (NMAR).
For the columns where imputation was possible, the values were imputed using Mean imputation. Imputation are essential to improve the predictive power of an algorithm.

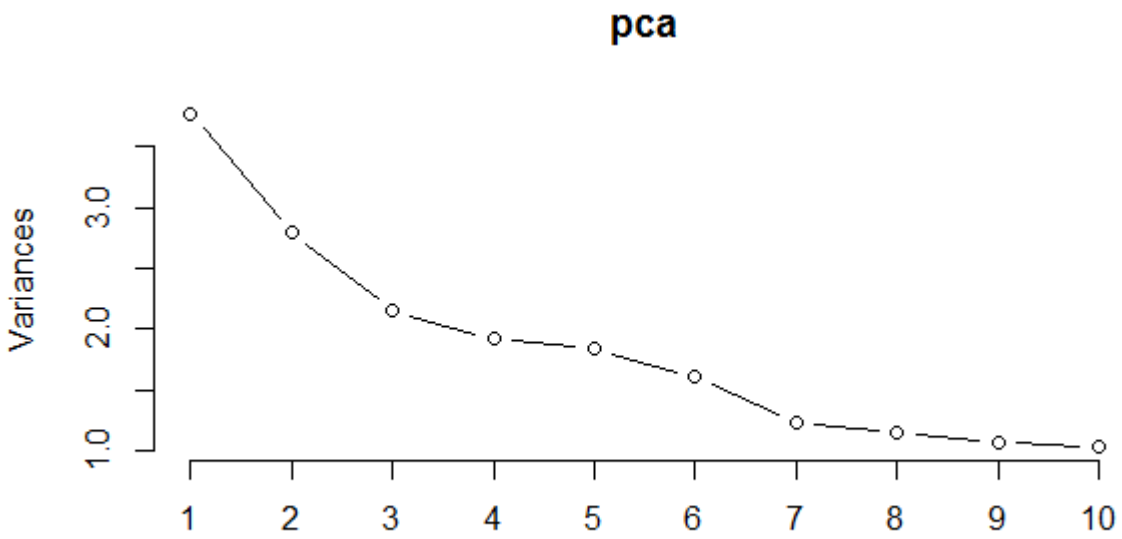
3.3 Principal Component Analysis

Principal Component analysis is a popular technique used to do dimensionality reduction of the features. Since in our case more than 130 features were there, it won't make intuitive sense to feed all the features inside the model as this may lead to a poorly fitted model. With PCA, we essentially want to capture the maximum variance explained without losing out on too much information.
Advantages of PCA:

1. Better Visualization
2. Space Efficiency
3. Computational Efficiency

PCA is usually calculated by computing the covariance vector using which the Eigen values and Eigen vectors are derived, these represent the magnitude and direction of the principal components.

Result of PCA in Terrorism Dataset:



This visualization depicts the share of variance captured by each principal component. A total of 35 principal components got generated

```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9
Standard deviation     1.9412 1.6722 1.46525 1.38514 1.35373 1.26672 1.10732 1.07446 1.03256
Proportion of Variance 0.1077 0.0799 0.06134 0.05482 0.05236 0.04585 0.03503 0.03298 0.03046
Cumulative Proportion  0.1077 0.1876 0.24891 0.30372 0.35608 0.40193 0.43696 0.46995 0.50041
                          PC10    PC11    PC12    PC13    PC14    PC15    PC16    PC17    PC18
Standard deviation     1.01666 1.00898 1.00516 1.00109 0.99499 0.99143 0.98948 0.98202 0.95815
Proportion of Variance 0.02953 0.02909 0.02887 0.02863 0.02829 0.02808 0.02797 0.02755 0.02623
Cumulative Proportion  0.52994 0.55903 0.58789 0.61653 0.64481 0.67290 0.70087 0.72842 0.75465
                          PC19    PC20    PC21   PC22    PC23    PC24    PC25    PC26   PC27
Standard deviation     0.94623 0.94205 0.93605 0.9183 0.90177 0.85354 0.80074 0.75705 0.7341
Proportion of Variance 0.02558 0.02536 0.02503 0.0241 0.02323 0.02081 0.01832 0.01637 0.0154
Cumulative Proportion  0.78024 0.80559 0.83063 0.8547 0.87796 0.89877 0.91709 0.93346 0.9489
                          PC28    PC29    PC30    PC31    PC32    PC33    PC34      PC35
Standard deviation     0.64204 0.63854 0.60025 0.58453 0.38591 0.34089 0.05269 0.0008947
Proportion of Variance 0.01178 0.01165 0.01029 0.00976 0.00426 0.00332 0.00008 0.0000000
Cumulative Proportion  0.96064 0.97229 0.98258 0.99235 0.99660 0.99992 1.00000 1.0000000
```

3.4 Feature Engineering

Feature Engineering involves the process of extracting new features from the available features. In our case, feature engineering was done to convert a regression problem to a classification by binning the values and creating a new column as a result. The feature “nkills” which depicted the number of causalities in a terror attack was a sparse column with majority rows empty. This made the predictions difficult. To overcome this problem, its values were binned into Low and High kills.

3.5 Train test split and feature selection

A key step in the modelling process it to split the data into training and testing datasets, this is essential for generalizing the model. By segregating the dataset we essentially prevent our algorithm from over fitting.

We used K Fold Cross Validation to create K folds of training and testing datasets, the mean of the predictions resulting from the K folds of training and testing data is the final result.

Selecting the features for our model is probably the most important step as features exhibiting strong correlation with the dependent variable leads to better prediction compared to the features with weak correlation coefficients.

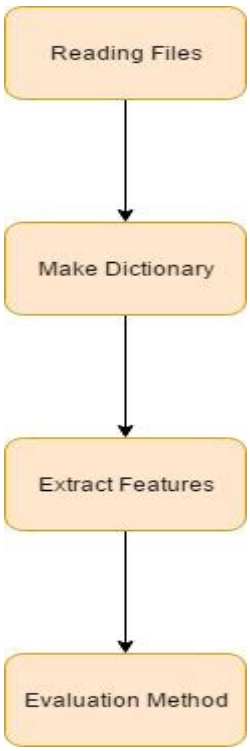
4 Modelling

Modelling refers to the process of training machine learning algorithms using the cleansed data we got after the data preparation step. Once the algorithm is trained using the training datasets, we make predictions using the testing datasets.

The following models were developed to answer the questions posed by the datasets.

1. Linear Model with Poisson distribution to predict the number of kills.
2. Model to predict the mortality rate associated with a terror attack.
3. KNN to predict region of attack.
4. Random forest to predict the terrorist groups
5. Predict the Motive and Claims of an attack using Multinomial NB and SVM
6. Predicting future terrorist attacks using ARIMA and ETS

General Procedure followed for Modelling



5 Evaluation

5.1 Model Performance

Once the models are fitted we can predict the accuracy using the validation data. Since this data is completely new for the model, it would give a true sense of the fit of the model.

The accuracy from the testing data is then compared to the training dataset, a significant difference between the two would indicate overfitting. Other metrics such as confusion matrix, ROC curves and Log loss are also used for determining the model performance.

1. Linear Model with Poisson distribution to predict the number of kills.

   Independent variables: The first 25 principal components after PCA.
   Dependent variable: nkill

   A linear model with Poisson family was trained to predict the number of kills, the model returned an accuracy of 78 %

```
Residual standard error: 5.116 on 156746 degrees of freedom
Multiple R-squared:  0.7878,    Adjusted R-squared:  0.7877
F-statistic: 2.327e+04 on 25 and 156746 DF,  p-value: < 2.2e-16
```

   Linear regression is better suited for continuous values, so we tried to improve the model by using a Generalized Linear Model with Poisson Family, but the model was not fitted properly which was confirmed with a chi square test.

   High Null Deviance and Residual deviance is an indicator of poor fit, Dispersion parameter is also quite high

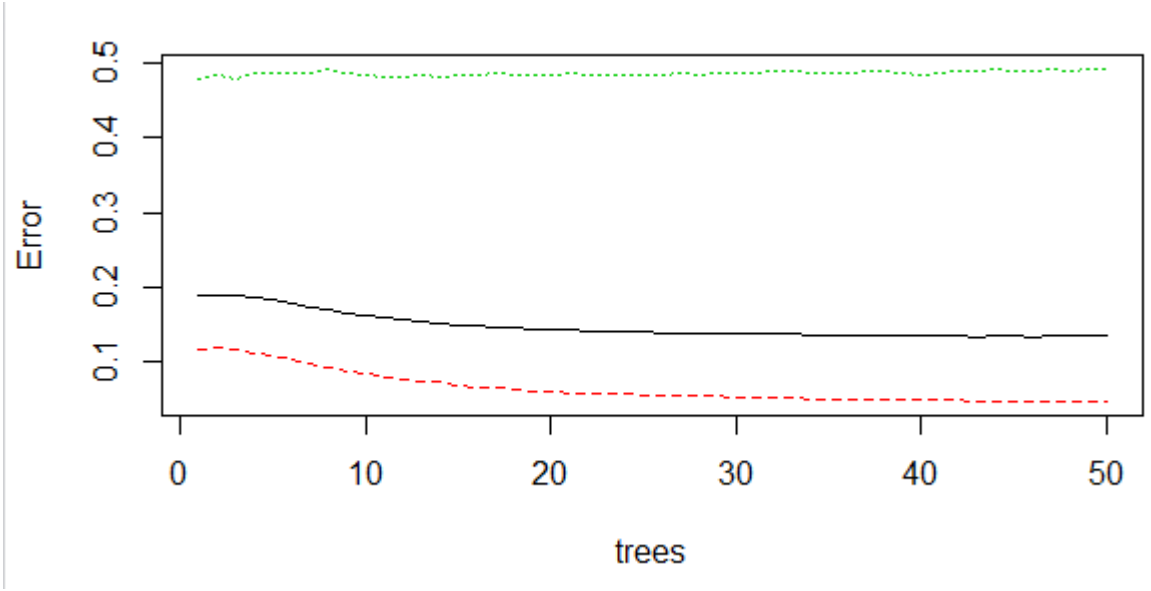
```
(Dispersion parameter for quasipoisson family taken to be 39.81687)

    Null deviance: 1245972  on 156771  degrees of freedom
Residual deviance: 1158915  on 156766  degrees of freedom
AIC: NA
```

   The chi square test on the model returned a value of 0, indication misfit.

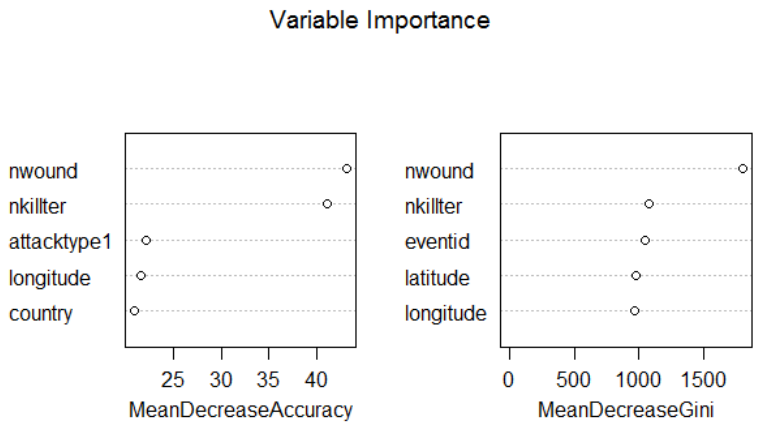
2. Since it was difficult to predict the number of kills, we converted it to a classification problem.
   Feature Engineering was done to create a new feature with Binary outcomes, High and Low for mortality rates. The problem now transformed to predict the rate of mortality (High or Low) based on the available features.

   A random Forest classifier with 50 trees was used to fit the model.

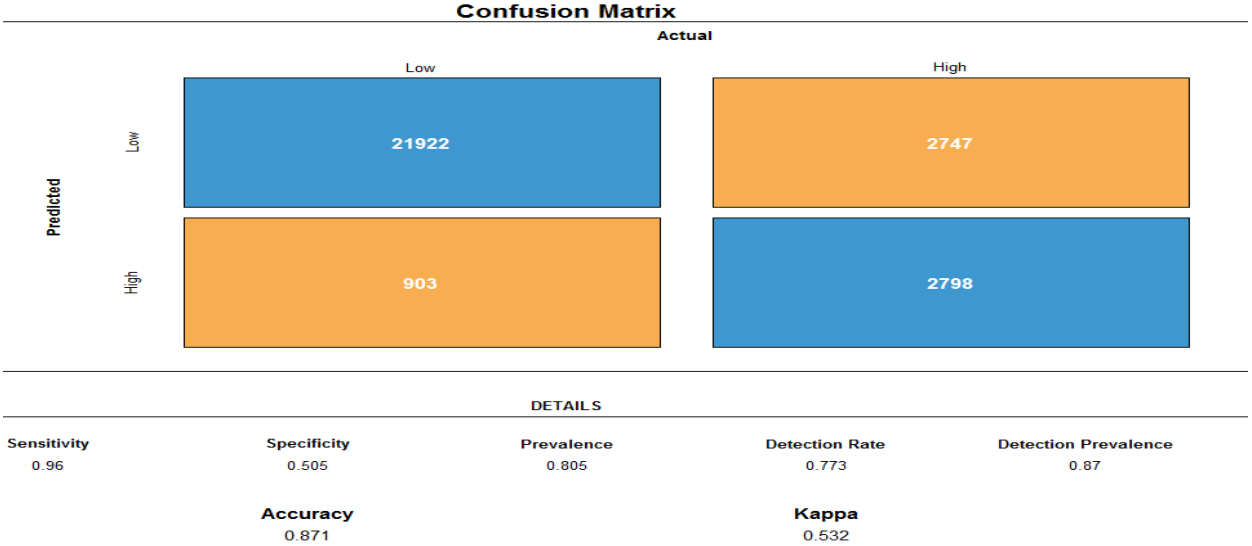
Error Rates: The errors normalized after 30 trees.



Feature Importance: The features which dominates the prediction



Confusion Matrix was then plotted to list the accuracy, specificity, sensitivity and Prevalence.

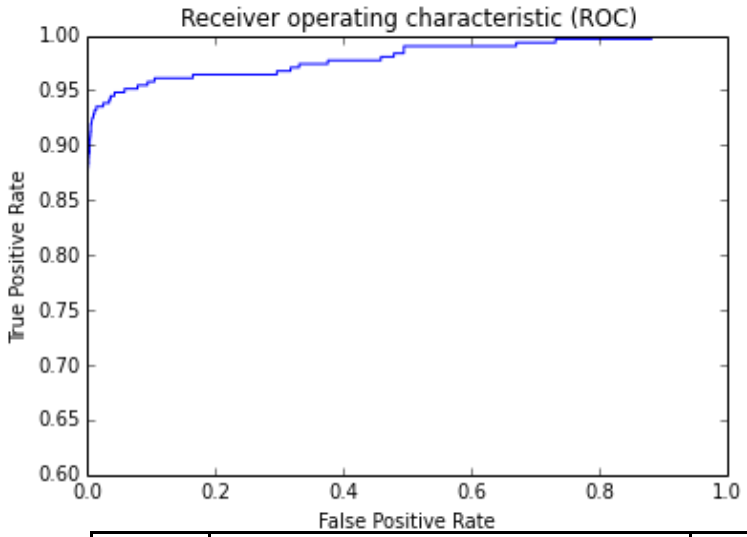


3. Predict the Motive, Claims and success of an attack using Multinomial NB and SVM

The motive files that were stored, were read in two different arrays, where, .success words were stored in one and .failure words were stored in another array. There were some words like conjunctions, connecting words that were taken out from the bag of words. In feature extraction face, dataset was split into 70:30 in series and model was set to predict success of the attack with motive. The model was fit according to MultinomialNB and SVM.

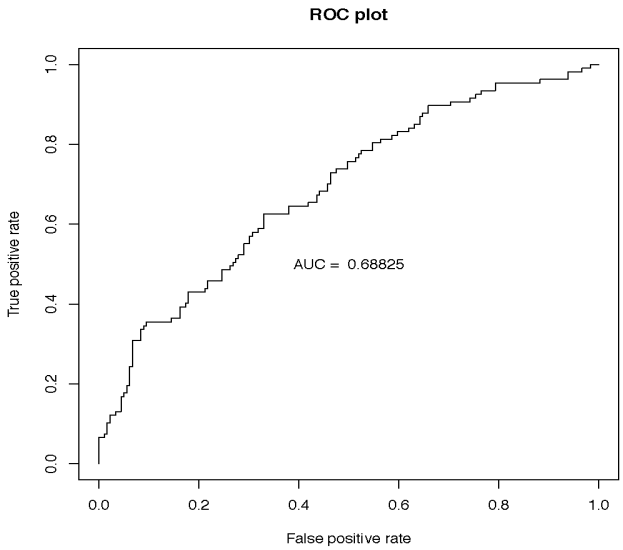
MultinomialNB	**Success**	**Failure**
Success	4805(True Positive)	127(False Positive)
Failure	59(False Negative)	5124(True Negative)

Table I - Multinomial Confusion Matrix

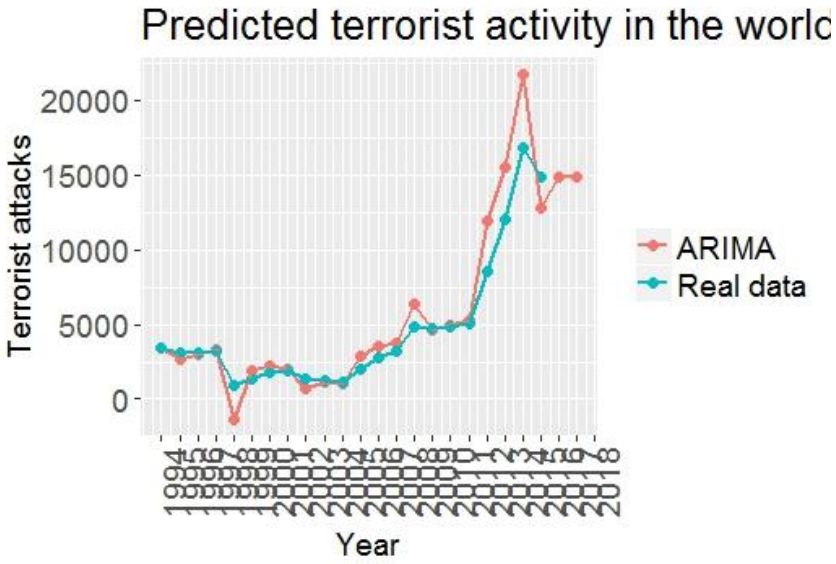


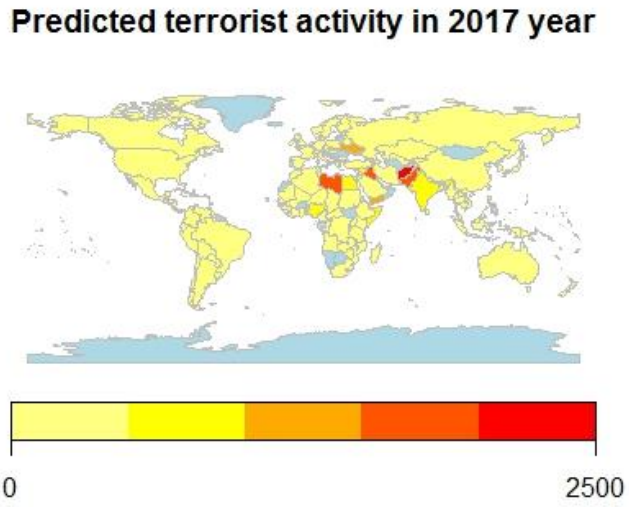
SVM	**Success**	**Failure**
Success	4778(True Positive)	154(False Positive)
Failure	116(False Negative)	5076(True Negative)

Table II - Support Vector Machines Confusion Matrix



4. Predicting future terrorist activities using ARIMA and ETS





5. K-Nearest Neighbour gave 60 % accuracy in predicting the country of attack.

6. Random Forest was again used to predict the terrorist groups with an accuracy of 72 %.

Conclusion

The analysis done for this dataset can be extended to provide real time insights to predict and prevent future terrorist attacks from happening. Following the CRISP-DM model helped us understand how to approach a Data Analysis task. The various machine learning algorithms we applied equipped us with the prediction power required to analyse the problem.

References

I. Sebastian Raschka,"Predictive modeling, supervised machine learning, and pattern classification",Aug 25, 2014 [Online Access].

II. Banhi Guha and Gautam Bandyopadhyay, "Gold Price Forecasting Using ARIMA Model", Journal of Advanced Management Science Vol. 4, No. 2, March 2016.

III. https://www.start.umd.edu/gtd/

IV. https://www.analyticsvidhya.com/blog/2015/08/common-machine-learning-algorithms/