Terrorism is Happening: What Are the Consequences?

Jaeyeon Won

DS 301 Final Report

November 20, 2020

Introduction

Terrorism is "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" (GTD). Whatever the purpose of terrorism is, terrorism is one of the severe threats to the whole world. It has resulted in devastating casualties and profoundly influenced the quality of lives of innocent civilians.

This project is aimed to find the model that best predicts the number of casualties of terrorism. Specifically, the casualty of a terrorist incident would be defined by the total number of the fatalities and injuries from the victims as well as the perpetrators. As predicting the number and magnitude of the casualties is a complicated task, it is important to develop a model that accurately estimates these parameters. This project can especially be helpful for governments in terror-prone regions to understand the behavior of terrorist activities and the factors behind it. The model I would develop can also be used for the governments to effectively design the strategies and be more prepared to deal with the consequences of terrorist incidents ahead of time.

To accomplish the purpose of this project, supervised multiple regression machine learning algorithm was used. The root mean squared error (RMSE), mean absolute error (MAE), R-squared (R^2), and standard deviation were used to diagnose the model performance. Out of these metrics, the RMSE was used as the main metric since it was assumed for the error in the model to be non-linear.

The data are assumed to follow a polynomial regression model. That is, the relationship between the dependent variable and the features is expected to be more complex than a simple linear line. I also assume that the features are independent of each other, and the errors are independent and normally distributed with a mean of 0.

The rest of this machine learning project was developed by following the main six steps:

1. Get the data
2. Explore the data
3. Prepare the data for Machine Learning algorithms
4. Explore different models and shortlist the best ones
5. Fine-tune the model
6. Present the solution

Description of Data

For this analysis, the globalterrorismdb_0919dist dataset was used, which comes from the Global Terrorism Database (GTD), an open-source database that provides systematic information on terrorist activities occurred around the world since 1970. This dataset contains 135 variables for 191,465 global terrorism cases that happened between 1970 and 2018, except for the information for 1993. Due to the issues occurred in data compilation process, the data for 1993 are not present. With the features available, a wide range of exploration is available, including detailed information on the incident, attacks, weapons, targets, perpetrators, and casualties.

Data Cleaning

With the purpose of predicting the number of overall casualties of a terrorist activity, in this report, 2 numerical variables (*nkill* and *nwound*) were aggregated into the casualty variable. The aggregation process simply included adding the total number of fatalities and injuries for each incident.

Rather than including the entire 133 attributes as features, 12 variables that were deemed to be related to casualties were selected as the explanatory variables. For the convenience of analysis, the attributes were renamed into reader-friendly names. Also, in the *weaponType* variable, the level ‘Vehicle (not to include vehicle-borne explosives, i.e., car or truck bombs)’ was shortened into ‘Vehicle.’

There were several missing values for the *victNationality*, *nKilled*, and *nInjured* variables. The missing values were dropped since it did not drastically decrease the number of observations in the original dataset. After this step, the dataset includes 171,626 observations.

The *multiple, success, suicide,* and *individual* variables are binary variables, which are encoded as 0 and 1. To avoid the confusion in the later steps, 0’s were renamed into ‘No’, and 1’s were renamed into ‘Yes.’

Exploring the Data

This section shows the insights of the data by visualizing the trends of terrorist activities in association with the first feature-selected 12 attributes. This step is helpful in detecting potential issues with the data and preparing the data for machine learning algorithms.

General Terrorist Activity Trends

The total number of terrorist incidents from 1970 to 2018 in Figure 1, except for 1993, demonstrates that the terrorist activities have gradually increased from 1970 to 1992. After 1992, it decreased around the world approximately until 2003. However, it started growing sharply after 2004, and especially in 2014, the rate of incidents has reached the peak.

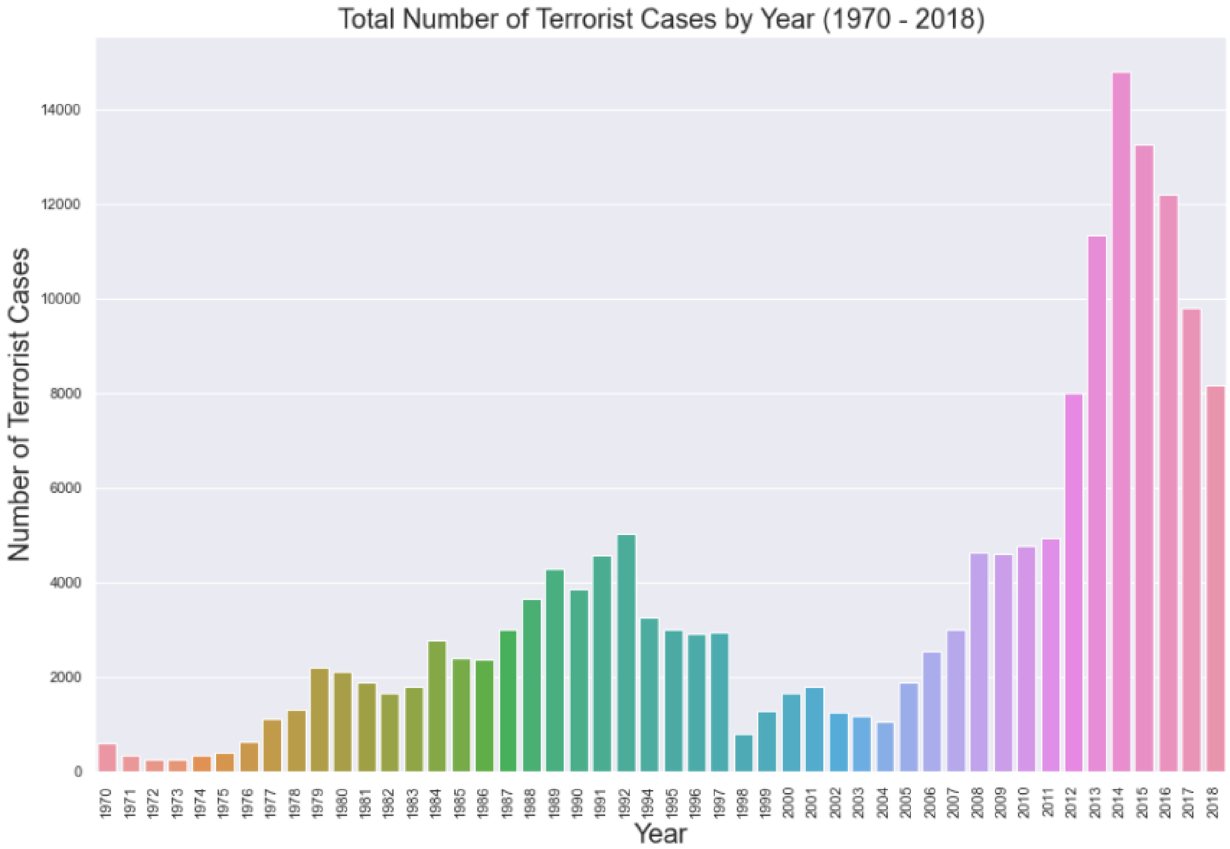

Figure 1: Total Number of Terrorist Cases by Year (1970 – 2018)

Figure 2 shows the total number of casualties from 1970 to 2018, except for the terrorist cases in 1993. Generally, the number of fatalities and injuries have had a very similar trend as the number of terrorist cases. Interestingly, however, there were around 35,000 casualties in 2001, which is an exceptionally high rate compared to the number of incidents.

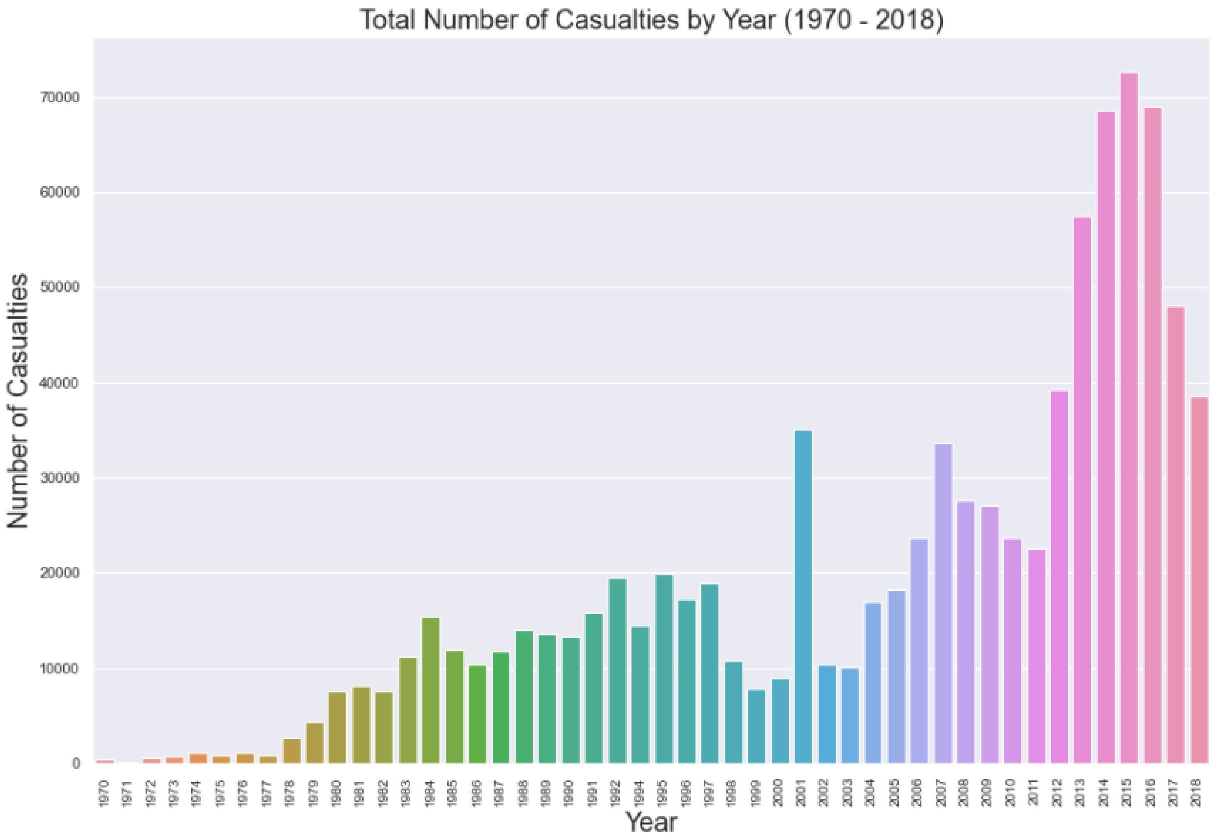

Figure 2: Total Number of Casualties by Year (1970 - 2018)

Numeric Attributes

In this analysis, there is only one numerical attribute, the *casualty.* All features are categorical attributes.

The histogram of the number of casualty (Figure 3) shows only one bar in the entire graph. This is because the occurrences of the casualties have a very high intensity around the range between 0 and 100 while there are several outliers in a very high range. For example, the maximum number of casualties is 12263. With this information, it is presumable that the casualty is heavily skewed to the right.

Applying the log scale to the y-axis, we can see the pattern of casualty better. Based on Figure 4, the casualty still has a heavily right-skewed and unimodal distribution with several outliers between 4000 and 6000 and over 12000.

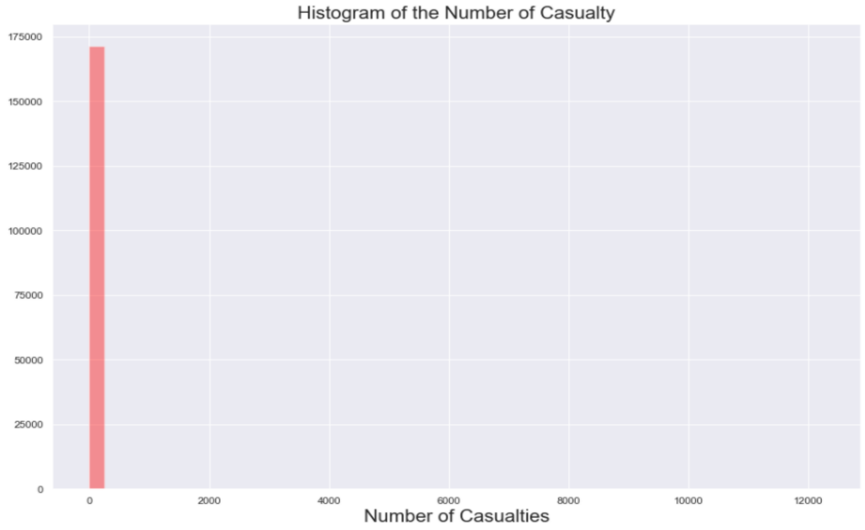

Figure 3: Histogram of the Number of Casualty

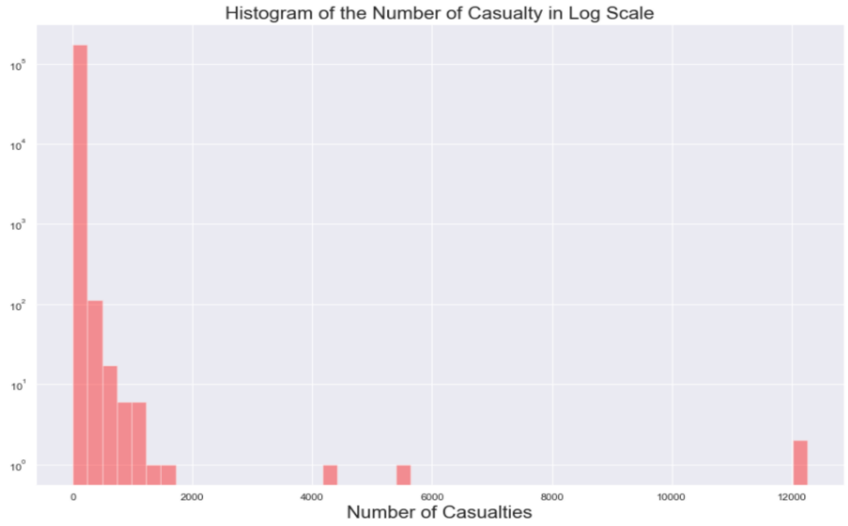

Figure 4: Histogram of Number of Casualty in Log Scale

Categorical Attributes

The distributions of binary attributes are illustrated in Figure 5. While the *multiple*, *suicide*, and *individual* variables have much dense distribution for the *No* level, the success variable has more observations for the *Yes* level.

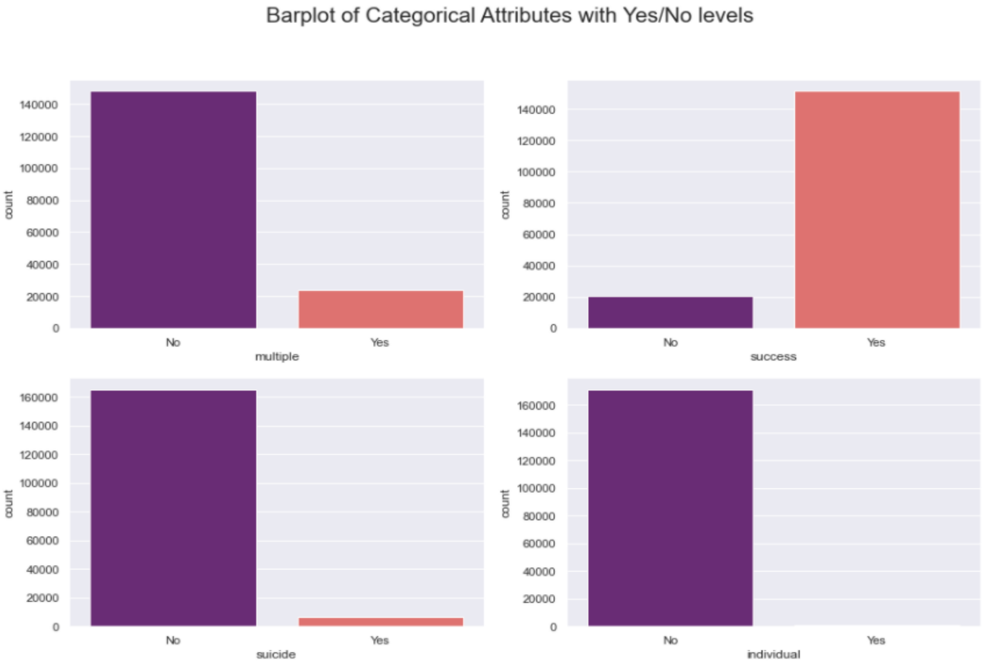

Figure 5: Barplot of Categorical Attributes with Yes/No Levels

There are originally 202 countries in the dataset. The following bar graph (Figure 6) indicates the number of terrorist activities by the top 20 countries. Iraq has had an incomparably high rate of incidents compared to other countries. Pakistan, Afghanistan, and India follow Iraq, in order, with a relatively large number of terrorist attacks. Up to top 7 countries show a distinctly higher number of cases.

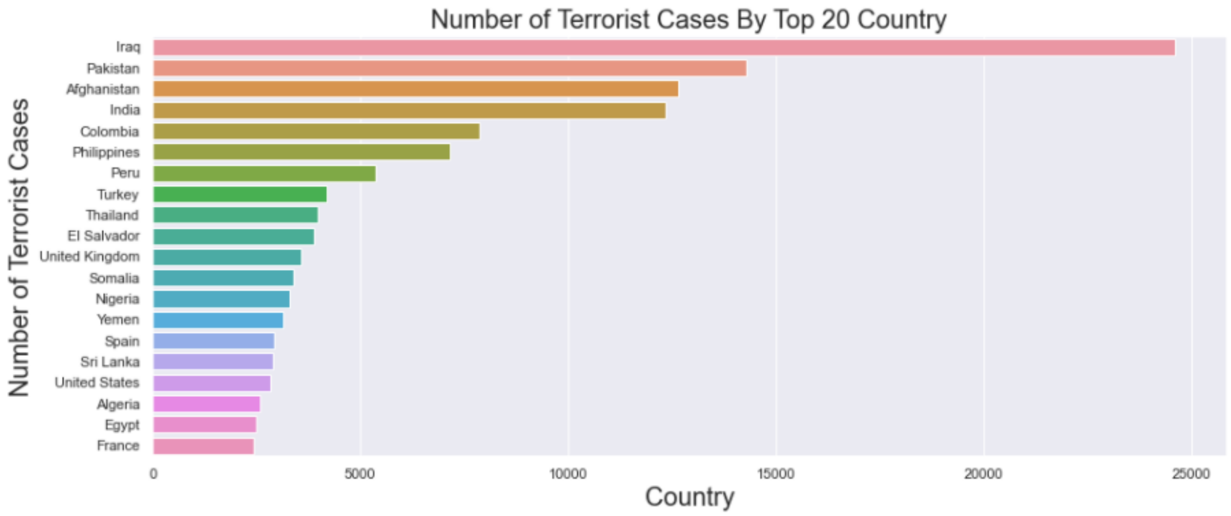

Figure 6: Number of Terrorist Cases by Top 20 Country

Out of 213 victim's nationalities, Figure 7 includes the analysis for the top 20 nationalities. The top 20 victim nationalities have a very similar pattern of the number of terrorist cases as in the previous graph. For example, the victims of the terrorist activities were majorly from the top 6 nationalities, including Iraq, Pakistan, India, Afghanistan, Colombia, and Philippines.

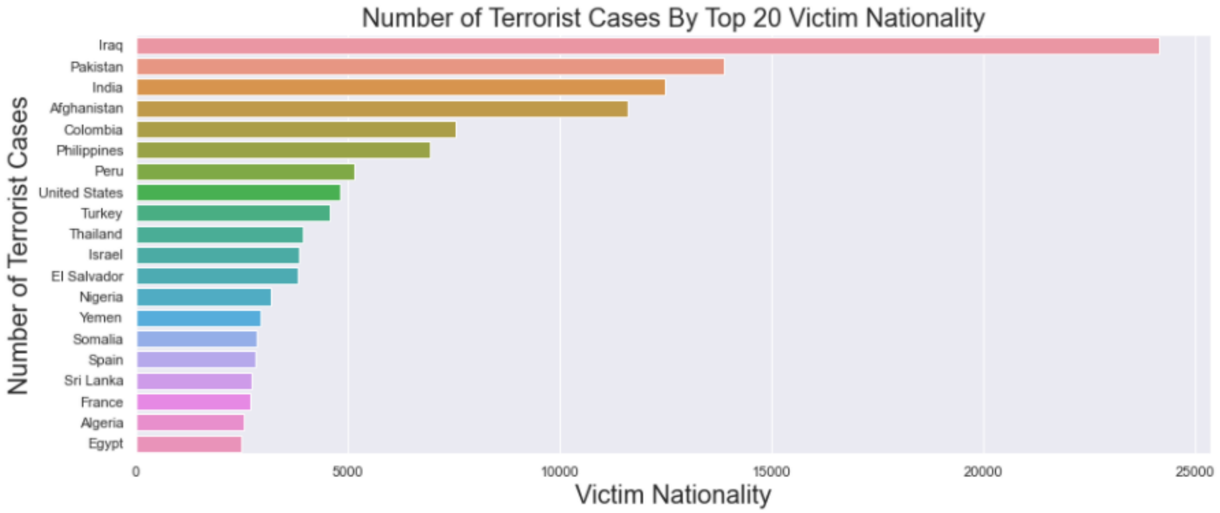

Figure 7: Number of Terrorist Cases by Top 20 Victim Nationality

The plot below (Figure 8) demonstrates the number of terrorist incidents by each region. A large number of terrorist cases occurred in Middle East & North Africa and South Asia, whereas there were relatively very little incidents in East Asia, Central Asia, and Australasia & Oceania.

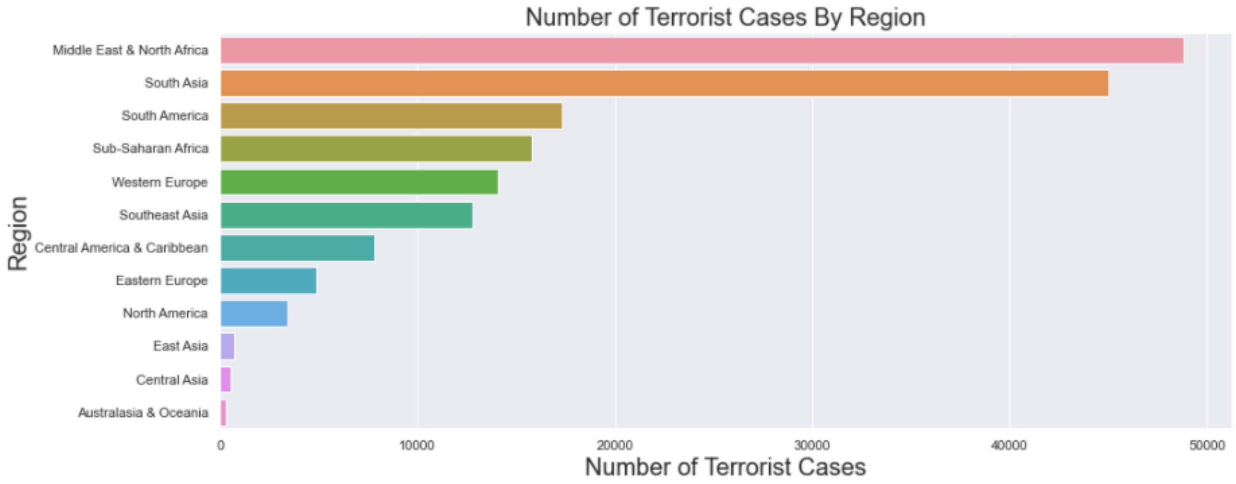

Figure 8: Number of Terrorist Cases by Region

Amongst 3426 different perpetrator groups, information of most of the groups were not available and coded as ‘Unknown’ according to Figure 9. Since it is difficult to explore the distributions of other groups due to the high density in the *Unknown* group, second graph that excludes the *Unknown* category is shown in Figure 10.

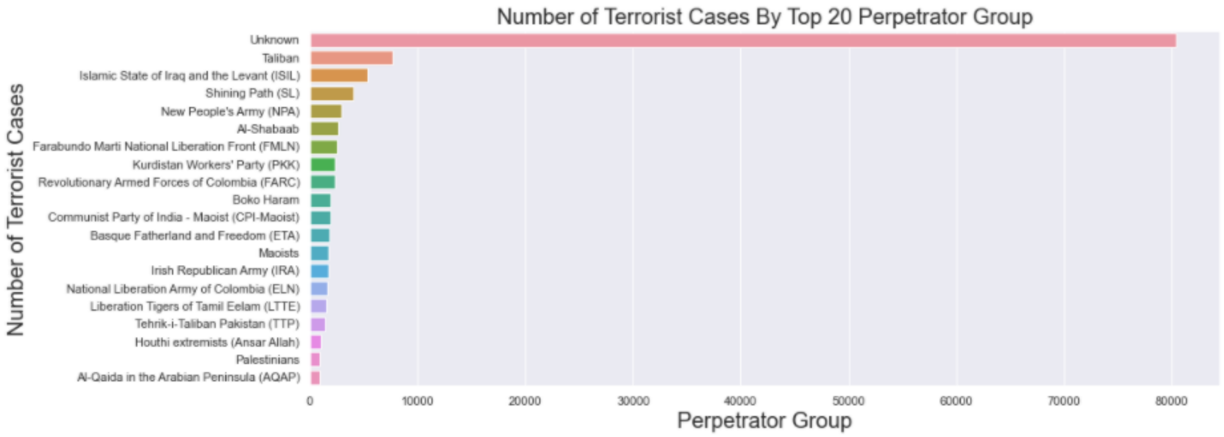

Figure 9: Number of Terrorist Cases by Top 20 Perpetrator Group

Excluding the *Unknown* level, it is clear that Taliban, Islamic State of Iraq and the Levant (ISIL), and Shining Path (SL) are the leading perpetrator groups with relatively higher frequencies of terrorist attacks.

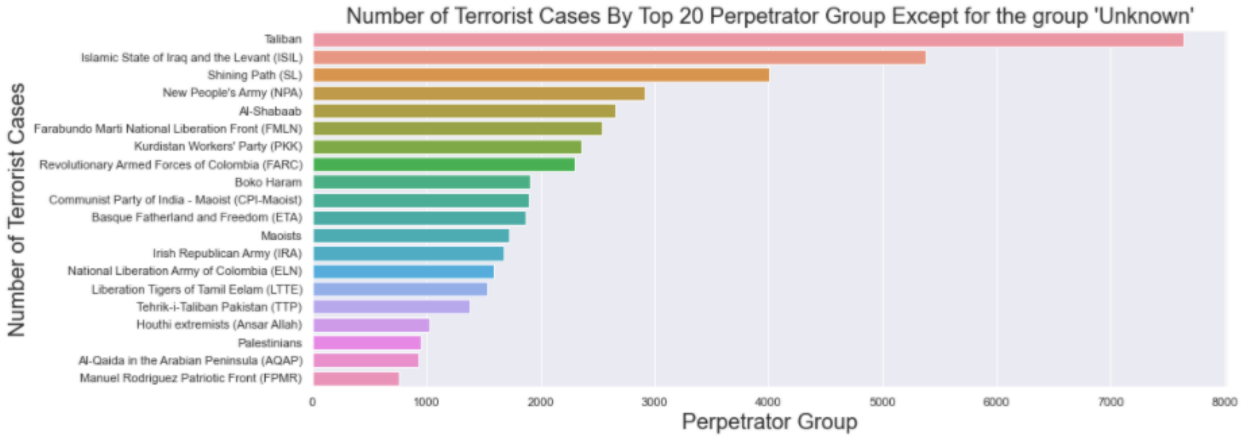

Figure 10: Number of Terrorist Cases by Perpetrator Group Except for the Group 'Unknown'

The three following graphs (Figure 11, Figure 12, and Figure 13) indicates the number of terrorist activities by types of target, attack, and weapon, respectively. Based on Figure 11, private citizens and property were the most targeted subject followed by the military, police, government (general), and business. Most of the terrorist attacks were occurred on the top 5 target types while other target types were targeted relatively less.

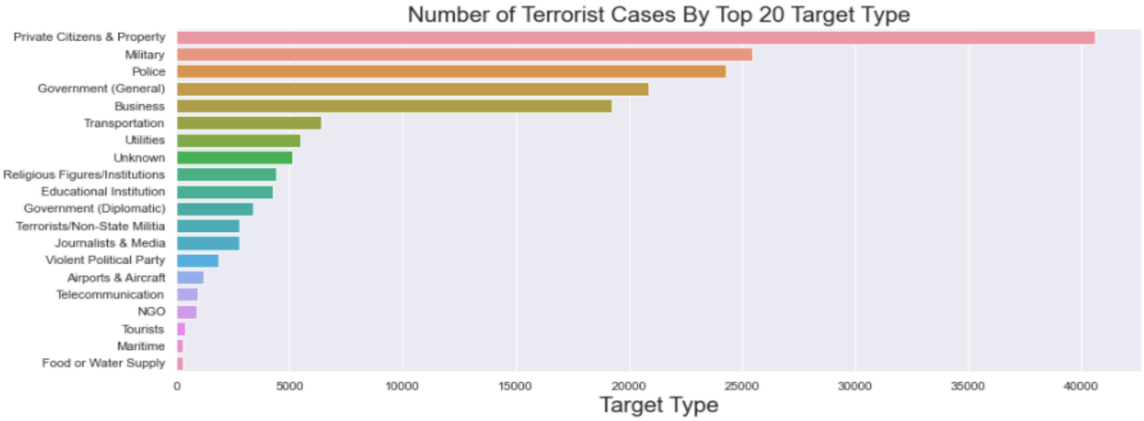

Figure 11: Number of Terrorist Cases by Top 20 Target Type

Mostly, terrorist attacks were carried out by bombing or explosion caused by an energetically unstable material. Convsersely, unarmed assult, hostage taking that resulted in incidents, and hijacking were merely used to attack the victims.

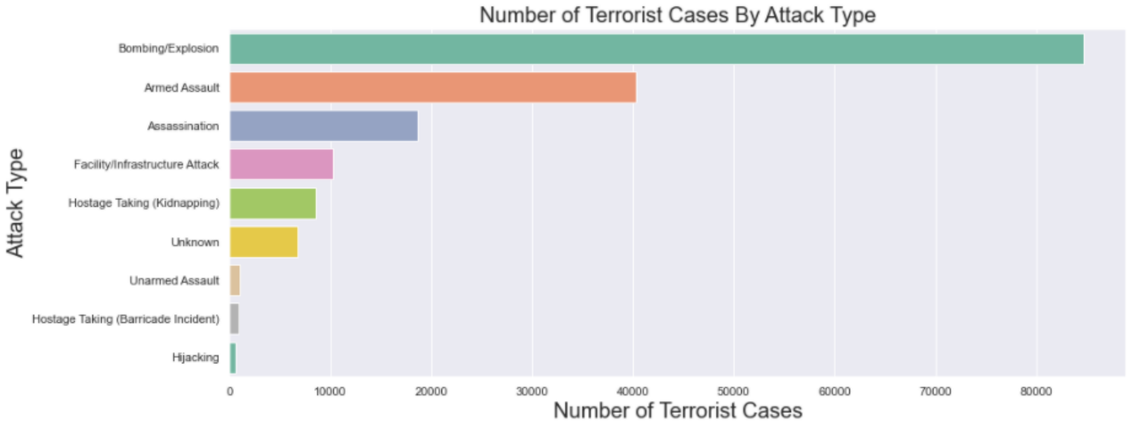

Figure 12: Number of Terrorist Cases by Attack Type

The mainly used weapon or tactic types include the explosives composed of energetically unstable material and firearms is capable of firing a projectile. Sabotage equipment, vehicle, fake weapons, weapons composed of biological origins, weapons composed of radioactive materials, and other were barely utilized in terrorist attacks.

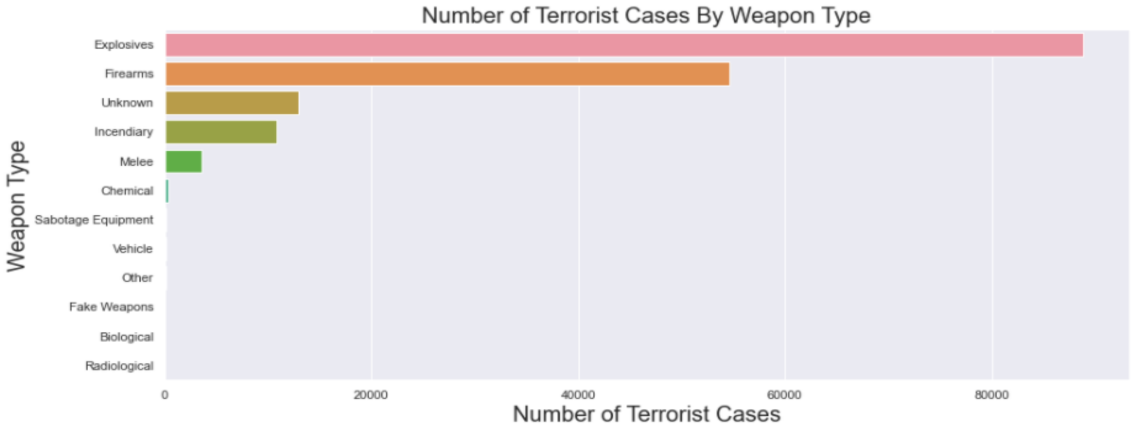

Figure 13: Number of Terrorist Cases by Weapon Type

Correlation

Prior to proceeding the correlation analysis, level grouping method was used in order to avoid producing an enormous size of correlation matrix. For example, several variables, including *country*, *victNationality*, *perpGroup*, and *targetType*, originally consist of over 200 categories. Thus, many of the levels with less frequencies were grouped into 'Other' based on the visualizations shown above. In specific, all other countries after the top 7, all nationalities after the top 6, all perpetrator groups after the top 4, and all target types after the top 5 now belong to the category 'Other.' That is, now the *country*, *victNationality*, *perpGroup*, and *targetType* variables consist of 8, 7, 5, and 6 levels, respectively.

After the completion of level grouping process, the categorical attributes were encoded as dummy variables. And the correlation between the dummy variables and the *casualty* variable was calculated. The most-highly associated dummy variables were the *suicide_No* and *suicide_Yes* with a correlation of 0.1071. The next one was the *weaponType_Vehicle* with a correlation of 0.1005. All other variables had a correlation below 0.05.

Preparing the Data for ML Algorithms

Based on the observations obtained in data exploration process, the data went through the preparation step for machine learning algorithms. With the missing values and naming convention issues already resolved in data cleaning process, a few more data-tidying methods were used to prepare more appropriate data.

In terms of the label, the original *casualty* variable will be used throughout the rest of the machine learning process for the following reasons. Firstly, in social science context, outliers are specifically more important than other observations to be analyzed because it is significant to study the unusual cases to figure out what happened and reflect them in the future analysis. Secondly, even with the logarithmic scale applied, the data still shows a strong skewness with several outliers. With these reasons, there was no modification or scaling applied to the *casualty* variable.

From the correlation analysis, the attributes with an absolute correlation over 0.1 were selected to be included in the model, which are the *suicide* and *weaponType* variables. Since there was no numerical attribute included in the model, no feature scaling was used. The categorical attributes were handled through one hot encoding.

Lastly, with the 'x', defined to include the *suicide* and *weaponType* variables that went through one hot encoding process, and the 'y', defined to be the *casualty* variable, the training set and test set were split in a ratio of 80% and 20% of the dataset. Here, to make the test set always remains consistent across multiple runs, random state of 42 was used.

With all the potential issues resolved in data cleaning and data preparation process, there is no more problems to be handled in the data. Thus, many different models were explored with the train set produced here without further improvements.

Exploring Different Models

Now that we have everything ready, 6 different models, including Polynomial Regression, Ridge Regression, Lasso Regression, SVM Regression, Elastic Net, and Decision Tree Regressor, were explored.

For the polynomial regression, the x train set was transformed using the *PloynomialFeatures()* function. Here, the *degree* was set to be 2, and the *include_bias* was set to be False. After this, the newly transformed x train set was fit into the *LinearRegression()* model. For the ridge regression, the *alpha* of 1 and the 'auto' *solver* was used as the hyperparameters. In the lasso regression, the *alpha* value was set to be 0.1. The hyperparameters in the SVM regression include the 'poly' *kernel*, *degree* of 2, *C* of 100, and *epsilon* of 0.1. The elastic net used the *alpha* of 0.1 and *l1_ratio* of 0.5. Lastly, for the decision tree regressor, I used the *max_depth* of 2.

After fitting all the models, the model performance was measured by the evaluation metrics explained in the introduction section. To reduce the overfitting of the models, a 3-fold cross validation was used in this step. Table 1 provides the performance of the 6 different models.

We observe here that the models produce very similar results in terms of the RMSE and MAE values. However, looking closer, we can see that the polynomial regression and ridge regression have relatively smaller RMSE values than other models. These two models also have the highest R^2 values and decent values for the MAE and standard deviation. The lasso regression has smaller standard deviation, but its RMSE and R^2 are not the best. Based on the MAE score and standard deviation, the SVM regression is deemed the best model with the smallest values. However, it has the smallest R^2 value of 0.0010. That is, only 0.1 percent of the variance in the casualties is explained by the features in the model, which is very low. Regarding the elastic net, even though it produces smaller standard deviation, the performance on the RMSE, MAE, and R^2 are worse than the polynomial and ridge regressions. The decision tree regressor performs the worst amongst all the fitted models based on all the performance measures. It results in one of the higher RMSE and MAE values and the smaller R^2 value. Moreover, its standard deviation is incomparably higher than other models' performance.

Based on the comparison of these numbers, we can narrow down the 4 good models: polynomial regression, ridge regression, lasso regression, and elastic net. It is not necessarily easy to find one specific model that performs distinctly better than the other. Focusing on our main metric, the RMSE, however, we draw the following conclusion that the polynomial regression and ridge regression are the best models as they result in the smallest RMSE with the highest R^2 values.

Training Model	RMSE	MAE	R^2	Standard Deviation
Polynomial Regression	53.4026	6.3488	0.0181	8.4799
Ridge Regression	53.4013	6.3488	0.0181	8.4287
Lasso Regression	53.5647	6.3385	0.0121	5.8548
SVM Regression	53.8653	4.9374	0.0010	2.3281
Elastic Net	53.6604	6.3386	0.0086	3.1146
Decision Tree Regressor	53.7124	6.3619	0.0067	36.5305

Table 1: Performance of Different Models

Fine-Tuning the Model

To improve the polynomial regression and ridge regression, hyperparameter tuning as well as ensemble method were applied. The performance of the fine-models are summarized in Table 2.

As the polynomial regression model does not have any hyperparameter available to be tuned, no hyperparameter tuning was applied for this model. For the ridge regression, two different alpha values (50 and 100) were tested. By comparing these two models, we observe that the ridge regression with the alpha of 50 performs slightly better than the one with the alpha of 100, according to the RMSE and R^2 values. We also see here that the Adaboost ensemble method critically worsened both the polynomial and ridge regressions. Not only are the RMSE, MAE, and standard deviation all skyrocketing compared to the original values, but the R^2 values are also now below 0. According to these experimental results, the ridge regression with the alpha of 50 was selected as the final system.

Fine-Tuning	**Training Model**	**RMSE**	**MAE**	$\boldsymbol{R^2}$	**Standard Deviation**
Hyperparameter Tuning	Ridge Regression with $\alpha = 50$	53.3946	6.3343	0.0184	7.0379
	Ridge Regression with $\alpha = 100$	53.4174	6.3267	0.0176	6.4328
Ensemble	Polynomial Regression Adaboost	85.0294	25.1447	-1.4893	81.6044
	Ridge Regression Adaboost	100.5837	32.3648	-2.4834	99.0716

Table 2: Performance of Fine-Tuned Models

Evaluating the final model on the test set, we have the model performance listed in Table 3. While the RMSE and MAE scores got better than the pure ridge regression algorithm before fine-tuning, the R^2 score and standard deviation got slightly worse. This, however, is expected because the system is fine-tuned to perform well on the validation data, not necessarily on the unknown datasets.

Final System	**RMSE**	**MAE**	$\boldsymbol{R^2}$	**Standard Deviation**
Ridge Regression with $\alpha = 50$	18.4634	6.1581	0.0046	7.2722

Table 3: Performance of the Final System on Test Set

Presenting the Solution

As it is indispensable to have a system that predicts to what extent terrorist attacks would cause physical fatalities and injuries, the machine learning algorithm developed in this project reveals its significance here. However, as the final model does account for the unusual terrorist cases with no log transformation applied in the data preparation step, it may prevent the system from generalizing the predictions to the usual and overall data. Also, the features in this algorithm explains only 0.4 percent of the variance of the casualties.

These limitations could be improved in the future works with the following suggestions. First, when preprocessing the dataset, it may be helpful to investigate the relationships between the attributes and casualties more carefully by incorporating the up-to-date terrorist activities information rather than just relying on the simple visualizations and correlations. Second, developing the model using more advanced machine learning methods, such as deep learning, may build more effective algorithm that predicts the consequences of the terrorist incidents more accurately. Lastly, the current project does not use the terrorist data occurred in 1993. Considering the potential effects of the 1993 data, it is suggested to integrate this data for future machine learning research.

Works Cited

“Global Terrorism Database.” GTD, The National Consortium for the Study of Terrorism and Responses to Terrorism, start.umd.edu/gtd/.