Predicting Future Terrorist Attacks with State-of-Art Machine Learning Techniques

Takuya Ando, Si Young Byun, Kunyu He, Ziyu Ye
Harris School of Public Policy Studies
University of Chicago
Chicago, IL 60615
takando, syb234, kunyuhe, ziyuye@uchicago.edu

Abstract

Our goal is to generate accurate and robust predictions for future terrorist attacks using machine learning techniques. In this report, we provide two different problem formulations, with corresponding methods, implementations, evaluations, bias and caveat analysis. One method is to predict the possibility of terrorist attack in the map grid by classifying them on whether they had terrorist attacks or not. The other is to predict the spatial-time series of terrorist groups' future attacks using recurrent neural network (Note that this approach should be viewed as a bonus in this report and will only be briefly discussed.). Policy recommendations are provided w.r.t. both approaches.

1 Introduction

1.1 Background

Although there is no universally accepted definition of terrorism, the Global Terrorism Database defines terrorism as satisfying three necessary components [1]:

1. Intentional,
2. Violence and/or threat of violence,
3. Sub-national perpetrators.

Terrorist attacks can take various forms including assassination, kidnapping, bombing, and infrastructure attack. After its first appearance in the 1960s, terrorism has been used by perpetrators to achieve some religious or political aims. The September 11 Attacks were a turning point in the history of terrorism and shifted the paradigm of terrorism [2]. Terrorist attacks that took place before and after 9/11 look very different from one another. For example, the geographical concentration of global terrorist activity shifted almost completely from Latin America to the Middle East after the September 11 Attacks. More importantly, since the September 11 Attacks, the battle against terrorism has dominated the national security agenda of the United States [3]. Just over a decade after the attacks, the United States spent over 600 billion US dollars on defense and homeland security [4].

1.2 Motivation

Despite the large amount of resources spent on counter-terrorism, unfortunately, the number of deaths and non-fatal injuries has been continuously increasing over the years [2]. Intelligence agencies were ineffective in preventing the September 11 Attacks partially because they rarely dealt with non-state entities and political objectives driven by extremist ideologies [5]. In response to the devastating failure to prevent terrorist attacks such as the September 11 Attacks and 2015 Paris Terror

Attacks, there have been an increasing interest in using predictive modeling to counter terrorism. In other words, if intelligence agencies can accurately predict when and where a terrorist attack will occur, they would be able to respond to terrorist attacks more effectively.

2 Literature Review

Though there have been a number of literature on terrorism, it is mostly concentrated in theoretical models about the causes and consequences of terrorism. Traditional attempts at predicting where terrorists will strike next are educated guesses at best. For example, Senior Adviser to the RAND President Brian Jenkins discussed four traditional ways of predicting terrorism and how they failed to predict terrorism in the past [6]:

1. extrapolating from trends in terrorism,
2. forecasting from the relationship between the world affairs and terrorism,
3. Examining how terrorists could use new technologies
4. Predicting of potential scenarios of future attacks.

Furthermore, because theoretical models for terrorism often cannot be empirically tested, it is hard to examine which models are more robust than others. Since the recent heightened interest in machine learning and big data, a number of researchers attempted to predict terrorist attacks using machine learning methods. For instance, Ding et al. used machine learning approaches such as SVM to predict terrorism with the historical data [7]. However, given the limited number of research done on this, much is left to be desired.

3 Problem Formulation

3.1 The Classification Approach

Based on previous studies, machine learning is a good approach in terms of predicting when and where terrorist attacks would happen. Given the historical data on time and locations of past attacks, we can learn which types of regions terrorists are likely to attack in the coming years and take preventive measures to stop it from happening or to mitigate the damage.

We formulate this policy problem as predicting whether terrorists would attack a specific location in the next two years, given features of locations that were known to be attacked at least once in the past six years. We look at the locations that were attacked at least once in the past six years with the assumption that terrorists are likely to launch attacks on locations that are similar and close to each other over a relatively long period of time. We predict whether an attack would happen in the next two years, as predictions over longer periods might become less precise. Given the information of historically attacked locations, we want to build a model that can warn policy makers of the risk of terrorist attacks at a specific set of locations in either of the next two years.

3.2 The Deep Learning Approach

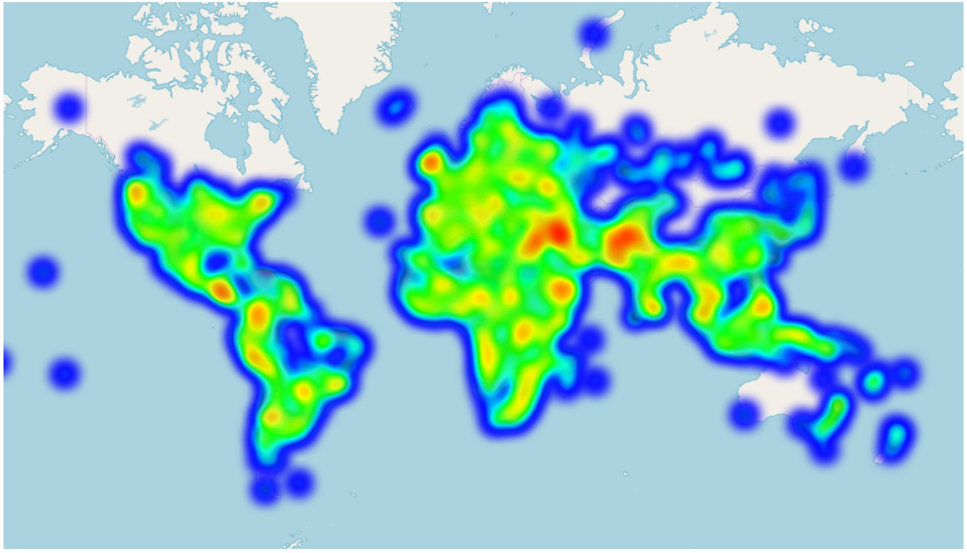
In this approach, we aim to uncover the attack patterns of terrorist groups, and form the problem into a temporal spatial prediction one. More specifically, we regard each terrorist group as an adaptive agent, whose behaviors could be modeled by its previous attack history (i.e. time series data with spatial context) through recurrent neural network, which will allow us to predict when and where a terrorist group will attack.

In this project, we adapt the Recurrent Neural Network model with spatial and temporal contexts by [11] and modify it to better suit our problem. Basically, we model sequential elements in an almost fixed time period in each recurrent layer. Such a recurrent structure is used to capture the periodical temporal contexts. In each layer, we employ the **time transition matrix** and **distance transition matrix** to capture the dynamic properties of terrorist attack in terms of time and geography. To efficiently create and calculate these transition matrix, we discretize the spatial and temporal value into multiple bins. Therefore, for each value inside a bin, we could approximate its transition matrix through linear interpolation via the lower and upper bound of the bin. A more comprehensive technical notes of the network structure will be provided in a separate appendix in this file's directory.

4 Data Description, Exploration, and Descriptive Statistics

The main data source we use for prediction is the ***Global Terrorism Database (GTD)***. It includes the records for over 180,000 terrorism event all over the world from 1970 to 2017, including data such as types of terrorist attacks, dates (day) and locations (longitude, latitude) of the events, numbers of causalities and names of the suspicious groups.

Figure 1: Terrorist Events around the World



We made some brief exploration for GTD data. **Figure 1** is a heatmap of distribution of terrorist attacks from 1970 to 2017. We can see that terrorist events are distributed globally, though focusing on a set of regions such as Middle East and Central America.

Figure 2: Regional Trend of Terrorist Attacks

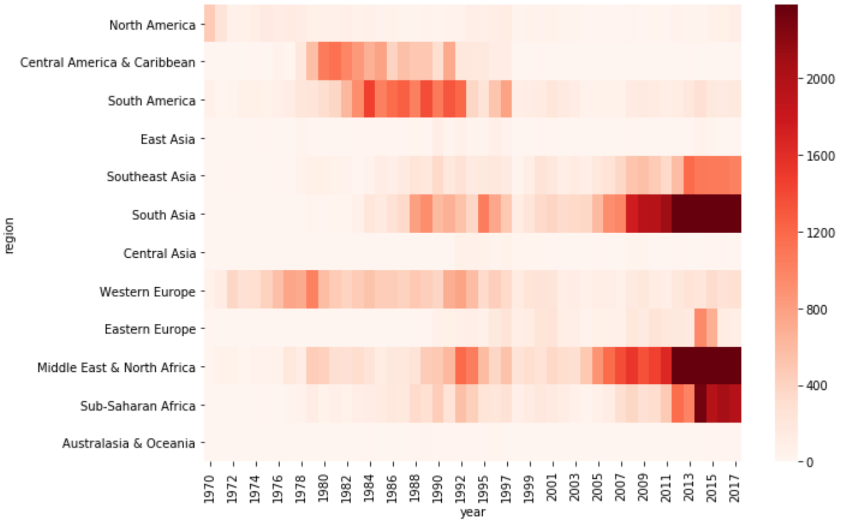


Figure 2 reveals the change of regional trend of terrorist attacks. We can see that in early period, regions such as Central America and Western Europe had a lot of terrorist attacks. However, in these days, there are more attacks in areas like South Asia and Middle East.

Figure 3: Number of Terrorist Attacks by Groups

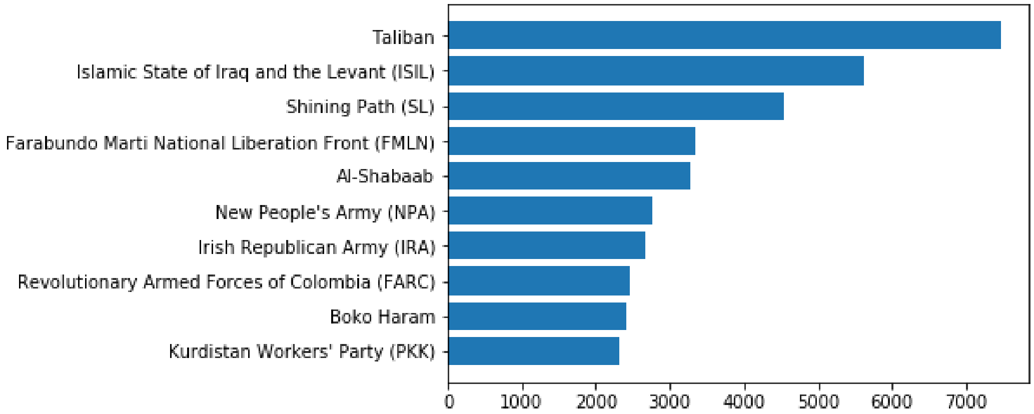


Figure 3 shows top 10 terrorist groups who had committed the highest numbers of terrorist attacks throughout the period. Overall, groups based in middle-east regions, such as Taliban and Islamic State of Iraq and the Levant (ISIL) are the reported groups responsible for the highest counts of terrorist attacks. However, we are also able to find groups from other areas like Shining Path (SL) and Irish Republican Army (IRA).

Figure 4: Trend of Terrorist Attacks by Group

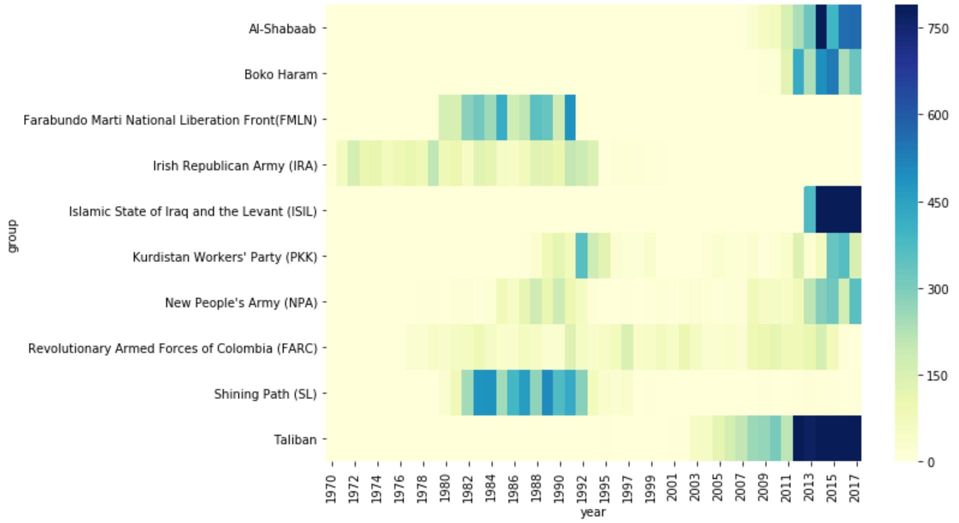


Figure 4 below visualizes the trends of attacks for these groups. We can see that, terrorist attacks were distributed more broadly around the world before 2000, as attacks launched by groups such as IRA and SL happened more frequently. Now, most of terrorist attacks are reported mainly by groups based in Middle East, such as Taliban, ISIL and Al-Shabaab.

To predict whether terrorist attack would take place at certain locations, we used data sources including GTD data. From GTD data, we used variables such as the time and location of the event, number of people killed or wounded, types of weapons used, and name of terrorist group name. We also used additional data sources to include the characteristics of the locations attacked historically. Such data include geographical information and social economic information. ***G-Econ (Geographically based Economic)*** data include geographical data such as distance to navigable water area, elevation and weather, as well as social economic information such as gross cell product and population of the area. These data are combined with 1x1 degree of latitude and longitude grids on earth. ***Geo-EPR (Geo-referencing Ethnic Power Relations)*** and ***World Religion Map*** provide ethnic and religious diversity data for each geographical region in the world.

We aggregated these data via latitude and longitude information in GTD data. Although latitudes and longitudes in GTD data have more than two-digit precision, those in G-Econ data have zero-digit precision. Therefore, we converted latitudes and longitudes in GTD data to 1 degree scale. Although it depends on which location on earth we are looking at, the length of one side of the grid in 1×1 degree scale is about 100 kilometers. Geo-EPR data and World Religion Map have a geographical polygon feature (in unit of province level) and we used it to combine them with GTD data.

5 The Classification Approach

5.1 Pipeline Overview

The **level of granularity** we use for our predictions is one year on time, and a grid defined by 1×1 latitude-longitude scale in terms of space. We implemented a **temporal train-test split** for model training and evaluations, with a fixed training window size of 6 years and test window size of 2 years in the period of 2002 to 2017. There is no gap between training and test sets. In terms of **model evaluations**, we are most concerned about the Recall of our models at a population level of 10%. We also report the training and testing time, and test performances on other metrics of our models. For **model tuning**, we apply 3-fold cross validation on each train-test pair and optimize over the whole training set to find the best set of hyperparameters to predict on the test set. For **model selection**, we admit the best model on each batch in terms of Recall at 10% in to our candidate set, and compare their test performances on all the batches from the year 2008 to 2017. We consider models in the candidate set comprehensively and recommend one model to the policy makers. Then we analyze the most **important features** contributing to the prediction over time to learn how our recommended model works. We then analyze the **bias and fairness** of our recommended model and check if it is fair in terms of space and religion. At last, provide the **list of locations** that are predicted to be attacked by year from the year 2008 to 2017, for policy makers to take measures accordingly and counter the "future" terrorist attacks.

5.2 Evaluation

We applied a temporal train-test split to find the best model for predicting coming terrorist attacks based on historical data. Since we believe that terrorist attacks follow certain patterns in terms of both space and time and are more related only when they are close in time, we employed a fix-length training time window of six years.

Each row in our training set is a location that was attacked at least once in the past six years. For a full list of generated features, please check this table. As we obtain the training set based on terrorist attacks in the GTD data, we aggregate the events at a specific location in a certain year by adding up the number of people killed or wounded, and summing up the counts of unique types of weapons, unique groups, and unique types of targets. Then we fill in the gaps for these locations with years in the training window of six years when they were not attacked. Each row is labeled one ("positive") when the location was attacked in the specific year and zero ("negative") otherwise.

Table 1: Descriptions of Training and Test Sets

Batch	Training Set	Observations (n)	Features (m)	Test Set	Observations (n)
0	2002 - 2007	4458	39	2008 - 2009	1338
1	2004 - 2009	5316	39	2010 - 2011	1288
2	2006 - 2011	5808	40	2012 - 2013	1757
4	2008 - 2013	6882	39	2014 - 2015	2440
5	2010 - 2015	8892	39	2016 - 2017	2252

As indicated in the literature review, the pattern of terrorist attacks have changed since the year 2002, and the lethality of terrorist attacks has increased rapidly, too. Thus, we mainly looked at terrorist attacks after the year 2002. Descriptions of the train-test sets are listed as below. Note that 2008 - 2009 refers to the period of 2008-01-01 to 2009-12-31, with a length of two years.

We use a test span of two years to cut the number of batches, since we have limited time to train and evaluate our models. Each row in the test set is a location that is known to be attacked at least once in the "next" two years. We also fill in the gaps for these locations with years in the two-year test window when they were not attacked. Each row is labeled in the same manner as that for the training set.

The metrics we care most about is **Recall**. As we are making predictions on terrorist attacks, life and public assets are at stake. We are more concerned about the coverage of our predictions, in other words, we want to catch as many future attacks as we can. However, as the interventions we would recommend consume a lot of resources to implement, we have to make decisions under constraint. Therefore, we plan to take actions on the 10% of locations at the highest risk of being attacked in the next two years. In other words, we optimize our models on the metrics at the level of 10% of the population. Thus the metrics we care most about is **Recall at 10%**. We would also optimize on other metrics, Accuracy, Precision, F-1 Score, and AUC score on the same population level. Evaluations of that model across all other metrics are also reported (check this example).

5.3 Bias and Fairness Analysis

Under the concern that our model would tend to predict grids where Muslims is the most popular religion, we examined bias and fairness of our model by the majority religion for each area. **Figure 5** shows the disparity of all metrics between the area where Muslims is the major religion and the other regions. We can see that for some metrics, there are considerable disparity between Muslim and others.

Figure 5: Disparity by Religions

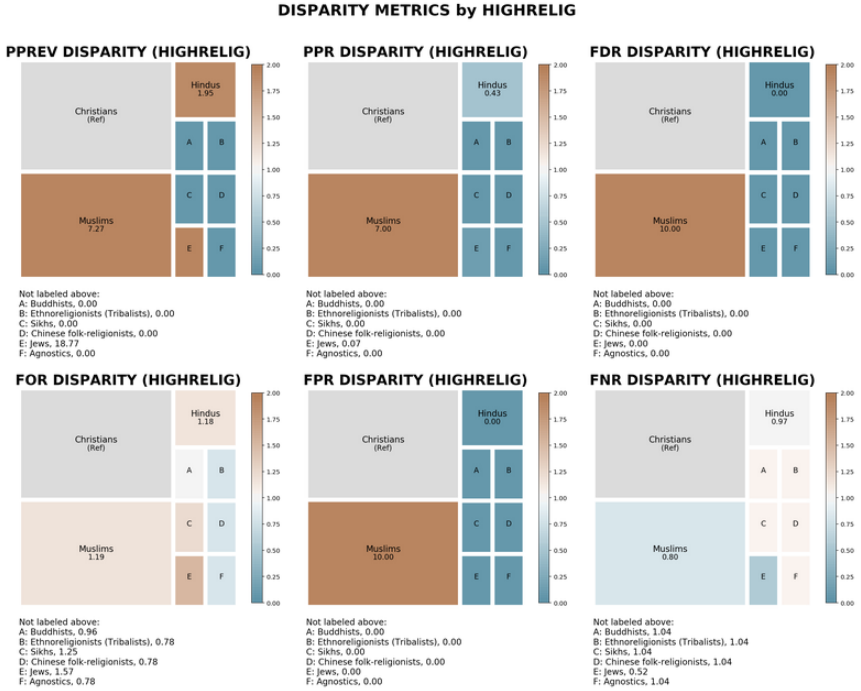
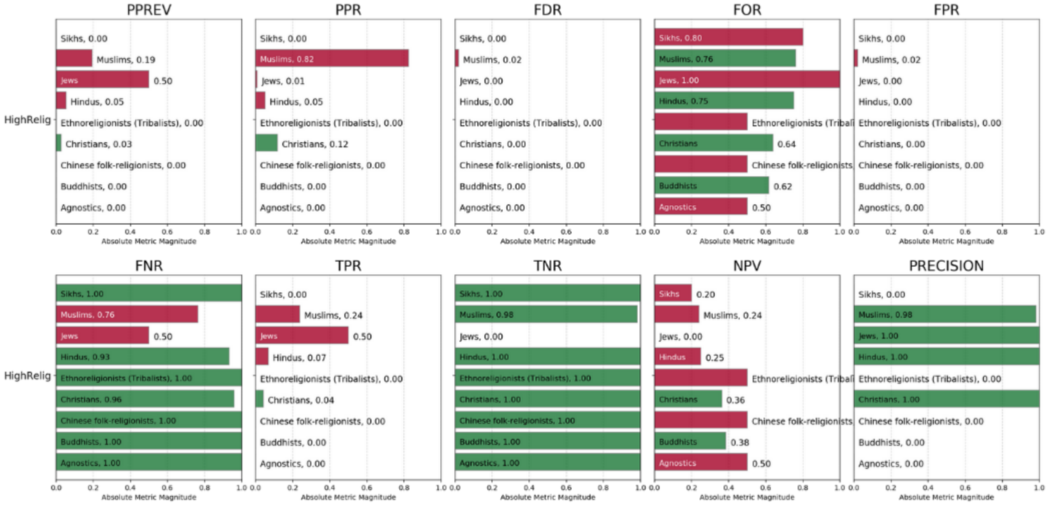


Figure 6 reveals the fairness based on the disparity. From some metrics such as FOR (False Omission Rate) and NPV (Negative Predictive Value), we can see that a considerable number of religions are categorized as "non-fair". Overall, there might be a possibility that our model has a bias in terms of religion.

Figure 6: Fairness for All Metrics



5.4 Results and Feature Importance Analysis

For the binary classification task, we applied the following models:

1. Logistic Regression and Decision Tree;
2. Random Forest, Bagging, Adaptive Boosting;
3. Naïve Bayes, KNN, Linear SVM

In terms of model tuning, for each pair of training-test sets, we apply **3-fold cross validation** on the training sets and optimize each type of model under a certain metrics on a specific population level, like Recall at 10%, to find the best set of hyperparameters. We then fit each type of model with its "best" set of hyperparameters on the whole training set and make predictions on the test set. At last, we select our "best" model of the batch based on Recall at 10%. When a tie happens, we make decision based on AUC score, as it indicates how well the positives are separated from the negatives in general.

For a partial list of "best" models on each train-test pair, check **Table 2** below. The full list of "best" models can be found here.

Table 2: Best Models Across Test Sets

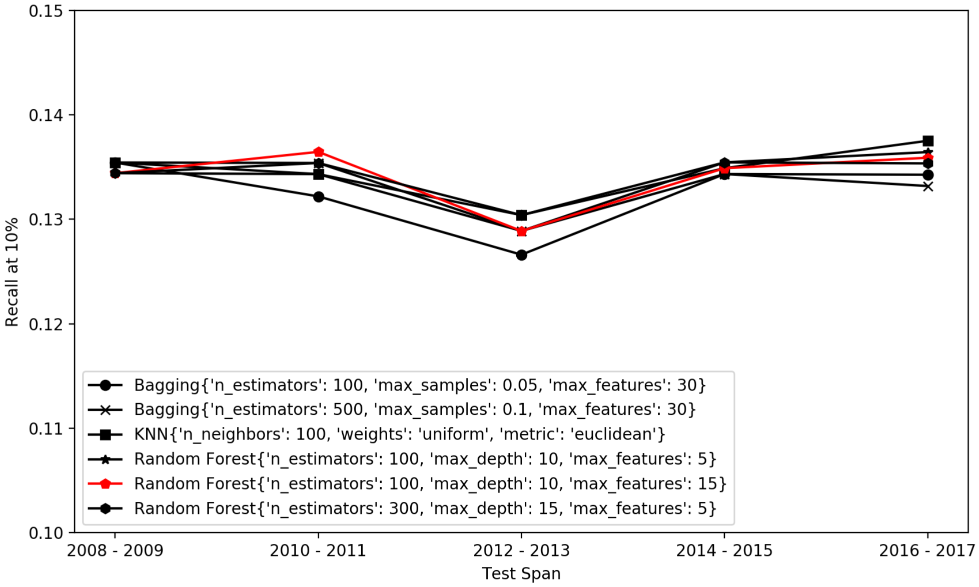
Batch	**0**	**1**	**2**	**3**	**4**
Test Span	2008 - 2009	2010 - 2011	2012 - 2013	2014 - 2015	2016 - 2017
Model	Bagging	Random Forest	Random Forest	Random Forest	KNN
Recall at 10%	0.1354	0.1365	0.1304	0.1355	0.1375
AUC	0.6506	0.7057	0.7002	0.7055	0.6908

It seems that Random Forest is our best model across test sets. Although the three best models are different from each other on hyperparameter sets. We make a plot of test performances of the best models as in **Figure 6** below.

While there is no model that is consistently the best across all test sets, **Random Forest with 100 estimators, maximum depth of 10 and maximum features of 15 for each base learner** is recommended for its fair performance in all the test sets. Its test performance on Batch 2 is visualized below.

As indicated in the panel above, beside its Recall at 10%, Precision at 10% and AUC score of our recommended model are fairly satisfying. Based on the feature importance in the bottom right, total number and diversity of the terrorist attacks in the past two years and in the past five years are contributing most to the recommended Random Forest.

Figure 7: Test Performance across Test Sets



6 The Deep Learning Approach

Please be aware that as we already provide a comprehensive machine learning pipeline on the prediction (i.e. the previous classification approach), this part should be viewed as a **bonus** and a **future work direction**, and we will only briefly discuss the following sections so that the report will not be too lengthy. Unlike the first approach, this part aims to provide a short introduction to a potential work direction for **non-professionals and policy audience**. Please refer to our corresponding repository if you want more implementation details. Our code is provided with rich annotation for you.

6.1 Pipeline Overview

As mentioned, this approach should be viewed as a bonus in this report and will only be briefly discussed. This pipeline of the deep learning approach is composed of the following parts: 1.data loading, 2. train/test/validation set split, 3. data pre-processing and RNN module formulation, 4. model training. Please see the following subsections for more details.

Basically our model is built upon the ST-RNN structure proposed by Liu et al (2016). The fundamental model structure has been discussed in section 3.2. If you are a math-inclined reader, please refer to the paper for the more details.

To implement the model, we should firstly do the data loading and train/test/validation set split. In the original GTD data set which contains terrorist events from 1970 to 2017, as discussed, each row indicates a terrorist event, and each column indicates the attributes of the terrorist event. We transform the data to suit our model as follows. For each terrorist group, the history of where it has attacked is given as $Q^u = (q^u_{t_1}, q^u_{t_{2,...}})$, where $q^u_{t_i}$ denotes where terrorist group u attacks. It should be noted that we set a attack frequency threshold of 30, which means that we would only consider those terrorist groups who have initiated terrorist events for more than 30 times. Also, we discard the data with unknown attributes. This could be problematic and we discussed it in the caveat part. Below is a sample transformed data. Then, given the time series data, we will make the initial train/test/valid split on the data. 70% elements of the behavioral history of each user are selected for training, 20% for the testing, and 10% as the validation set.

Figure 8: Performances of the Recommended Model on Test Set 2010-2011

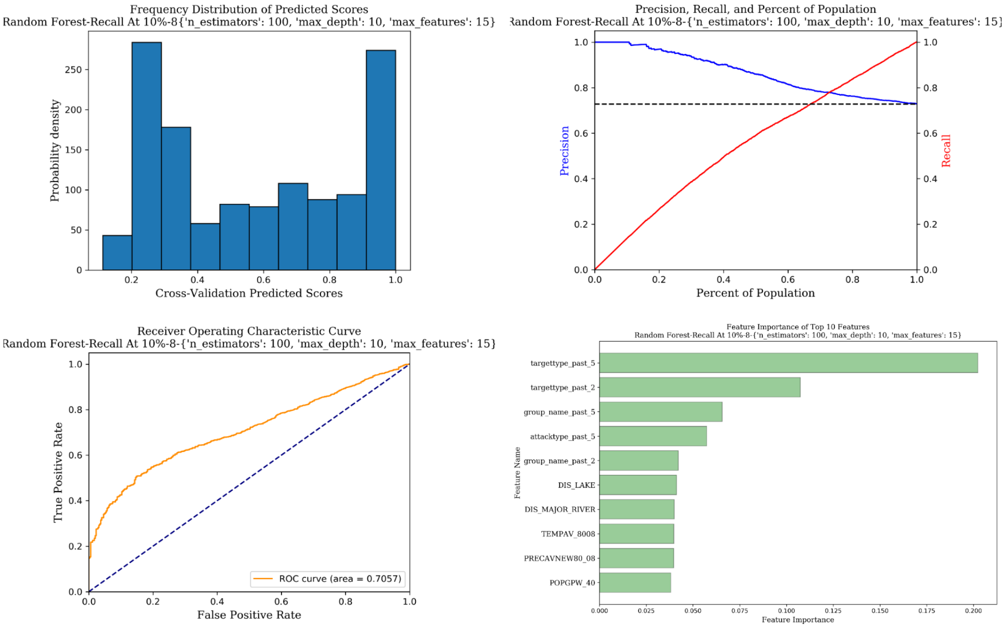


Table 3: Sample Data

Terrorist Group ID	Attack Time	Latitude	Longitude	Location ID
0	1989-01-23	37.99749	23.76728	17437
0	1989-04-10	37.99749	37.99749	17437
0	1989-11-01	37.99749	37.99749	17437
...	...	...	...	
3524	1979-02-12	-16.78805	28.85397	18356
3524	1979-02-13	-17.82839	31.052986	19897

Secondly, the pipeline contains data pre-processing and RNN module formulation. By far, the training, testing and validation set is still in the form of python lists, and we shall extract their attributes (i.e. terrorist group id, terrorist attack time, latitude, etc.), define their dimensions, and transform them into tensors which can be feed into the RNN module. The 'preprocessing.py file' in the corresponding directory. Aside from the above operations, it also transforms the original training, testing and validation set into batches, and for each batch, the attributes of time, latitude, longitude and location id has been extracted and then be encoded as tensors and fed into the RNN module.

Then, the pipeline contains a formal training phase built upon modified RNN cells and employs stochastic gradient descent, and a evaluation phase, which will be further extended in the following.

6.2 Evaluation

The evaluation score for our experiment is computed according to where the next selected location appears in the ranked list. Please refer to the code for a more comprehensive illustration of how the ranked list is generated. Basically, we use the learning process of Bayesian Personalized Ranking (BPR), which assumes for a terrorist group should prefer to attack a selected location than a negative one. Then what we need to maximize is the following probability: $p(u,t,v \succ v') = g(o_{u,t,v} - o_{u,t,v'})$, where v' denotes negative location, and $g(x)$ is the sigmoid function. Please be aware that we are still unsure about the correctness of the calculation of the recall rate as it is adapted from another source trying to re-implement the ST-RNN structure, and also we use a different approach to define time and space transition matrices than the linear interpolation approach mentioned in the theory part. We will definitely work towards for a better and correct code for the recall rate.

Anyway, the recall@k in our experiments is reported as follow for reference. Note that the code compares the predicted location id to the true predicted location id, which may increase the granularity, and reduce the value of recall@k due to the geographical content encoded in the location id. This is because we generate each location id for each different latitude and longitude combinations in the GTD data. Normally in the GTD data, each city is assigned with the same latitude and longitude combinations, but there are cases that different latitude and longitude combinations correspond to the same city in the data. So our next step is to generate the same location id for same city/states/countries in the data at the loading step, and use the new location id to train the model. The following table is obtained using a time width of 15 days, 1 month, 2 months, 3 months, 4 months and 5 months with the original location id definitions.

Table 4: Recall on Varying Time Window

Time Width	**Recall @1**	**Recall @5**	**Recall @10**
15d	0.1456	0.4287	0.6021
1m	0.1598	0.4567	0.6314
2m	0.1487	0.4679	0.6401
3m	0.1447	0.4762	0.6702
4m	0.1501	0.4612	0.6405
6m	0.1477	0.4700	0.6515

6.3 Bias and Fairness Analysis and Discussion

Due to time limit, and we have already provided the bias analysis in the classification part, we will leave this part for future work. A possible way is to link our result with other sociological attributes and check the bias distribution. For example, we could generate the result with one-year time window width, and in turn generate a religion distribution according to the predicted location at different time.

6.4 Results and Feature Importance Analysis

A basic result analysis has been provided in the evaluation part. As this is only the bonus approach, we will work in future for more meaningful result analysis. To be specific, we plan to 1. experiment with different dimensionality to check the sensitivity of the neural network; and 2. calculate the normalized values of convergence results of different evaluation metrics (e.g. recall@1, recall@5, etc.), and report the corresponding convergence rate for different evaluation metrics. As we simply use the time series data of terrorist groups w.r.t. their attacks and skip all the other features, we shall not conduct the feature importance analysis here.

7 Policy Recommendations

Ideally, a predictive model would be able to precisely predict the location and the time of potential terrorist attacks so that law enforcement and intelligence agencies can incorporate this information to strengthen the security of the predicted location to prevent the terrorist attacks. Unfortunately, due to the limitations of our classification model, our prediction is restrained by a large time window of two years for limited locations. However, considering the fact that a large number of terrorist attacks post-9/11 occur in the Middle East region, we argue that the location and the time window predicted by our machine learning model when used with other information collected by intelligence agencies could be still informative in preventing future terrorist attacks.

Considering that our relatively long prediction time span, the policy recommendation would be more long-term based and preventive approach, such as development of necessary infrastructure and human resources, rather than deployment of security forces to the dangerous areas. Infrastructure building includes introducing necessary equipment such as advanced security system and reinforcement of important buildings as well as preparing enough medicine and medical appliances in case of happening of attacks. Human resource development includes training for people such as police officers, military officers and medical professionals. These long-term action both in hard and soft aspect would be important for preventing terrorist attacks and decreasing the damage even if it happens.

8 Caveats

First of all, GTD relies on media coverage for data collection.The records in GTD data are reported incidents in open source media such as electronic news archives, existing data sets, secondary source materials such as books and journals, and legal documents. Therefore, there might be missing minor incidents which have not been reported to these sources.

Another point is that, geographical data and social economic data from G-Econ are not real time data. G-Econ includes data only every 5 years from 1990 to 2005. Therefore we might not be able to analyze the effect of time variant features especially in recent 10 years. Actually We also considered using country level data. There were some other sources which include annual social economic data for each country. However, we thought it would be too large to make a prediction in country level. Therefore we chose to use G-Econ data for this project.

Regarding the unit of prediction, we had to make prediction in large unit of grid(1 degree for latitude and longitude) because G-Econ adopted that scale. Although it would be much useful than country level prediction, it is still too large in unit to help policy makers prepare for future terrorist attack.

Finally, as we analyzed in the bias and fairness section, we have to take it into consideration that our prediction might be affected by the religion factor.

For the bonus deep learning approach, the linear Interpolation for transition matrix could introduce some errors and is hard to code, and it could only make predictions for existing terrorist groups which has an attack history of more than 30 events. It also lacks the ability to predict unknown terrorist groups as we filtered all the unknown data out during the data loading process. Therefore, it would only be a good tool if the government would like to learn about the possible future attacks of some certain terrorist groups. It could not function as a comprehensive prediction system as it is blind to the new or unknown terrorist groups.

References

[1] Codebook: Inclusion Criteria and Variables. 2018. The Global Terrorism Database.

[2] Roser, M., Nagdy, M., Ritchie, H. (2013). Terrorism. Our World in Data. Retrieved from https://ourworldindata.org/terrorism

[3] Kaczmarek, M., Lazarou, E., Guevara, M., Fogel, B. (2018). US counter-terrorism since 9/11. European Parliamentary Research Service.

[4] U.S. Security Spending Since 9/11. Retrieved June 11, 2019, from National Priorities Project website: https://www.nationalpriorities.org/analysis/2011/us-security-spending-since-911/

[5] ORourke, S. (2010). The Emergent Challenges for Policing Terrorism: Lessons from Mumbai. *Australian Counter Terrorism Conference*.

[6] Perry, W. L., Berrebi, C., Brown, R. A., Hollywood, J. S., Jaycocks, A., Roshan, P., Kraus, L. (2013). Predicting Suicide Attacks. RAND Corporation. Retrieved June 12, 2019, from https://www.rand.org/pubs/monographs/MG1246.html

[7] Ding, F., Ge, Q., Jiang, D., Fu, J., Hao, M. (2017). Understanding the dynamics of terrorism events with multiple-discipline datasets and machine learning approach. *PLOS ONE*, 12(6).

[8] Bahadori, M. T.; Yu, Q. R.; and Liu, Y. 2014. Fast multivariate spatio-temporal analysis via low rank tensor learning. In *NIPS*, 34913499.

[9] Bhargava, P.; Phan, T.; Zhou, J.; and Lee, J. 2015. Who, what, when, and where: Multi-dimensional collaborative recommendations using tensor factorization on sparse usergenerated data. In *WWW*, 130140.

[10] Zhong, Y.; Yuan, N. J.; Zhong, W.; Zhang, F.; and Xie, X. 2015. You are where you go: Inferring demographic attributes from location check-ins. In *WSDM*, 295304.

[11] Liu, Q., Wu, S., Wang, L., Tan, T. (2016, February). Predicting the next location: A recurrent model with spatial and temporal contexts. In *Thirtieth AAAI Conference on Artificial Intelligence*.