MEMORANDUM

TO: U.S. Department of Homeland Security Staff

FROM: Jonathan Hudgins, Lead Analyst, Joint Artificial Intelligence Center

SUBJECT: Terrorist Attack Prediction using Machine Learning Methods

**Issue:**

How should the U.S. Department of Homeland Security utilize machine learning methods and big data analytics to predict terrorist events better?

**Interests:**

Terrorism remains a critical national security challenge for the United States and is of particular importance to the Department of Homeland Security. There is no job more important than the safety and security of the American people.

**Analysis:**

I’ve developed a machine learning model utilizing the Global Terrorism Database (GTD), which is an open-source database including information on domestic and international terrorist attacks around the world from 1970 through 2018. The model includes 150,000 cases of terrorism and utilizes 13 different variables to predict both the probability of a “successful” terrorist attack along with the number of U.S. citizens killed.

* **The Data Set**: The original data set includes 136 different variables that could be used as predictors to help better predict the success and number of U.S. citizens killed. However, after many hours of analysis, I came to the determination that only 13 variables are necessary to achieve the best performance on the model. This determination was made through a mixture of conceptual understanding of the problem (terrorism) and my background in dealing with this issue, along with machine learning principles. Overall, the data set is extensive and required scaling back many variables that did not make sense for the desired outcome. For example, many variables were redundant information such as location variables like country, city, province, and state. With other variables like "Doubt Terrorism," I took the conservative approach. I assumed that if it's on the list, it is worthy enough of being included in the model for public policy purposes.
* Additionally, many variables with several "N.A.'s” I omitted altogether as these N.A.'s are not useful to the result. For the number of perpetrators and the number of perpetrators captured, this information was not relevant enough for this model, and there were often discrepancies in information on this value. In general, I tried to pick features that were most likely to correlate to better prediction outcomes, while being mindful of the interpretability of the model and the dataset. Some tradeoffs were made, but overall I think this new, clean dataset is most useful in attempting to predict a highly unpredictable event.
* **F Score on Success:** Utilizing Excel first to get a better sense of the data, I was able to narrow down the most pertinent variables to the problem at hand. Once the unnecessary variables were deleted, I was left with 36 most relevant variables. From there, 13 variables were used to both simplify the data set and attempt to achieve the best prediction possible. Once the variables were narrowed down, I looked at the rows in an attempt to clean the data further. In doing so, I decided to omit the “N.A.'s" from the rows, which still left 93,000 observations. Next, I split my data into 80% training and 20% test. I used logistic regression methods first to get a sense of the data and test the error rate. I then moved to KNN, both with a k=5 and a k=10. KNN proved to be challenging to utilize on this data. I then moved to a Random Forest model, which proved to be the best, both for ease of use and F Score. I was able to achieve an F Score of 0.75 with a Random Forest model utilizing the default 500 setting. This was the best F Score I was able to achieve even after adding and deleting certain variables. I also crowd-sourced this score among other students, and it appeared that this score was one of the highest. Based on this, I chose to stick with the 13 predictors and the Random Forest model.
* **MSE on the number of U.S. Citizens Killed:** When I first ran the data, I got a mean squared error of 170 for “nkillus," so I knew I needed to change something. I added and deleted predictors, which would alter the MSE some but not a lot. When thinking more about the context of the data, I decided to experiment with removing the terrorist events of September 11, 2001. This event is a “Black Swan” event and very much an outlier when viewed in context with the rest of the dataset. After doing so, my MSE went down to 3.7. However, I realized how I was deleting the event was a bit unorthodox. I was removing the row numbers instead of a value associated with the event. I then chose to update my model to delete nkillus > 100. This increased my MSE to 5, but I think it provides an overall better model as it is easier to interpret and replicate on new data.

**Conclusion:**

The most challenging aspect of this assignment is proper data cleaning to provide the most effective model possible. With 136 variables, there are at least 136 different decisions that need to be made about the content of the dataset. Picking and training an appropriate model is the next hardest step as each model comes with its' advantages and disadvantages, given the unique problem set at hand. Overall, it's important to remember that these models are useful to inform and augment human thinking and decision making, but should be used in combination with other methods of decision-making.