Introduction

Terrorism is open to different interpretation. Depending on who you ask, the responses you get would be different. Most will probably say it is a negative foreign impact on lives and properties. However, recent incidents like in the case of Las Vegas shootings on October 1, 2017, and that of Texas church shootings a month after tells a different story. Could such be classified as terrorist attacks based on the nationality of the perpetrators? Added to that, the rate at which such occurrences are happening are becoming of great concern. One wants to be sure you are safe when going about daily endeavors. Businesses - small and large need to estimate their risk related to such occurrences and government establishments are looking for ways to efficiently allocate resources to help combat or prevent such activities. Knowing where the causes lie and being able to foresee future occurrences form the bases of this project

Materials and Methods

Data extraction and cleaning

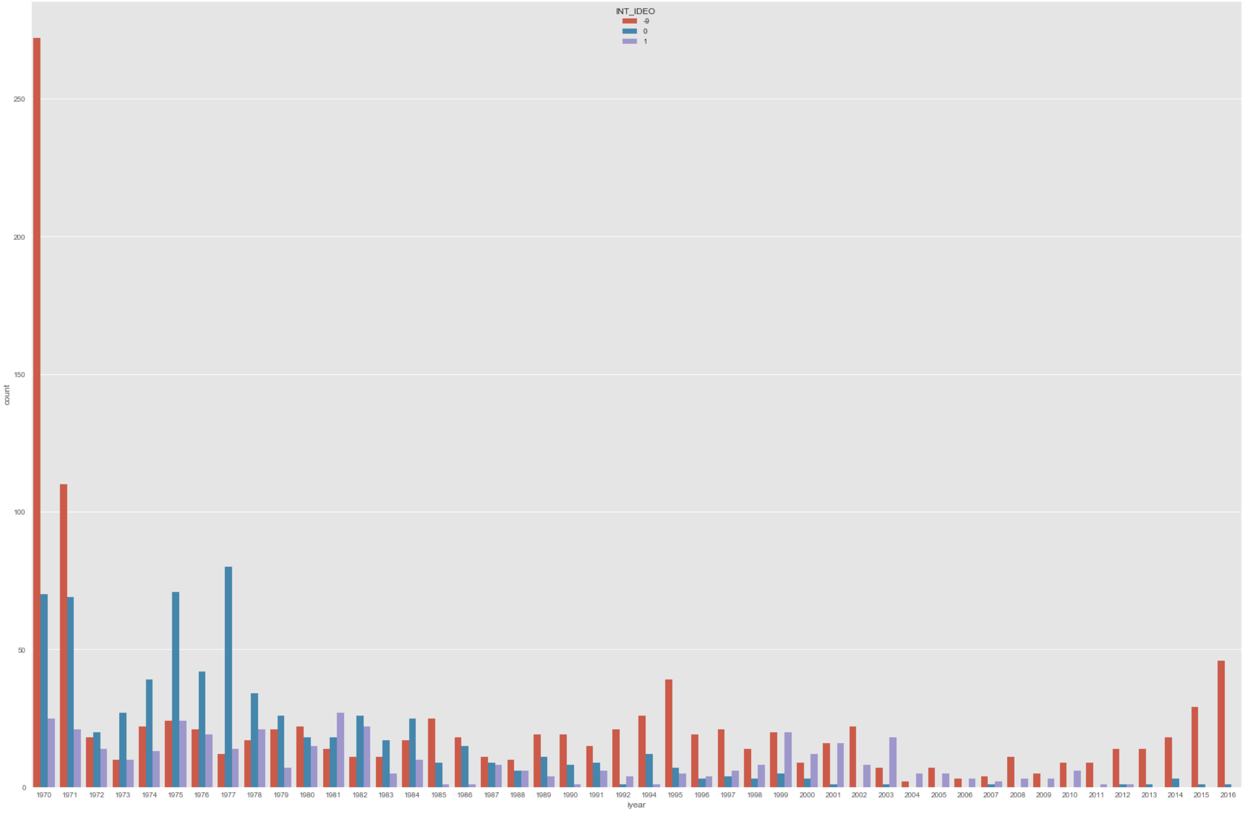
The dataset used in this analysis is from the Global Terrorism Database (GTD) maintained by the University of Maryland. It includes all reported and verified cases terrorist attacks in the world. The original dataset had about 170, 000 cases from a period of 1970 to 2016, but I restricted that to cases in the US because it was of more concern to me. There so many missing values mostly seen in a real-life dataset with lots of string commented values. It is also significantly imbalanced. With this comes data cleaning.

My initial step in the cleaning the data was to get an idea of what I was up against with an initial descriptive statistics and exploratory data analysis. Most of the variables, especially the dummy variables seem right skewed, but nothing severe since the spread of the data points is not overly high. Skewness translates to the presence of some outliers in the distribution, which sometimes can affect results of the analysis, and may need further investigation. It was also essential to eliminate commented string features and features with a lot of missing values. There is no rule of thumb here, but my approach was to remove any factor that has up to 80% missing values because it is probably not useful because of its insignificance.

Next up was to create, since we already have a target variable, it is now advisable to create an out of sample test set data. This act usually prevents what is called an information snooping bias. The premise is that the human brain is known to be a fantastic pattern detection system, which means that it is highly prone to overfitting. So, we set the test set data aside that we will not touch at all, and continue with the training set data. The test data set is used as an out of sample dataset to validate the model.

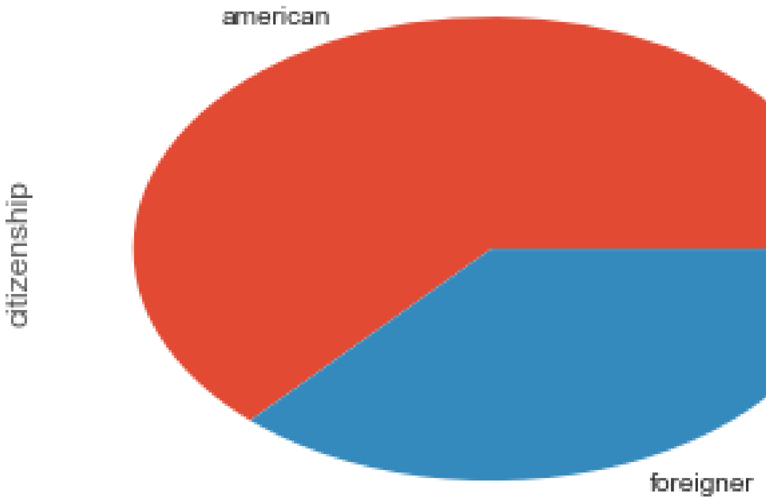
Exploratory Data Analysis

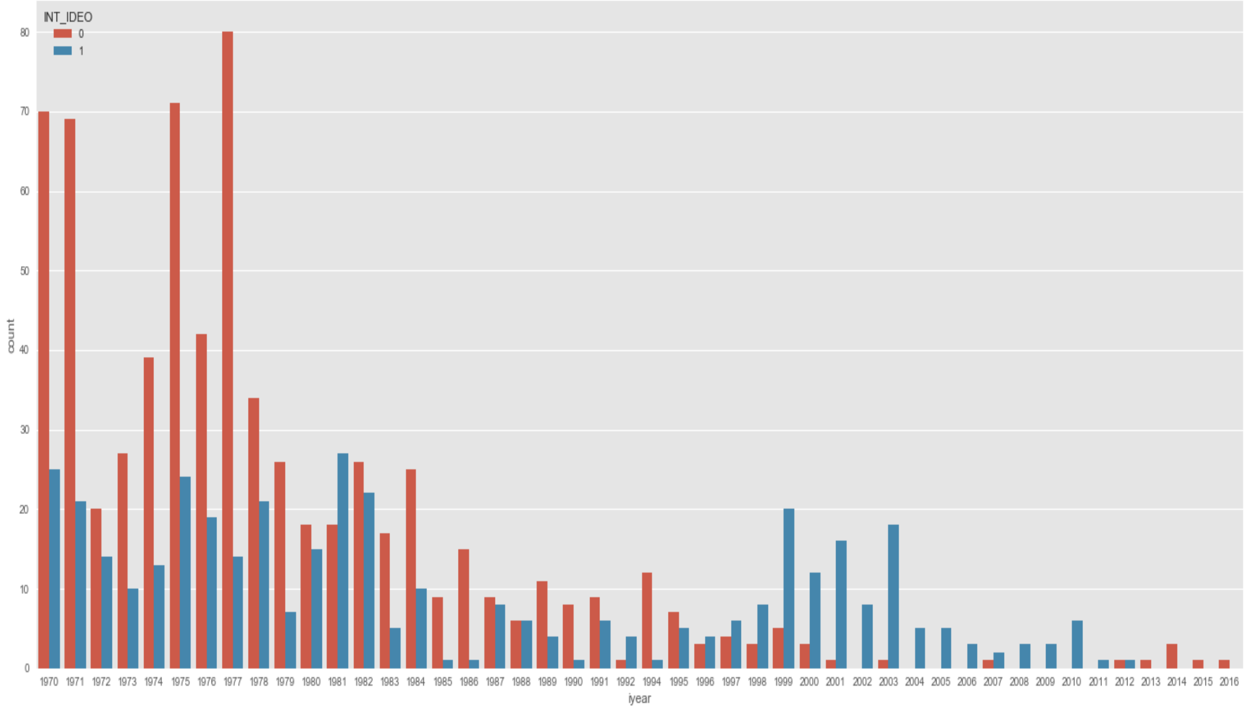
Exploring the dataset a little bit deeper using some plots, we see some trends in our datasets for the classes of the target variables and proportional differences as shown below answering some of our questions.



The whole dataset had cases for 18 % Americans, 32% foreigners and those whose perpetrators are unknown of 50%. The 2-dimensional chart above shows the overall trend of attacks in the US, with zero (0) representing Americans and a one (1) being a foreign citizen. This plot shows overall attack trend was much higher 30 to 40 years compared to what we have now, probably because of social awareness and technological advancement in curbing such occurrences.

The pie plot below shows 35% of the known cases are Americans. This plot shows attacks by foreigners is about 35% of known attacks in the USA. It makes sense to think that it is possible for the unknown instances to swing the pendulum either way. However, we demonstrate in the inferential statistics part of the overall tasks that if one takes the known cases as a sample of the whole population that comprises all instances, ensuring we meet certain statistical conditions, we can generalize findings from the known to the unknowns.





For cases that are known, Americans perpetrated more attacks against fellow citizens from 1970 up till about 1997. Around 1997, foreign attacks on home soil equaled and surpassed that of local attack. This alien attack on home soil peaked in 1999, went down a little around 2001 and then dropped significantly in 2004. After declining in 2004, it remained somewhat constant at that level till 2010 before falling dramatically again around 2011. After this drop, it has continued at this level to date. While attacks by fellow Americans remained at the same level as that of foreign nationals as of 2015, nothing is known yet about years after that. It suffices to say that going by recent occurrences after 2015; I guess that attacks by Americans on fellow citizens have surpassed that of foreigners on the home soil. This analysis will be updated with that information as soon as the University of Maryland updates the Global Terrorism Database.

In conclusion, this analysis shows a correlation between terrorist attacks and nationalities of individuals. Correlation does not necessarily mean causation without a designed experiment. However, residents in the USA - both citizens and non-citizens, should change their mindset in that a terrorist attack is as likely to be perpetrated by an American as a foreigner. Even more so an American on home soil.

Statistical Inference

From the exploratory data analysis already carried out, we know that INT_IDEO - nationality of attackers, which is our target variable, has some association with time series, "iyear." The correlation matrix computed shows association between some predictors, and also a strong a very strong positive linear relationship between ‘INT_LOG’, ‘INT_ANY’ and the target variable. This may need further investigation if model created does not meet required outcomes. Verifying if the relationship is genuinely significant and checking if time series is the most significant feature out of the bunch, we start off with a hypothesis test and then double check our work with a linear regression model.

The two-sample t-test should be appropriate for this because we are comparing assumed independent measures on observational units where the standard deviation is either unknown or not given.

Central Limit Theorem (CLT) also applies because the t-statistics follow a distribution which is the t-distribution. This distribution arises from a plot of every t-statistics of every sample randomly selected from the measured population measured. Population measured being each variable of the observational units. And, this distribution is assumed to be normal following a bell curve with the t-statics values in the horizontal and the probability density function (PDF) on the vertical. The more random samples one selected from the population, the more the curve tend to a finer and perfect normal distribution. This theory exemplifies the CLT in that even if a population is not normal; the sample mean distribution itself, which is the plot of the sample mean from randomly selected samples from the population tends to a normal distribution.

My finding was that the p-value of the 2-sample t-test was not significant at the significant level α = 0.01 and the confidence interval of the portion of both nationalities did not include zero, so we reject the null hypothesis Ho. And, report that there is a statistically significant difference between the overall average number of Americans perpetrating terrorist attacks compared to foreign nationals. I also discovered that the time series variable is probably not the most significant predictor in the overall model. Some other features are slightly better. For example, we see that "iday," which is time of the day, and the "natlty1" are somewhat more significant. This does not tell the whole story, because they are summary statistics. Other factors need to be taken into consideration to make a firm conclusion.

Building the machine learning algorithm.

My approach to this was comparing different algorithms to see how they performed regarding accuracy and going with best. Since our dataset is significantly imbalanced, care was taken to balance the training dataset before training the model. We now carried out the prediction on the imbalanced out of sample data we created at the start of the analysis. This approach shows if the model generalizes well to any dataset it has not seen before, balanced or imbalanced. The model was also cross-validated to ensure we are not overfitting the training dataset using the StratifiedKFold and KFold validation techniques.

In addition to checking for accuracy, because of the imbalanced nature of the dataset, accuracy alone is not enough to measure how well a model will do on out of sample dataset. Performance metrics like the Confusion Matrix allows one see the likelihood of the model distinguishing between classes of different observations. In other words, if we randomly select an example from each class, what's the probability that the model will be able to "rank" them correctly?

Results and Conclusion

With the Random Forest algorithm, we achieved about 95% accuracy cross-validated in comparison to other tried models, showing that our model did not overfit the training dataset. Regarding model performance, we see that:

- for cases of all nationality of terrorism perpetrator group same as the nationality of the target(s)/victim(s), represented as 0 with index 0, we see that the model got the prediction of 184 observations correctly, and seven wrong.
- for cases of the nationality of terrorism perpetrator group differing from the nationality of the target(s)/victim(s), represented as 1 with index 1, we see that the model got the prediction of 80 observations correctly, and 11 wrong.

These results are useful knowing sufficiently well that there are factors that can still improve it that were left out at the start of my analysis due to the need for re-engineering because of the high cardinality and multicollinearity and outlier issues in the model. I will need to revisit those to see if we can improve on this model. Few things I will be looking at are:

- grouping classes in each predictor to about five or six, so that a feature like 'provstate' will be now be divided into regions: Northeast, West, South, Midwest, and Pacific, and 'gname' will be grouped into political, racial, religious or social causes.
- I would also engineer a feature by combining the date into a datetime format, so the 'iyear,' 'imonth,' 'iday' will just be one column.
- doing some research that help create this features, so domain knowledge is important here
- looking closely at outliers and multicollinearity issues. Questions have to be asked if removing some data points make sense or not.
- Considering hyperparameter tuning.