Trends and Determinants of Terrorism

Data Science 1 Final Project Report (Word Count = 2,931)

Mary Kryslette C. Bunyi

Introduction

What drives terrorism? How does the Philippines compare to the global terrorism landscape? I intend to answer these questions in my project. This report begins with the motivation for the study and a background on terrorism in literature, followed by a discussion of the terrorism and macro data sources sourced in this project. A walkthrough of the steps and thought process behind the data wrangling (especially to address inconsistencies and missingness), pre-processing, and model building follows. I will then discuss a descriptive analysis comparing the Philippines to the world's terrorism situation as well as insights from the predictive models. The report concludes with a stocktaking against success indicators set at the beginning of the project and future directions that the project may take.

Problem Statement and Background

Following the collapse of the Islamic State of Iraq and Syria's final territorial foothold in 2019, fears arise that they might move to other geographical locations or use the internet to influence localized attacks. This is a concern especially in the Philippines, where there are fears of increased radicalization due to foreign terrorist fighters. However, terrorism studies specific to the Philippines is scant. I hope to address this by analyzing trends in the Philippines against the global backdrop. Moreover, existing terrorism literature on the determinants of terrorist incidents exhibit mixed results, where diverse assertions are made on the significance (or lack thereof) of a variety of indicators such as economic growth, GDP per capita, inflation, population size, poverty, inequality, education/literacy, unemployment rate, religion, terrain ruggedness, military spending, political stability, government effectiveness, rule of law, and human rights protection score [1-6]. Thus, I also intend to take a machine learning approach in determining which variables matter most in predicting country-level terrorist incidents in a year.

Data Collection and Methodology

Data collection involved a variety of sources. Terrorism data came from the Global Terrorism Database (GTD) [7]. For the machine learning model that aims to determine the relationship of socioeconomic, political, and geographical variables with terrorism, I sourced macro-indicators based on factors cited in existing terrorism literature [1-6]. The following are my supplemental data sources: (1) World Bank DataBank for socioeconomic variables [8]; (2) Harvard Dataverse for human rights protection scores [9]; (3) The Correlates of War Project for religious composition [10]; (4) Our World in Data for terrain ruggedness [11] and military-related indicators [12-13]; (5) Center for Systemic Peace for polity variables [14]; and (6) the Varieties of Democracy Project for democracy indices [15]. All data sources can provide downloadable spreadsheets through their websites. To supplement this, I used the World Bank wbdata API [19] to immediately combine the indicators I needed in a readily manipulable dataframe.

The GTD's unit of observation is at an event level. On the other hand, the terrain ruggedness index is static at a country-level and is a geographical indicator that does not change through time, while all the other variables are at a country-year level. However, data for the latter are not necessarily available annually nor available for all countries. For most of the variables, their coverage is not wide enough to coincide with the 1970-2018 period covered by the GTD.

The main variable of interest that I wanted to predict is the number of terrorism incidents that occurred in a country in a year. I also intended to create a new country-year dataset consolidating variables from the GTD and the supplemental data sources. As such, data wrangling occupied a significant chunk of my time, which included the reshaping of data, group manipulations, and joining of multiple datasets. To read in the data sources, I used the pandas package [16]. Since the GTD does not have a column for civilian casualties and injuries, I created variables based on total and terrorist casualties and injuries. However, total and/or terrorist casualties/injuries would be inconsistent or missing, so I clipped my computations of civilian figures such that they will not go below zero. The GTD also coded unknown data as negative numbers in fields such as number of perpetrators, among others. To prevent these negative numbers from unduly affecting my analysis, I recoded them as null values. I then created dummies from the categorical variables in order to prepare my data for the machine learning pipeline. I created dummies per attack type, target type, and weapon type. Since the introduction of dummies per category would lead to numerous variables, I binned items that were similar or immaterial (e.g., low-frequency targets). After fixing the variables of interest, I summed them up at a country-year level via group manipulation. This yielded a terrorism dataset with a country-year unit of observation, where the data referred to counts/totals of incidents, certain attack types, casualties, and others. To prepare this dataset for consolidation with the other data sources, I standardized all country names via the country-converter package [17]. For simplicity, I dropped the countries that were not found in the module. These would likely have minimal effect as they likely will not have matching variables from the other data sources as well. For the remaining data sources, I applied a similar data wrangling process. I read files in via pandas, turned variables of interest to numeric/binary and binned if applicable, standardized country names via the country converter package.

After completing data cleaning and standardization of all data sources, I conducted multiple dataframe merges based on country-year and took a cursory look at missingness, subsetting for 1990 onwards when terrorism data was more reliable. I noticed that there were several country-year data points from non-GTD sources for which there was no corresponding GTD data. It is safe to assume that the non-inclusion in GTD for 1990 onwards means that there is no terror incident in the country for the year, so I assigned all GTD indicators for that country-year as 0. This made GTD data complete across the board. On the other hand, there also were countries which did not have data points for each year of the coverage period. A cursory look at one of the countries (Guadeloupe) explains the lack of data points – it exists only in the GTD. Other data sources do not have it in their database. Thus, we can safely drop these countries. After preliminary fixes and subsetting, the dataset was still noticeably sparse in various non-GTD data, of which missingness in the religion data was most evident.

Analysis of Tools and Methods

In terms of machine learning tools and methods post-data wrangling, I intended to predict the number of annual domestic terrorist incidents using macro-explanatory variables via scikit-learn algorithms covering K-fold Cross Validation, Linear Regression, Decision Trees, and Random Forest [18]. However, further data processing was needed before the dataset could be subjected to machine learning methods. First, in predicting the number of terrorism incidents based on macro indicators, we would expect that the latter will not have a synchronous effect or link to terrorism. For instance, poverty or government repression now should cause discontent to simmer until individuals are compelled to join terrorist organizations or carry out attacks. As such, I predicted the number of terrorism incidents based on 1-year lags of the macro indicators. Although missingness is a major issue, I noticed that the available data is distributed across countries through time, so it appears that the disparities may lie in the data frequency. Since one could reasonable expect that macro indicators (such as religious composition and population) are generally enduring and do not usually fluctuate through time, I interpolated linearly across time by country. This greatly improved missingess, especially in the religion dataset. However, there still remained null values which were impossible to interpolate due to non-existence of data. Since estimating missing values in this case would be a challenge, I subsetted the data to country-year observations which had actual or interpolated data across all the variables.

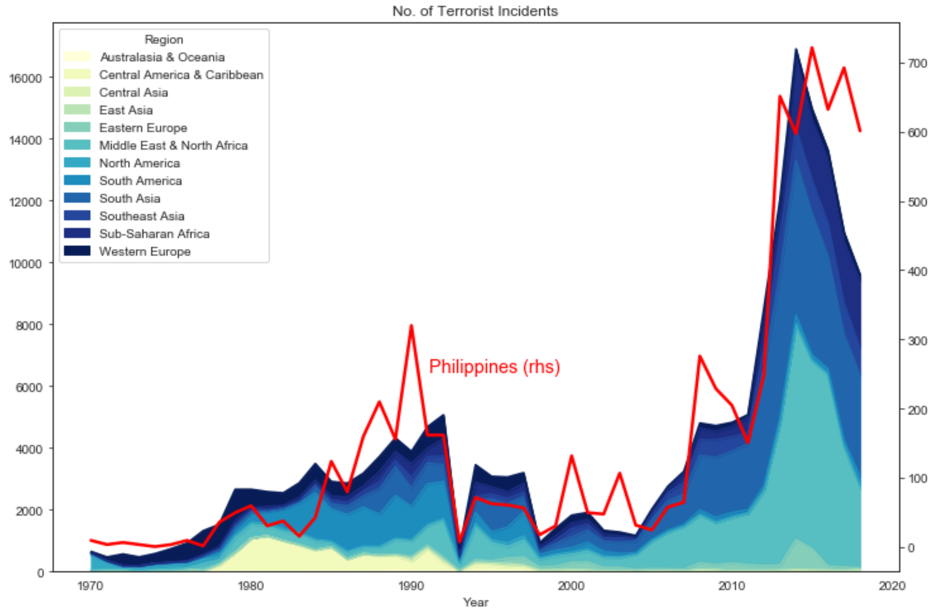
After completion of data pre-processing, I split the dataset into training and test data at a 75%-25% ratio of randomly selected observations. At cross validation, I set a random seed for the folds index to ensure comparable samples, so that better fit will not be due to a "lucky" sample. I then initialized the pipeline through MinMaxScaler data pre-processing [18]. This ensures that the predictors will be scaled accordingly and large value predictors will not unduly appear to have a more substantial effect than smaller-value predictors. After this, I selected models and tuning parameters which shall compete for the best fit of the training data. The selection of models are Naive Bayes, K Nearest Neighbors at different numbers of neighbors, Decision Tree at different maximum tree depths (branches), and Random Forest at different maximum tree depths (branches), number of estimators, and numberof maximum features [18]. Combining these in a GridSearch, I used a scoring metric of R^2 in order to facilitate interpretability of the results [18]. I then took the best fit score and assessed in-sample performance (on training data) against out-of-sample performance (on test data). This serves as a check on whether overfitting occurred on the training data, in which case it would perform badly on test data. Moreover, it is not enough to find the best-fitting model for our data. It is also important to analyze variable importance, which would allow us to determine which features are the most important in predicting the outcome. One way to do this is through permutation importance, which generates intervals showing how the scoring metric changes with permutations in the respective predictors. I also plotted partial dependency plots to further study the relationship between the top features and the outcome. Finally, I checked ICE plots for the most important feature to check whether there is heterogeneity across the observations as our most important feature changes.

For the other models that I ran for the project, changes were mostly in the data pre-processing stage while the main deviation in the machine learning pipeline was the manner of traning-test data split. In my second model, instead of randomly assigning data points to the training and test datasets, I divided the data chronologically and assigned the 25% latest data points (i.e., 2015 onwards) as test data while sticking to a 75%-25% training-test data ratio. In my third model, I initially wanted to drop all columns with missing data. However, there are only 20 observations with no missing values. This is insufficient to run a model. I also noticed that dropping incomplete data points may cause our dataset to drop areas with high terrorism incidence, since these countries are usually developing countries which tend to have less country-level statistics compared to more advanced countries. As such, I pivot to subsetting the data for high-incidence areas. There were 59 countries with at least 200 terror attacks from 1990 to 2018. To solve the missingness issue (which still persisted even after my earlier linear interpolation), the scikit-learn KNNImputer [18] algorithm imputed values for all data that remained missing. The KNNImputer computes its imputation based on the neighbors most similar in terms of the other non-missing features.

Results

On the descriptive analysis front, in terms of number of terrorist incidents, the Philippines' broader pattern appears to follow that of the global space (Figure 1). However, in more recent years, terrorism events in the Philippines don't appear to have subsided as much as it has globally. Even in the Middle East and North Africa, the number appears to have waned. This would be worrisome in relation to fears of the Philippines emerging as a terrorism hotspot.

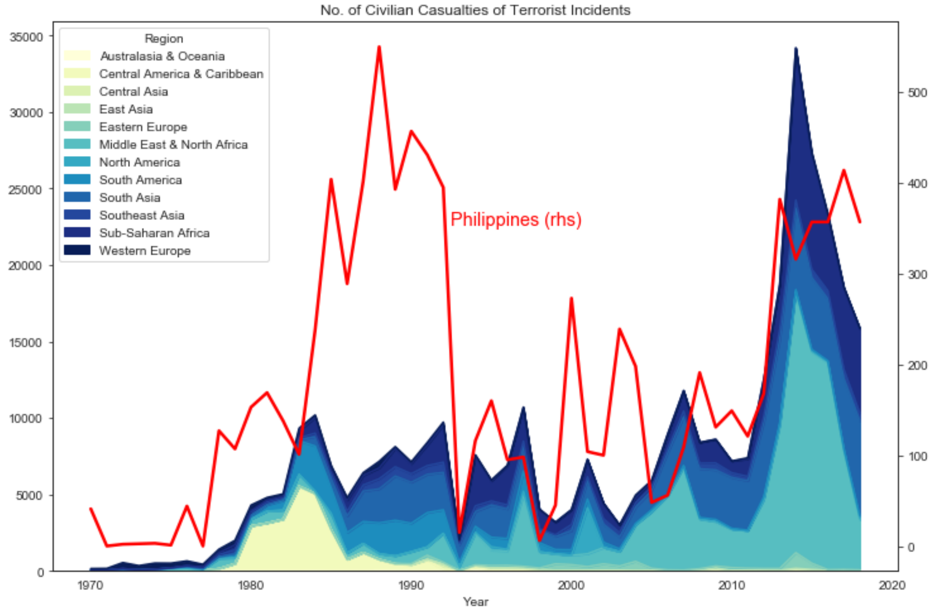
In terms of number of civilian deaths, what's striking is that the global trend appears to



No. of Terrorist Incidents

Source: Global Terrorism Database

Figure 1: Number of Terrorist Incidents



No. of Civilian Casualties of Terrorist Incidents

Source: Global Terrorism Database

Figure 2: Number of Civilian Casualties from Terrorist Incidents

be roughly the same as the number of incidents but in the Philippines, the pattern has been more erratic (Figure 2). Nonetheless, relative to Figure 1, we could see that despite increasing terrorism incidents in the Philippines, the number of casualties has lessened.

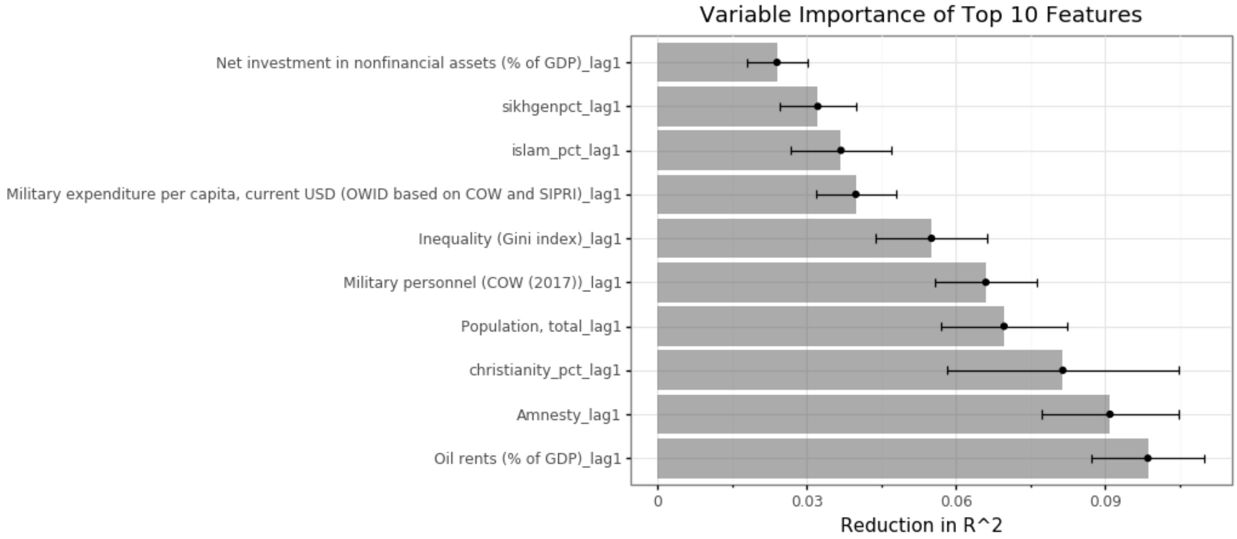


Figure 3: Variable Importance of Top 10 Features for Model 1

In my first model, which incorporated 1-year lags of the macro indicators, I interpolated data linearly, subsetted only the countries without missing data, and split into training and test data randomly. The best model is a Random Forest Classifier with maximum depth of 5, maximum features of 10, and 500 estimators. Its in-sample R^2 is 0.9090, while its out-of-sample R^2 is 0.498. In making predictions, the model relies the most on the following 1-year lagged variables: Oil rents (as % of GDP), Amnesty ordinal variable from the Political Terror Scale, and Percentage of Christians in the Population (Figure 3).

My second model is the same, except that the training-test data split is based on chronology (with the 25% latest data assigned as the test data). Once again, the best model is a Random Forest Classifier with maximum depth of 5, maximum features of 10, and 500 estimators. Its in-sample R^2 is 0.9201, while its out-of-sample R^2 is merely 0.1768. This could be a sign of overfitting and/or the significant change in the dynamics of terrorism in recent years. In making predictions, the model relies the most on a different set of 1-year lagged variables:

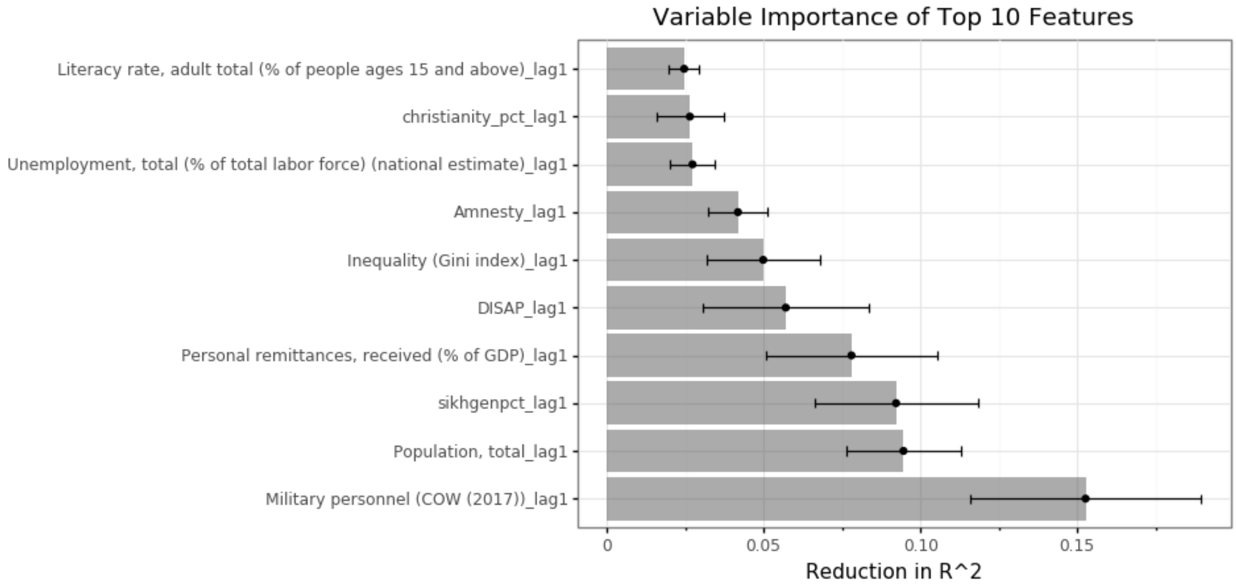


Figure 4: Variable Importance of Top 10 Features for Model 2

Military Personnel, Population, and Percentage of Sikhs in the Population (Figure 4).

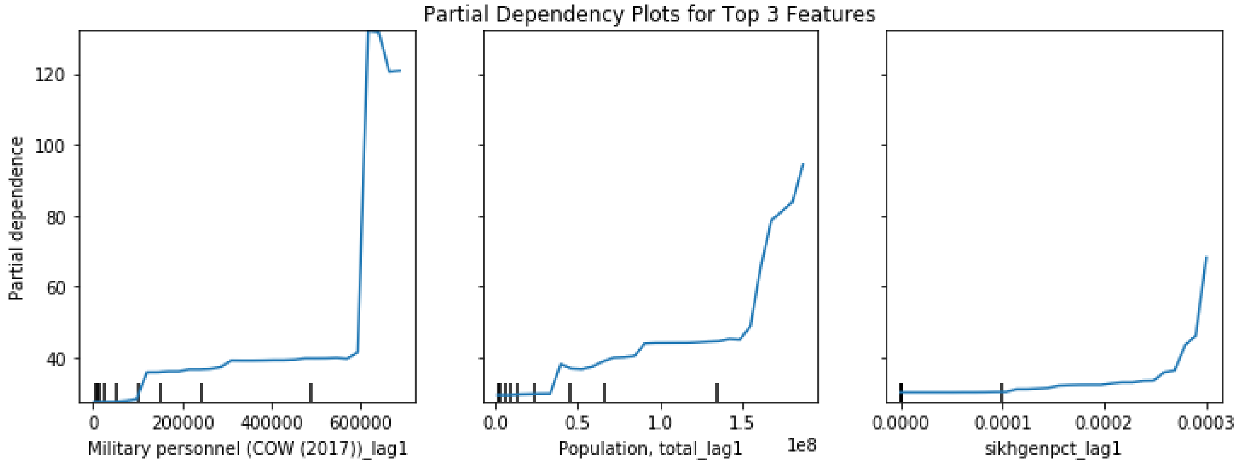


Figure 5: Partial Dependency Plots (PDPs) for Top 3 Features for Model 2

The marginal relationships reflected by the PDPs (Figure 5) are interesting in that the number of terrorism incidents spikes from 0 to 40 as military personnel reach 100,000 and 40 to 130 as military personnel reach 600,000, but decreasing after that. A similar behavior appears in the PDP of population, where terrorism incidents spike from 0 to 40 as population increases to 50 million, steadily increasing from there and sharply rising as population exceeds

150 million. This may be reflective of the larger populations in less developed countries where terrorists take advantage of weaker law enforcement density/capacity. Finally, terrorism incidents were found to spike from 40 to 70 as the percentage of Sikhs in the population increased to 0.0003%. While this is a small number, the fact that the share of other religions to the population also appear significant in other models could point to the importance of religious diversity as a predictor for terrorism. Further analysis may be warranted in this subject.

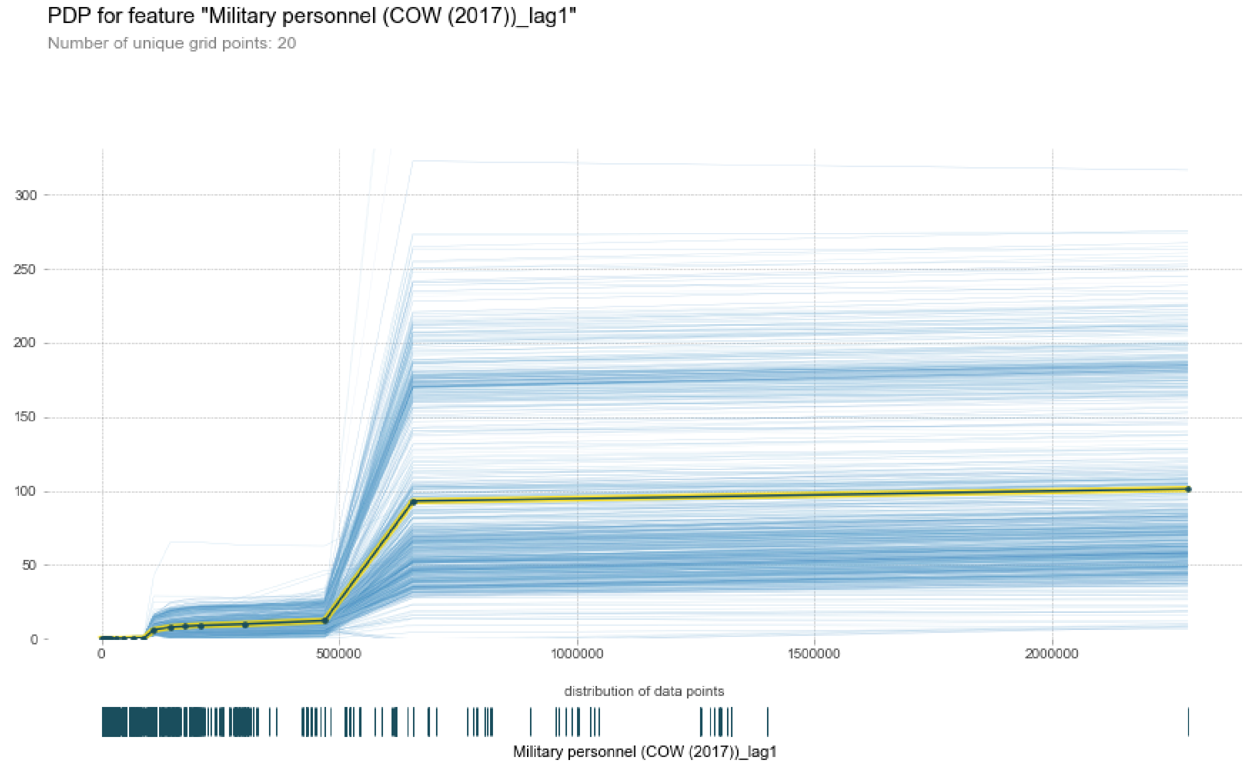


Figure 6: ICE Plot for Top Feature for Model 2

The ICE plot for the number of military personnel (Figure 6) shows significant heterogeneity as we move from 500,000 to 700,000 personnel. The effect on several data points is highly dispersed, with some predicted to soar from less than 50 to more than 300 terrorism incidents while others deviate from the average and report lower or minimally increasing terrorism incidents near 0, with the increase of military personnel. This signals to us that there is a strong interaction and that there are other factors pushing our observations to go on trajectories that we do not expect. Majority of the data points are clustered below the mean line, illustrating a smaller increase in terrorism incidents, while there is also a sizeable number of points clustered far above the mean line. Nonetheless, the interaction likely does not cause differing results as our observations are still generally moving in the same direction (i.e., more terrorism incidents with an increase in military personnel). The difference is in the magnitude.

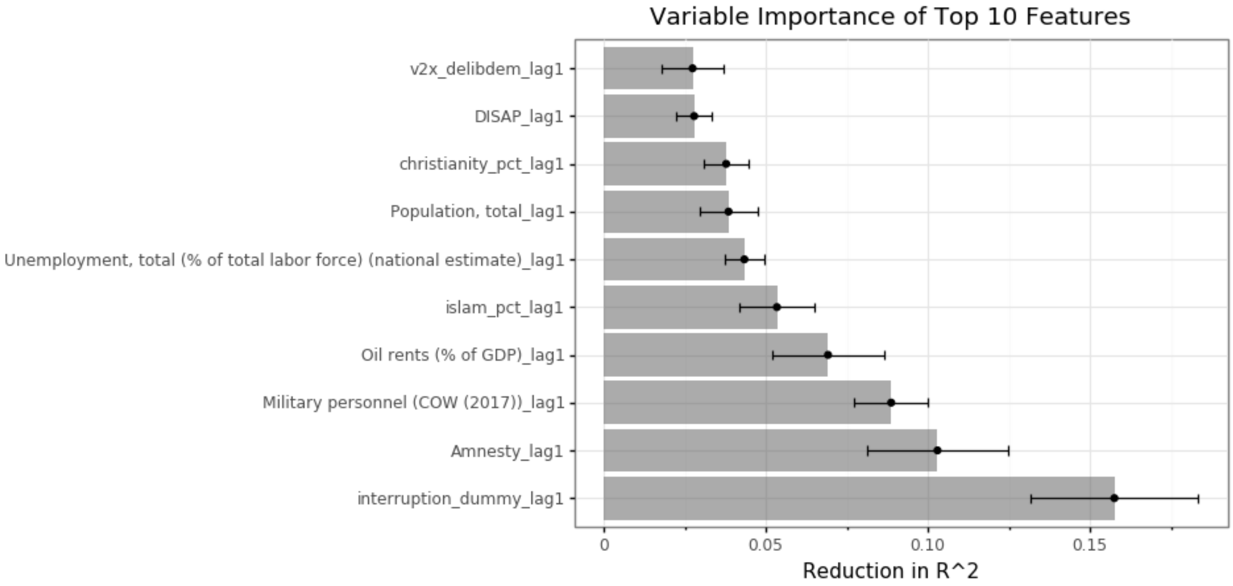


Figure 7: Variable Importance of Top 10 Features for Model 3

In the two models, the priority was data availability, so in the third model, I focused on the inclusion of data points. I noticed that outcomes were heavily skewed to zero (given that terrorism is a rare occurrence in other countries, which tend to be more advanced and provide complete statistics). On the other hand, dropping incomplete country-year data points forces us to drop high-incidence areas which may suffer from data unavailability. Thus, I dropped variables with more than 25% missing data and subsetted the data to 59 countries which listed at least 200 terrorism incidences from 1990 to 2018. Since linear interpolation could not address further missingness in the data for countries, the KNN Imputer filled out all remaining missing values based on all the predictor variables as well as the year. Based on a chronological test-training split, the best model is a Random Forest Classifier with maximum depth of 5, maximum features of 10, and 1500 estimators. Its in-sample R^2 is 0.7518, while its out-of-sample R^2 is again substantially lower at 0.0769, signalling overfitting. Nonetheless, it's interesting that the top features are similar to the previous models (with the exception of the dummy variable on foreign power interruption) i.e.: grant of amnesty, number of military personnel, share of oil rents to GDP, and religious composition (Figure 7). The ICE plot shows reasonable results which align with the reality that terrorism incidents indeed tend to happen along with foreign intervention, although the direction of causality cannot be determined based on this data.

Discussion of Project Success

I believe that I was able to broadly achieve the success indicators that I set at the start of the project: (1) the creation of a consolidated country-year dataset; (2) descriptive information assessing the Philippines against the global backdrop; and (3) the generation of a predictive model incorporating country-level factors to determine the variables most important in predicting terrorism. Nonetheless, given the breadth of information embedded in the GTD and other sources, as well as possible overfitting and issues with missingness and reliability, there is significant room for expansion in terms of both descriptive and prescriptive analysis. For instance, my initial plan was to add a geographic mapping of terrorism hotspots, but time constraints prevented me from exploring that. Also, I would like to consider other ways to impute or subset the data. Finally, the significance of the share of different religions to the population was constant across the models; I'd like to refine this and further analyze the importance of religious diversity as a predictor for terrorism. I would also like to improve feature engineering, as I surmise that combining some of the related indicators or dropping confounding variables would improve model accuracy.

Works Cited

[1] Orlandrew E. Danzell, Yao-Yuan Yeh & Melia Pfannenstiel (2019) Determinants of Domestic Terrorism: An Examination of Ethnic Polarization and Economic Development, *Terrorism and Political Violence, 31:3, 536-558*, DOI:/underline%7B10.1080/09546553.2016.1258636}

[2] Muhammad Nasir, Amanat Ali & Faiz Ur Rehman (2011) Determinants of Terrorism: A Panel Data Analysis of Selected South Asian Countries, *The Singapore Economic Review, 56:2, 175-187*, DOI:/underline%7B10.1142/S0217590811004225}

[3] K. Peren Arin, Oliver Lorz, Otto F.M. Reich & Nicola Spagnolo (2011) Exploring the dynamics between terrorism and anti-terror spending: Theory and UK-evidence, *Journal of Economic Behavior and Organization, 77, 189-202*, DOI:/underline%7B10.1016/j.jebo.2010.10.007}

[4] Mohammad Nurunnabi & Asma Sghaier (2018) Socioeconomic Determinants of Terrorism, *Digest of Middle East Studies, 27:2, 278-302*, DOI:/underline%7B10.1111/dome.12139}

[5] Andreas Freytag, Jens J. Krüger, Daniel Meierrieks & Friedrich Schneider (2011) The origins of terrorism: Cross-country estimates of socio-economic determinants of terrorism, *European Journal of Political Economy, 27, S5–S16*, DOI:/underline%7B10.1016/j.ejpoleco.2011.06.009}

[6] Raul Caruso & Friedrich Schneider (2011) The socio-economic determinants of terrorism and political violence in Western Europe (1994–2007), *European Journal of Political Economy 27, S37–S49*, DOI:/underline%7B10.1016/j.ejpoleco.2011.02.003}

[7] National Consortium for the Study of Terrorism and Responses to Terrorism (START). (n.d.) *Global Terrorism Database (GTD)* https://www.start.umd.edu/gtd/about/

[8] World Bank. (n.d.) *DataBank | The World Bank.* https://databank.worldbank.org/home

[9] Fariss, Christopher, 2019, "Latent Human Rights Protection Scores Version 3", https://

doi.org/10.7910/DVN/TADPGE, Harvard Dataverse, V1, UNF:6:0sWy9tSpQVpzz2xGoGLtkA== [fileUNF]. Retrieved from: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/TADPGE)

[10] Zeev Maoz and Errol A. Henderson. 2013. “The World Religion Dataset, 1945-2010: Logic, Estimates, and Trends.” *International Interactions, 39: 265-291.* Retrieved from: https://correlatesofwar.org/data-sets/world-religion-data)

[11] Hannah Ritchie and Max Roser (2013) - “Terrain Ruggedness Index”. Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/grapher/terrain-ruggedness-index [Online Resource]

[12] Max Roser (2014) - “Human Development Index (HDI)”. *Published online at OurWorldInData.org.* Retrieved from: ‘https://ourworldindata.org/human-development-index’ [Online Resource]

[13] Max Roser and Mohamed Nagdy (2013) - “Military Spending”. *Published online at OurWorldInData.org.* Retrieved from: ‘https://ourworldindata.org/military-spending’ [Online Resource]

[14] Marshall, Monty G. (2020). “Polity5 Project”, Center for Systemic Peace.

[15] Coppedge, Michael, John Gerring, Carl Henrik Knutsen, Staffan I. Lindberg, Jan Teorell, David Altman, Michael Bernhard, M. Steven Fish, Adam Glynn, Allen Hicken, Anna Lührmann, Kyle L. Marquardt, Kelly McMann, Pamela Paxton, Daniel Pemstein, Brigitte Seim, Rachel Sigman, Svend-Erik Skaaning, Jeffrey Staton, Agnes Cornell, Lisa Gastaldi, Haakon Gjerløw, Valeriya Mechkova, Johannes von Römer, Aksel Sundtröm, Eitan Tzelgov, Luca Uberti, Yi-ting Wang, Tore Wig, and Daniel Ziblatt. 2020. “V-Dem Codebook v10” Varieties of Democracy (V-Dem) Project.

[16] Wes McKinney. **Data Structures for Statistical Computing in Python**, Proceedings of the 9th Python in Science Conference, 51-56 (2010)

[17] Stadler, K. (2017). The country converter coco - a Python package for converting country names between different classification schemes. The Journal of Open Source Software. doi: 10.21105/joss.00332

[18] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[19] Sherouse, O. (n.d.). wbdata: A library to access World Bank data. Retrieved from: https://pypi.org/project/wbdata/#description