Stakeholder Report

Problem statement:

We want to use the ‘Global Terrorism Database’ dataset to gain insights into terrorism attacks. Using unsupervised learning we hope to find that some features says more about attacks, and are more defining when it comes to predicting certain features in the acts, f.ex. The number of casualties or how bad/successful its probable to be. Using supervised learning we wish to predict the number of casualties given the features we have selected.

Description of data acquisition:

The data we used is aquired trough Kaggle, and is called, “Global Terrorism Database. The data was collected through news articles by the publisher, and the author of the data has the following to say about it.

The following text is taken from “https://www.kaggle.com/START-UMD/gtd”

“Context

Information on more than 180,000 Terrorist Attacks

The Global Terrorism Database (GTD) is an open-source database including information on terrorist attacks around the world from 1970 through 2017. The GTD includes systematic data on domestic as well as international terrorist incidents that have occurred during this time period and now includes more than 180,000 attacks. The database is maintained by researchers at the National Consortium for the Study of Terrorism and Responses to Terrorism (START), headquartered at the University of Maryland. More Information

Content

Geography: Worldwide

Time period: 1970-2017, except 1993

Unit of analysis: Attack

Variables: >100 variables on location, tactics, perpetrators, targets, and outcomes

Sources: Unclassified media articles (Note: Please interpret changes over time with caution. Global patterns are driven by diverse trends in particular regions, and data collection is influenced by fluctuations in access to media coverage over both time and place.)

Definition of terrorism:

"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."

See the GTD Codebook for important details on data collection methodology, definitions, and coding schema."

Data cleaning:

Chosen features:

Timestamp	Year-Month-Day
Duration	For those that lasted more than a day, how long? Calculated from resolution and extended.
Longitude	
Latitude	
Success	Was it successful y/n
Nkill	Number of killed victims and attackers
Nkillter	Number of killed attackers
Nwound	Number of wounded
Property	Was there property damage y/n
Vicinity	Was it within or near a city y/n (n= inside a city)
Suicide	Was it a suicide attack y/n
Claimed	Was it claimed by a known group y/n
Gname	Group name if any
Individual	Was it a lone wolf attack y/n
Crit1	Was the act aimed at obtaining a political, economic, religious, or social goal
Crit2	Was the act aimed at, intention to coerce, intimidate or publicize to larger audience(s)
Crit3	Was the act aimed, outside international humanitarian law
Region	World region
Attack type	How was the attack carried out
Weapon type	What was used
Target type	What was attacked

The four last variables have been turned in to dummy variables in our dataset.
The above features, have been chosen by us, because they appear to be, the ones most likely to say something interesting about the attacks, and therefore, be useful in modeling of the data.

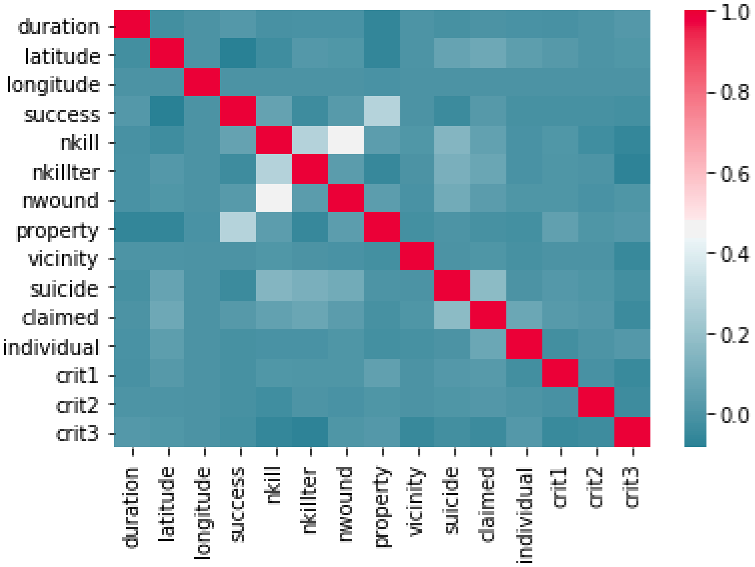
Through the analysis, its hoped, that something in the attacks, are so defining of the act, that it becomes possible to predict or determine, certain features of the attacks, that they might be, better responded to. Gaining more insight into the terrorist attacks, and the style and targets they chose, may give states, and private citizens the insights to better guard themselves.
The aim is to do so through visualizations and prediction models. One last thing to note, on this part, is that the definition of a successful terrorist attack, is not that, there were casualties, but merely, that they succeeded in carrying out their intention. As such, a terrorist wanting to sow

fear, by attacking the police, will be counted as successful, as long as he fires at them or their buildings.

Data exploration

By manipulating and sorting the data we ended up discarding a lot of it since it didn't match the criteria. We then did some exploration and analysis on the remaining data.
In this we found a few interesting points, and some points, that might not come as a surprise to most people.

The first thing we wanted to do with this, was to see if there were any correlations in the data that had been collected.



By looking at this heat map that has been generated over the entire data, we can see that the data doesn't really have any strong correlations through out. There are a few very light correlations around between some of it, but mostly it doesn't correlate significantly. The ones that are interesting are worth mentioning are, the correlation between

1. The killed and the wound.
2. Killed terrorist and killed victims
3. Suicide attacks and killed,wound,killedterrorist
4. If its was claimed and number of killed, wounded
5. Claimed and suicide
6. Individual and claimed.

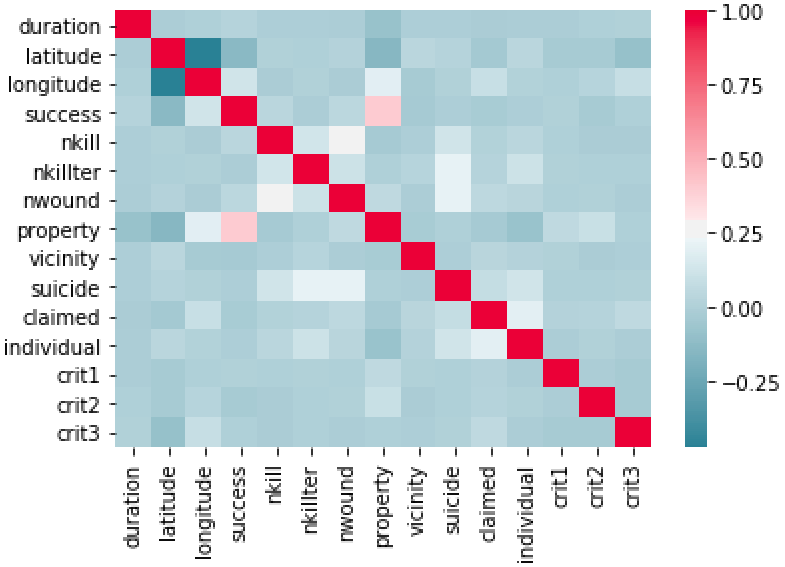
Now the first 2 correlations seems pretty logical given that there are probably going to be more people dying, the more that people are wounded or vice versa. Also the number of killed terrorists in the attacks correlation with the number of killed people could be explained by the fact that if terrorist die in the attack, it is more likely to have been a bigger event.

The third correlation with the suicide attacks and killed,wounded and killed terrorists is intresting, due the fact, that it seems to be, that when the attacker doesn't intend to get out alive of the attack, it becomes more dangerous.

The fourth one appears, to indicate,that when the attacks are claimed by a group, they have a tendency to have been more deadly than if they are not.

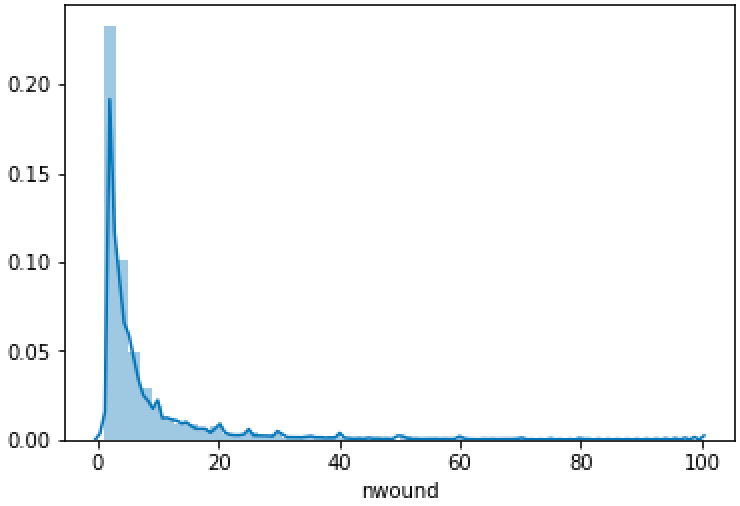
This also ties into the next one, that looks at claimed and suicide attacks.

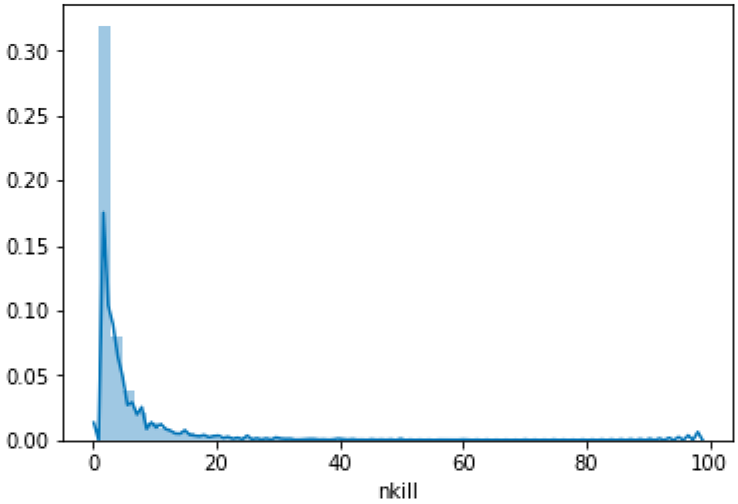
The last one is whether or not it is and individual or a lone wolf attack. Wich of course correlates.



When looking at the same data, for Western Europe only, it can be seen, that it's almost the same, but that some of the correlations become a little less clear as to what is going on.

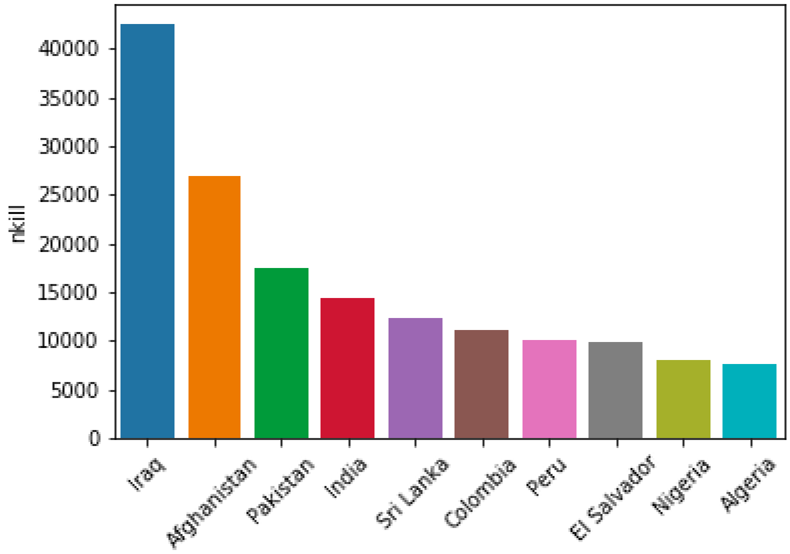
The next thing that becomes interesting to look at, is the distribution of the number of wounded and number of killed.



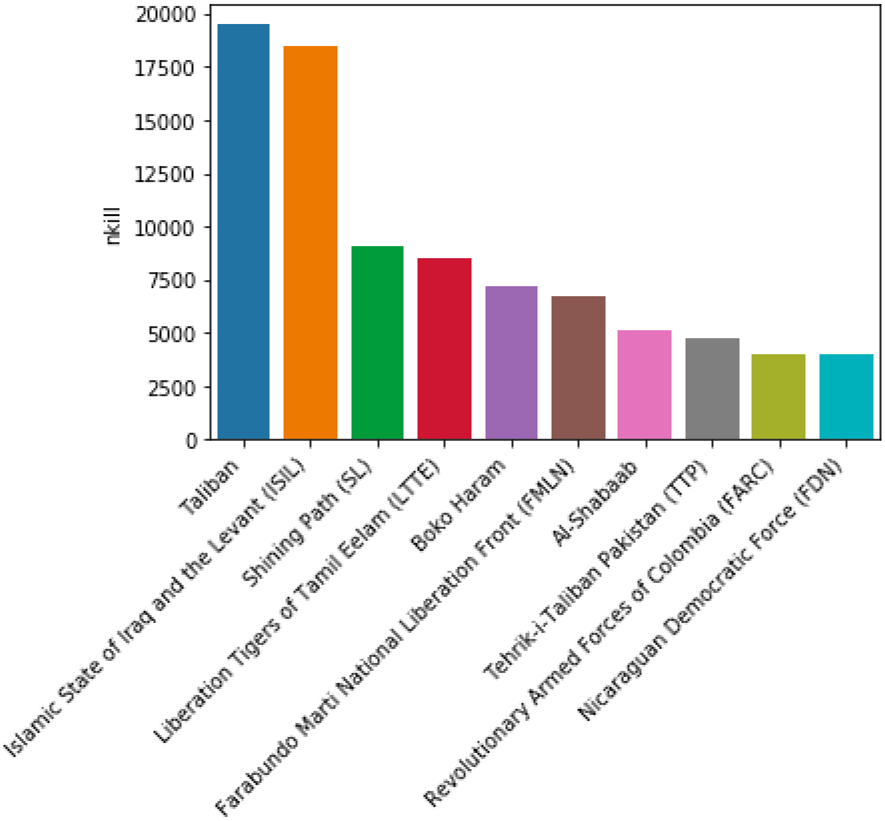


By looking at the two graphs it can be seen, that there neither are normally distributed, and that, most of the attacks, have very few wounded or killed. As mentioned in the colab files, this graph only include attacks with between 0 and 100 killed/wounded to get the most easily interpretable images of most of the attacks that cause casualties. A lot of the attacks in the full data have 0 wounded or killed. Meaning that most terrorist attacks don't cause a lot of damage.

When looking at the top 10 countries for people killed in attacks, it becomes clear that these countries are placed in regions of great geopolitical interest or unrest. It is therefore not a surprise that these countries have the most people killed in attacks.

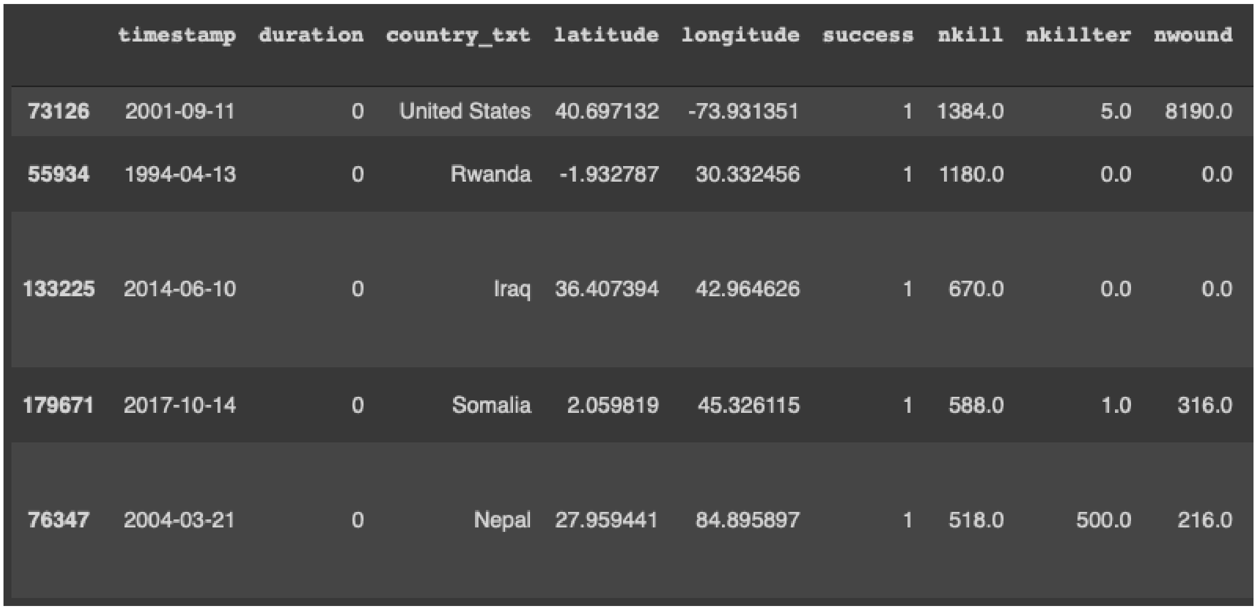


The image remains the same, when we go further and look at, what groups have killed the most people. The groups with most people killed mainly operate in one or more of the top ten countries. With the exception of some of the last ranks in the deadliest terror organisations.



What we can tell from this quick look at the data, is that not a lot of people are necessarily killed in a single attack, and that groups operating in unstable regions seem to be deadlier.

This points to the fact, that deadly attacks on the western world are rare (outliers), but also that some of the attacks such as 9/11 are among the deadliest. Here we see that the deadliest terror attack in the data was in fact 9/11.



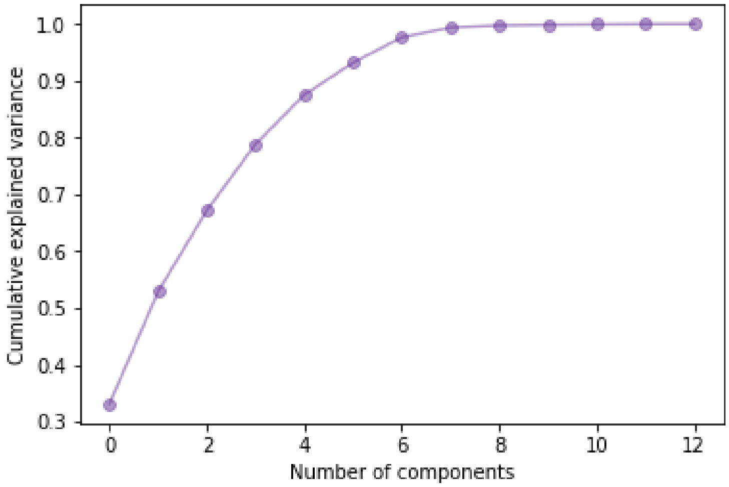
	timestamp	duration	country_txt	latitude	longitude	success	nkill	nkillter	nwound
73126	2001-09-11	0	United States	40.697132	-73.931351	1	1384.0	5.0	8190.0
55934	1994-04-13	0	Rwanda	-1.932787	30.332456	1	1180.0	0.0	0.0
133225	2014-06-10	0	Iraq	36.407394	42.964626	1	670.0	0.0	0.0
179671	2017-10-14	0	Somalia	2.059819	45.326115	1	588.0	1.0	316.0
76347	2004-03-21	0	Nepal	27.959441	84.895897	1	518.0	500.0	216.0

Unsupervised Machine Learning (Dimension Reduction and Clustering)

In order to make sense of the data, we use unsupervised machine learning to reduce the amount of features we need to describe the data. In order to preserve the information in the data while reducing features, we need to create new features which contain all the variation. To do this we use two different dimension reduction methods: Non-negative Matrix Factorization (NMF) and Principal Component Analysis (PCA).

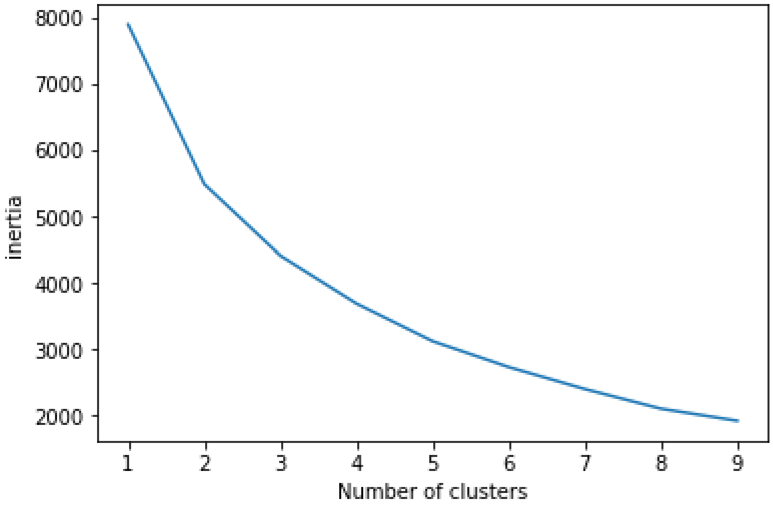
NMF is recommended to use on Categorical variables i.e Gender, Country or in our case AttackType. NMF creates new features which are characterized by patterns in the old features. Since the NMF variables explain patterns in the old categorical variables, the old variables become redundant and can be removed. We then run all the variables through a PCA to find, the most important components and rank them.

The PCA shows that 87% of the variance can be explained by just five principal components.

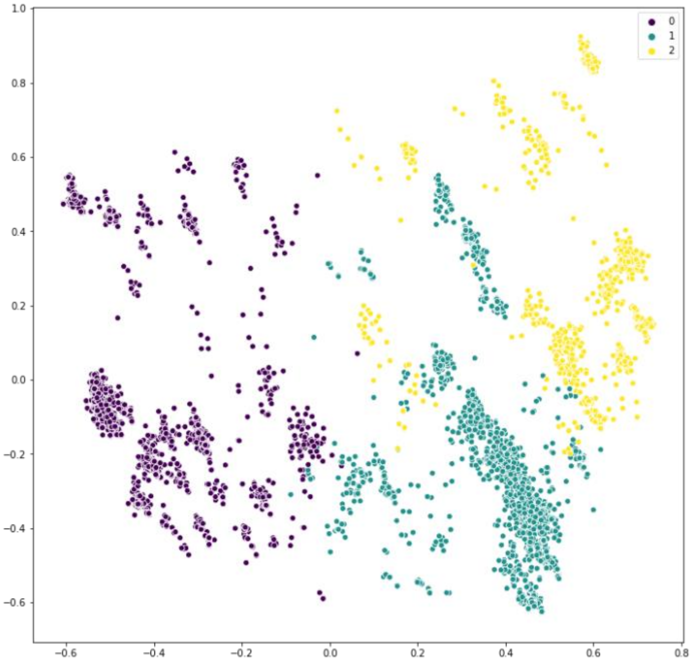


PCA unfortunately loses some interpretability, and these features don’t directly translate into our original features, but by clustering how different data points fall within our two most important features, it becomes possible to give all our observations a cluster and thus describe how important they are to our most important features.

We do this with the K-means clustering. By testing the inertia of the model we can find the best number of clusters using the elbow test. The elbow test looks at when the inertia of the model starts to flatten after adding more clusters.



Looking at the graph the is a visible elbow at 2, meaning 3 clusters should separate the data somewhat nicely. Plotting the data with the two first principal components at the x and y axis, and coloring the data points according to their cluster, we get this plot:



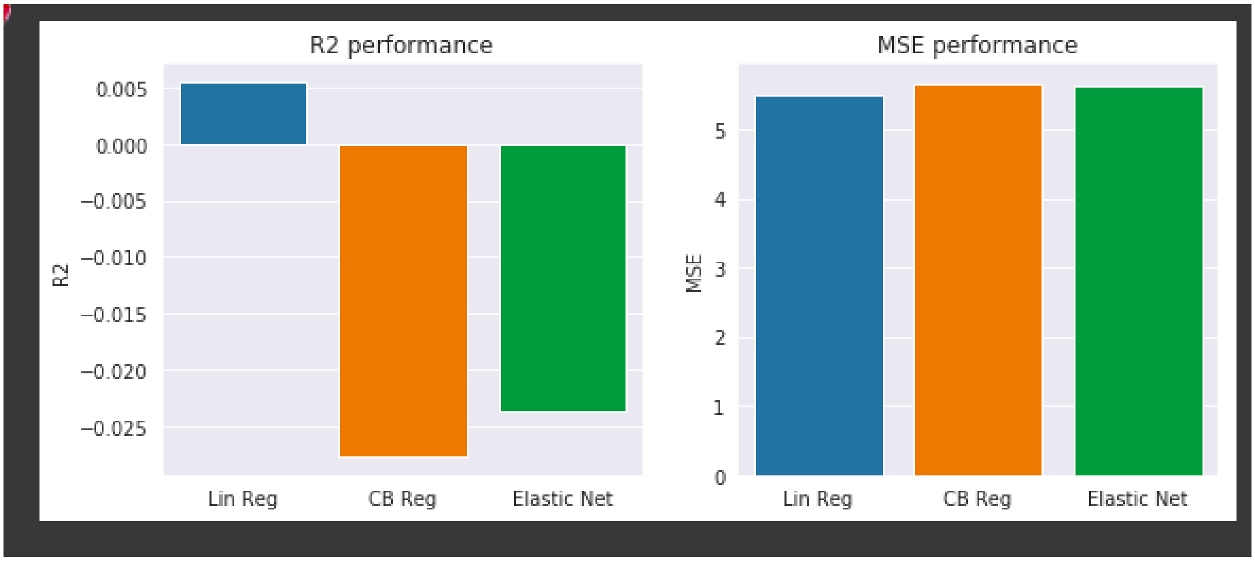
The clustering reveals that one cluster is characterized by having the largest amount of wounded with some people killed. This cluster is also characterized by bombings and property damage.

The second cluster is characterized by having the largest amount of killed, with some wounded. Its is also characterized by Successful Assassinations and firearm attacks. This cluster also has the longest durations of the patterns seen in the data.The final cluster is characterized by low killed and wounded. It is also characterized by successful incendiary attacks where facilities were hit.

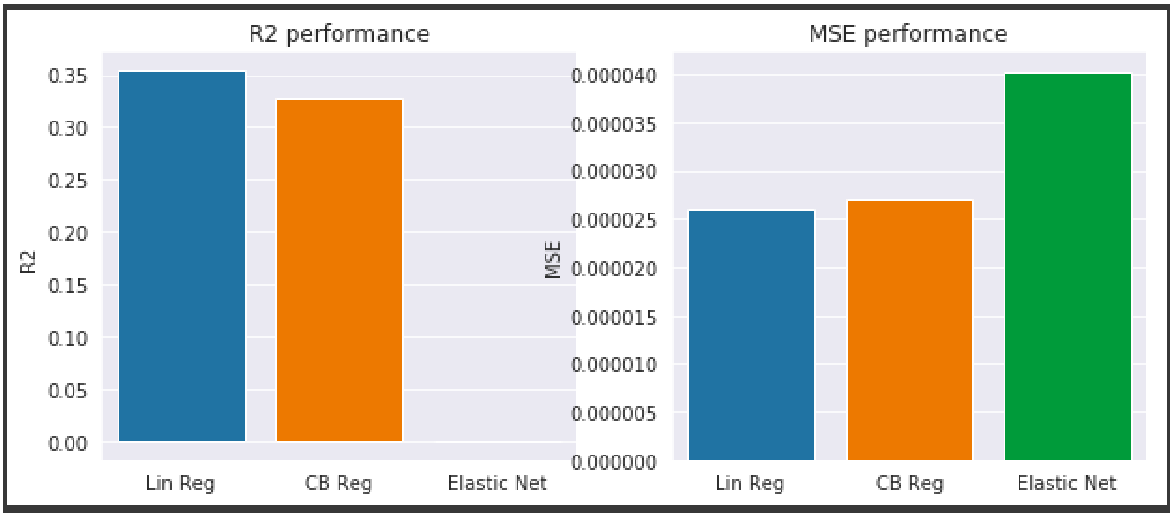
Supervised learning

During the supervised learning, we learned that the data, we were using, did not perform well when used as input in any of the three supervised learning models we tried. The different models all showed poor performance, when it comes to predicting the number of killed people.

The image below shows the r2 and mse of the three different models when fittet on data that has been reduced using NMF. It can be seen that the different models all performed poorly. Negative R2 values are generally an indicator that the model is not appropriate to use on the data, so the Liniar regresson seems to be the best model here.



When we look at the data without that has not been reduced, we see a drop in the MSE indicating the models work better, and now the r2 are positive (or approximately 0).
But the r2 are still a long way from 1, so the models still aren’t great, and elastic net should still be disregarded.



Conclusion

From a combination of our dimension reduction and clustering we were able to find that our data could be characterized by three main types of attacks, which resulted in either many wounded(1), many killed(2) or few casualties(3).

In the supervised learning part we were not as successful, we did not succeed in fitting any model great predictive qualities. We suspect that this is in part due to the fact that terrorist attacks are a very complex and chaotic events, that have many more features or parts than we have looked at here.