Evaluating and Analysing Global Terrorism Database using Data Mining Methodology

Amit Kumar
Department of Computing
Dublin City University
19210716
amit.kumar8@mail.dcu.ie

Nikhil Mittal
Department of Computing
Dublin City University
19210509
nikhil.mittal2@mail.dcu.ie

Yashaswi Verma
Department of Computing
Dublin City University
19211007
yashaswi.verma2@mail.dcu.ie

***Abstract*—The objective of the paper is to follow a data mining method and do an analysis of the Global Terrorism Database. CRISP-DM approach is used to process the data and understand it. Data Preparation is done and is shown via heatmap. Feature importance is taken out so that predictions becomes easier. Use of SVM, and Logistic Regression is used for correlation between weather and terrorist attacks. For seconds part of the prediction, Random forests tell what features help in determining the number of casualties.**

Keywords-CRISP-DM,Random Forest, SVM, and Logistic Regression

I. INTRODUCTION

The Global Terrorism Database (GTD) lists about 190,000 foreign and domestic terrorist activities that have occurred since 1970 worldwide. With information on the various aspects of each attack, the GTD is familiarizing researchers, politicians, academics and journalists with terrorist trends. The GTD describes terrorist attacks as:-The threatened or real use by a non-stop unlawful force and violence

Some general results obtained from the GTD include the existence of terrorist attacks and their distribution. For example, over half of all terrorist attacks in the GTD are non-lethal, and only one percent of the attacks include 25 or more deaths, between 1970 and 2018, these highly lethal attacks killed over 140,000 people altogether. The GTD attacks are linked to more than 2,000 designated perpetrator organizations and more than 700 additional generic groupings such as "Tamil separatists." However, two-thirds of these groups have been involved for less than a year and are carrying out less than four total attacks. Likewise, only 20 classes of attackers are responsible for half of all the 1970 to 2018 attacks for which a attacker was responsible. Examples of terrorist attacks are typically very varied over time and location and the GTD supports a detailed study of these trends [5]

In this paper, we'll answer the problem of estimating the number of kills and wounded from the different characteristics of the attack. We should adopt the CRISP-DM methodology by first understanding the various phases and , then going through the pre-processing of the data and how it was achieved. Firstly, by looking into the research issue, the business understanding of the dataset and understanding the problem statement. Then we look through the data to get a better understanding of what features and then take useful columns for our analysis. Then we did the preparation of the data, i.e. cleaning up our data set by removing null values and removing its values in each section of the heat map.

We will now go through the transformation phase.i.e. e after the data preparation i.e making our model. First, we'll use an extra tree classifier that helps to get what features to take that tells what are the essential features that play a part in our research problem.

We then implement our model, so that we can easily answer the research problem. First, we use Random Forest, Logistic Regression, and Support Vector Machines to answer the research problem, i.e. the association between terrorist attacks and weather conditions that is achieved by combining weather database and terrorist database. Then we go through the second part of the issue, i.e. knowing important features that have more impact on the number of casualties in the worldwide terrorist attacks and introducing Random Forest. We plot the ROC curve, obtain the F1 score and the matrix of uncertainty. We predict using logistic regression and SVM, using the same problem, and plot the confusion matrix, F1 score.

II. RELATED WORK

There is no much work done on the global terrorism database but there are few which are below:

[1] José V. Pagán's aim in this paper is to analyse various pre-processing methods for extracting the Global Terrorism Database in order to improve the classification of terrorist attacks by the perpetrator in Iraq. Four methods for dealing with missing values , three discretionary methods and three separate classifiers are tested using a 10-fold cross-validation error calculation. The authors conclude that pre-processing data may significantly reduce the rate of error classification for this data set, and that applying the Global Positioning System coordinates to the location of events may further reduce the rate of error classification.

[2]. Semeh Ben Salem, Sami Naouali and Moetez Sallami have described that the increasing amount of collected data has limited the performance of the current analyzing algorithms. Thus, developing new cost-effective algorithms in terms of complexity, scalability, and accuracy raised significant interests. In this paper, a modified effective k-means based algorithm is developed and experimented. The new algorithm aims to reduce the computational load without significantly affecting the quality of the clustering. The algorithm uses the City Block distance and a new stop criterion to ensure convergence. Real world data set observations show its high performance compared to the original k-means version..

[3]. Semeh BEN SALEM, Sami NAOUALI provides an approach to pattern recognition in multidimensional databases. The technique is based on a clustering method using the distance calculation between the reference profile and the observation of the database. Two distance measurements are proposed: adaptation of the Khi2 formula to the multidimensional sense, derived from the Multiple Correspondence Analysis (MCA) and the Euclidean distance. For retention, the comparison between the two distances would be the most effective for the multidimensional clustering sense. The suggested methodology will be applied to a particular case study describing armed attacks stored worldwide in the Global Terrorism Database (GTD)

[4]. LI Guohuia, b, LU Songa, CHENG Xudonga, YANG Huia and ZHANG Hepinga have done a deeper understanding of terrorist attacks, and the relationship between fatality and factors of control is helpful in enhancing the decision to allocate resources in the war against terrorism. Killing is divided into four levels: 1-2, 3-9, 10-29 and ≥30. The article uses a novel approach to communication

research to investigate the connection between terrorist attacks and causes. The level of fatality of terrorist attacks is affected by nations, areas, weapons, forms of attacks and targets. Factors that appear to result in high rates of fatality are established. Attacks in developing countries and regions are associated with a fatality level of ≥30, while in developed countries such as North America and Western Europe they appear to be associated with low fatalities. Attacks by explosive and nuclear weapons (Biological, Toxic, Radiological, Nuclear) continue to precipitate fairly large deaths. Attack categories like Hijacking, Barricade Accident, and Facility / Infrastructure Assault have close relationships with fatality rates of ≥30. Airports & Airlines and Maritime are more likely to cause exceptionally high casualties compared to other targets. The findings are important for recognizing the root cause of death in terrorist attacks.

III. DATA MINING METHODOLOGY

For a proficient execution with needed outcomes, the research project venture must follow a data mining methodology that is most appropriate to take care of the issue. Some of the Data mining methodology used across the world is KDD, CRISP-DM, and SEMMA.Crisp DM(cross-industry standard process for data mining) is a Data mining technique or approach for a detailed examination of facts. It is a robust and established methodology in data analysis[6]. The figure shown below involves the necessary six steps for the interpretation of records.

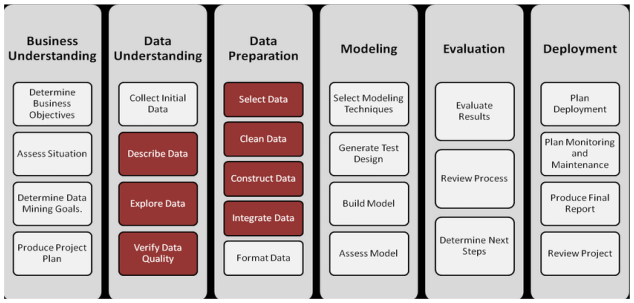

Figure 1: CRISP-DM Methodology

The CRISP-DM approach is an open standard, often used in the market and technological Industries, big data projects, and it is quite similar to other approaches. The approach helps the industry to overcome the requirement of the customers.

A. Business Understanding

The first step includes defining our problem statement, Understanding the research questions which we are going to solve with this methodology, as our research question or problem statement is about understanding the significant features which have more affect the no of casualties in the terrorist attacks over the world by looking at the datasets provided on the internet. The other problem is whether we can predict the success of the attack by cross-referencing weather conditions. The research question plays a crucial part in data mining technology as it tells specifically about the problems we have and the solution to them step by step.

B. Data Understanding

The Second step involves Data, which plays a significant role in the whole data mining technique as we need to understand and select the Data, which is useful for our Research Question(Business Understanding).

The dataset has been collected from Kaggle, which is a known website for data repositories of different domains. The name of the dataset is GTD(Global terrorism Data)is an open-source database storing data for terrorism from 1970 through 2017. The GTD defines a terrorist attack 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 and now incorporates more than 180,000 assaults. The database is kept up by analysts at the National Consortium for the Study of Terrorism and Responses to Terrorism (START), headquartered at the University of Maryland.[7]

The weather dataset was collected and merged with the GTD using longitude, latitude, and time with the specific column of the terrorist data. The weather dataset is

The dataset is large, so we need to take columns that would be useful for our analysis. Many columns are not useful and have null values. The column's importance is understood by the description given in the codebook of the dataset provided by START. The dataset is in CSV format having 181962 rows and 135 columns.

C .Data Preparation

After the analysis of columns in the Data Set, at this step, we clean the data or pre-process the large dataset of Global terrorism Data into a useful set of information, which is useful for applying models and get insights from it.

Data Preparation comprises different steps, for example, evacuating pointless traits(attributes), exceptions, invalid qualities, outliers, missing qualities, uproarious information. The cleaning task is done in python using various packages. Each task is explained below for cleaning the data below:

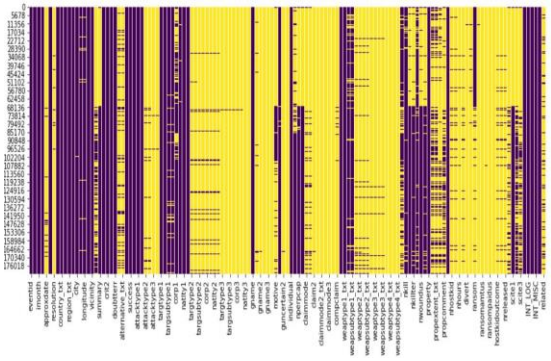

Figure 2: Uncleaned Dataset

The dataset visualization by heat map function helped to get the idea of null values in the columns, which were treated then by dropping the entire column having the null values higher than 50 percent and also the column which was not contributing to the analysis.

We are removing missing value, renaming the columns, filling null value, merging two columns of "killed" and "wounds" to one "Casualties". The weather dataset was combined with the terrorist dataset. The weather data is after 2012, so the model applied to this data shows the result of attacks after 2012.

GTD contained other violent acts and outliers, which should not be part of the terrorist data. To resolve this, we put constraints on criteria selection.

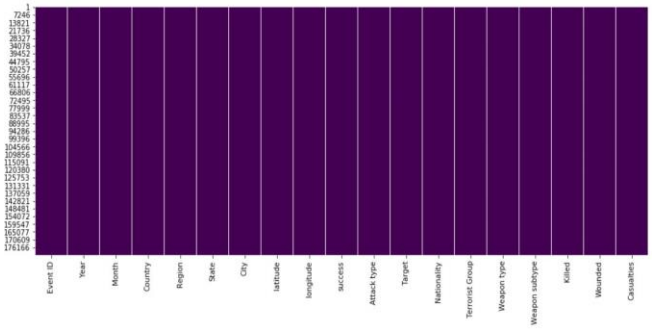

Figure 3:Cleaned Dataset

D. Modelling

1. Transformation

Now that we have a clean dataset which is pre-processed, it can go to the transformation phase. At this stage, we are trying to figure out the important features and analyze it with a model. Important features are chosen to serve the aim of the project.

An ExtraTreesClassifier is used to build a forest to find out the feature importance which is an ensemble learning method which is shown below in Figure 1.

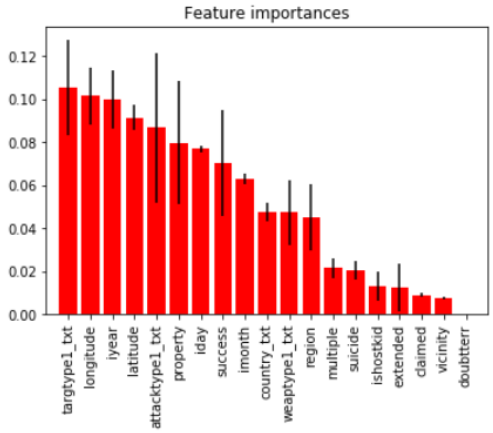

Figure 4: Feature Importance

The Plot clearly states the importance of each feature to predict whether it will lead to causalities or not. The features with an accuracy score of at least 0.05 are used to prevent overfitting in the model.

```
feature_cols = ['longitude', 'targtype1_txt', 'latitude', 'attacktype1_txt', 'success', 'property', 'country_txt', 'weaptype1_txt', 'region']
```

A. Data Mining

Implementation of a model is very important in data mining to serve the problem statement. We have to split the dataset into test and training set before we jump on to model implementation. We have used 30% test data and 70% training data. As mentioned earlier, The first step includes defining our problem statement, Understanding the research questions which we are going to solve with this methodology, as our research question or problem statement is about understanding the significant features which have more effect on the no of casualties in the terrorist attacks over the world by looking at the datasets provided on the internet. The other problem is whether we can predict the success of the attack by cross-referencing weather conditions. The research question plays a crucial part in data mining technology as it tells specifically about the problems we have and the solution to them step by step. We have a classification problem to deal with and we have used models like Random Forest, Logistic Regression, Support Vector Machine which are discussed below.

Our first prediction which is understanding whether we can predict the success of the attack by cross-referencing weather conditions for which we downloaded a weather database and merged it with our data to check whether any correlation exist between attacks by terrorist and conditions of weather. We have taken our target/dependent variable as has_casualities and independent/predictor variable as t2m, tcc, vidgf, sp, v10 which are 2 metre temperature, total cloud cover, vertical integral of divergence of geopotential, surface pressure, 10 metre v wind component respectively.

1. Random Forest

Here, the prediction is about understanding the significant features which have more effect on the no of casualties in the terrorist attacks over the world. As mentioned above, An ExtraTreesClassifier is used to build a forest to find out the feature importance which is an ensemble learning method. For this prediction, target/dependent variable is kept same as has_casualities and independent/predictor variable as longitude, targtype1_txt, latitude, attacktype1_txt, success, property, country_txt, weaptype_txt, region. Evaluation of Random Forest model is done using confusion matrix plot with the help of Scikit learn library which is shown in the below figure 2. The accuracy when compared with 10 cross validation is 77%. We have received an accuracy of 83% with Random Forest.

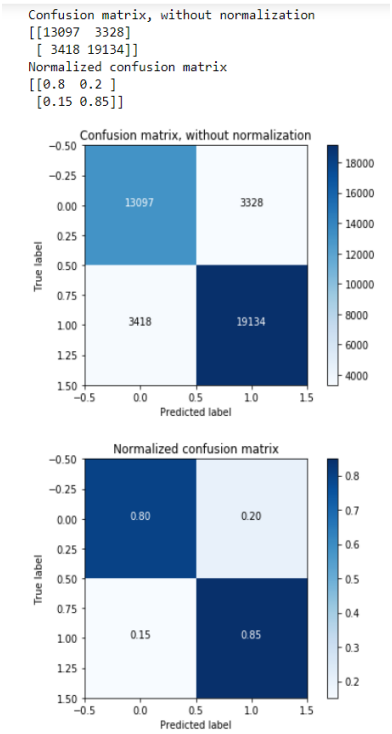

Figure 5: Confusion Matrix for Random Forest

Accuracy and other results like precision, recall, f1 score etc are shown below in Figure 3.

```
True Negatives:  13097
False Positives:  3328
False Negatives:  3418
True Positives:  19134
Precision 0.85
Recall 0.85
F1 Score 0.85
F2 0.85

F0.5 0.85
Specificity 0.80
Accuracy 0.83
```

Figure 6: Accuracy and other Results

We have tried to draw a ROC (Receiver Operating Characteristics) Curve which is shown in the below figure 4.

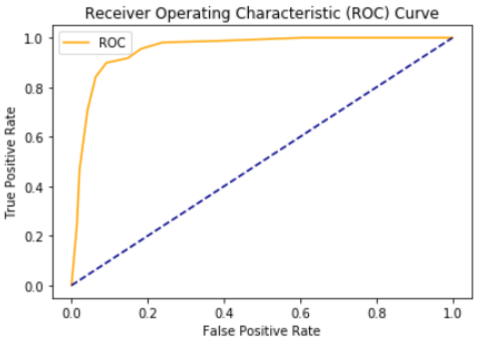

Figure 7: ROC Curve

2. Logistic Regression

We have chosen Logistic Regression this time to know whether we can predict the success of the attack by cross-referencing weather conditions. Dependent and independent variables are same in this situation too. We have received an accuracy of 68% with Logistic Regression. Logistic Regression is useful to find out the probability of an event like pass/fail, win/lose.[8] In our project, target is a binary value and we need to figure out whether there was any causality in the attack or not. Evaluation of Logistic Regression model is done using confusion matrix plot with the help of Scikit learn library which is shown in the below figure 5.

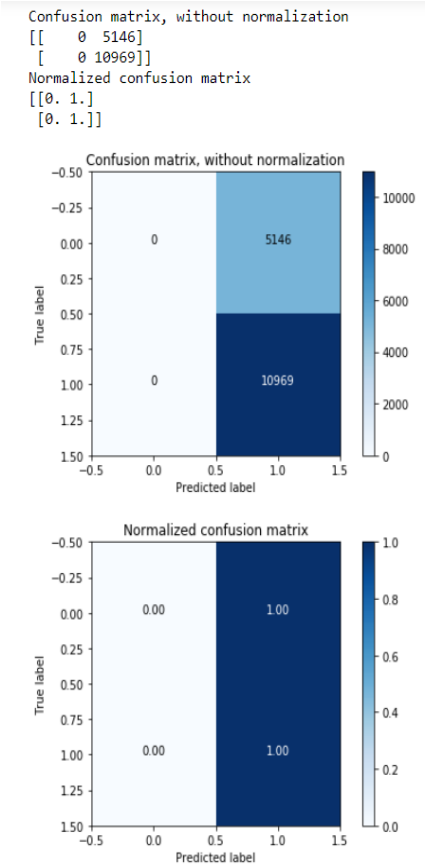

Figure 8: Confusion Matrix for Logistic Regression

Accuracy and other results like precision, recall, F1 score etc are shown below in figure 6.

```
True Negatives:  0
False Positives:  5146
False Negatives:  0
True Positives:  10969
Precision 0.68
Recall 1.00
F1 Score 0.81
F2 0.91

F0.5 0.73
Specificity 0.00
Accuracy 0.68
```

Figure 9: Accuracy and other Results

We have tried to draw a ROC (Receiver Operating Characteristics) Curve which is shown in the below figure 7.

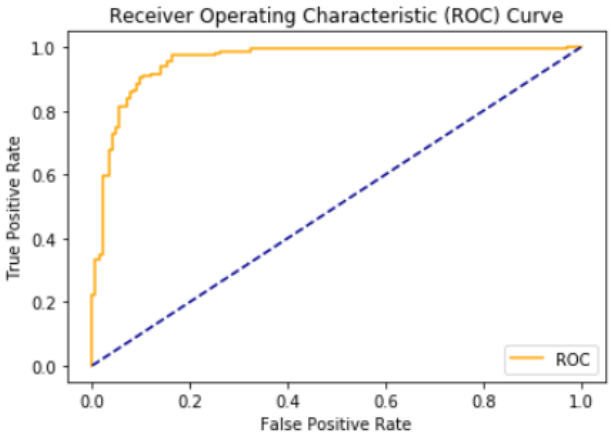

Figure 10: ROC Curve

3.SVM- Support Vector Machine

We have taken into consideration Support Vector Machine this time to know whether we can predict the success of the attack by cross-referencing weather conditions. Support Vector Machine is a supervised learning model which can be used for both classification regression analysis We have received an accuracy of 73% with Support Vector Machine . Evaluation of Logistic Regression model is done using confusion matrix plot with the help of Scikit learn library which is shown in the below Figure 8. and regression analysis [9].

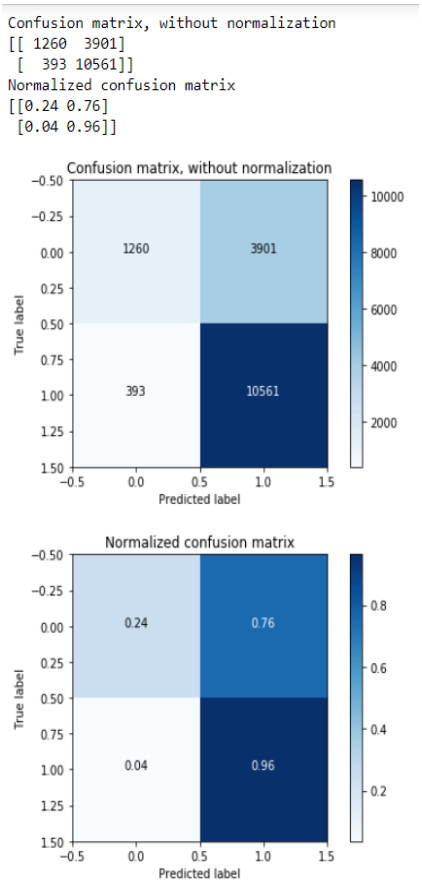

Figure 11: Confusion Matrix for SVM

Accuracy and other results like precision, recall, F1 score etc are shown below in figure 9.

```
True Negatives:  1260
False Positives:  3901
False Negatives:  393
True Positives:  10561
Precision 0.73
Recall 0.96
F1 Score 0.83
F2 0.91

F0.5 0.77
Specificity 0.24
Accuracy 0.73
```

Figure 12: Accuracy and other Results

E. Evaluation

Now that we have our models implemented, Evaluation of our result is important for the analysis. As mentioned, our first objective which is about whether we can predict the success of the attack by cross-referencing weather conditions was achieved by several models like Logistic Regression, Support Vector Machine. We have received an accuracy of 68% with Logistic Regression and an accuracy of 73% with Support Vector Machine. The results show that there is some role of weather conditions to predict the success of attack.

Our second objective is about understanding the significant features like longitude, targtype1_txt, latitude, attacktype1_txt, success, property, country_txt, weaptype_txt, region which have more effect on the no of casualties in the terrorist attacks over the world. We have received an accuracy of 83% with Random Forest. The accuracy when compared with 10 cross validation is 77%. We can clearly say from the result that these features are affecting the causalities in the terrorist attacks.

V. CONCLUSION & FUTURE SCOPE

The objective of this paper was to analyse GTD data with crisp-dm methodology and involving the steps for the prediction of casualties based on the features provided in the dataset. Secondary, the aim of the paper was to merge GTD data with weather data to predict if there was a factor of spatial feature affecting those success attacks that took place.

The model was evaluated based on the data used for modelling.

The future work on this project can be done by combining other datasets like GDP, Development index, to see the effect of this terrorist attacks on the growth of the country. Regression models can be applied to predict the number of casualties to accurate precision.

VI . GITHUB REPOSITORY OF THE PROJECT

https://github.com/amscool007/DataMining

VII. REFRENCES

[1] J. V. Pagán, "Improving the classification of terrorist attacks a study on data pre-processing for mining the Global Terrorism Database," 2010 2nd International Conference on Software Technology and Engineering, San Juan, PR, 2010, pp. V1-104-V1-110.

[2] Salem, S.B., Naouali, S. and Sallami, M., 2017, April. A computational cost-effective clustering algorithm in multidimensional space using the Manhattan Metric: application to

the Global Terrorism Database. In 19th International Conference on Machine Learning and Applications (ICMLA).
[3] Salem, S.B. and Naouali, S., 2016. Pattern recognition approach in multidimensional databases: application to the global terrorism database. International Journal of Advanced Computer Science and Applications (IJACSA), 7(8).
[4] Guohui, L., Song, L., Xudong, C., Hui, Y. and Heping, Z., 2014. Study on correlation factors that influence terrorist attack fatalities using Global Terrorism Database. Procedia Engineering, 84, pp.698-707
[5].Global Terrorism Database (GTD)
https://www.start.umd.edu/research-projects/global-terrorism-database-gtd
[6] A. F. Fahmy, H. K. Mohamed, and A. H. Yousef, "A data mining experimentation framework to improve six sigma projects," 2017 13th International Computer Engineering Conference (ICENCO), Cairo, 2017, pp. 243-249.
[7]Global Terrorism Database Code Book
https://www.start.umd.edu/gtd/downloads/Codebook.pdf
[8] Logistic Regression
https://en.wikipedia.org/wiki/Logistic_regression
[9] Support-vector_machine
https://en.wikipedia.org/wiki/Support-vector_machine