the Prosecution Project
Final Report

Athena Chapekis, Jing Lin, Ruoqi Tan, James Wieneck
April 17th, 2019

Non-technical Summary

Introduction

We are presented with a data set involving individuals who were indicted and prosecuted for crimes which have socio-political motivations and/or crimes that have rendered them as designated terrorists in the United States. These cases involve various *identity* variables of the defendants (age, race/ethnicity, gender, "othered" status, religion, citizenship status, and veteran status), as well as various criminal *activity* variables (people vs. property, number injured, number killed, physical target, ideological target, and tactic). The question we seek to answer is: how do aspects of a defendant's identity play a role in their criminal activity?

Results

The statistical result shows unbalanced levels among identity variables. For gender, the vast majority of offenders are male. For race and religion, 'Muslim' appear more frequently. Most cases have civilian status. The most common tactic is 'Providing material/financial support to terrorist organization', 'Unspecified' appear most frequently as ideological target, 'Online' appear most frequently as physical target.

All identity variables have significant relationships with activity variables, however the actual size of the effect varies across different variables. Gender and othered status affect the number of persons killed or injured significantly, with men and othered defendants having a higher injury count. Age was a consistently influential variable when examining how trends in criminal activity are influenced by one's identity across the board, though it almost always had some interaction with citizenship status, veteran status, and/or othered status. Othered status was also a highly influential variable in predicting different trends in criminal activity.

Conclusions

This report finds that the identity variables which have the greatest prediction effect of criminal activity are Othered Status, Religion, Ethnicity/race, Citizenship Status, and Veteran Status. Gender is a significant predictor of the number of killed and injured by a crime but is *not* a significant predictor of other criminal activity variables.

The models we built in predicting trends in criminal activity based on the identities of the defendants had poor predictive power, in part because of unused scenarios and unspecified cases for multiple variables. The data set used for the analysis may likely need more information provided to give a more complete picture of how criminal activity is linked to a defendant's identity.

Technical report

Introduction

The definition of what constitutes "terrorism" is not a unanimous one. Different sources report different standards for what an act of terror entails. Because of this, there has not been a thorough body of research built on terrorism in all its forms. Issue-specific groups like the Department of Justice (DOJ)/Federal Bureau of Investigation (FBI), the Center for Biomedical Research (CBR), and the National Abortion Federation (NAF) have collected their own databases of terrorism and terrorists over time, but they generally focus on one specific ideological group – whichever is of the greatest concern to them.

The Prosecution Project (tPP) is a large-scale project out of Miami University that seeks to construct a database of all acts of terrorism and socio-politically motivated crimes ending in felony prosecutions in the United States 1990-present. Each case in tPP's database is coded across 44 variables, including demographic information on the defendant, details of their affiliations, details of the crime they committed, and details of the legal proceedings.

This report seeks to investigate the connection between a defendant's identity (i.e. their demographic information) and their criminal activity and provide an answer to the question of how *who* someone is relates to *what* they do.

Methodology

The first step in approaching this analysis is to clean the data. Categorical variables which have many levels are reduced to allow for better comparison and analysis. Much of this reduction was done using the classification provided by the Prosecution Project codebook. For example, in the variable Physical Target, the levels of 'Federal site: non-military non-judicial', 'Federal site: military', 'Federal site: judicial', and 'Federal site: non-U.S. embassy or consulate' are combined and recoded simply as 'Federal site'. Furthermore, the levels for 'State site' and the levels for 'Municipal site' are combined with 'Federal site' to make one unified level of 'Governmental site'. This is done for the variables of Physical Target and Ideological Target. Due to the low representation in many of the levels for the variable 'Tactic', many levels were combined into an 'Other' level. Other categorical variables that were not recoded but included in this report in their original state are People vs. Property, Gender, Ethnicity, Religion, 'Other' Status, Citizenship Status, and Veteran Status. For each categorical variable, a bar chart is generated to compare frequencies of levels.

To conduct an analysis, this report begins with T-tests to determine the influence binary predictor variables Gender (male v. female), Othered Status (othered v. non-othered), and Veteran Status (citizen v. non-citizen) may have on number of people killed and number of people injured in socio-politically-motivated crimes. A significance level of 0.05 is used. Furthermore, Analysis of Variance (ANOVA) tests are used to test for significant differences in the number of people killed and the number of people injured between demographic groups for the identity variables of Race/ethnicity, Religion, and Citizenship Status. As well, ANOVA tests are used to see if a defendant's age differs significantly between the types of things that are targeted in socio-political crimes (both physically and ideologically) and if age differs significantly between types of tactics. On top of the ANOVA tests, Eta Squared values are calculated to test for effect size in the relationships (Brown). To investigate relationships between categorical identity

variables (e.g. Religion, Citizenship Status, etc.) and categorical activity variables (e.g Tactic, Physical target, etc.) Chi-Squared Tests of Independence are used. As well, Cramer's V is used to calculate effect size for the respective relationships between these categorical variables.

Initially, this report sought to use linear regression to create a predictive model of trends. However, we have found that due to the categorical nature of many of the variables (often with many levels) and given there are different trends among differing variables related to the crime, it is not advisable that we attempt to build regression models based on a singular response variable. Instead, we will want to use classification tree modeling for the categorical variables whose trends we want to analyze and ANOVA tree modeling for the numerical variables whose trends we want to analyze.

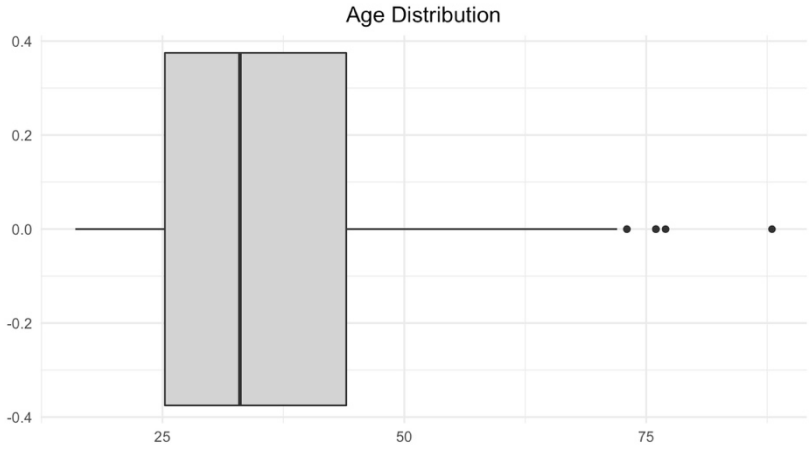
We will be using classification trees for the following variables: People vs. Property, Physical Target, Ideological Target, and Tactic; we will be using ANOVA/regression trees for the following variables: Number Injured and Number Killed. These will be considered as our criminal activity variables for this portion of the analysis. The identity variables we are using in this portion of the analysis are age, gender, race/ethnicity, religion, othered status, veteran status, and citizenship status. The purpose of this portion of the analysis is to see which aspects of a criminal's identity are most often associated with various aspects of criminal activity, and also how these aspects interact or intersect. To validate the results from our classification and regression trees, we will also be using random forests for each model to see which variables are most significantly linked to each criminal activity variable, and to see which variables the most significant contributors were to differences in criminal activity trends (Liaw). For each random forest, 1,000 classification trees will be generated.

Results

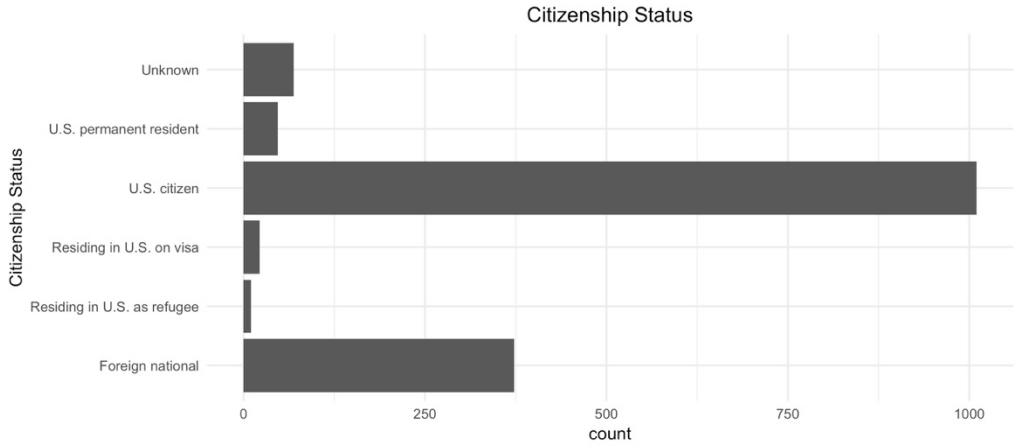
For most of the categorical variables, there are a number of levels which appear in the data very infrequently.

Identity variables

Looking at the demographics of the data, we see fairly uneven representation among levels for almost all of the variables. As far as gender, the data is overwhelmingly male, and the levels of 'Non-binary/gender non-conforming' and 'Unknown/unclear' are used virtually never.



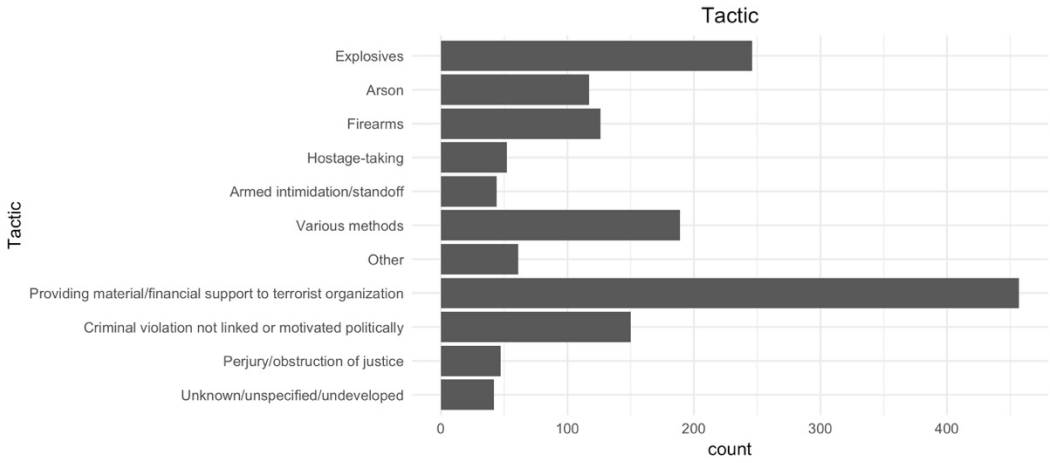
Ages range from 16 to 88 with a median age of 33 and a mean age of 35.9. The ethnicities of 'Biracial' and 'American Indian/Alaskan Native' hardly occur, and for Religion, 'Jewish' and 'Other' appear very infrequently. As well, 'Christian' and 'Christian Identity', while occurring somewhat more often, do not occur in the data nearly as often as 'Muslim' and 'Unknown'.



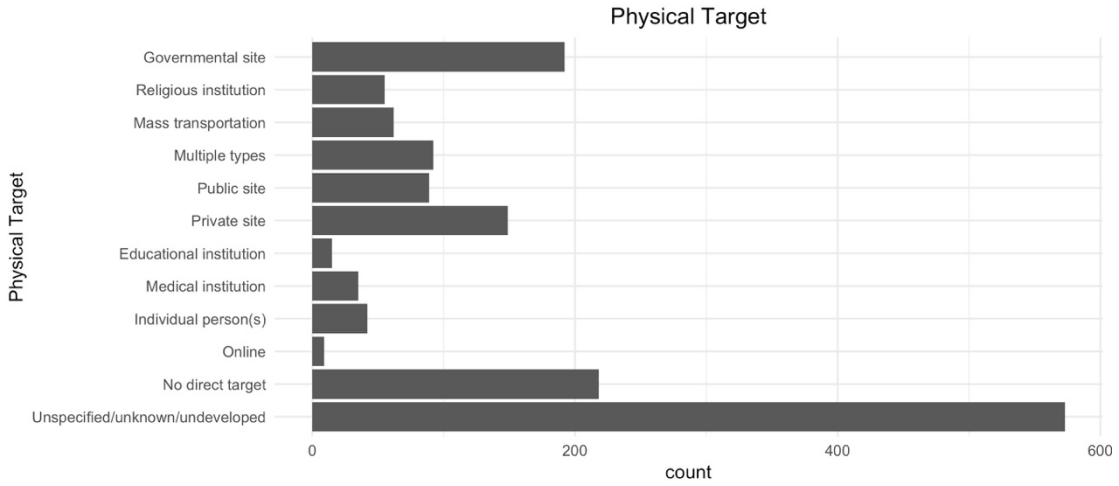
Regarding Citizenship Status, all levels are relatively infrequent compared to 'Civilian' and 'Foreign national'. There are more cases marked as 'Othered' than 'Non-othered', but both are well-represented in the data. Lastly, when looking at Veteran Status, almost all cases are coded 'Civilian'. All othered statuses are fairly uncommon and combined make up only about 16% of the data.

Criminal activity variables

The most commonly occurring tactic by far is 'Providing material/financial support to terrorist organization'. After that, 'Explosives', 'Criminal violation not linked or motivated politically', 'Various methods', 'Arson', and 'Firearms' occur most frequently.



All levels in the People vs. Property variable are fairly well represented. Regarding targets, for Ideological Target, 'Unspecified' is the most frequently occurring level in the data followed by 'Government', but all levels aside from those do appear to occur at similar rates. For Physical Target, the levels of 'Online', 'Educational institution', and 'Municipal site' do not occur frequently.



Analysis of Variance (ANOVA)

From the results of ANOVA test, the F test shows that race, religion, and citizenship have significant influence on number of killed and injured. The identity variable age has significant relationship with the activity variables people or property, physical target, ideology target, and tactic. The eta squared test shows that citizenship has larger effect on number of killed and injured than race and religion, and ideological target has the largest effect on age.

	ANOVA & Eta Squared (η^2)	**Response Variables**	
		Number killed (F(p), η^2)	**Number injured (F(p), η^2)**
Predictor Variables	**Race/ethnicity**	2.677(*0.009*), 0.012	2.752(*0.008*), 0.013
	Religion	2.629(*0.023*), 0.009	2.76(*0.017*), 0.009
	Citizenship	5.105(<*0.001*), 0.017	6.046(<*0.001*), 0.02
		Age (F(p), η^2)	
	People v. Property	5.994(<*0.001*), 0.017	
	Physical target	5.112(<*0.001*), 0.076	
	Ideological target	11.76(<*0.001*), **0.106**	
	Tactic	3.818(<*0.001*), 0.051	

Student's T-test

Regarding the number of people killed by a crime, we can be 95% confident that, on average, for each death caused by a woman's crime, men's crimes kill between 0.08 and 8.76 more people. For the differences in the number of people injured, we can say with 95% confidence that, on average, men injure anywhere between 16.11 and 52.71 more people than women in the course of a socio-politically motivated crime. There is no statistically significant difference in fatalities between crimes committed by othered and non-othered defendants, however, we can be 95% confident that othered defendants injure between 20.15 and 76.3 more people in the course of their crime than non-othered defendants. As well, there is no statistically significant difference found in the number of people killed or the number of people injured between the those who are civilians and those who were not.

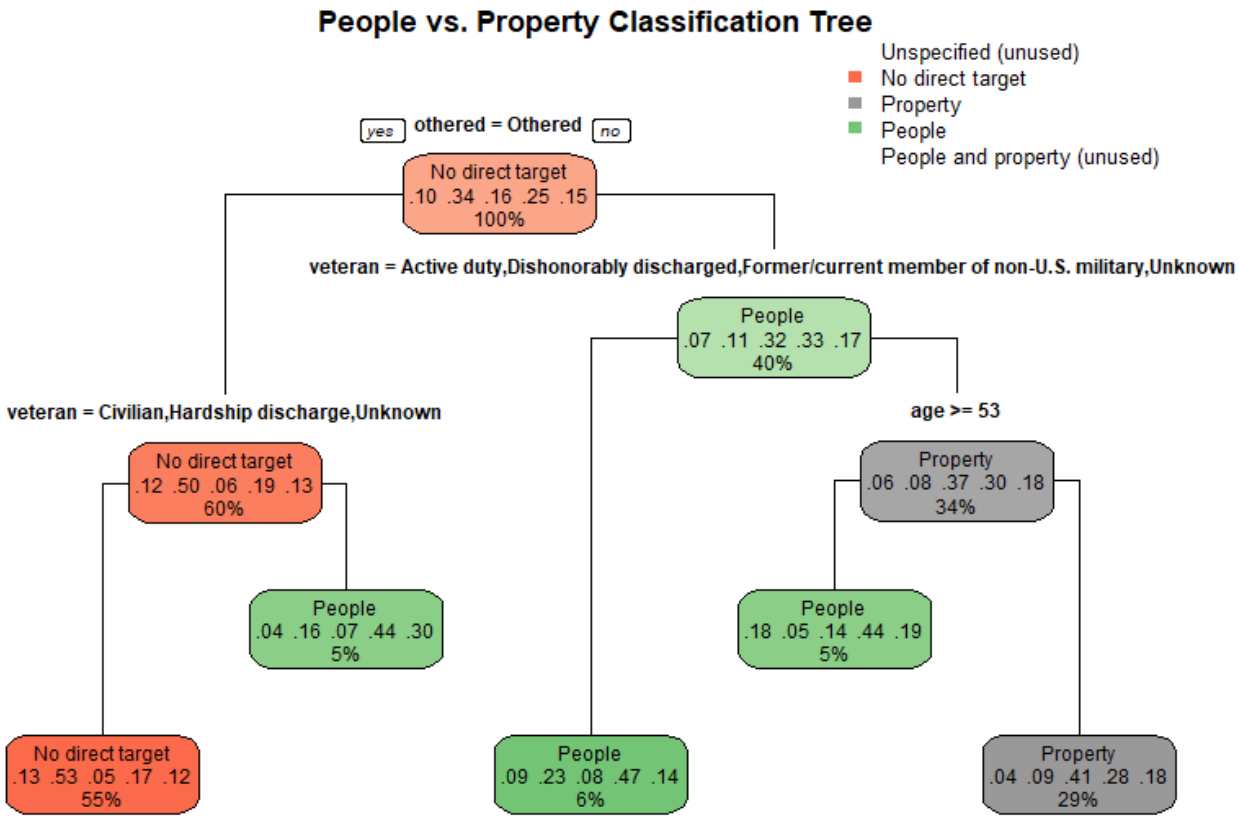
Chi-Squared and Cramer's V

The results of the Chi-Squared Test of Independence showed widespread statistical significance between all identity variables and all criminal activity variables. When Cramer's V is calculated for effect size, however, it appears that many identity variables have a weak effect on criminal activity. Specifically, gender seems to have the least effect on criminal activity. Othered Status has a particularly significant effect on criminal activity, so much so that Cramer's V indicates Othered Status may be measuring the exact same trends as the criminal activity variables.

Chi-Squared (χ^2) & Cramer's V		**Response Variables (T(*p*), *V*)**			
		Tactic	**People v. Property**	**Physical target**	**Ideological target**
Predictor Variables	**Veteran**	69.46(*<0.001*), ***0.22***	49.83(*<0.001*), *0.19*	81.1(*<0.001*), ***0.24***	65.11(*<0.001*), ***0.22***
	Religion	665.5(*<0.001*), ***0.29***	256.2(*<0.001*), *0.2*	557.7(*<0.001*), ***0.27***	546.3(*<0.001*), ***0.27***
	Race/ethnicity	776.8(*<0.001*), ***0.27***	420.1(*<0.001*), ***0.26***	569.2(*<0.001*), ***0.23***	485.5(*<0.001*), ***0.21***
	Citizenship	397.3(*<0.001*), ***0.23***	228.8(*<0.001*), *0.19*	341.7(*<0.001*), ***0.21***	232(*<0.001*), *0.17*
	Gender	67.6(*0.001*), *0.12*	24.0(*0.02*), *0.07*	67.49(*<0.001*), *0.12*	57.43(*<0.001*), *0.11*
	Othered	541.5(*<0.001*), ***0.59***	375.25(*<0.001*), ***0.5***	489.7(*<0.001*), ***0.57***	484.4(*<0.001*), ***0.56***

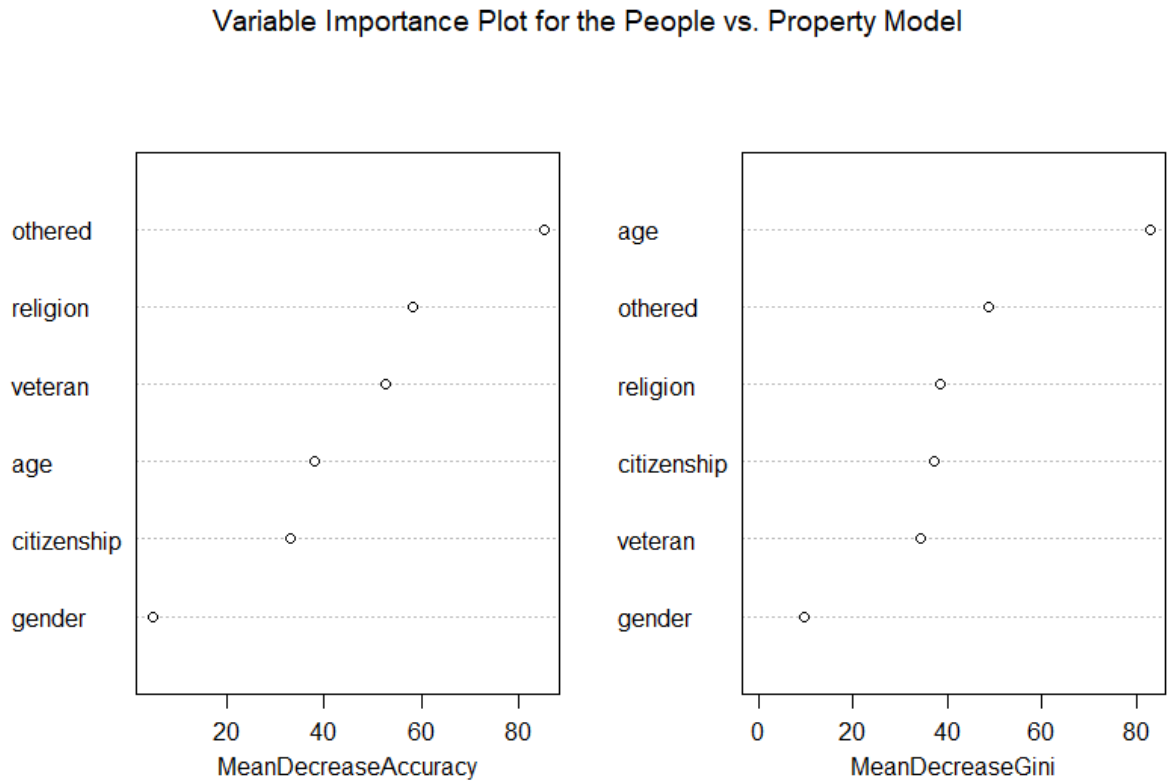
Classification/Regression Trees and Random Forests

Figure 1. The classification tree for the people vs. property variable. At least 50 cases were required for each split, and each final outcome required at least 50 cases.



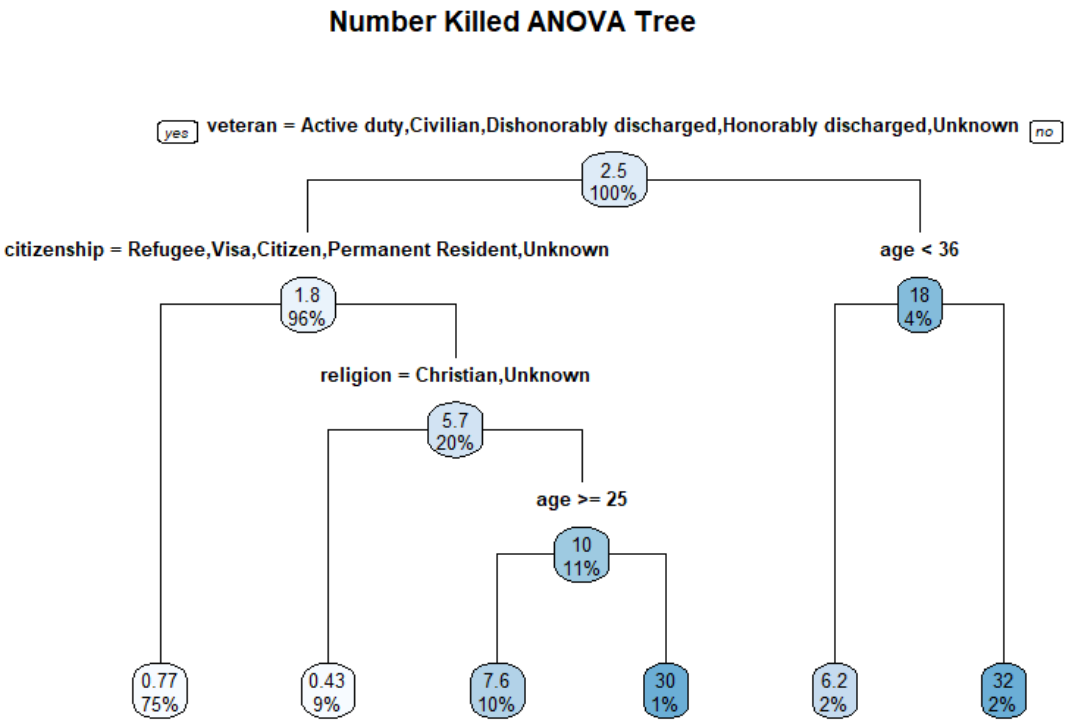
What we have been able to see is that for predicting the trends in whether a target is human or property, othered status appears to interact with veteran status and age. Othered defendants are more likely to either have targeted people or have no direct target (Figure 1). Of othered defendants who were of civilian status, released on hardship discharge, or whose veteran status was unknown, no direct target was identified; otherwise, people were more likely to be targeted. Among those of non-othered status, those whose veteran status was active duty, dishonorably discharged, belonging to a non-U.S. military, or unknown were more likely to target people. Among those who were not of those veteran statuses, age was an additional factor; those and who were 52 and under were more likely to target property, and those 53 and over were more likely to target people (Figure 1). We can see that the most significant variables which made a difference in the trends in which type of target was involved were othered status, veteran status, and age, in this order.

Figure 2. The variable importance plot for the people vs. property random forest model.



After conducting a random forest on the data used to build the classification model and plotting the importance of each variable, we find that veteran status, othered status, and age are the largest contributors to the differences in which types of targets defendants tend to target. Age and othered status were particularly strong determinants in these patterns (Figure 2).

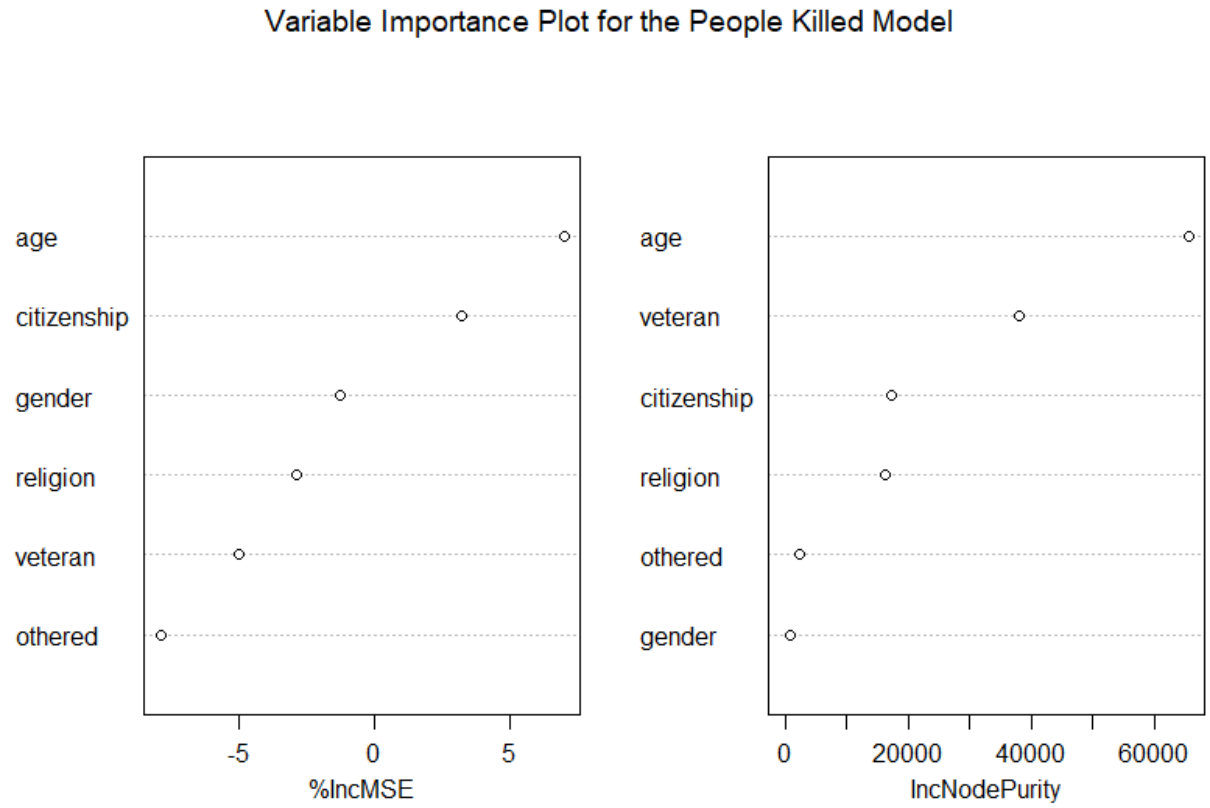
Figure 3. The ANOVA/regression tree for the number killed variable. At least 20 cases were required for each split, and each final outcome of the tree required at least 15 cases.



What we can see for the number of fatalities in each crime is that there is a split at veteran status. Those whose veteran status was either active duty, civilian, dishonorably discharged, honorably discharged, or unknown had an average of 1.8 fatalities (Figure 3). Among that group, the average number of people who were killed as a result of a defendant whose citizenship status was either refugee, residing on a visa, a citizen, a permanent resident, or unknown had a fairly low average of 0.77 (Figure 3). Among defendants who were not of these citizenship statuses, there was an average of 5.7, and another split at religion (Figure 3). Those whose religion was identified as Christian or unknown had fairly low average fatalities at 0.43, which was lower than for those whose religions fell outside of these 2 categories at 10 (Figure 3). From there, age was a major determinant in the number of fatalities. Those who were under 25 had, on average, the second-most fatalities at 30, and those who were 25 or older only had 7.6 fatalities on average (Figure 3).

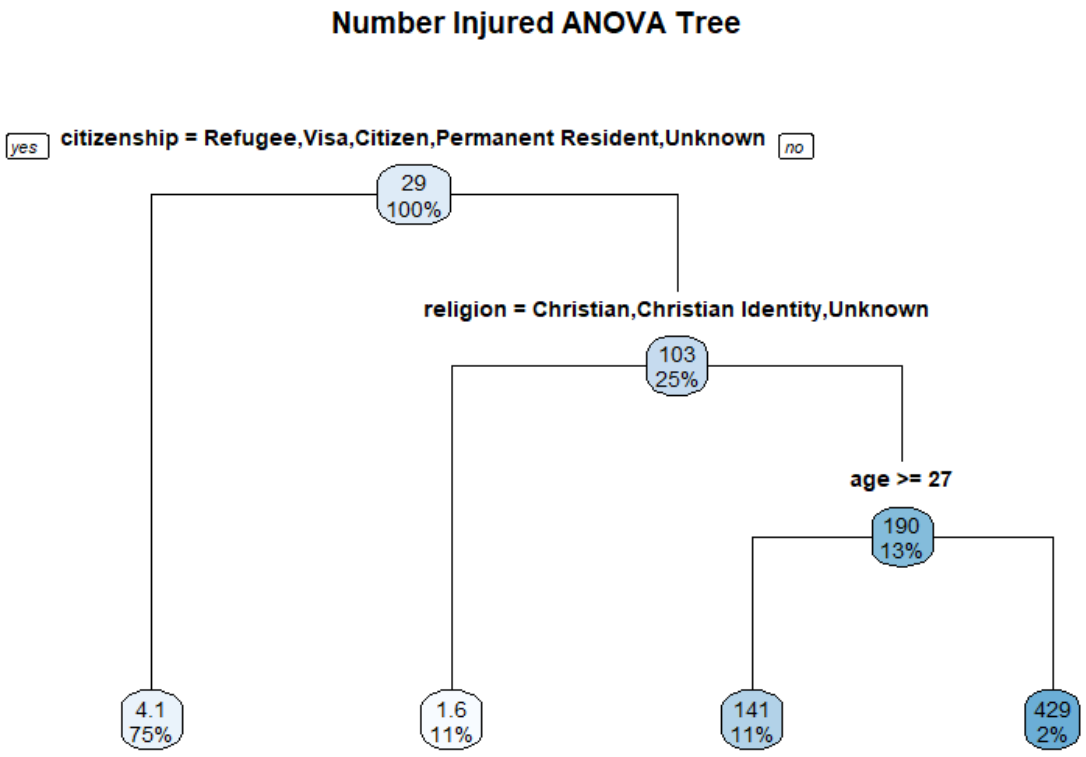
For defendants who were a former or current non-U.S. military member or who were discharged on the basis of hardship, the average number of fatalities was 10 times higher than defendants not of these veteran status categories at 18 fatalities (Figure 3). We notice that, from here, there is a split at age; those who were 35 or younger had an average fatality count of 6.2, whereas those who were 36 or older had an average fatality count of 32 (Figure 3). We can see that the most significant variables in predicting differences in the number of people killed were veteran status, citizenship status, religion, and age.

Figure 4. The variable importance plot for the people killed random forest model.



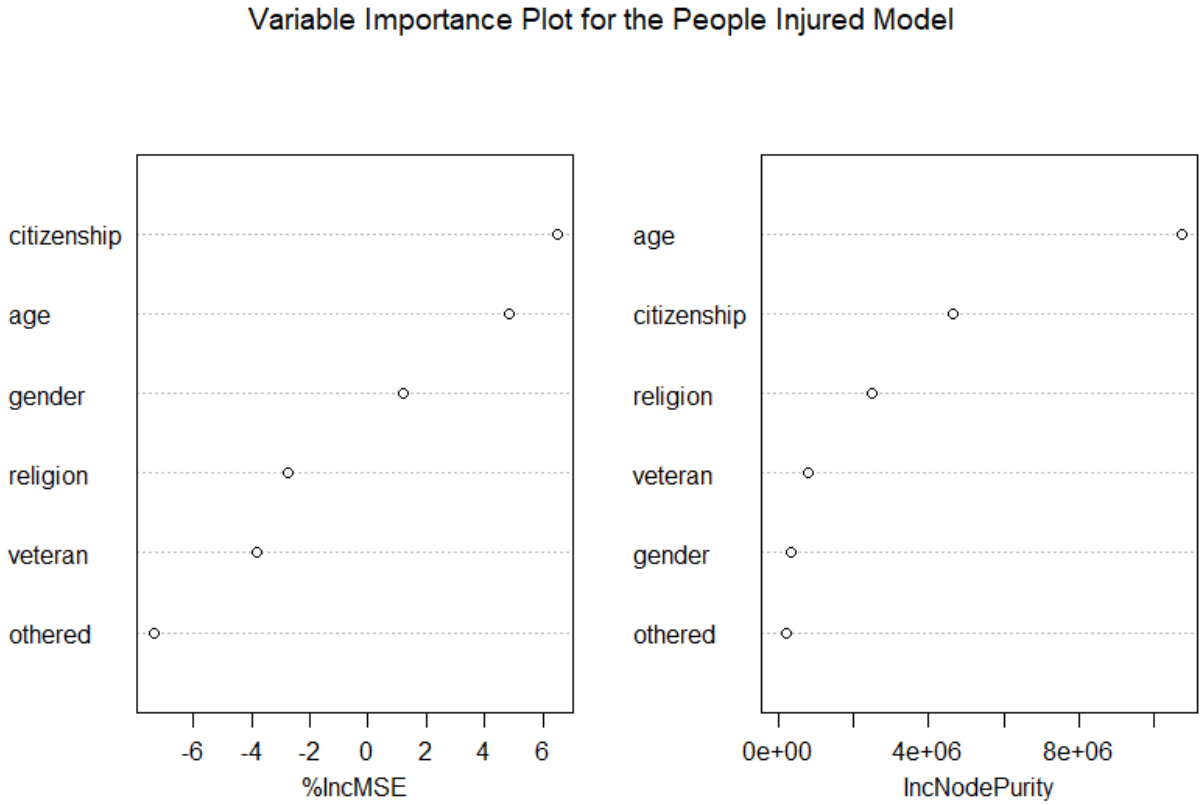
After conducting a random forest on the data used to build the regression/ANOVA model and plotting the importance of each variable, we find that age is a very significant predictor in determining the differences in fatalities among each case of crime (Figure 4). However, we cannot ignore the influence of veteran status or citizenship status, as they were significant variables on which the regression trees were split, and the variable importance plot also reflects this (Figure 4).

Figure 5. The ANOVA/regression tree for the people injured variable. At least 50 cases were required for each split, and each final outcome of the tree required at least 25 cases.



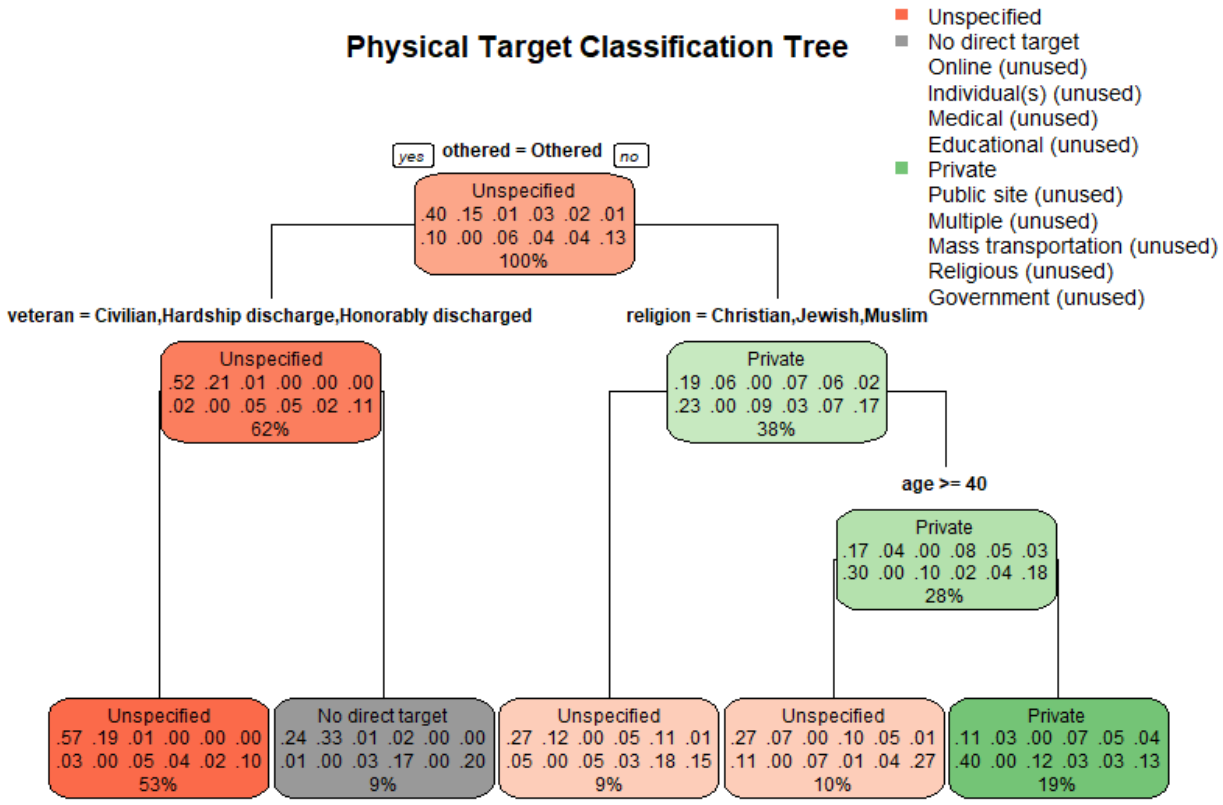
Looking at our results in Figure 5, we find that among defendants who were U.S. citizens, refugees, residents on a visa, permanent residents, or of unknown citizenship status, the average number of people injured was 4.1. For defendants who were not, there was a split at religion; those whose religion was identified as Christian, Christian Identity, or unknown had an average of 1.6 injuries (Figure 5). Among those whose religions were not in those categories, there was a split at age. For those who were 27 or older, the average number was 141, and for those who were 26 or under, the average number was 429 (Figure 5). We can conclude from this tree that citizenship status, religion, and age were important factors in predicting the differences in the number of people injured.

Figure 6. The variable importance plot for the people injured random forest model.



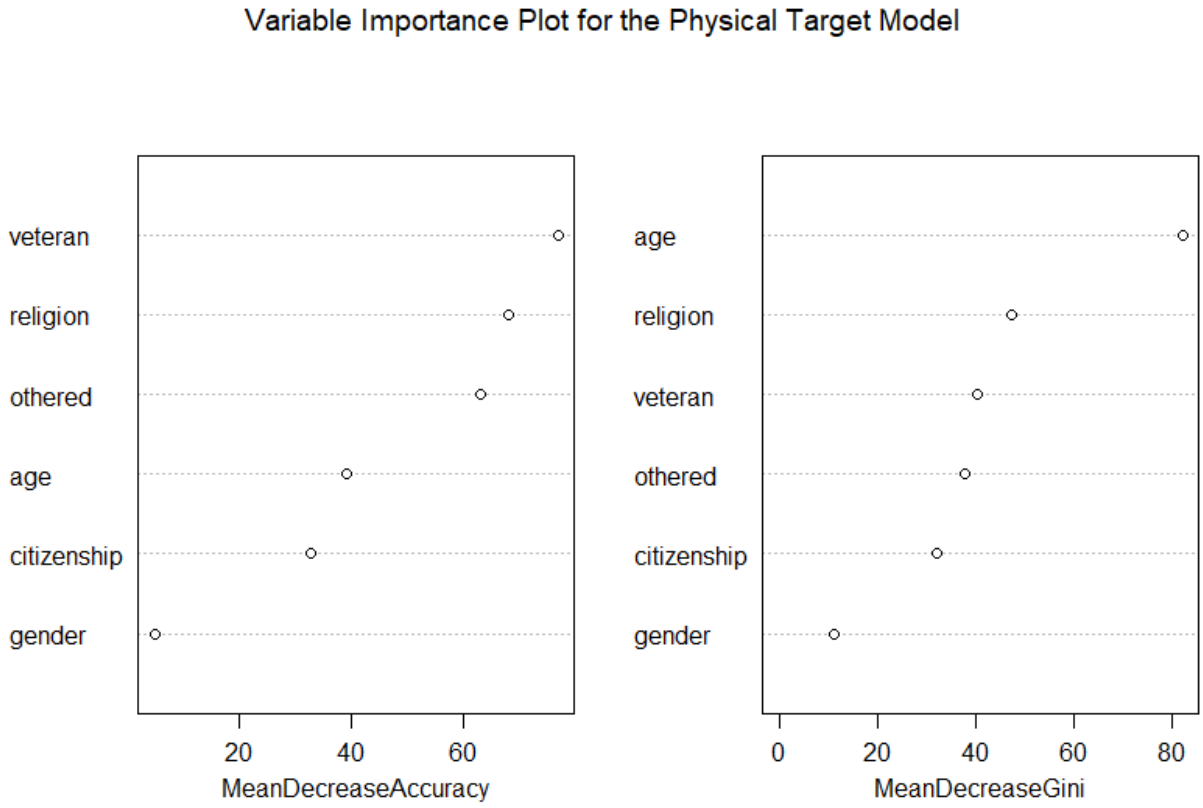
After conducting a random forest on the data used to build the regression/ANOVA model and plotting the importance of each variable, we find that citizenship status and age are particularly important in determining trends and predicting differences in the number of people injured as a result of a crime (Figure 6).

Figure 7. The classification tree for the physical target variable. At least 75 cases were required for each split, and each final outcome required at least 75 cases.



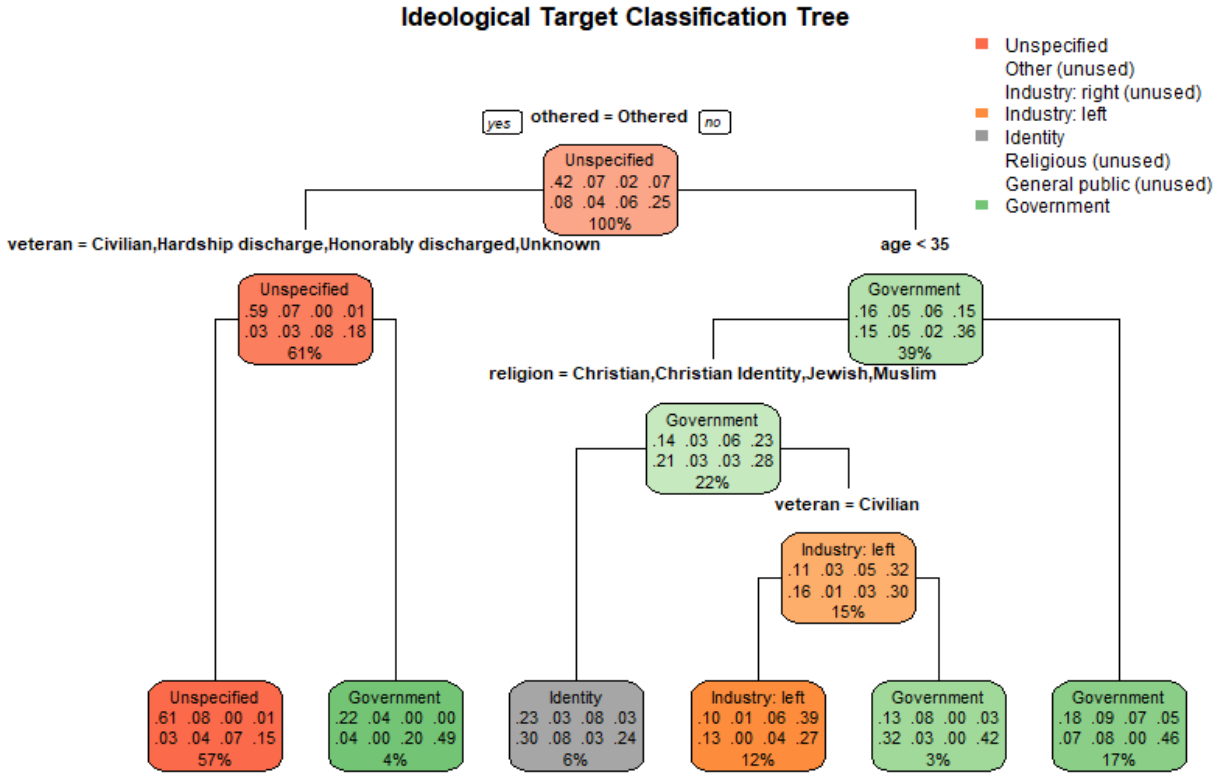
What we can see in this classification tree is that there is an initial split for othered status (Figure 7). Among those of othered status, we can see a split for veteran status. Among defendants who were civilians, former veterans released on hardship discharge, or former veterans who were honorably discharged, the physical target was more likely to be unspecified; among defendants whose veteran status did not fall in these 3 categories, no direct physical target was found (Figure 7). For those of non-othered status, private sites were more likely to be attacked, and there was a split for religion. Defendants whose religion was identified as Christian, Jewish, or Muslim were more likely to have an unspecified target, and those whose religion was not one of those 3 were more likely to attack private property (Figure 7). There is a further split in age; defendants who were 40 or older often had an unspecified physical target, whereas those under 40 tended to attack private sites (Figure 7).

Figure 8. The variable importance plot for the physical target random forest model.



After conducting a random forest on the data used to build the classification model and plotting the importance of each variable, we find that age, religion, othered status, and veteran status are important in predicting differences in physical targets (Figure 8). Age and veteran status appear to be particularly important in determining the differences between physical targets (Figure 8).

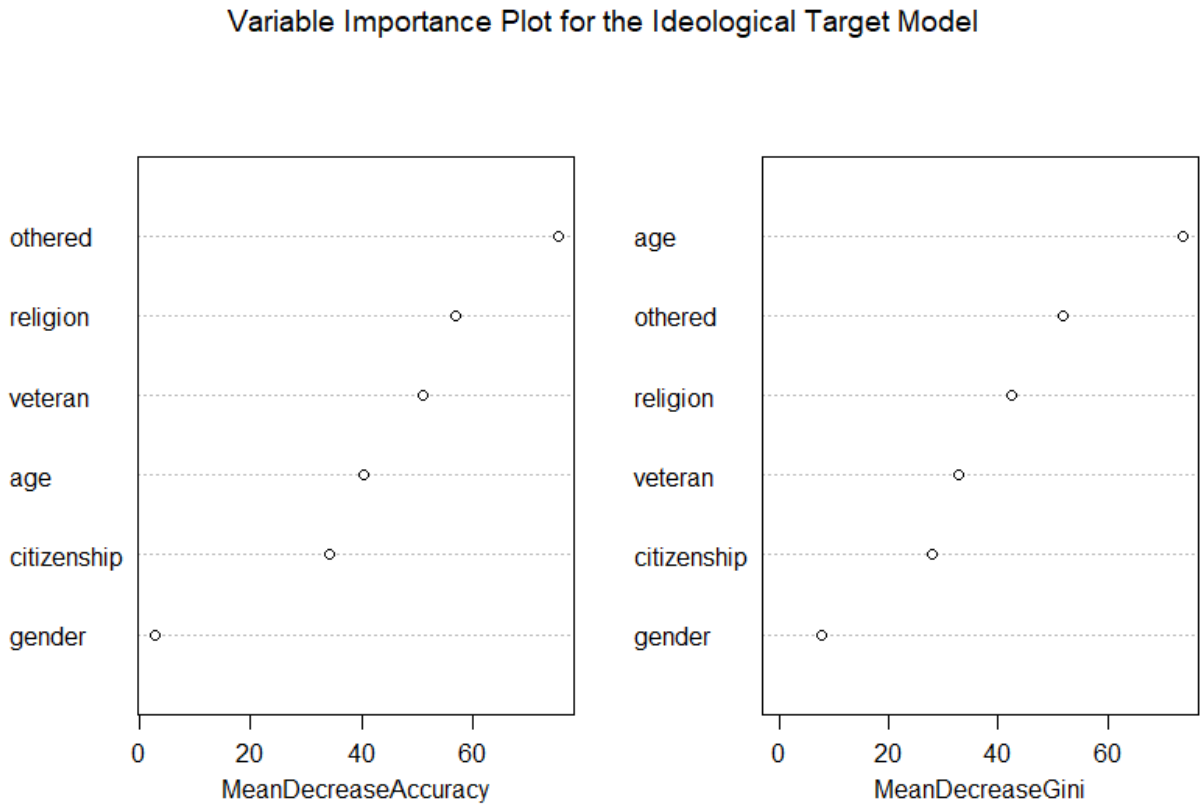
Figure 9. The classification tree for the ideological target variable. At least 50 cases were required for each split, and each final outcome required at least 25 cases.



We notice that the first split of this classification tree is at othered status (Figure 9). Of defendants who are of othered status, there is a split at veteran status. For defendants who are civilians, were honorably discharged, were discharged on the basis of hardship, or whose veteran status is unknown, there was an unspecified ideological target; for defendants whose veteran status is not one of those 4 categories, government was the most likely ideological target (Figure 9). For those of non-othered status, there is a split on age; those who were 35 or over were more likely to attack government targets on the basis of ideology (Figure 9).

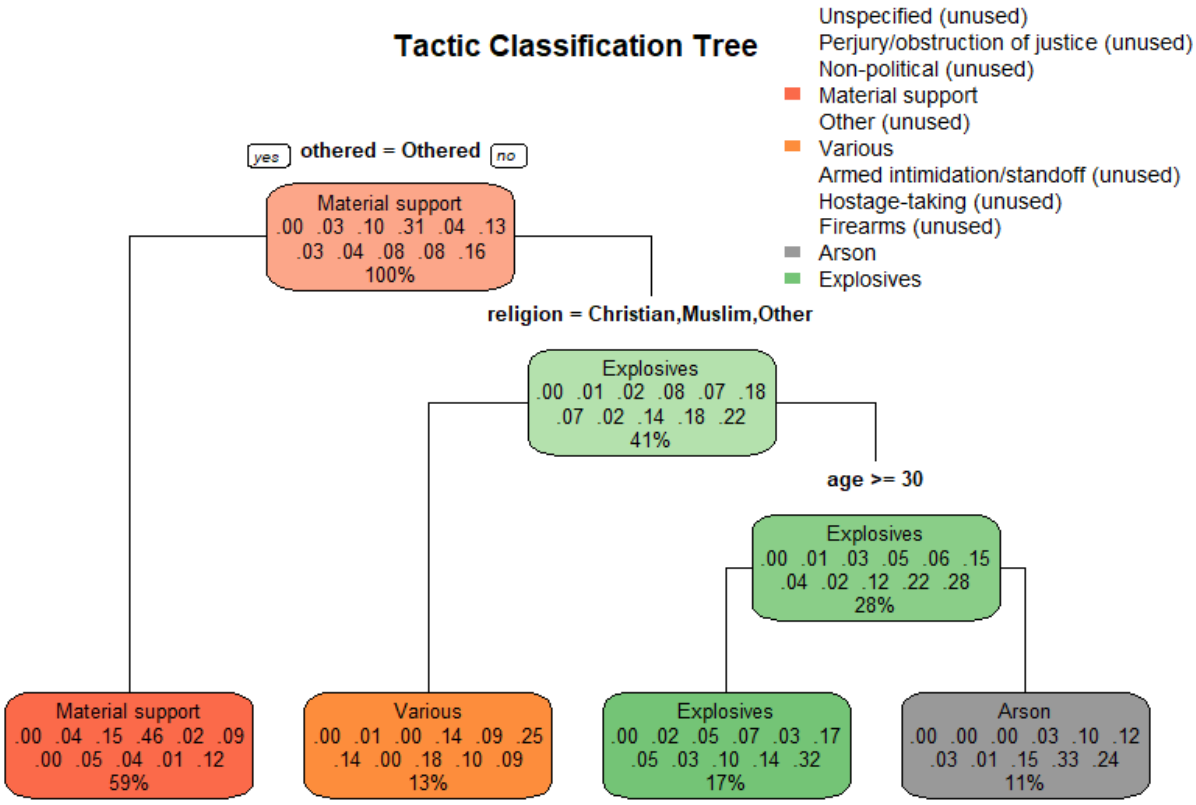
For non-othered defendants who were under 35, there was a split on religion; those whose religions were identified as Christian, Christian Identity, Jewish, or Muslim tended to attack on the basis of identity (Figure 9). Among those whose religions were not one of those 4 categories, veteran status was a significant predictor; civilians were more likely to attack left-leaning industries, while non-civilians were more likely to attack government on an ideological basis (Figure 9). In general, we have found that othered status, veteran status, age, and religion were significant variables in predicting ideological target.

Figure 10. The variable importance plot for the ideological target random forest model.



After conducting a random forest on the data used to build the classification model and plotting the importance of each variable, we find that age, othered status, and religion are important in predicting differences in ideological targets (Figure 10). Age and othered status appear to be particularly important in determining the differences between ideological targets (Figure 10).

Figure 11. The classification tree for the tactic variable. At least 100 cases were required for each split, and each final outcome required at least 100 cases.



Othered status appears to be very significant in predicting the tactic that a defendant used in committing a crime (Figure 11). Among those who are of othered status, the most common tactic, by far, was providing material support to a terrorist organization (Figure 11). Among those of non-othered status, religion is a significant predictor of tactic; defendants whose religion was identified as Christian, Muslim, or "Other" were more likely to employ multiple (or various) methods (Figure 11). Among defendants whose religion was not Christian, Muslim, or "other", age is a significant predictor of tactic; those who were 30 or over were more likely to use explosives when committing a terrorist act, and those who were under 30 were more likely to use arson (Figure 11).

Conclusions

This report finds that while all interactions between variables that define a defendant's identity and variables that define a defendant's criminal activity are significant, the variables which have the greatest prediction effect in terms of criminal activity are whether a defendant is othered or non-othered and the factors which contribute to that differentiation (religion, ethnicity/race, citizenship status), and a defendant's veteran status. A defendant's gender, while a significant factor in terms of the number of victims that result from a socio-politically-motivated crime, is generally not a significant predictor in other factors of criminal activity (tactic, target, etc.).

The results from our classification/regression trees and random forests appear to show that the most significant identity variables associated with different trends in criminal activity were related to age, citizenship status, veteran status, religion, and othered status. For the classification trees and their associated random forests, the variables that were particularly of importance were age and othered status, and for the regression trees and their associated random forests, the variables that were particularly of importance were age and citizenship status. Overall, age proved to be a very significant predictor in explaining differences in trends in criminal activity.

Some limitations of these random forests and classification/regression trees was the large number of unspecified or unknown cases, as well as a sizable number of unused levels for tactic, physical target, ideological target, and people vs. property. We noticed that for the classification tree models, the general error rate generally ranged from 46-55%, and for the regression/ANOVA tree models, the percentage of variability explained by the model was in the negatives. Thus, because of the poor predictive power of these models, we must exercise caution in assuming that the identity variables we found to be significant have any causal effect.

References

Brown, James D. 2008. "Effect size and eta squared." JALT Testing & Evaluation SIG News.

conjugateprior. 2013. “Formulae in R: ANOVA and other models, mixed and fixed.” Blog. Accessed February 27, 2019. Retrieved from http://conjugateprior.org/2013/01/formulae-in-r-anova/

Liaw, A., and M. Wiener 2002. Classification and Regression by randomForest. R News 2(3), 18-22.

Loadenthal, Michael, et al. 2019. “The Prosecution Project (tPP)” (Version March 2019) [Dataset]. Miami University Sociology Department. https://tpp.lib.miamioh.edu.

Loadenthal, Michael, Athena Chapekis, Lauren Donahoe, Alexandria Doty, and Sarah Moore. 2019. “The Prosecution Project (tPP) Codebook” (Version 2) [Code book]. Miami University Sociology Department. https://tpp.lib.miamioh.edu.

Loadenthal, Michael, Athena Chapekis, Lauren Donahoe, Alexandria Doty, and Sarah Moore. 2019. “The Prosecution Project (tPP) New Member Guidebook” (Version 1) [Instructional Manual]. Miami University Sociology Department. https://tpp.lib.miamioh.edu.

Milborrow, Stephen. 2018. rpart.plot: Plot 'rpart' Models: An Enhanced Version of 'plot.rpart'. R package version 3.0.6. https://CRAN.R-project.org/package=rpart.plot

Navarro, D. J. 2015. Learning statistics with R: A tutorial for psychology students and other beginners. R package version 0.5. University of Adelaide. Adelaide, Australia.

Salvatore S. Mangiafico. 2015. “Student’s *t*–test for Two Samples”. http://rcompanion.org/rcompanion/d_02.html

Therneau, Terry, and Beth Atkinson. 2018. rpart: Recursive Partitioning and Regression Trees. R package version 4.1-13. https://CRAN.R-project.org/package=rpart

Wickham, Hadley. 2017. tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse