Assignment-3

Data Analysis on Global Terrorist Database

By

Devaneek Sharma

(29781965)

Table of contents:

1. Introduction:

Terrorism is the biggest threat to world peace, with the growing geological instability situation has become alarming and demands the study or analysis about the factors affecting the terrorism, their impact on the human beings, economy. Predicting or forecasting the parameters well in advanced will help us preparing well in advances in future and save many human life's and property.
As of today, there are many terrorist groups in the world each terror groups ready to take credit of their miss deeds, sometimes their claims are false and try to advantage of the situation. By this study I will try bridge this gap and predict the terrorist group behind each attack.
In the second case, I will try to study the effect of the population growth on the terrorism, has terrorism gone up from last few years and what will be the effect of population on terrorism in the future years. By this study I intend to help the security or investigation agencies, researchers and people fighting the terrorism.

Key Business Objective:

1. Data analysis about how Terror group are related to the terror attack, what are the characteristic of each terror group and how we could identify the terror group responsible for terror attack. There are about 50% unclaimed terror attack in GTD. In this study targets to predict all the unclaimed terror attack in the global terrorist dataset.

2. Analysis the effect of the population growth on terrorism, check if they are related to each other and forecast how about the terror attack will happen in future based on the future population.

2. Description of Dataset:

Below two are the dataset which are implemented in the MongoDB and used in this study.

1. Global Terrorism database (GTD) (http://www.start.umd.edu/gtd) (180k rows X 135 columns)
   Dataset is in tabular format, which is having the many categorical, numeric, geospatical and date and time.
   Issue with the data:
   Almost all the data in the data set were in string value, numeric, integer so data is cleaned in the R and converted to the desire datatype. Null or na value are imputed using preprocess ML.
   Mains columns are:
   - Eventid- id of the attack :numeric
   - Iyear,imonth,iday- Date and time
   - Attack_type1 -type of attack like bombing etc :categorical
   - Target-type1-type of target: categorical
   - Nwond- no of wounded: numeric

- Nkill- no of killed: numeric
- Claimed- weather attack is claimed – Boolean
- Weapon_type1- type of weapon:categorical

2. World Population dataset(https://esa.un.org/unpd/wpp/Download/Standard/Population/)(280946 rows X 9 columns)
   Tabular data, columns are numeric and characters types.

3. Why this study is done?

Global terrorism is the threat human population, government, Infrastructure and global peace. This area still lacks the comprehensive study will could answers all the question or the parameters, which can help in preventing the terror attack or solve the terror cases and neutralizing them thus making the world peaceful place. There are about 50% unclaimed terror attack in Global Terrorist database by this study I tried to fill the gaps and provide the insights about how we could identify which terror group is behind the terror attack which might help global investigation or security agencies in determine the culprit.

Secondly, by forecasting I intend to show how the population is affecting the terrorisms' and therefore various factors can be take well in advance to curb terrorism.

4. Methodology:

GTD dataset is huge and contain about 400k records due to the processing constraint on the data from the year 2007-2017 is taking for prediction, similar is the case with population dataset.

1. **Security Feature in mongo DB:** Security feature in the mondo db is been setup, admin account is been created based on the steps given at the mongoAPI.
   In the tableau connection string of *"mongodb://admin:admin@localhost" is* been pass to connect to mongo db, both the username and password for connecting to mongo db is admin.
2. **Data preprocessing**: Data is preprocessed in the R, first the subset of the data for both global terrorism dataset for the year 2007 and 2017 for prediction is been done. All the required columns are taken out from the data frame. Columns are converted to the desire datatype; categorical columns are change to the numeric value that they could be used for classification. Then merging the two-dataset population and GTD by the column ***year*** and ***country*** is done.

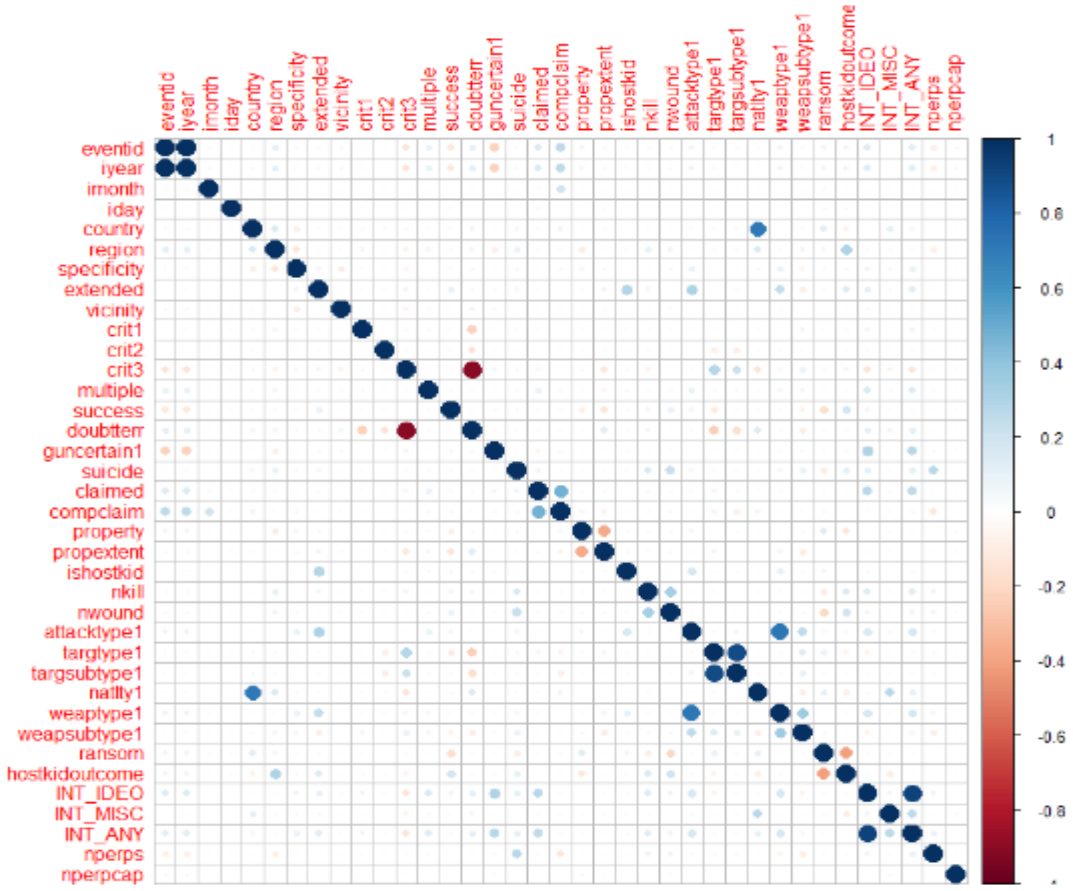
3. **Imputation of Null Values**: there were many Null values in the dataset thus, those null values are imputed by using *Ipred* library preprocess and then predicting the null values.

4. **Feature selection:** This is one of the most important part of the data analysis, it helps to find all the important features in the data set which will be used further correspondence to the target which we are trying to find. Below are the some of techniques and algorithms which were used in finding features in the dataset.

- **Correlation matrix-** relations between the different features is been shown by this matrix. The highly corelated features are.

```
[1] "INT_ANY"     "iyear"       "doubtterr"   "attacktype1" "targtype1"
"natlty1"
```

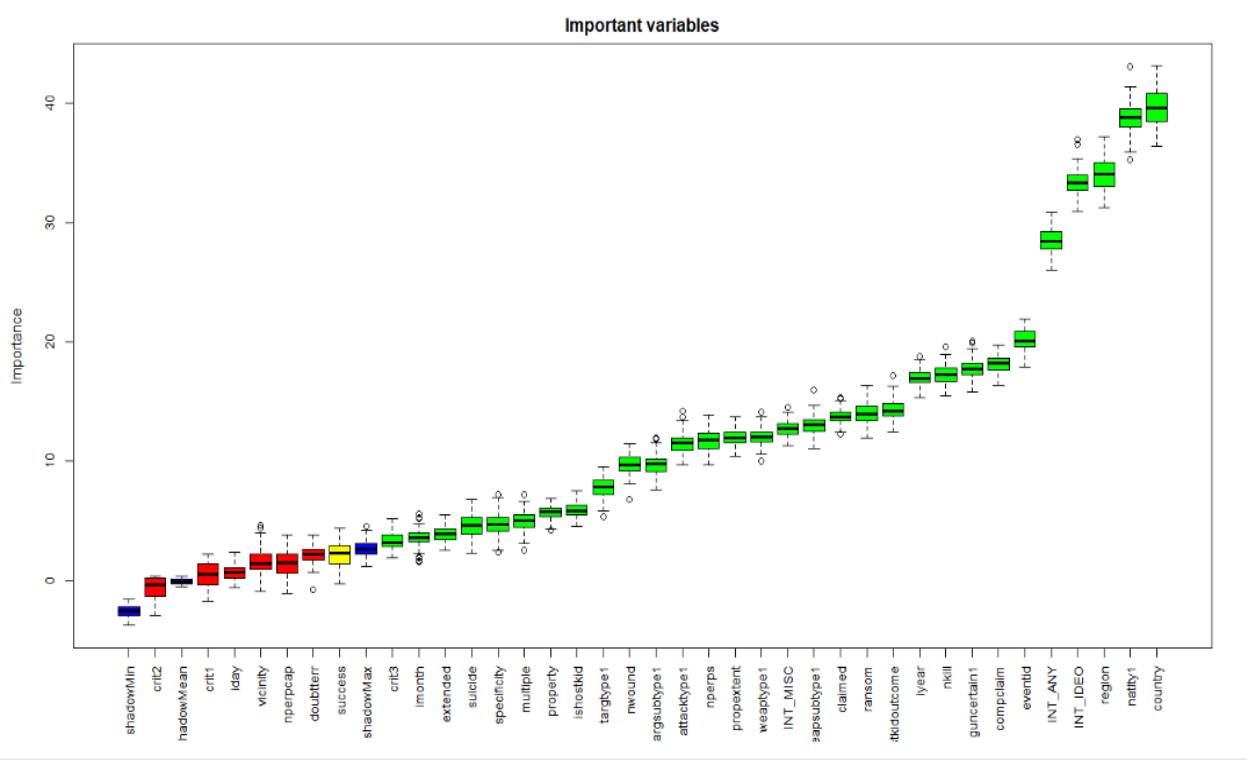
Carrot is also used but didn't provided the sufficient output as there are lots of categorical column in the data. Another method Boruta algorithms is used find the important features.



Corelation matrix

- **Boruna Analysis**- it works as the wrapper algorithms around the rainforest. It has capture all the below important features.

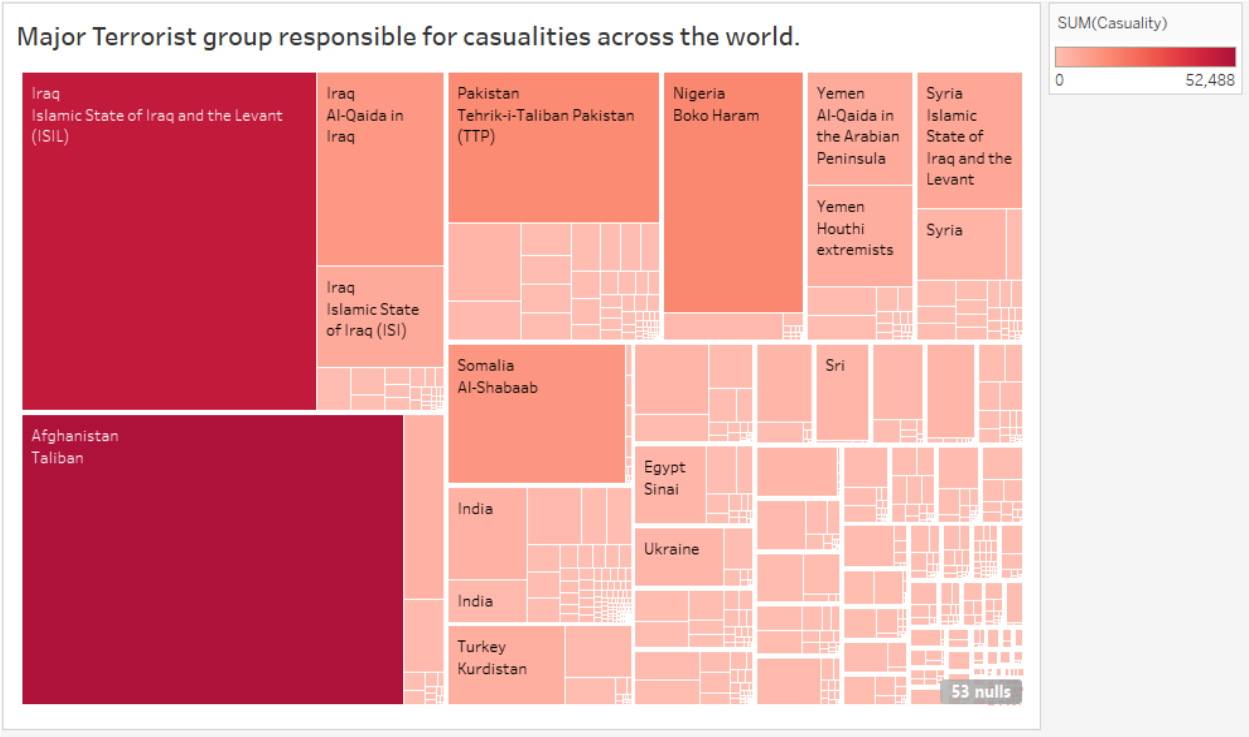
```
[1] "eventid"        "iyear"          "imonth"         "country"
"region"         "specificity"
 [7] "extended"       "crit3"          "multiple"       "success"
"guncertain1"    "suicide"
[13] "claimed"        "compclaim"      "property"       "propextent"
"ishostkid"      "nkill"
[19] "nwound"         "attacktype1"    "targtype1"      "targsubtype1"
"natlty1"        "weaptype1"
[25] "weapsubtype1"   "ransom"         "hostkidoutcome" "INT_IDEO"
"INT_MISC"       "INT_ANY"
[31] "nperps"
```



Important features

5. Modelling for predicting Terrorist group in the data set.

Below are some of the visualization graphs helps use understands the data.

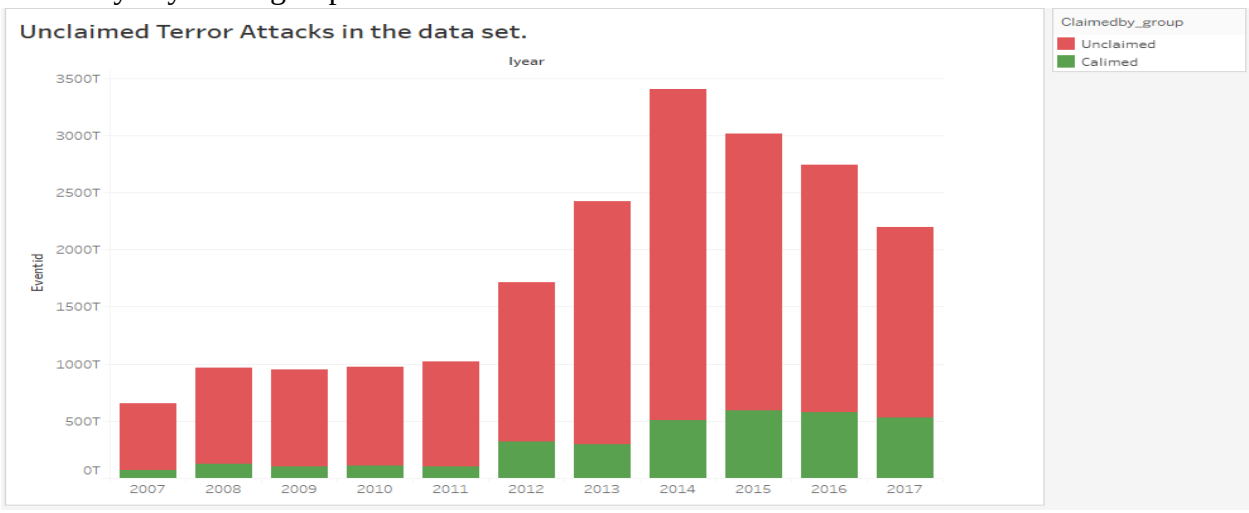


Terrorist group responsible for attack across globe.

- From the graph we could see that Taliban tops the charts with maximum numbers of attacks, followed by ISIS.

- Iraq and Afghanistan are the deeply affected country, maybe because of the ongoing war.

- Below graph shows there are more than 48% attacks which are not been claimed by nay terror group.



Unclaimed attacks.

Data Modelling:

After the feature selection, building of the model is started in order to predict which terrorist group cause the attack. Since we have more than 1200 terrorist groups only top 20 groups have been considered and rest are taken into the category of others.

For this I tried to build the multiple classification model with the 20 classes which could help in predict whether we are able to predict the correct terror group or not.

Different classification model like rainforest classifier, naïve baiyes and bagging tress are used to classing the attacks. After than accuracy of the model is compared.

Abu Sayyaf Group (ASG)	Al-Qaida in Iraq
406	558
Al-Qaida in the Arabian Peninsula (AQAP)	Al-Shabaab
1015	3288
Boko Haram	Communist Party of India - Maoist (CPI-Maoist)
2418	1853
Donetsk People's Republic	Fulani extremists
624	510
Houthi extremists (Ansar Allah)	Islamic State of Iraq and the Levant (ISIL)
1061	5613
Kurdistan Workers' Party (PKK)	Maoists
1175	1388
Muslim extremists	New People's Army (NPA)
549	1696
Others groups	Palestinian Extremists
67279	418
Revolutionary Armed Forces of Colombia (FARC)	Sinai Province of the Islamic State
862	445
Taliban	Tehrik-i-Taliban Pakistan (TTP)
7074	1351

Class for Classifications

Creating the test and train data: sampling of the data is done due to the processing isuues and sample of 15k – 20k is used. dataset is split into 75% : 25% train/test

Random forest classifier:

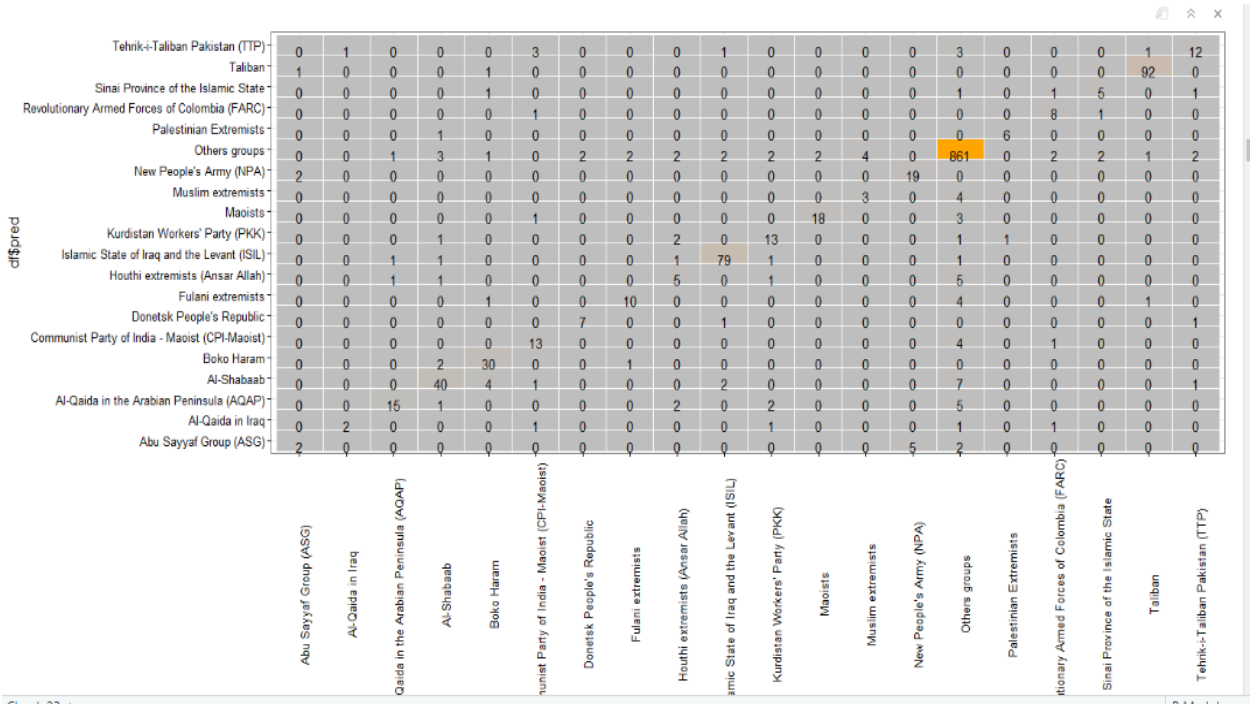
After running the random forest for the ntree=20 we got the accuracy of 99.2 which is quite high. Below is the statistics of the random forest.

```
Overall Statistics

               Accuracy : 0.8814
                 95% CI : (0.8631, 0.8981)
    No Information Rate : 0.6603
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7853
 Mcnemar's Test P-Value : NA
```

Random forest shows that we are successfully in predicting the terror group with accuracy of **0.8814**.

AUC which represents the area under the curve is calculated from the ROC the more the value of AUC close to 1 the better our prediction are: AUC for randomforest is **0.7** which means our prediction are very well predicted.

Confusion matrix:



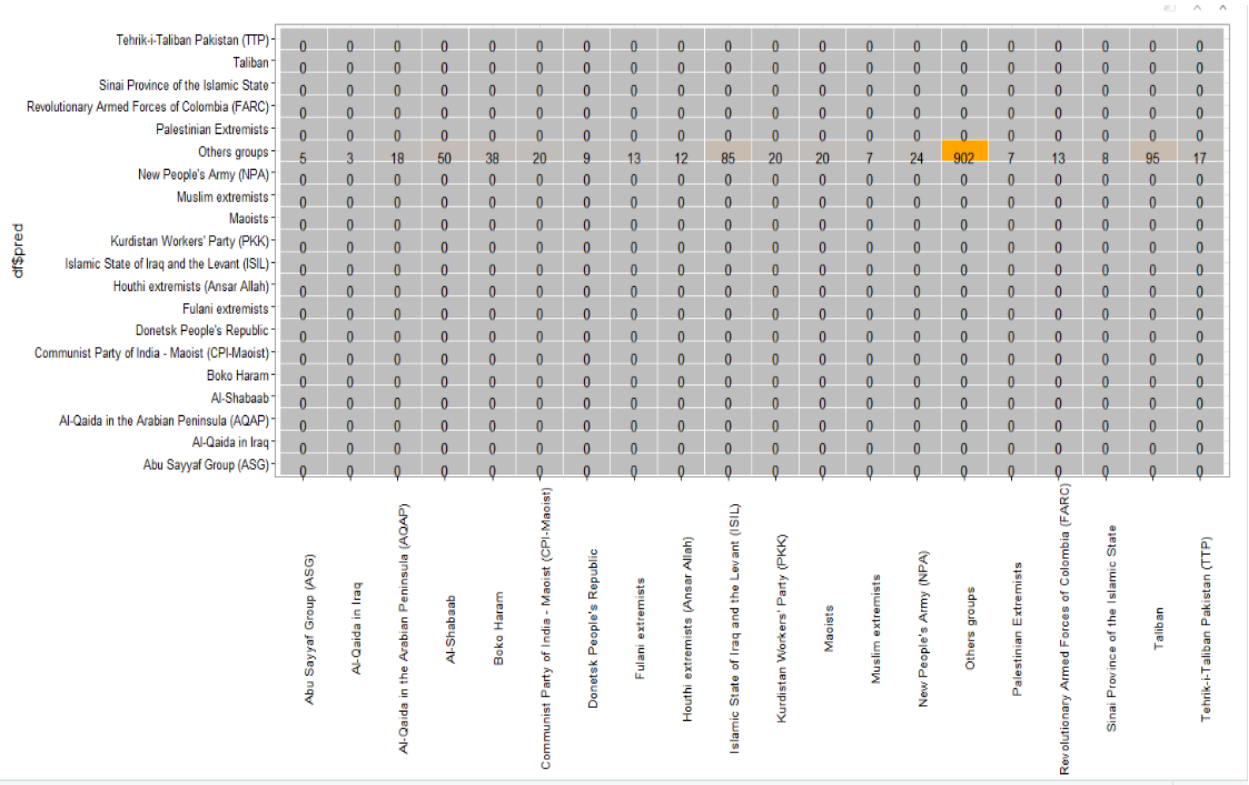
Confusion matrix shows how the true values are related to the false predicted values, diagonal of the matrix shows the true predicted values. “Others groups” is been dominated and have successfully predicted the other group values, followed by Taliban and ISIS.

Naïve Baiyes Classification:

```
Overall Statistics

               Accuracy : 0.6603
                 95% CI : (0.6345, 0.6854)
    No Information Rate : 0.6603
    P-Value [Acc > NIR] : 0.5126

                  Kappa : 0
 Mcnemar's Test P-Value : NA
```

Classification by naïve baiyes shows that that we can predict by the accuracy of **0.66** and AUC which represents the area under the curve is **0.5**

Confusion Matrix:



Confusion matrix shows model predict high false values we don't have any true values other than other group.

Comparing the different models:

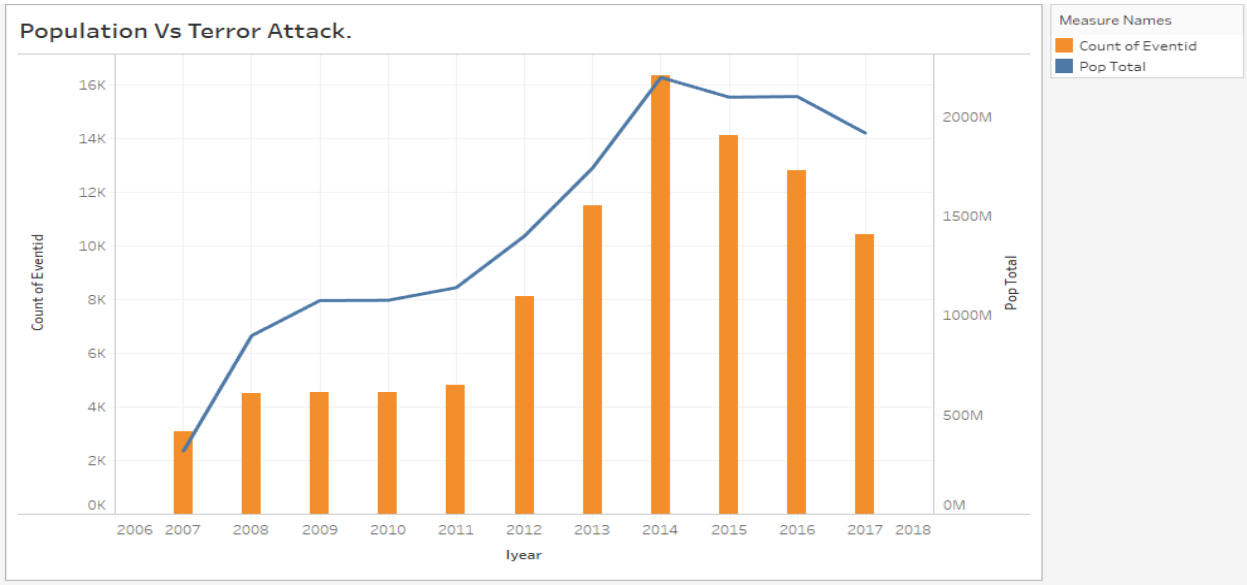
Model	AUC	Accuracy
RandomForest	0.7	0.8814
naiveBayes	0.5	0.6682

Results:

Results from the two Model shows that randomforest predict more accurately than naivebayes and we are successfully able to predict the value of all the terror groups.

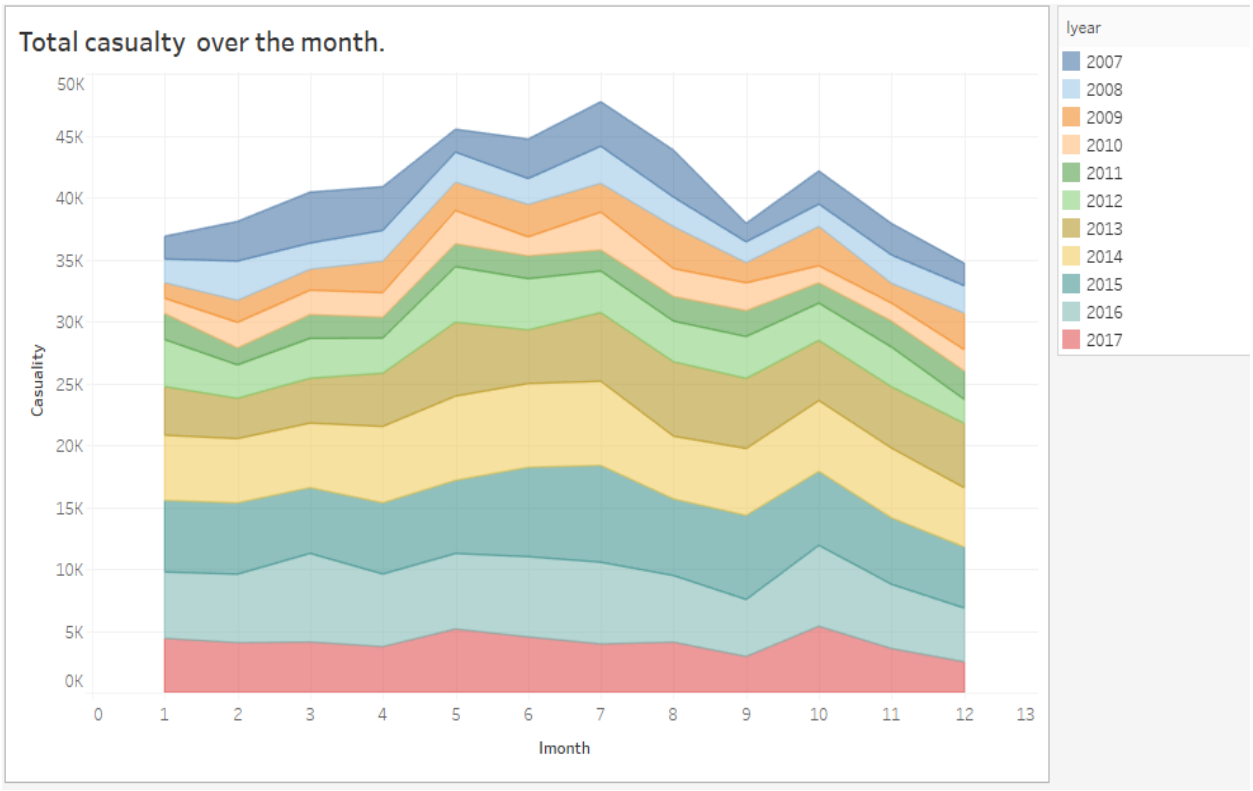
Effect of the population on terrorism:

Now we will be discussing the effect of the population on terrorism, below is the visualization of the how the population changes with terrorism over the period of time.



From the above graph we could infer as population increase than number of attacks increase and year 2014 record maximum number of attack and population.

Below graph shows the total casualty over the month form the year 2007 to 2017.

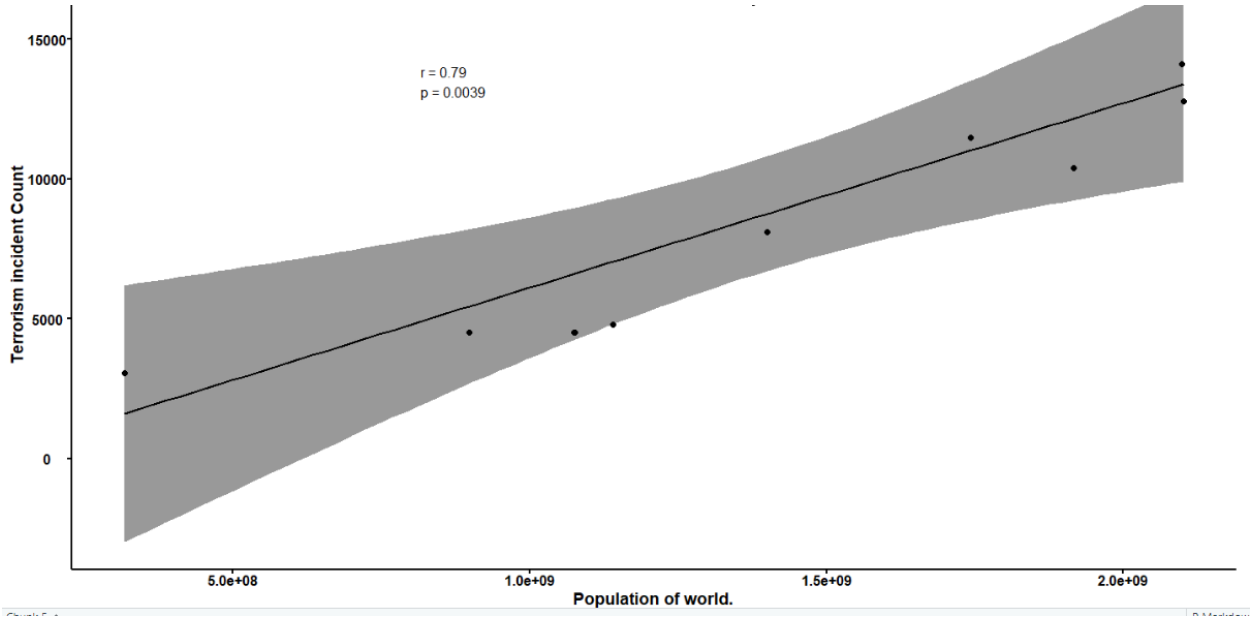


Is there is Correlation between the population and Terrorism?

Correlation is been calculated by using the Pearson method and corr function.

```
        Pearson's product-moment correlation

data:  Totalpop_terror$PopTotal and Totalpop_terror$terror_attack_count
t = 8.7192, df = 9, p-value = 1.106e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7988650 0.9861097
sample estimates:
      cor
0.9455935
```

Corr values 0.94 shows they are closely corelated.

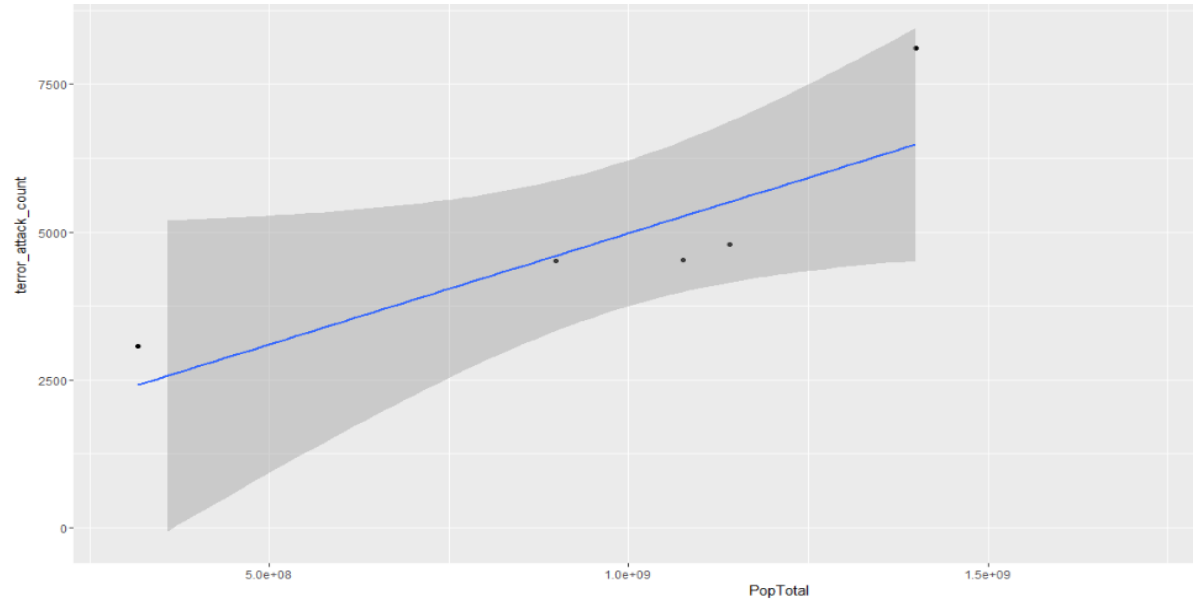


Scatter plot show how the terrorism and population are related to each other.

Here, first we formed the hypothesis where linear model can be used to find the relationship between the population and terrorrism.

Linear model for forecasting:

First the linear regresstion model is been used as we are having only 10 years of data data points will be less and contains the aggreagate value. The fit for linear regression is not so good and most of the points are outside of the confidance interval and not aligned to the straight linear thus linear model poorly explain the realstionship between the population and terrorism.

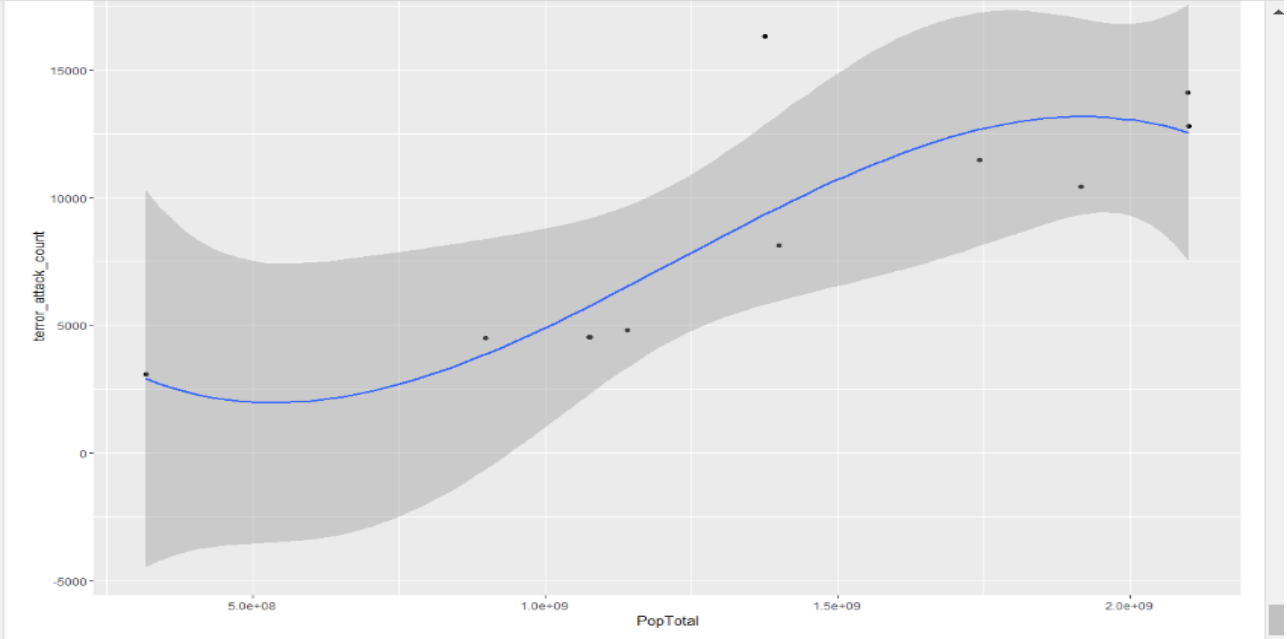


`R-squared: 0.6221`

For the linear regression

Ploynomial regression:

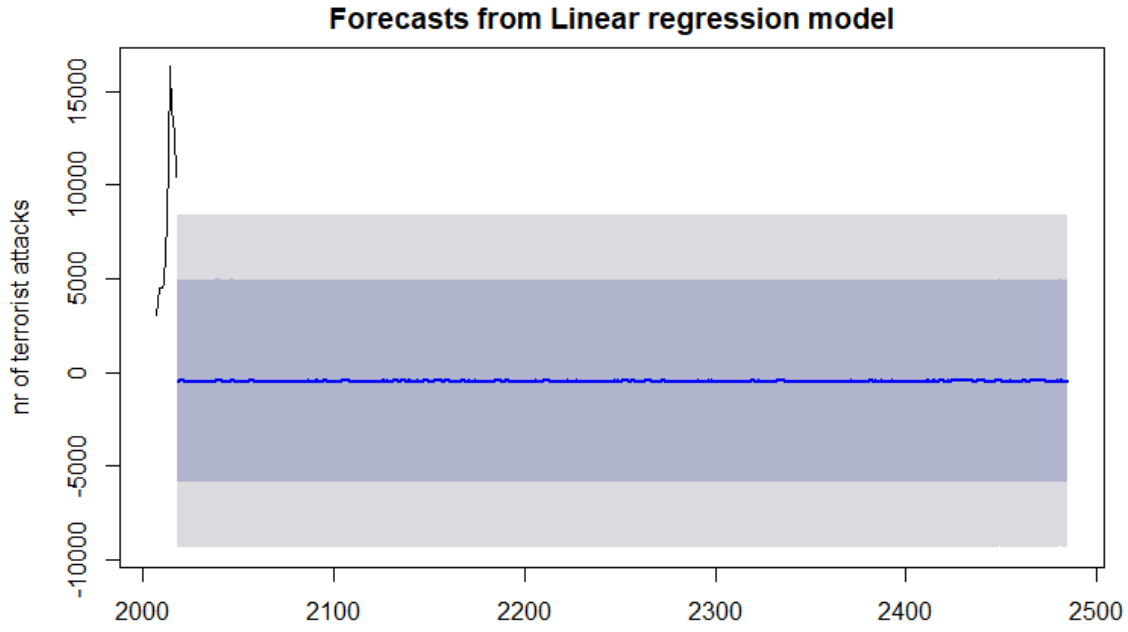
Polynomial regression of order 3 is been used to predict the values. The fit shows that model is good and value of `R-squared: 0.6795 which is greater than linear model.`



R square which is used to determine closeness of the variable about the mean, ploy regression shows they are more closely related than linear.

Forcasting:

Due to the processing constraint and a larger dataset the graph is visible clearly. But by using the polynominal regression of the order 3 we are successfully able to determine the relation between the terrorism and population. The different color shows the confidence interval of 50, 70 and 95%. This plot shows that rate of terrorism will not change and remain less than 15k for the 95% confidence interval.



Conclusion:

Through the data analysis done in this study we are successful in predicting the terror group responsible for the terror attack by using randomforest classifier with the accuracy of 88% and AUC~0.7 thus, this model is highly accurate and further could be used as a pilot project to get the more insights from the Global Terrorism Dataset. In case of forecasting we tried forecasting the terrorism and population relation by using polynomial order 3 regression and it fits wells for the smaller dataset, due to the processing power constraints it was not tested for the larger dataset hence their relationship between them cannot be established with the confidence. Maybe in future studies with more resources we might be able to establish it clearly.