Samuel Luxenberg
SlideRule
Foundations of Data Science Workshop
Capstone Project

Source:
National Consortium for the Study of Terrorism and Responses to Terrorism (START). (2014). Global Terrorism Database [gtd_06to13_0814dist]. Retrieved from http://start.umd.edu/gtd

Research Question:
Given various characteristics of terrorist attacks, can we accurately predict which terrorist organization is responsible for the act? This could help identify behaviors of specific organizations and typical consequences of their attacks.

Study:
I've decided to only include data from 2011 to 2014 because I was originally motivated to work with this dataset as a consequence of the recent ISIS beheadings. It was just too hard for me to sit back and not try to do anything to help. ISIS has really become prominent in the last few years. In addition, because the years 2011 to 2014 have seen over 42,000 terrorist attacks with 562 different organizations responsible for these attacks, the goal to classify according to all of these organizations would have been computationally expensive and rather difficult. To simplify this problem I've only considered data from the top 5 most common terrorist organizations; these are the ones most people have heard about in the news. They are: Al-Qa`ida in the Arabian Peninsula (AQAP), Al-Shabaab, Boko Haram, Islamic State of Iraq and the Levant (ISIL – otherwise known as ISIS), and the Taliban.

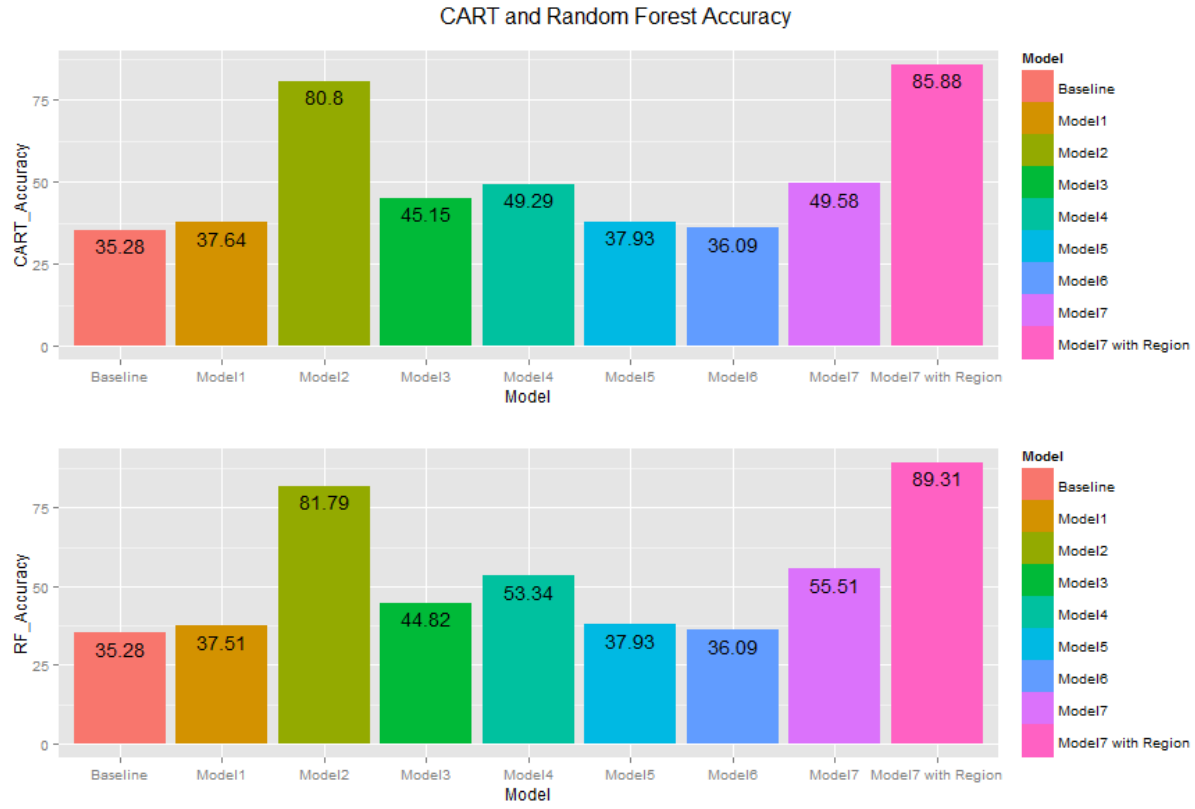
In order to classify the attacks, I used classification trees and random forests to model these outcomes. I created 9 models for both classification trees and random forests. These are based on the following variables:

-Whether or not the incident occurred within a city or within the immediate vicinity of a city
-Whether or not the attack was a success
-Whether or not the attack was a suicide attack
-Whether or not there was property damage
-The year, month, and day of the incident
-Whether or not the incident had extended past 24 hours
-Region in which the incident occurred
-Whether or not the goal was political, religious, economic or social
-Whether or not the goal was to coerce, intimidate, or publicize to larger audiences
-Whether or not the incident was outside international humanitarian law
-Whether or not the incident was part of several separate but connected attacks
-Whether or not there was a possibility that incident was not actually a terrorist attack
-Method of attack
-Target/Victim type
-Whether or not the information reported by sources about the terrorist organization name is based on speculation or dubious claims of responsibility
-Type of Weapon used in the attack
-Whether or not there was a claim of responsibility
-Whether or not victims were taken hostage
-Whether or not victims were of multiple nationalities

-Whether or not there were multiple types of targets/victims
-Whether or not there were multiple types of weapons used
-Whether or not there were multiple methods of attack
-Number of U.S. citizens injured
-Number of U.S. citizens killed
-Number of perpetrators involved in the incident
-Number of perpetrators captured

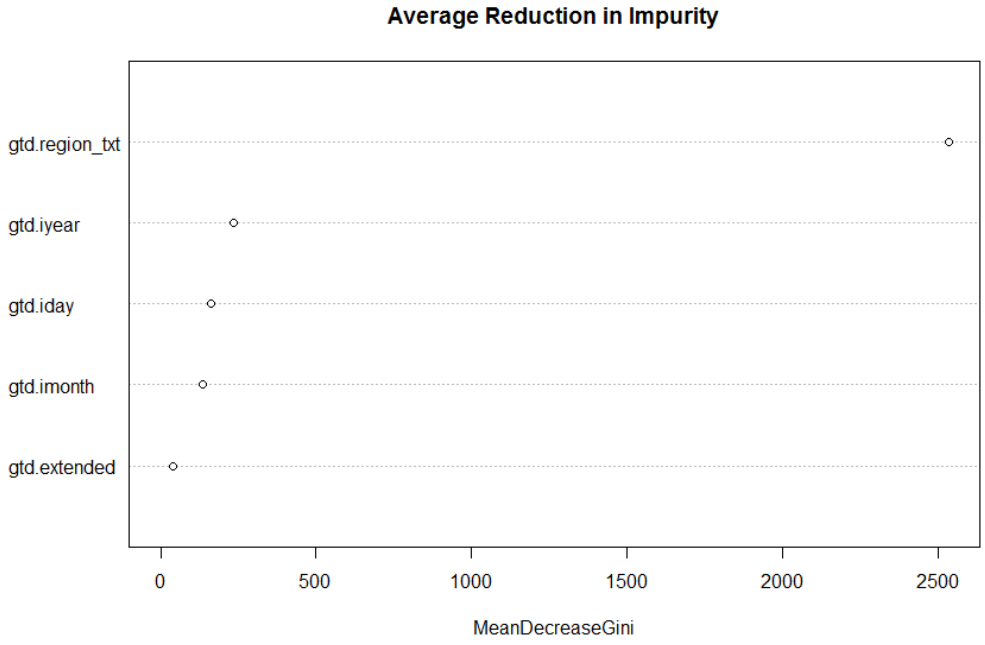
Models:
The classification trees and random forests were compared to a single baseline model which always predicts the most common outcome. In this case the model always predicts that the Taliban is responsible for an attack, which is correct about 35.28% of the time.

In the graphic below, the bar graphs show the accuracy percentage for the classification tree models (top) and the random forest models (bottom).

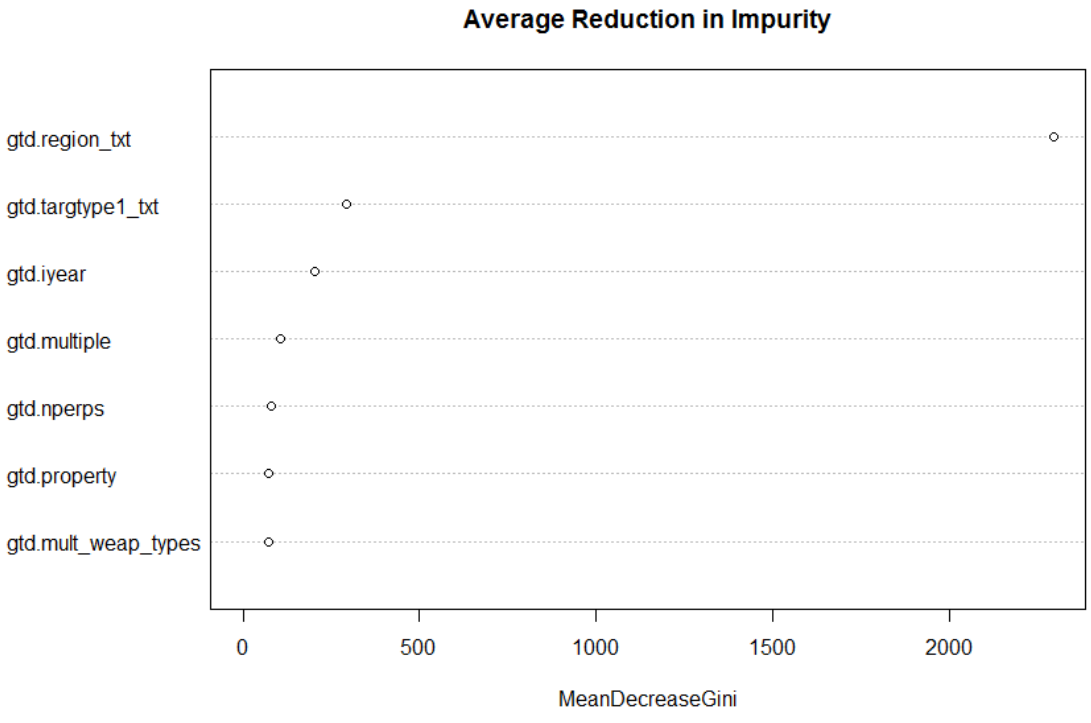


We can see here that for most of the corresponding models, the random forests give at least as good accuracy as their classification tree counterparts. Also, notice that Model 2 and Model 7 with Region are significantly more accurate than any of the others. This jump in accuracy occurs because both have the variable, region in which the incident occurred, built into their models. Since the 5 terrorist organizations I considered are dominant in certain parts of the world, Al-Qa`ida and ISIS in the Middle East and North Africa, the Taliban in South Asia, and Al-Shabaab and Boko Haram in Sub-Saharan Africa, it is obvious that region is probably the most defining characteristic of attacks between these groups. If I

were to continue this project, I would expand the list of terrorist organizations so that there isn't such a clear divide between regions. In the graphics below, we can see that the region variable is responsible for the highest average reduction in impurity, a measure used to evaluate the importance of variables within a model.



(Random Forest Model 2)

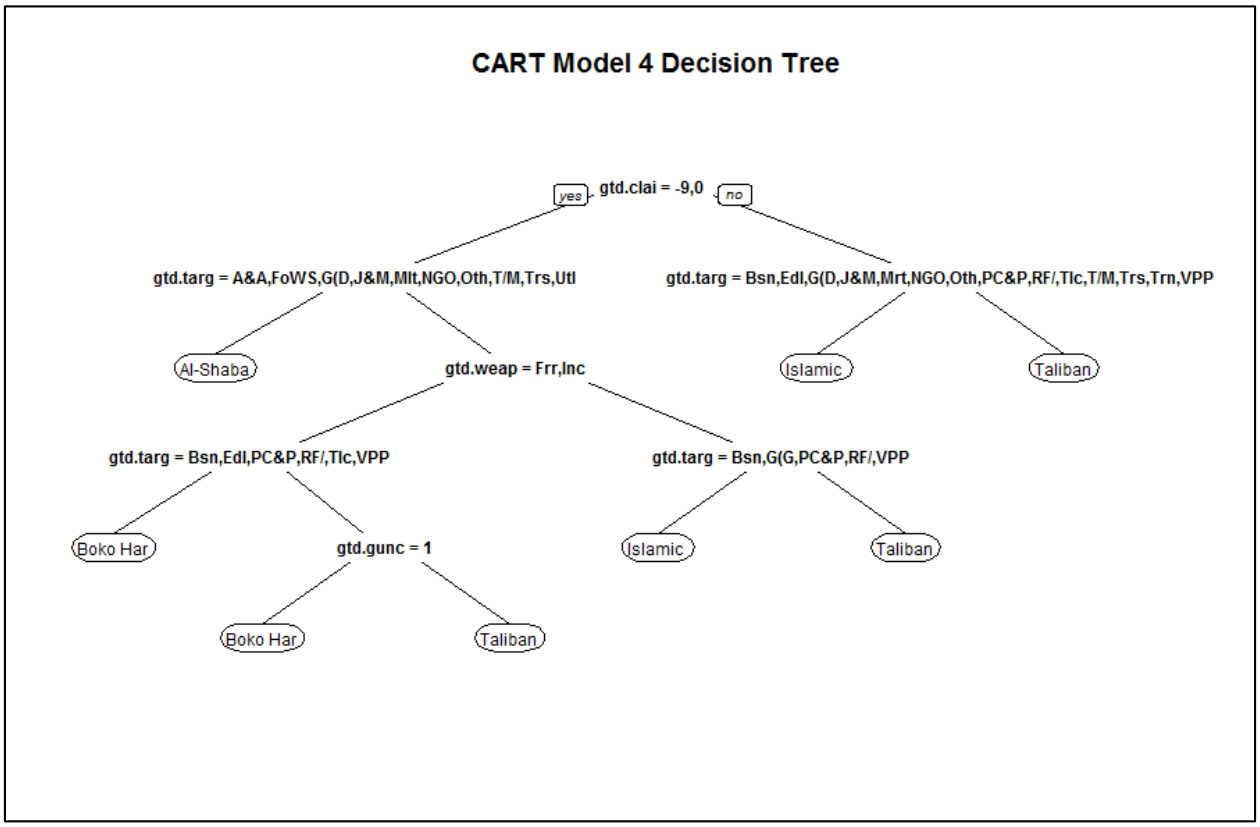


(Random Forest Model 7 with Region)

If we take region out of the model, the models do become significantly less accurate, but some may still be useful in directing us towards relatively important characteristics of terrorist attacks. Let's consider our two best models that don't incorporate region.

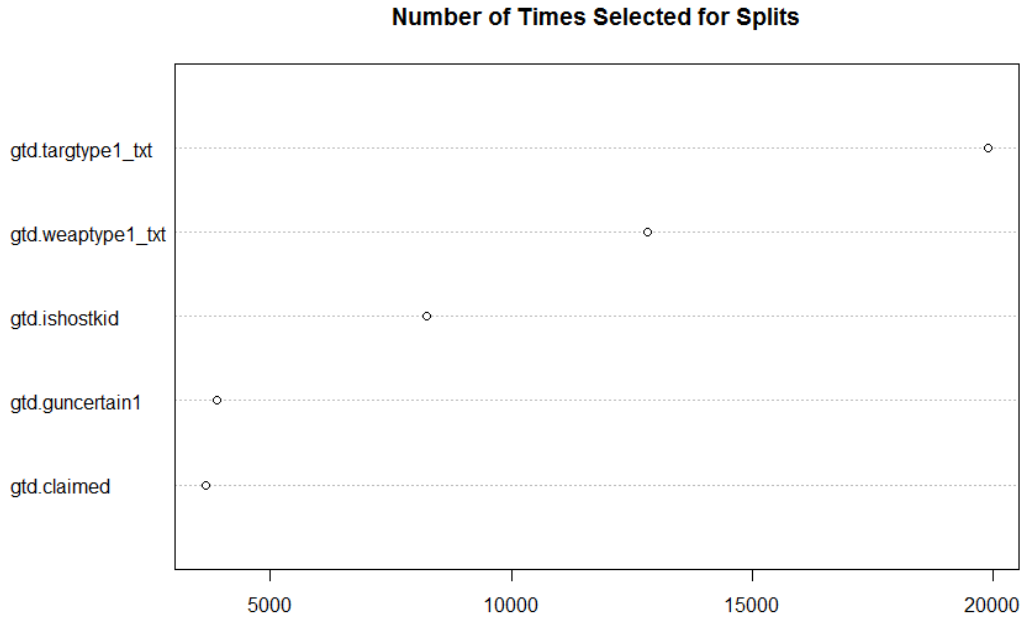
The accuracy of the classification trees of both Model 4 and Model 7 is just under 50%, whereas the random forest of Model 4 and Model 7 are both a little more than 50% accurate.

Consider Model 4, which has variables: type of target, whether or not the information reported is based on speculation, whether or not the responsible organization claimed responsibility, type of weapon used, and whether or not victims were taken hostage.



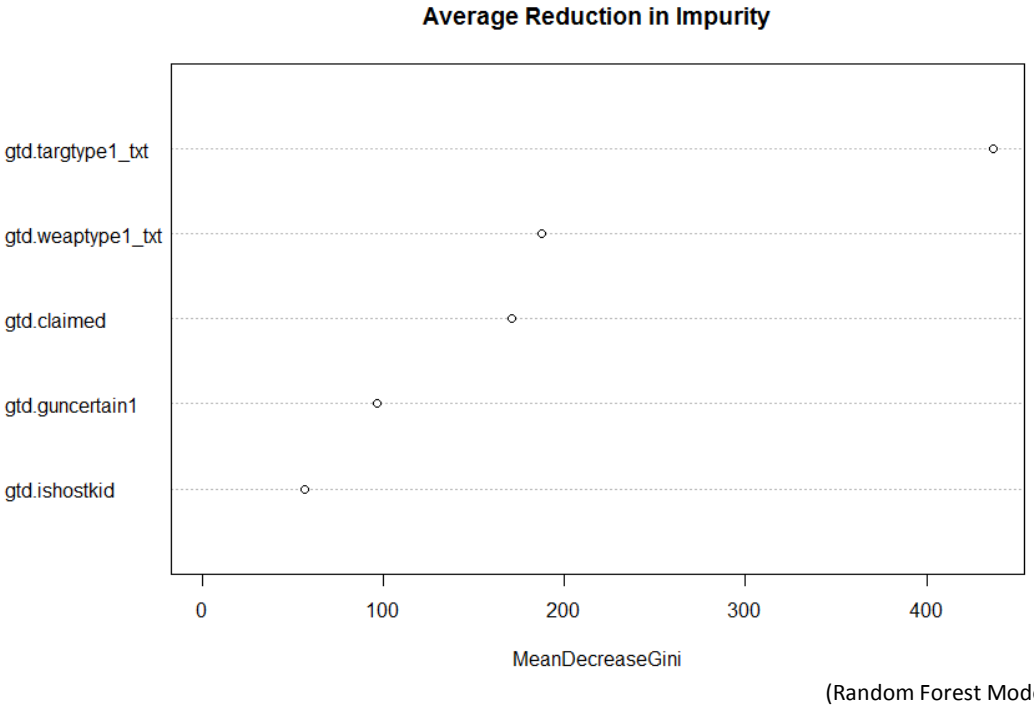
We can see here that while this tree is somewhat difficult to interpret, the most important variables are the type of target and whether or not the terrorist organization claimed responsibility for the attack. If the organization did claim responsibility then this model classifies this attack as either being carried out by ISIS or the Taliban depending on the type of target. If the target was a business, an educational institution, the government, the media, maritime, NGOs, private citizens and property, religious figures and institutions, telecommunication, other terrorists and non-state militias, tourists, transportation, or political parties, then the attack would be classified as an attack by ISIS, otherwise by the Taliban. In the event that either it is unknown whether or not the responsible organization claimed responsibility or there was no claim, this models predicts any of the organizations except for Al-Qa`ida.

As evidenced below, in the random forest Model 4, the type of target is still the most important characteristic of an attack, but now whether or not the organization claimed responsibility is less important and is replaced with the type of weapon used to facilitate the attack.



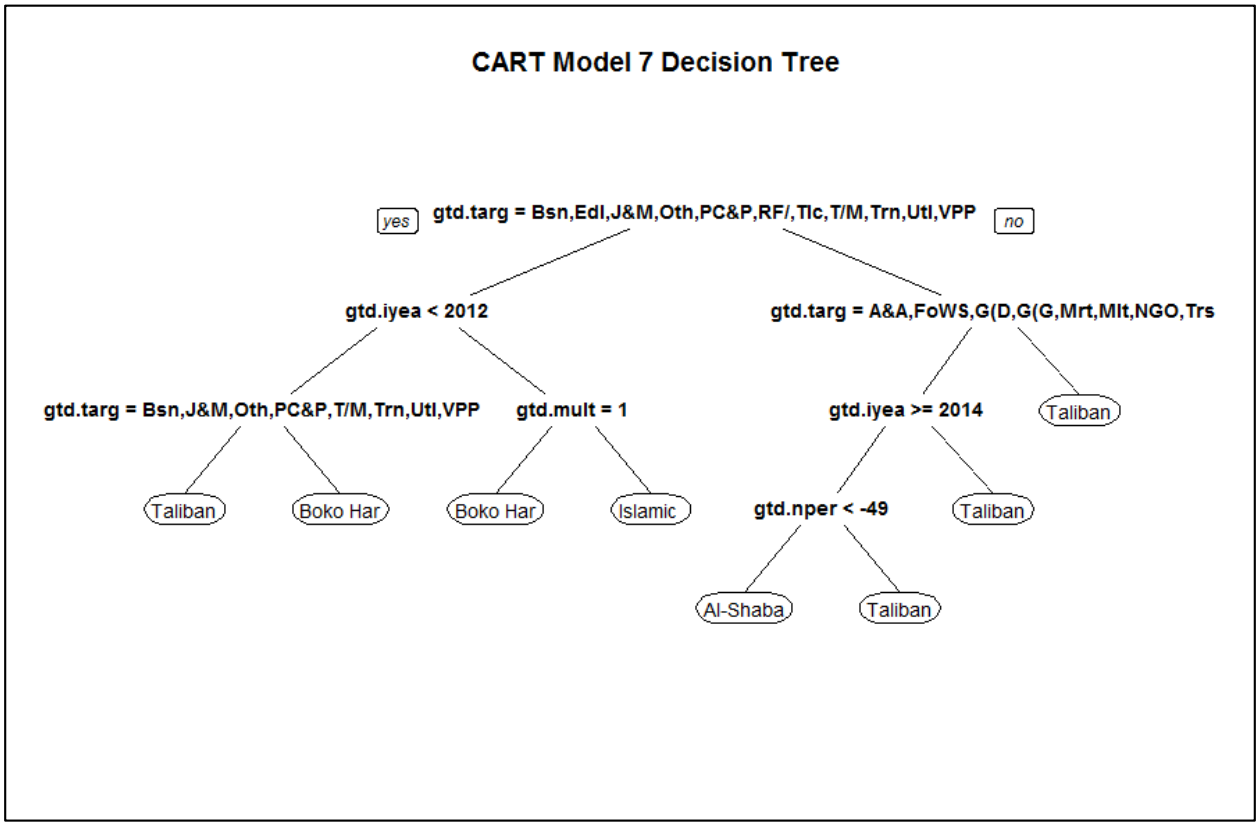
(Random Forest Model 4)

Another measure of variable importance is the number of times a variable is selected for a split in the random forest.



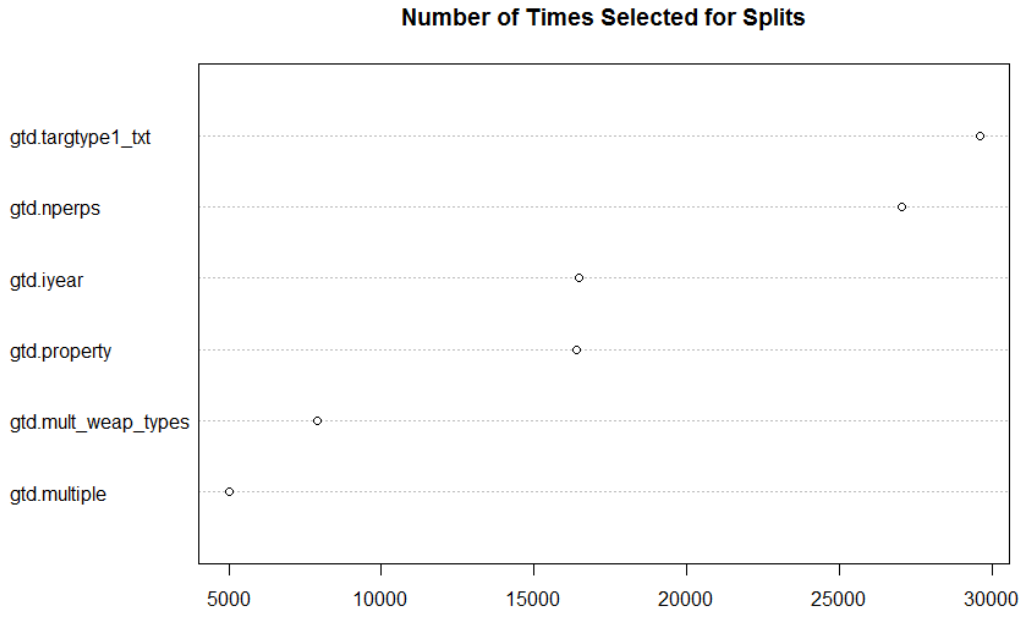
(Random Forest Model 4)

Finally, let's consider Model 7 (without region). For this model, I decided to use all of the most important variables throughout all the other classification tree models (other than region). These included: the number of perpetrators involved in the attack, whether or not there were multiple types of weapons used, the type of target, whether or not the attack was part of a series of separate but connected attacks, the year of the incident, and whether or not there was property damage.

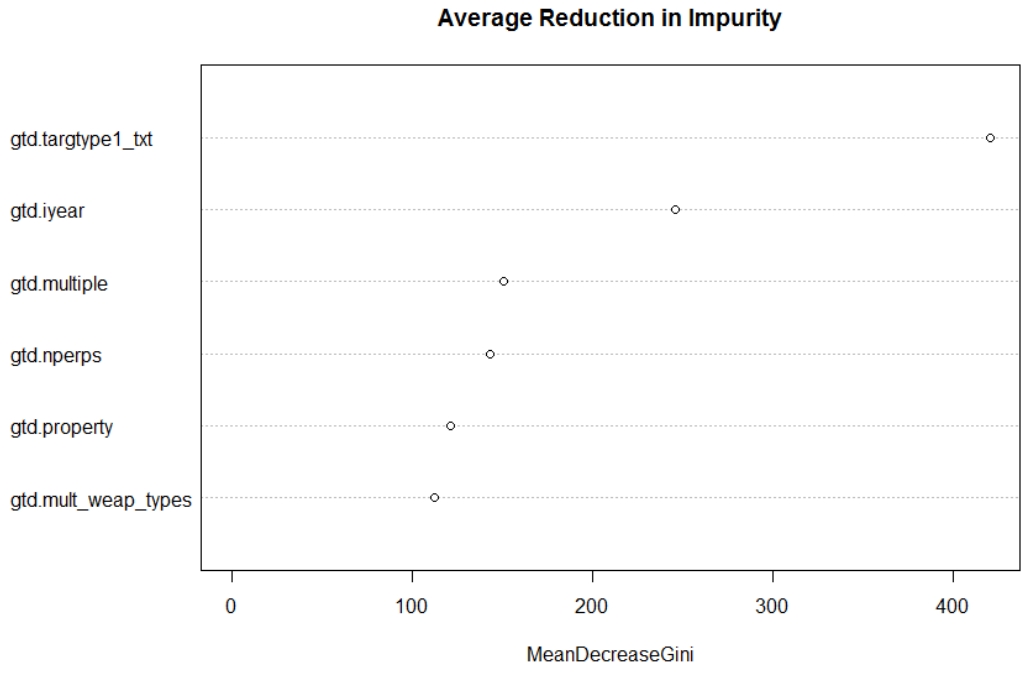


So, of all the most important variables according to classification tree models, the type of target and year (along with region) bear the most significance when classifying according to these five terrorist organizations. It should be noted that in the most accurate of these models, Al-Qa`ida doesn't even show up as one of the leaves on the trees. Perhaps this means it's more difficult to classify Al-Qa`ida's actions and there is no pattern. More research about this specific issue should be done. I'm no foreign policy expert, but maybe this is a reason for their success in carrying out terrorist attacks.

If we look at the random forest Model 7, we can see that except for the type of target, the order of the variables in terms of importance changes when looking at both the average reduction in impurity and the number splits. This suggests that the type of target may indeed be the most important characteristic of an attack (besides region) in order to classify according to these organizations.



(Random Forest Model 7)



(Random Forest Model 7)

Conclusion:

Using classification trees and random forests, I was able to predict with high accuracy the terrorist organization responsible for the attack given the region the attack occurred. Since these organizations are largely restricted to certain geographic locations, it's obvious that the region variable would do the

best job at classification. If I had included the other 557 terrorist organizations who committed attacks between 2011 and 2014, it's likely that region would not have been as important as it is here. The better models without the region variable were able to classify attacks correctly about half the time. While this isn't good enough for any government to take any action based on these models, they do serve as good starting points, noting that the types of targets of terrorist attacks do reflect some tendencies of the five most prominent organizations. In future research, hopefully the models will be significantly more accurate and could perhaps assist governments in preparing likely victims for potential attacks.