Machine Learning Engineer Nanodegree

Capstone Project

Context of Terrorism in Somalia

Ahmed Ga'al June 16, 2021

I. Definition

Project Overview

Somalia is a coastal country that lies at the Horn of Africa from the Gulf of Aden to the Indian Ocean. It has a coastline of more than 3, 300 km making Somalia the country with the longest coastline in Mainland Africa.

Somalia's population had **75%** literacy rate and a booming tourism industry hence earning the name **'The White Pearl of the Indian Ocean.'**

Somalia has immense natural resources and mineral reserves such as oil and natural gas, uranium, iron ore, tin, gypsum, bauxite, copper and many more. Somalia once possessed the most lethal Air Force in Africa and as such it trained **Burundi's**[1] air force, protected borders of **Tanzania** & **Uganda** and flew fighter jets for **Zambia**.[2]

It defended **Mozambique** from the Portuguese, trained **South Africans** fighting apartheid and supplied **Eritrea's**[3] war of independence.

Somalia sent troops to **Angola**, supported **Egypt** with naval logistics, supported **Djibouti's** independence movement and also supported **Zimbabwe** and **Namibia's** forces in the war against apartheid.[4]

1. Burundi Air Force
2. Somali Air Force aids Zambia
3. Somalia is a key ally during the Eritrean War of Independence
4. Somalia role in Africa's Modern Warfare

SOMALI CIVIL WAR

In 1991, Somalia was ravaged with war which went on for almost the past 3 decades after the toppling of the military regime governed by Major Gen. Mohamed Siad Barre. This gave birth to Somalia's Civil War and clan factions began fighting for power. Many civilians were annihilated and many more fled the country as refugees in neighbouring Kenya and all over the world.

In 1993, the US Rangers conducted Operation Gothic Serpent which resulted in a failed mission after it lost 2 Sirkosky UH-60 Black Hawk helicopters, 19 US servicemen and at least 1,000 Somali civilians were killed and 4,000 were wounded.

In 2006, **the Islamic Courts Union (ICU)**, an Islamic organization, assumed control of much of the southern part of the country and promptly imposed Sharia' law. On the dawn of 2007, TFG President and founder Abdullahi Yusuf Ahmed, a former colonel in the Somali Army, entered Mogadishu with the Ethiopian military support and relocated the government to **Villa Somalia** in the capital from its interim location in Baidoa.

This marked the first time since the fall of the Siad Barre regime in 1991 that the federal government controlled most of the country.

Following this defeat, the Islamic Courts Union splintered into several different factions. Some of the more radical elements, **Al-Shabaab**, regrouped to continue their insurgency against the TFG and oppose the Ethiopian military's presence in Somalia.

Throughout 2007 and 2008, Al-Shabaab scored military victories, seizing control of key towns and ports in both central and southern Somalia.

At the end of 2008, the group had captured Baidoa but not Mogadishu.

By January 2009, Al-Shabaab and other militias had managed to force the Ethiopian troops to retreat, leaving behind an ill-equipped African Union peacekeeping force to assist the Transitional Federal Government's troops.

Problem Statement

Terrorism is defined as the threatened or actual use of illegal force and violence by a non-state actor to attain a political, economic, religious or social goal through fear, coercion, or intimidation.[1] Somalia has faced quite a strife for almost 30 years and most of those was between clans. Up to date, clans engage in armed conflict attributed to diverse motives that are not terrorism.

The objective of this project is to carefully examine, analyse, map and classify the events where the type of attack was an act of terrorism or not where the targets of attacks were non-combatants (collateral damage).[2]

I intend to use decision trees and ensemble method to tackle this problem efficiently Also, I will experiment with boosting which are effective in reducing model bias and support vector machines which is versatile, memory efficient and utilizes subset of training points in the decision function known as support vectors. This will provide a clear understanding of context of terrorism in Somalia.

Metrics

Since this is a binary classification problem, I have employed a variety of classification metrics including but not limited to precision score.

$$\textbf{Precision} = \frac{\textbf{TRUE POSITIVE}}{\textbf{TRUE POSITIVE + FALSE POSITIVE}}$$

To regulate the model from overfitting or underfitting since the target class is not balanced, I implemented a confusion matrix to help me determine the sensitivity and selectivity of the model.

Sensitivity:

$$\frac{\textbf{TRUE POSITIVE}}{\textbf{TRUE POSITIVE + FALSE NEGATIVE}}$$

Selectivity:

$$\frac{\textbf{TRUE NEGATIVE}}{\textbf{TRUE NEGATIVE + FALSE POSITIVE}}$$

It is imperative to classify incidents there is doubt is an act of terrorism. Sensitivity score will rule out the incident was an act of terrorism as defined below.

Selectivity will rule in the incident was indeed an act of terrorism

1. Terrorism According to National Consortium for the Study of Terrorism and Responses to Terrorism.
2. USAF Intelligence Targeting Guide

II. Analysis

Data Exploration

The data that will be used throughout the course of this project will be obtained from the **Global Terrorism Database**.

This database was granted by the **National Consortium for the Study of Terrorism and Responses to Terrorism** at the **University of Maryland**.

This is the most reliable dataset to accomplish the objective of this project.

The dataset contains 191,463 data points and 135 features.

The dataset contains numerous missing values in 104 features out the 135 features of the original dimensions.

Our target class, '**doubt_terrorism**' contains two classes:

- 0 = 2,969 data points
- 1 = 1, 687 data points

Exploratory Visualization

After performing intensive data cleansing, I went ahead and commenced Exploratory Data Analysis.

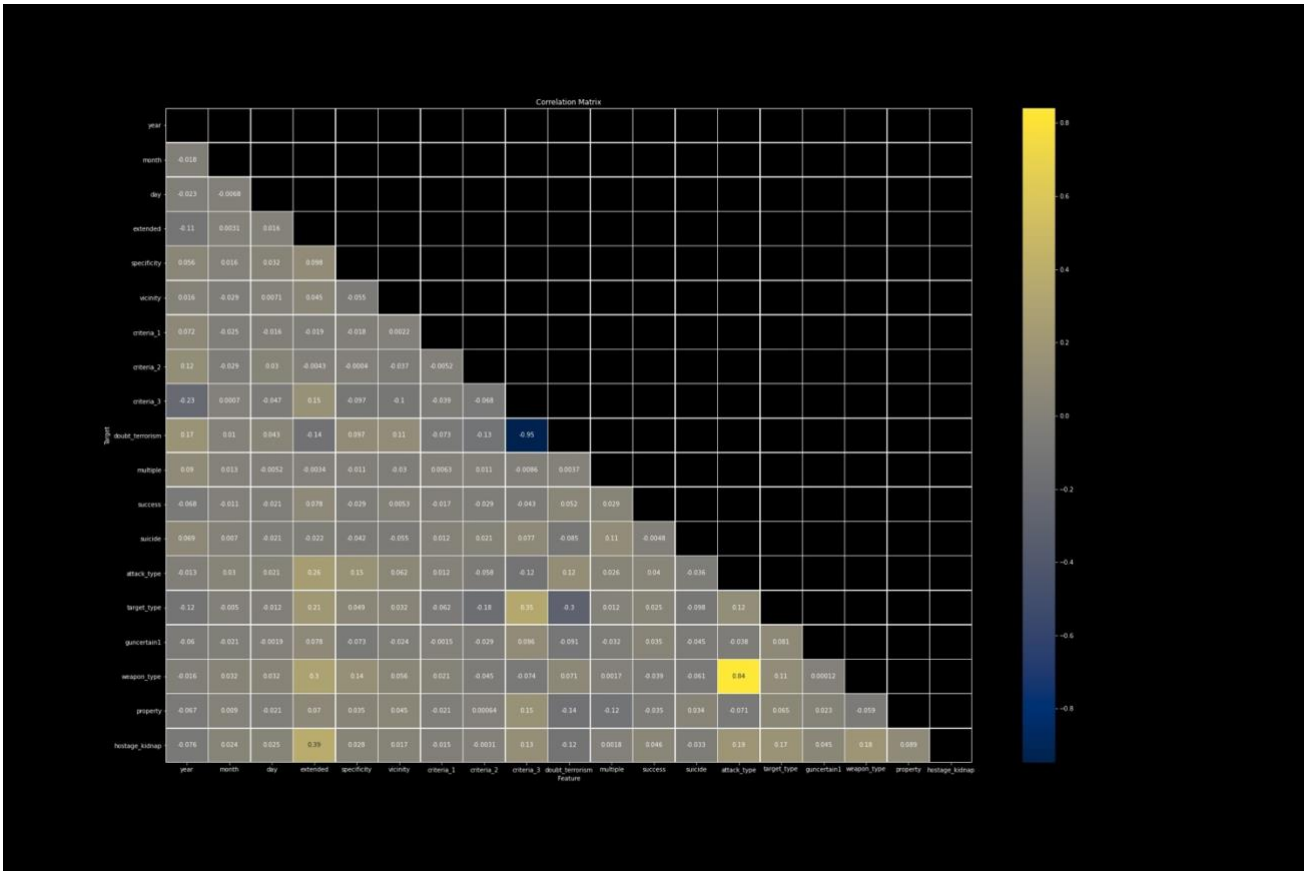
First, I computed the descriptive statistics as follows:

	year	month	day	extended	specificity	vicinity	criteria_1	criteria_2	criteria_3	doubt_terrorism	multiple	success	suicide	attack_type	target_type	guncertain1	weapon_type	property	hostage_kidnap
count	4656	4656	4656	4656	4656	4656	4656	4656	4656	4656	4656	4656	4656	4656	4656	4656	4656	4656	4656
mean	2013.385954	6.505154639	15.69179553	0.06056701	1.211554983	0.082474227	0.996993127	0.990979381	0.662371134	0.362328179	0.106958763	0.914304124	0.046821306	3.347293814	6.526417526	0.133591065	6.733891753	-2.18556701	0.079896907
std	4.884651016	3.329775712	8.893965738	0.238559994	0.657211232	0.275115403	0.054758338	0.094557746	0.472952066	0.480724561	0.309094008	0.279944503	0.211278628	2.221824465	5.14312344	0.340248963	2.662275417	4.25360579	0.477704489
min	1975	1	1	0	1	0	0	0	0	0	0	0	0	1	1	0	2	-9	-9
25%	2012	4	8	0	1	0	1	1	0	0	0	1	0	2	3	0	5	-9	0
50%	2014	7	16	0	1	0	1	1	1	0	0	1	0	3	4	0	6	0	0
75%	2016	9	24	0	1	0	1	1	1	1	0	1	0	3	12	0	6	1	0
max	2018	12	31	1	5	1	1	1	1	1	1	1	1	9	22	1	13	1	1

After that I employed a pairwise correlation heatmap to show linear relationships.

I found out that the target feature of the first problem has a very strong negative linear relationship with **criteria_3**. This means when the value of doubt terrorism is No, then the value of **criteria_3** is Yes which is the attack targeted non-combatants.

PAIRWISE CORRELATION HEATMAP

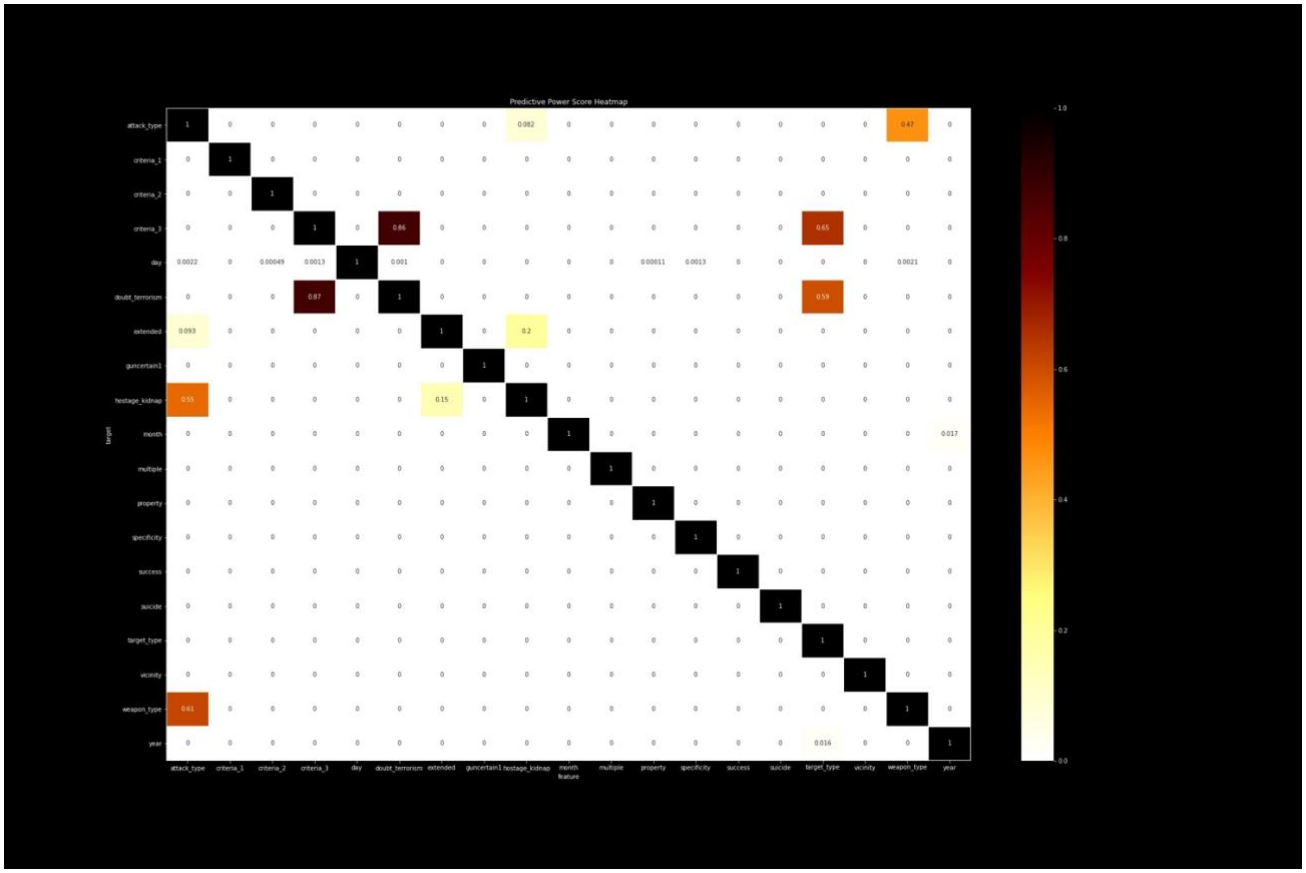


I also employed the Predictive Power Score to reveal relationships that are non-linear and were not revealed when I computed the pairwise correlation.

Here I learnt that **criteria_3** is a strong predictor of our target followed by **target_type** feature.
Hypothetically, to predict whether an attack of an act of terrorism or not the type of target will play role and also whether the target involved non-combatants will play a huge role.

For this reason, I find it suitable to begin modelling with these two attributes as our features and **doubt_terrorism** as our target.

PREDICTIVE POWER SCORE HEATMAP



ALGORITHMS AND TECHNIQUES

I have implemented a range of default algorithms in the course of this project. Initially I began using Decision Tree because it rarely overfits and is the most desirable given this is a binary classification problem. After that I implemented a few ensembles such as Random forests achieve a reduced variance by combining diverse trees hence generating a more robust model.

Both Gradient Boosting and Extreme Gradient Boosting are effective in reducing model bias. They combine weak models to produce a powerful ensemble.

Bagging classifiers improve with respect to a single model, without making it necessary to adapt the underlying base algorithm. As they provide a way to reduce overfitting, bagging methods work best with strong and complex models i.e. fully developed decision trees, in contrast with boosting methods which usually work best with weak models i.e. shallow decision trees.

I also went ahead and implemented Support Vector Machines which is versatile, memory efficient and utilizes subset of training points in the decision function known as support vectors.

BENCHMARK

Using the features I gathered during the Exploratory Data Analysis phases, I benchmarked the models I thought would be viable in my situation.

Algorithm	Accuracy Score
Decision Tree	0.8753
Random Forest	0.85

III. METHODOLOGY

DATA PREPROCESSING

The dataset contains 135 dimensions and 191,463 data points. Most of the classes in the dataset are imbalanced.

After calculating the dimensions of the dataset, I went ahead and printed more information about and I discovered there was 74 numerical features, 57 categorical features and 1 datetime feature.

The dataset contains a lot of missing values in 104 out of 135 features.

Due to the problem statement, I decided to group the data by country to retrieve dataset that is relevant to Somalia.

I thereafter renamed the columns to more descriptive names and dropped missing values column wise and dropped irrelevant columns.

I created a utility function to drop 0 in the day and month columns because in the occasion where the day or the month the attack occurred was not known, the complier assigned zero.

Before 1997, instances where there is essentially no doubt that the incident was an act of terrorism was recorded as -9, so I employed a for loop to loop through the column and replace -9 to 0.

I also introduced a new feature called Date using year, month and day features.

This is a datetime feature I engineered using pandas utility function **to_datetime**.

The final dataset I achieved had 4,656 entries and 26 columns ready for feature scaling.

I employed the StandardScaler class of the Pre-processing module found in the scikit-learn package. This scales the data to zero mean and unit variance making it standard normally distributed data.

Implementation

I created a utility method that makes use of scikit-learns pipeline module. I chose to employ this class method simply because it avoids test data to leak into training data.

The first step in the pipeline is scaling the data to zero mean and unit variance by using the StandardScaler class method.

The next step is to instantiate suitable algorithm and fit it to the train data.

After the model is created, the next step is making predictions and evaluate with the test data.

Then we generate metrics by computing:

- Accuracy score
- Precision score
- Recall score
- F1 score

I also experimented with a lightweight ML workflow framework developed by **Iterative Inc** known as DVC.

It is a lightweight, git compatible, reproducible, storage agnostic and end to end machine learning pipeline framework that is impeccable and easy to use and tracks metrics and failures over the course of the experiments

I created 4 stages of the DVC pipeline system:

- Data Exploration
  - Retrieves data from remote storage
- Feature extraction
  - Perform feature engineering and pre-processing
- Model training
  - Trains a model using a default machine learning algorithm
  - You can perform hyper-parameter tuning on the notebook and update the optimised parameters in the script.
- Model evaluation.
  - Evaluates model trained by previous script and produces metrics file

Using this method, I did not encounter any complication during the entire process.

Refinement

To optimise our model, we will employ GridSearch class method from scikit learns model selection module. This will conduct an exhaustive search on parameter grid provided so that it may reveal the most suitable parameters that will give the best accuracy score.

To make this happen, I created a utility function that performs the hyper parameter optimisation process.

Algorithm	**Optimised Score**
Decision Tree	0.9887
Random Forest	0.99
Gradient Boosting	0.98
Bagging Classifier	0.98
Support Vector Machines	0.98
Extreme Gradient Boosting	0.98

IV. Results

Model Evaluation and Validation

Finally, after careful modelling I made use of 20% of data I set aside for testing the accuracy of the model.

I also computed the precision, recall and f1 score of the model.

Furthermore, I implemented a confusion matrix to determine the sensitivity and selectivity of the model.

Using the confusion matrix, I was able to detect the miss rate (false negative rate) and fall out (false positive rate) of the model. This is was critical in detecting anomalies presented by the model.

Justification

After intensive and iterative modelling, I achieved the following results

Algorithm	**Final Score**	**Benchmark Score**
Decision Tree	0.966738	0.8753
Random Forest	0.987124	0.85
Gradient Boosting	0.984979	-
Bagging Classifier	0.982833	-
Support Vector Machines	0.986052	-
Extreme Gradient Boosting	0.982838	-
Nearest Neighbors	0.980687	-

I believe Random Forest was the model that performed the best compared to the Decision tree. Random forests achieve a reduced variance by combining diverse trees hence generating a more robust model.