Machine Learning Engineer Nanodegree

Frederico Guerra

October 2018

1 Domain Background

Terrorism is usually understood as the use or threat of violence to further a political cause[1] around the world. The attacks are increasing since the 70 decade and it is being called Modern Terrorism with the rise of nationalist movements around the globe after The Second World War.

Although this growth, the Institute for Economics and Peace's[2] appointed that in the last few years the number of deaths and injuries has decreased 22% to 25,673 deaths in 2016 compared to the peak of terror activity in 2014 when over 32,500 people were killed. Still it is a gain, Daniel Brown[3] showed that more countries experienced at least one terrorism-related death in 2016 than in any other year 2001 with 77 countries affected.

2 Problem Statement

This project aims to understand the behavior of incidents that occurred deaths, therefore know which features have most impact to an attack kills someone. In this way, the main goal is to predict whether occurred deaths in these terrorist events. Although terror attacks are becoming less deadly in the past years, as can bee seen in Figure 1, the number of deaths is still the highest since 1970. Therefore, is important to know what are the main reasons associated with deaths in these attacks.

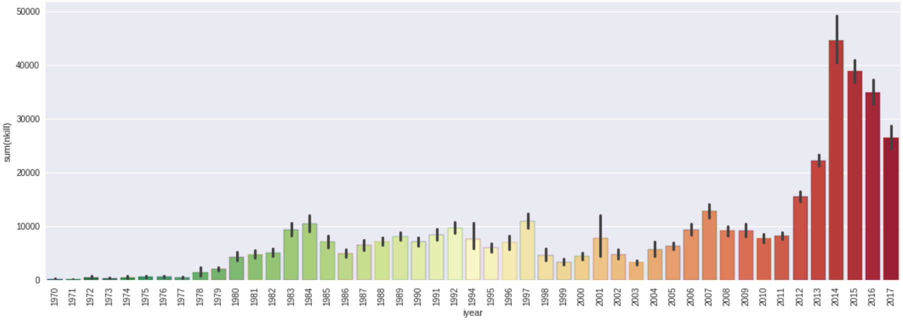


Figure 1: Number of deaths by year. (1970-2017)

[1] https://ourworldindata.org/terrorismdata-quality-definitions

[2] http://visionofhumanity.org/indexes/global-peace-index/

[3] https://www.businessinsider.com/global-terrorism-terrorism-increased-deaths-from-terrorism-decreased-2017-11

3 Datasets and Inputs

The Global Terrorism Database(GTD) is an open-source database including information and systematic data on domestic, transnational and international terrorist events around the world from 1970 to 2017. The dataset uded in this project will be caught from Kaggle API[4] At the moment, the dataset includes more than 180,000 cases and has a lot of information such as date, location, the weapons used, nature of the target, the number of casualties and - when identifiable - the group or individual responsible.
In total, there are 135 features and, as described in the GTD codebook[5], contains:

- Date
- Incident Information
- Incident Location
- Attack Information
- Weapon Information
- Target/Victim Information
- Perpetrator Information
- Casualties and Consequences
- Additional Information and Sources

The target variable death will be a boolean calculated column from the Total Number of Fatalities *nkill* feature in Casualties and Consequences group, where:

$$death = \begin{cases} 1, & if \quad nkill > 0 \\ 0, & otherwise \end{cases}$$

After a briefly exploratory analysis of nkill variable, can be inferred that there are 94% filled values - which corresponds to 171k valid data and in **48**% (83k) of them occurred at least one death. The other 6% data with missing values will be avoided.

4 Solution Statement

This project aims firstly to understand the correlation between the 135 features and the death column result, attending to the topics described above (Incident, Location, Attack, Weapon, Perpetrator, etc.). To avoid overfitting during model train step and to guarantee a good performance of the model chosen, will be done dimensionality reduction with PCA, missing values treatment and model evaluation. As said before, it will be a binary classification problem.

[4] https://www.kaggle.com/START-UMD/gtd
[5] https://www.start.umd.edu/gtd/downloads/Codebook.pdf

5 Benchmark Model

There is not a work in the literature of model evaluation to predict death occurrence in Terrorism Attacks around the world using GTD, so since this is a dataset commonly used in many studies and visualizations[6] this project it will be a great work to be used as reference for these studies. Furthermore it can be used to predict instantly after one attack whether it takes fatalities or not.

6 Evaluation Metrics

Considering that this is a binary classification problem, two metrics will be used to evaluate the model chosen: Confusion Matrix and Area Under the Curve (ROC curve). The main dataset will be splitted in other two for training and testing procedures.

6.0.1 Confusion Matrix

n = Test data size	**Predicted: NO**	**Predicted: YES**
Actual: NO		
Actual: YES		

To evaluate the model will be calculated its accuracy taking the number of attacks with death predictions (yes) by Total number of Samples:

$$accuracy = \frac{TruePositives + FalseNegatives}{TotalNumberofSamples}$$

where:

- **True Positives:** The cases in which will be predicted death (YES) and the actual output was also death (YES).
- **True Negatives:** The cases in which will be predicted no death (NO) and the actual output was no death (NO).
- **False Positives:** The cases in which will be predicted death (YES) and the actual output was no death (NO).
- **False Negatives:** The cases in which will be predicted no death (NO) and the actual output was death (YES).

6.0.2 Area Under the Curve

AUC of a classifier is equal to the probability that the classifier will rank a randomly chosen positive example higher than a randomly chosen negative example. To define AUC, should be understood about Sensitivity (True Positive Rate) and Specificity(False Positive Rate).

- **Sensitivity(True Positive Rate):** Number of deaths correctly predicted by the total number of deaths, in another words,

[6] https://public.tableau.com/en-us/search/vizzes/Global%20Terrorism%20Database

$$Sensitivity = \frac{TruePositives}{FalseNegatives + TruePositives}$$

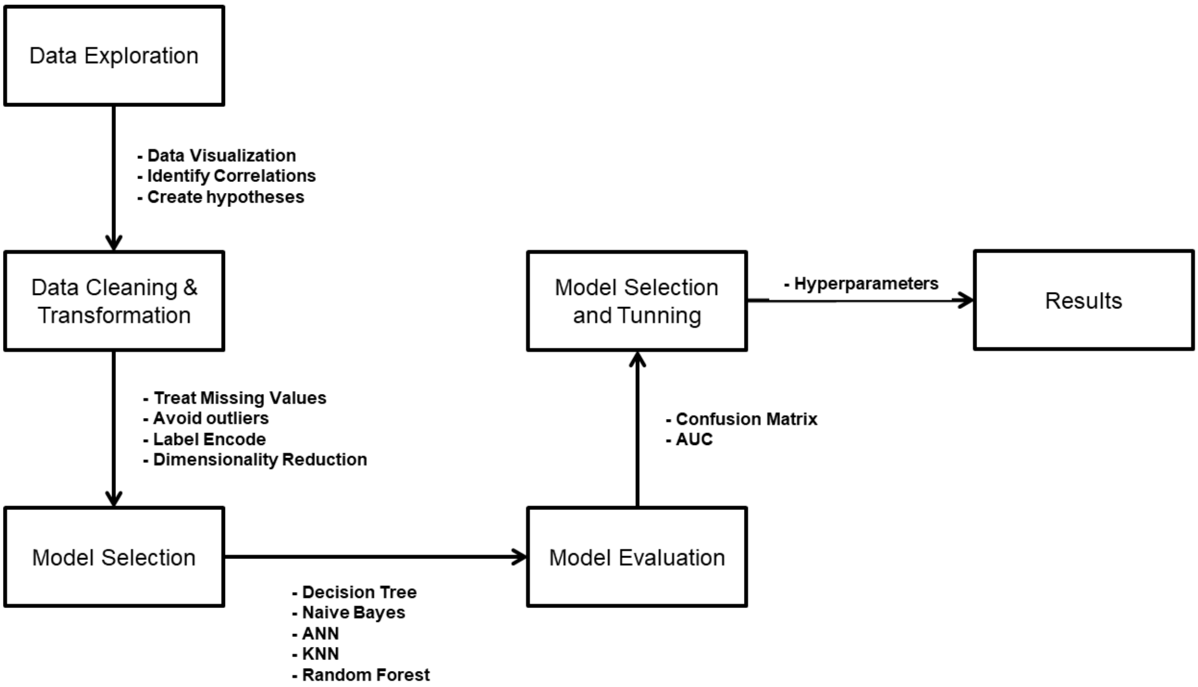
- **Specificity(False Positive Rate):** Number of deaths mistakenly predicted by the total number of attacks without fatalities.

$$Specificity = \frac{FalsePositive}{FalsePositive + TrueNegative}$$

The values of TPR and FPR are computed continuously between [0,1] and drawn in a line chart. The area under this curve corresponds to the AUC score. Most closely this value is to 1, better is the model.

7 Project Design

Project steps:


Data Exploration
- Data Visualization
- Identify Correlations
- Create hypotheses
Data Cleaning & Transformation
- Treat Missing Values
- Avoid outliers
- Label Encode
- Dimensionality Reduction
Model Selection
- Decision Tree
- Naive Bayes
- ANN
- KNN
- Random Forest
Model Evaluation
- Confusion Matrix
- AUC
Model Selection and Tunning
- Hyperparameters
Results