April 4th, 2019
STA 475

The Classification of Terrorist Attacks

The Prosecution Project, Team 5

Characteristic Tree Analysis

Team Members:
Brent Crist, Elena McDonald, Yuan Liu, Xinru Yu

Non-Technical Summary

Introduction

Classification of terrorist attacks is the main problem of the Prosecution Project. Terrorism is one of the hottest topics in the news today, due to its increasing prevalence. Looking at acts of terrorism or political violence from a case-to-case basis, it is interesting to see how the government classifies each of them. Having the only reason for inclusion being "State Speech Act" in comparison to a combination of State Speech Act with other reasons, or no State Speech is of interest. Determining factors for why and how the government labels these cases provides an opportunity for analysis. The data comes from The Prosecution Project (tPP) from the sociology department at Miami University and yields the Reason for Inclusion, Tactic, Number Killed, Number Injured, and Othered Status for each case. This tPP dataset looks into the taxonomy of felony criminal cases involving illegal political violence, occurring in the United States since 1990. Utilizing the tPP dataset will allow for an explanation of the government classifications and the effects these variables have on the decision and how it changes through time.

Results

The Lethality variable is split by Reason of Inclusion categories: State Speech (the motivation for the terrorist act is explicitly political), No State Speech (the motivation for the terrorist act does not involve political purposes), and Combination (a mixture of the two). For better examination of the distribution for the lethality, below is the mean and the standard deviation for each reason, along with the number of cases belonging to the Reasons. It is clear that mean and standard deviation of State Speech are the lowest and have a large variability in comparison to No State Speech and Combination. It also occurs in the same as number of cases.

Lethality per Reason for Inclusion

Reason	Mean	Stand. Dev	Cases
Combination	4.68	7.67	522 (34.1%)
No State Speech	4.42	7.72	811(53.0%)
State Speech	2.41	3.33	198 (12.9%)

Looking at the Methods attackers are using, the top three Methods per Reason for Inclusion are below. Providing Support to a terrorist organization is the top method for No State Speech and a Combination. Non-political Method is the most common for State Speech and

represents over half of all State Speech Cases. Generally, terrorist attacks in the news, in recent years, involve explosives, firearms, and/or vehicle ramming. The Explosives Method appears less frequently than one might expect, given the frequency of news articles.

Top Three Methods per Reason for Inclusion

Reason for Inclusion	Method	Count
State Speech (198 Cases)	Non-Political	116 (58.6%)
	Perjury/Obst. Of Justice	15 (7.6%)
	Explosives	12 (6.1%)
No State Speech (811 Cases)	Provide Support	214 (26.4%)
	Explosives	192 (23.7%)
	Various Methods	98 (12.1%)
Combination (522 Cases)	Provide Support	236 (45.2%)
	Various Methods	85 (16.3%)
	Hostage/Standoff	49 (9.39%)

The third variable of interest is Othered Status. The table below, once again, breaks down Othered Status into each Reason for Inclusion. For both State Speech and a Combination, Othered individuals heavily out number Non-Othered. In cases that are No State Speech, the two groups are almost perfectly split fifty-fifty.

Othered Status

Reason for Inclusion	Othered Status	Count
State Speech (198 Cases)	Othered	162 (81.8%)
	Non-Othered	36 (18.2%)
No State Speech (811 Cases)	Othered	405 (49.9%)
	Non-Othered	406 (50.1%)
Combination (522 Cases)	Othered	353 (67.6%)
	Non-Othered	169 (32.4%)

Conclusion

For Lethality, no state speech is the most common reason, where state speech is much lower. Interestingly, providing support to terrorists or terrorist organizations is the most frequently encountered category for both no state speech and combination. Given the size of both of these categories, the frequency of this providing support is of interest to researchers for its implications in both separate categories. In all cases, the othered status of an individual might help researchers better understand how the state labels these people as terrorists. Because the categories state speech and combination carry implications of a directed attack against the state, the juxtaposition of the othered status reveals data to researchers who might be studying the othered status of terrorists.

Technical Summary

Introduction

The Prosecution Project provides a chance to determine when and what factors cause the state to label a criminal act as terrorism. In this analysis, many different techniques aid the process of determination of how these acts make the list. Data manipulation and cleaning assist the analysis by creating convenient (and statistically viable) groupings. Summary statistics and data visualization further enhances the ability to better understand how these variables change over time and how they relate to one another. Creating a characteristic tree is a strong method for analyzing what factors cause the government to label criminal acts as terrorism. The random forest method allows for validation of pruned trees and aids the analysis in this paper.

Methods

Data cleaning and manipulation are the first two crucial steps to proper analysis. For the tPP data, the research question revolves around the following variables: Reason for Inclusion, Tactic, Lethality, Other Status, and Date. Lethality is not a variable present in the data set; construction of the Lethality variable consists of adding the total kills and injuries per case, resulting from an offense. To answer the time element to the research question, the use of presidential terms creates meaningful time intervals for comparison. Associating the Day, Month, and Year of an event with the Day, Month, and Year of the inauguration of each president (in the scope of the data frame) allows for this timeline to form. The earliest case in the data frame occurs during Bill Clinton's service, while the latest case occurs during Donald Trump's service, with George W. Bush and Barack Obama in between. By adding the political affiliation of each president, another layer of analysis and comparison comes into play.

For purposes of the characteristic tree analysis, reduction of the Tactic variable with twenty unique levels is necessary. Reducing the number of levels gives more splitting power in the characteristic trees, further in the analysis. The percentage of cases involving each tactic hints at how much information each unique tactic provides to the overall analysis. Having eight levels, seven without Other, rather than the original twenty levels strengthens the resulting analysis.

Creating Method from Tactic

Tactic → Method	
Criminal violation not linked or motivated politically	Non-Political
Explosives, Bomb threat/hoax	Explosives
Hostage-taking, Armed intimidation/standoff	Hostage/Standoff
Arson	Arson
Firearms: civilian, Firearms: military	Firearms
Various Methods	Various Methods
Providing material/financial support to terrorist organization	Provide Support
Blade/Blunt Weapon, Chemical/Biological, Vandalism/Sabotage, Other, Vehicle Ramming, Unarmed Assault, Animal Release, Blockading, Unknown/Unspecified	Other

Reason for Inclusion also must undergo manipulation. To look specifically at the prevelane of the State Speech Act, splitting of Reason for Inclusion reflects this act. The three groups become cases that are State Speech, Not State Speech, and a Combination of the State Speech Act and other reasons. With this new variable, along with the others, the data are ready for investigation. Working with the data, summary statistics for Reason, Method, Lethality, and Other Status show how the data behaves and what it looks like. Additionally, separating bar graphs for the same set of variables by President, shows how each of these are changing in time. The bar graphs for Reason, Method, and Other Status are proportions while the bar graph for Lethality represents a count.

Creation of a characteristic tree (Buntine, 1992) can help analyze what factors cause the government to include each case, and the reason for the inclusion. Building a characteristic tree is not enough, both cross-validation and building a random forest provide insight as to how well the tree fits to the data. Execution of this technique in R, by partitioning the data into a training and testing set, produces this information. Fitting a tree, using a cost element for each partition, creates the optimal tree which will undergo methods of cross-validation (Zhong, 2016).

Comparing the values of the predictions and the real data computes the accuracy of these models. Further testing of the accuracy comes from the Random Forest, in the creation of a large sample of random trees (Zhong, 2016). By creating a large number of random trees, which use a random selection of the variables to split on, provides more evidence of model accuracy. The random forest generalizes the process, as such, the comparing predictions from the testing data set gives a stronger accuracy measure.

Many R packages are essential for the methods of this analysis. These procedures require the *lubridate* (Grolemund and Wickham, 2011), *caret* (Kuhn and Others, 2019), *rpart* (Therneau and Atkinson, 2018), *rpart.plot* (Milborrow, 2018), and *randomForest* (Liaw and Wiener, 2002) packages in R.

Results

In order to properly understand the motive of terrorist attacks, the execution methods play a vital role in their inclusion to this dataset. The Prosecution Project includes an exhaustive list of methods detailing how the acts are committed; however, grouping methods with similar tactics allow for proper analysis. That is, all acts, including acts that effectively serve as the threat of committing another act, are in the same group for analysis (e.g. “Explosives” and “Bomb Threats” become “Explosives”). Additionally, tactics that are “Unspecified” are not useful to a deeper understanding and hence, do not appear in this analysis. Lastly, all tactics that comprise less than 1% of the total tactics and do not fit neatly into the aforementioned methods (Animal Release, Blockading, Unarmed Assault, Vandalism) do not appear in this analysis (see *Prevalence of Tactic* table in Appendix for more details). These categories, with the terrorists’ reasoning, offer more insight into how a terrorist attack carries out given their motivation. The table below shows the prevalence of each Method in the data in relation to each Reason for Inclusion.

\

Top 3 Methods per Reason for Inclusion

Reason for Inclusion	Method	Count
State Speech (198 Cases)	Non-Political	116 (58.6%)
	Perjury/Obst. Of Justice	15 (7.6%)
	Explosives	12 (6.1%)
No State Speech (811 Cases)	Provide Support	214 (26.4%)
	Explosives	192 (23.7%)
	Various Methods	98 (12.1%)
Combination (522 Cases)	Provide Support	236 (45.2%)
	Various Methods	85 (16.3%)
	Hostage/Standoff	49 (9.39%)

Interestingly, more than half the cases that are State Speech are Non-Political (e.g. James Tyler Williams who killed a homosexual couple because they were gay). The majority of State Speech cases are Non-Political which are non-violent crimes relating to assisting terrorism or denying the ability of the state to pursue these crimes. No State Speech's top three methods together account for 62.2% of the cases in this category. This means that there is a higher spread of types of crimes as opposed to State Speech's Non-Political or Combination's Provide Support which are more highly skewed to these crimes.

The summary statistics of lethality per method provides useful insight into how each of these crimes change by lethality. For instance, the mean lethality of Firearms should be different from the Provide Support method. The standard deviation also shows the spread of each of these methods.

Lethality per Method

Method	Mean	Stand. Dev	Cases (%)
Provide Support	2.23	3.08	457 (29.8%)
Explosives	6.44	10.28	248 (16.2%)
Various Methods	5.03	8.54	189 (12.3%)
Arson	2.26	1.43	177 (11.6%)
Non-Political	2.00	0.00	150 (9.8%)
Firearms	9.62	11.01	126 (8.2%)
Other	3.06	4.61	10 (6.6%)
Hostage/Standoff	7.93	10.57	96 (6.2%)
Perjury/Obst of Justice	2.00	0.00	47 (3.1%)

Most cases yield results that fit the narrative of terrorism. Notice the higher means in the Firearms and Hostage/Standoff categories and the lower means in Non-Political and Provide Support categories. Higher standard deviations in the Explosives, Firearms, and Hostage/Standoff categories create a level of uncertainty in how many people are likely to be killed or injured from one of these attacks.

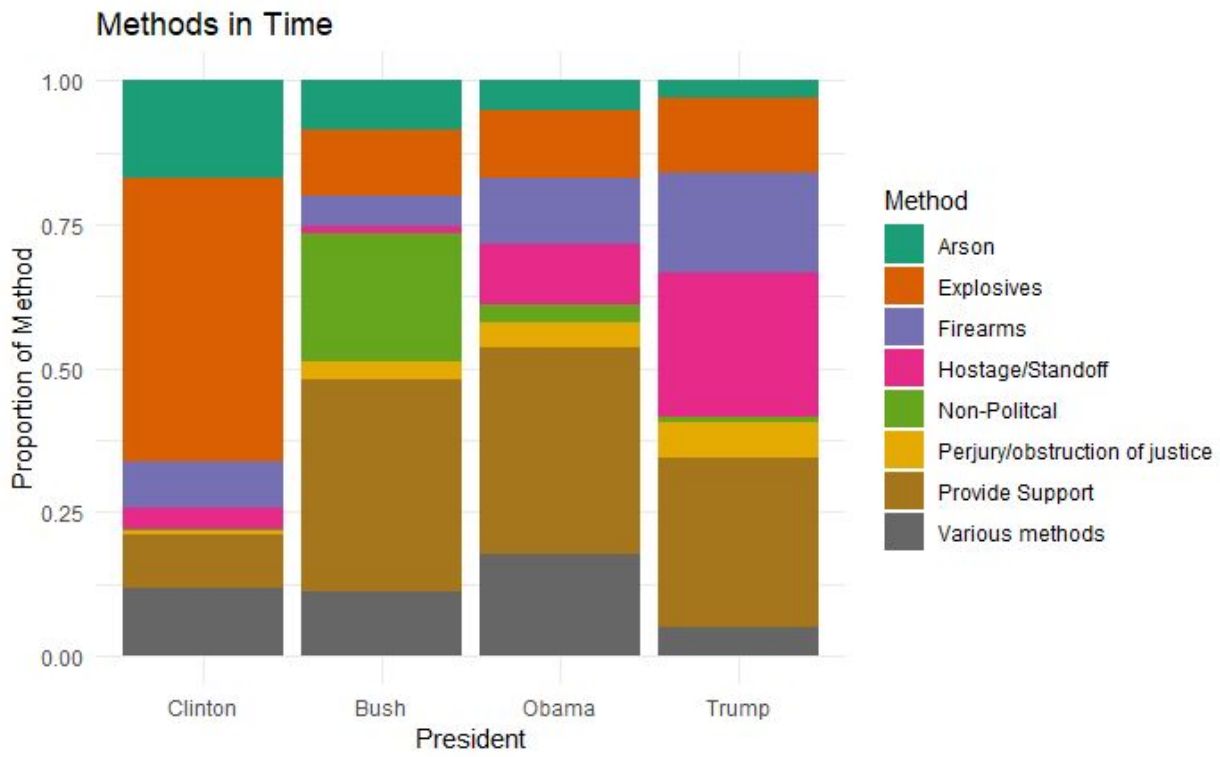
The Othered Status of an individual provides notable statistics for the Reason for Inclusion as well. It is critical to note that the Othered Status itself is quite subjective and is not a uniform label. That is, in no way are there exact criteria for a terrorist to be given an Othered Status. Mapping the Othered Status of a person to the reason their crime was included in the database allows for insight on how an othered person's crime might be perceived by the State.

Othered Status

Reason for Inclusion	Othered Status	Count
State Speech (198 Cases)	Othered	162 (81.8%)
	Non-Othered	36 (18.2%)
No State Speech (811 Cases)	Othered	405 (49.9%)
	Non-Othered	406 (50.1%)
Combination (522 Cases)	Othered	353 (67.6%)
	Non-Othered	169 (32.4%)

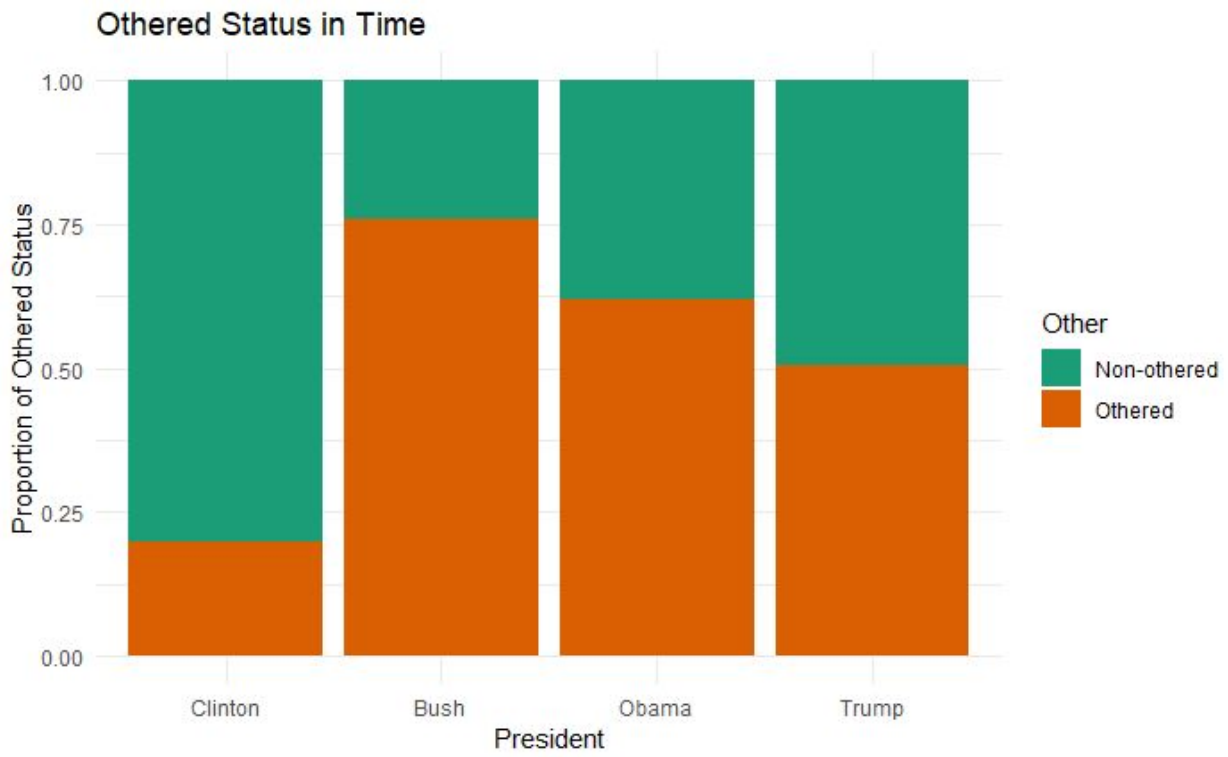
State Speech has the largest discrepancy between Othered and Non-Othered Status. This is to say that the vast majority of terroristic acts, when involving State Speech, are by Othered people. Whether or not this has any bearing to what period of time these acts happen appears later in this paper. No State Speech sees an almost even percentages by either Othered or Non-Othered people. As the guidelines for No State Speech are less specific than the other Reasons for Inclusion, there might be less cause for people of Othered and Non-Othered status to commit motivated terrorist attacks and more for the sake of senseless violence. The Combination Reason for Inclusion sees just over twice as many Othered people committing terrorist attacks for this reason as Non-Othered people.

As technology and geopolitical climates change with time, so too does the methodology of a terroristic act. Grouping these methods by their place in time relative to the President in office at the time of their happening gives way to visual representation of these statistics.



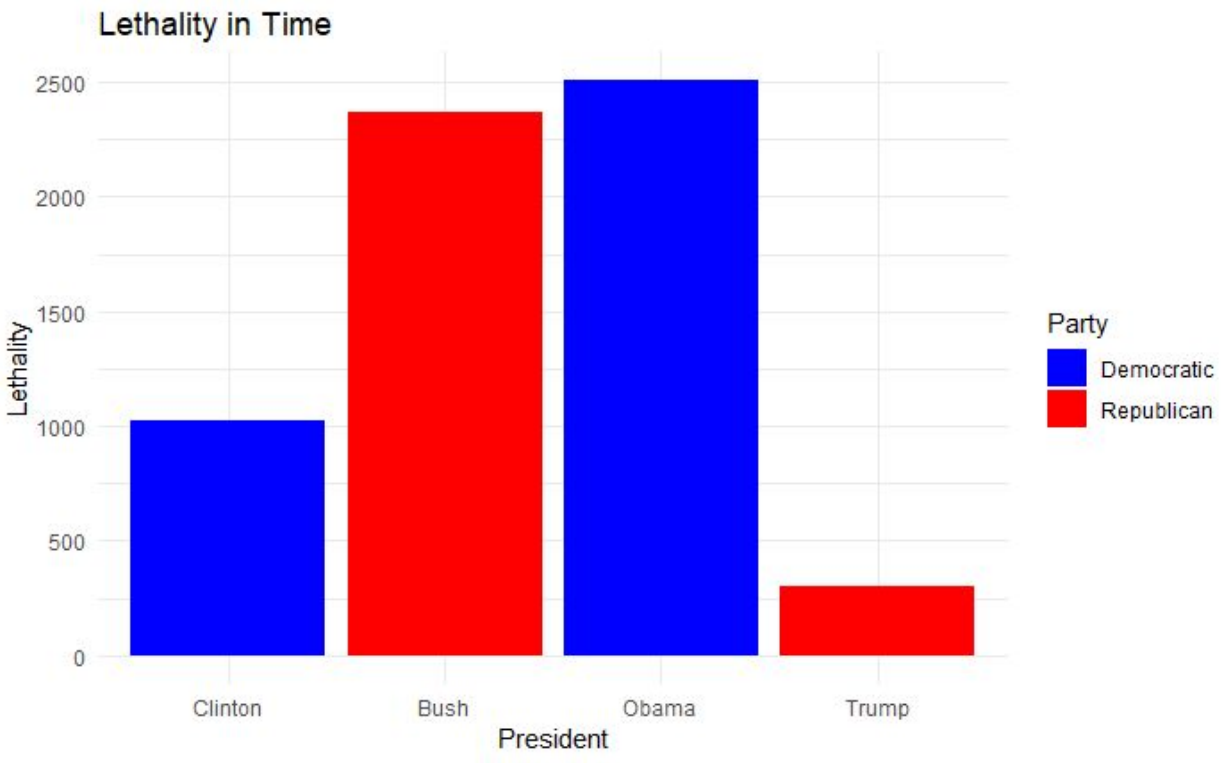
Drastic changes come over the years as the geopolitical climate changes. Notice there is a massive increase in the Provide Support (yellow) Method in the Bush, Obama, and Trump Administrations vice the Clinton Administration. This could be due largely to the fact that the Global War on Terror takes place during these Presidencies but not during Clinton's. It is not unreasonable to believe that the United States, as a strategy to deter violent terrorism, is labeling more non-violent crimes as terrorism than in years past. Since the United States is an economic superpower, its dollar has more buying power around the world. Because of this, terrorist sympathizers are able to accrue cash with much more buying power than in their home countries (assuming they support foreign terrorist organizations). This results in the ability of foreign terrorist organizations to acquire much higher numbers of supplies for violent terrorist attacks.

As the changing of methods through time offers insight into how the United States labels a crime as a terroristic act, the Othered Status of a person, too, changes in time. Different conditions in the United States during the four Presidencies included in this dataset might offer clues into how the status changes.



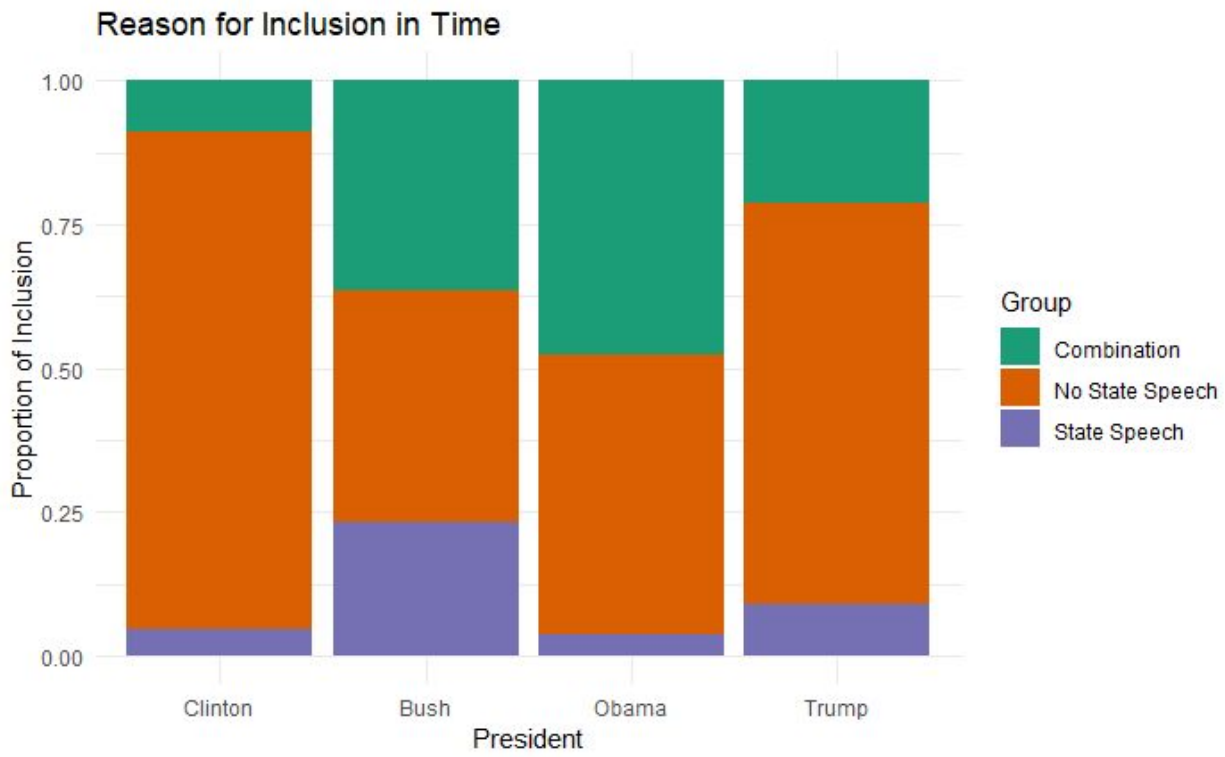
Notice, again, how the Othered status of terrorists changes drastically after the Clinton era. The United States, during this time period, could be experiencing higher sensitivity to terrorism due greatly to the loss of life from the September 11 attacks. As the Global War on Terror continues through the years, the Othered status of terrorists lowers. Whether this is due to a Liberal Obama Administration and a smaller sample size for the Trump Administration or that the United States and its citizens are becoming less skeptical of the people committing these crimes requires further study.

Seeing how the lethality of each of these acts changes in time can give clues as to how violent the crimes committed in these separate time periods are. Given the rise in non-violent methods in the past three Presidencies, studying the counts of lethality in their terms will shed light on how many people were killed in violent terrorist attacks in these time periods.



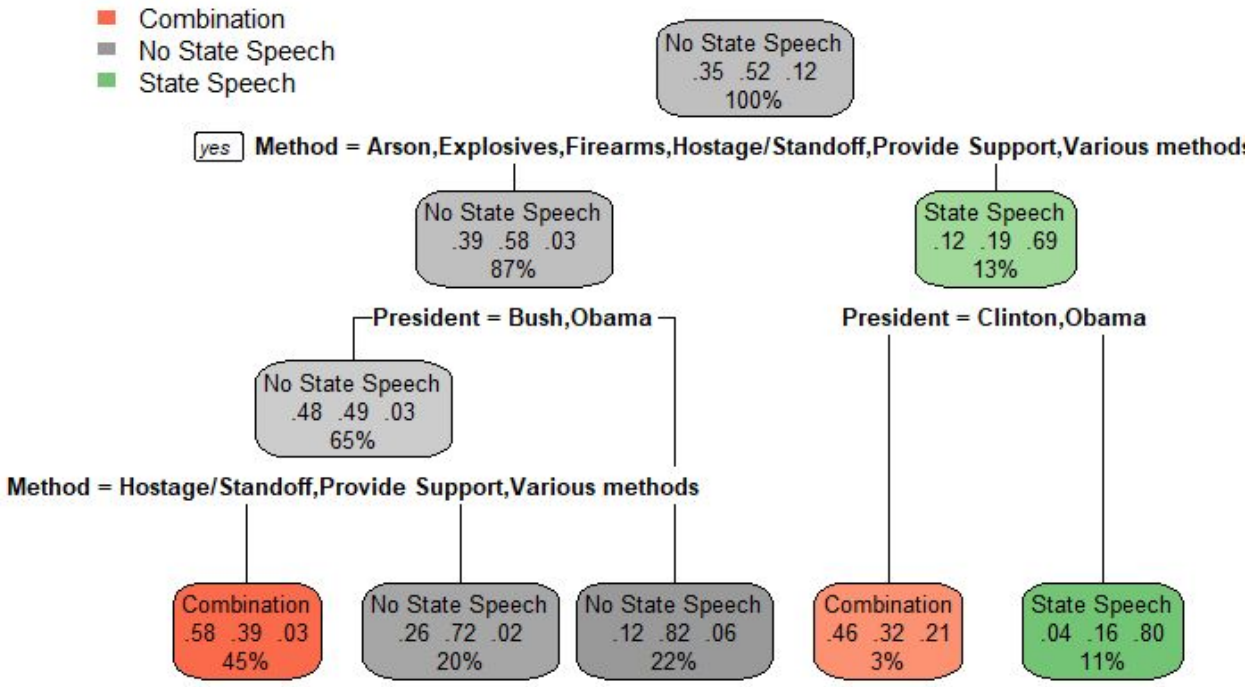
The lethality of these attacks again increases substantially during the years following the September 11 attacks. It is important to note that President Donald Trump has only been in office for just over two years at the time of conducting this analysis. The significantly lower lethality could be due mostly to the fact that the sample size is much smaller.

Seeing how the Method, Othered Status, and Lethality has changed through time then lends itself to studying how all terroristic acts included in the dataset has changed. Political moods and outside factors might play into how these crimes are included, and can be visualized by plotting them by the four Presidencies included in tPP.



A large change in No State Speech occurs from the Clinton to Bush Administrations. It is difficult to determine whether this is due to the United States' sentiment towards terrorism changing after the September 11 attacks or some other unknown variable. The Combination and State Speech groups constitute the largest change from the Clinton to Bush Administration. From the Bush to Obama Administration, a change in these two categories again occurs with State Speech becoming less prevalent and Combination becoming more prevalent. The increase in the Combination group might be a result of President Barack Obama being the first African American President in the history of the United States. A terroristic act due to this fact along with other racially charged motivations constitutes inclusion in the Combination group; however, this hypothesis requires further analysis and is not part of this study.

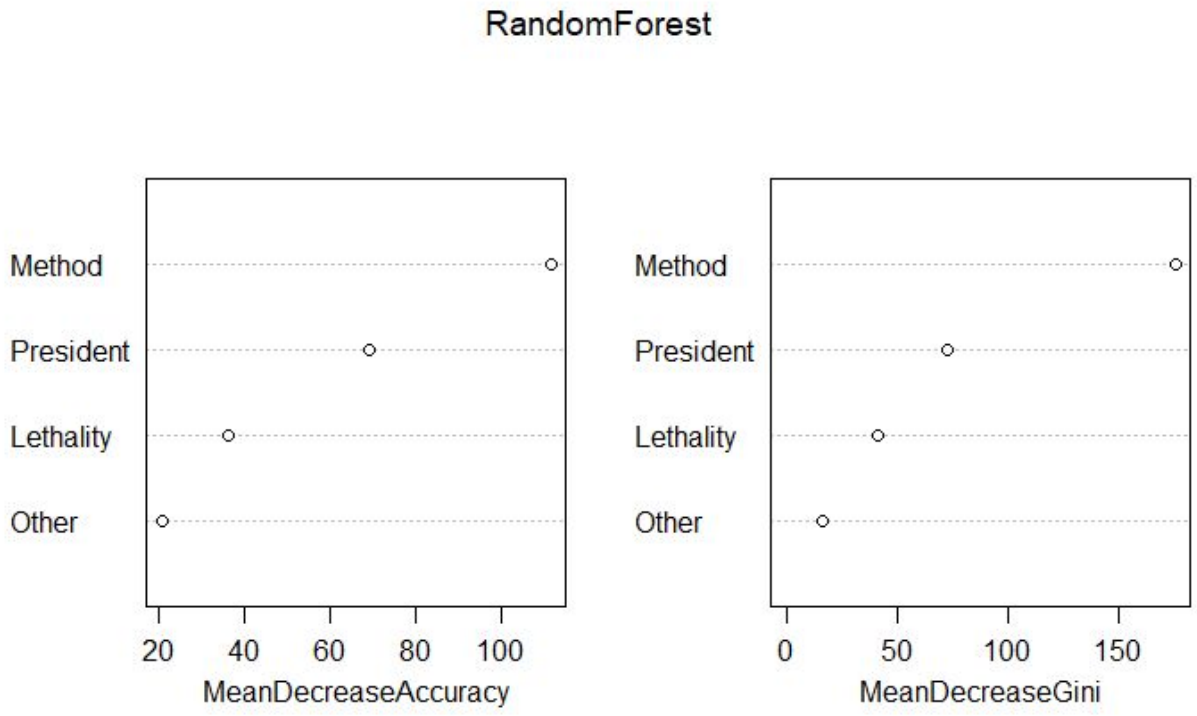
Machine learning processes can help to classify each case, with respect to their Reason for Inclusion, by the separate variables in the dataset. Splitting each of the nodes into various methods, Presidents, and lethality allows for the computer to decide where a case might fit based on the given factors and to create a Classification Tree from these splits.



The most important factor in this tree is the Method. From the first partition, all of the methods are present except for Non-Political and Perjury/Obstruction of Justice, which lends itself to the State Speech node on the right. The only Methods for which a case is likely to be State Speech are Perjury/Obstruction of Justice or Non-Political. Of note, President is the second partition on both the No State Speech and the State Speech nodes and that Obama appears in both of the positive splits for president. Only 10% of the entries fall under the criteria of Non-Political Method. Additionally, there is no partition that requires the Othered Status or Lethality in this tree. This tree shows a path of which to follow to see the categorization by the government of each type of case. The model accuracy rate of this optimized tree is about 65%, this comes from comparing the predicted values with those in the testing data set. A confusion matrix allows for the analysis of the performance of a Classification Tree. The model is the most accurate in predicting cases of State Speech and the least accurate for cases of a Combination.

		Predicted		
		Combination	No State Speech	State Speech
Truth	Combination	75	47	4
	No State Speech	57	126	4
	State Speech	3	9	31
	Percent Correct	55.6%	69.2%	79.5%
	Percentage Correctly Predicted (Overall)		65.2%	

The above procedure of obtaining a pruned tree involves using a training and testing data set. Splitting the data and training a model on part of it and then testing the model on the other part is a form of cross-validation. Another way to check the accuracy of the model is through a random forest. A random forest allows for validation of singular trees. Random forest importance plots show the validity of five hundred random trees from the data.



The Mean Decrease Accuracy and Mean Decrease Gini coefficients plots how important a variables is to the partitioning process in the creation of a characteristic tree. The further along a variable is on the x-axis (in both plots) signifies a greater presence in the partitioning process in randomly generated trees in the forest. As in the earlier singular characteristic tree, Method again is the most important variable for determining whether an act is State Speech, No State Speech, or Combination. Despite the large gaps in the variables (meaning the partitioning process becomes less accurate), it is worth noting that the variables in this order help to increase the validity of the singular tree. Lethality and Othered Status are the two least important predictors, according to the random forest data. Summarily, this means that the order of importance for determining the Reason for Inclusion is Method, President, Lethality, and Othered Status. The accuracy rate for the random forest is 70.2%, meaning the model for this data is predicting cases correctly 70.2% of the time.

		Predicted		
		Combination	No State Speech	State Speech
Truth	Combination	85	38	3
	No State Speech	51	133	3
	State Speech	3	8	32
	Percent Correct	61.2%	74.3%	84.2%
	Percentage Correctly Predicted (Overall)		70.2%	

Conclusion

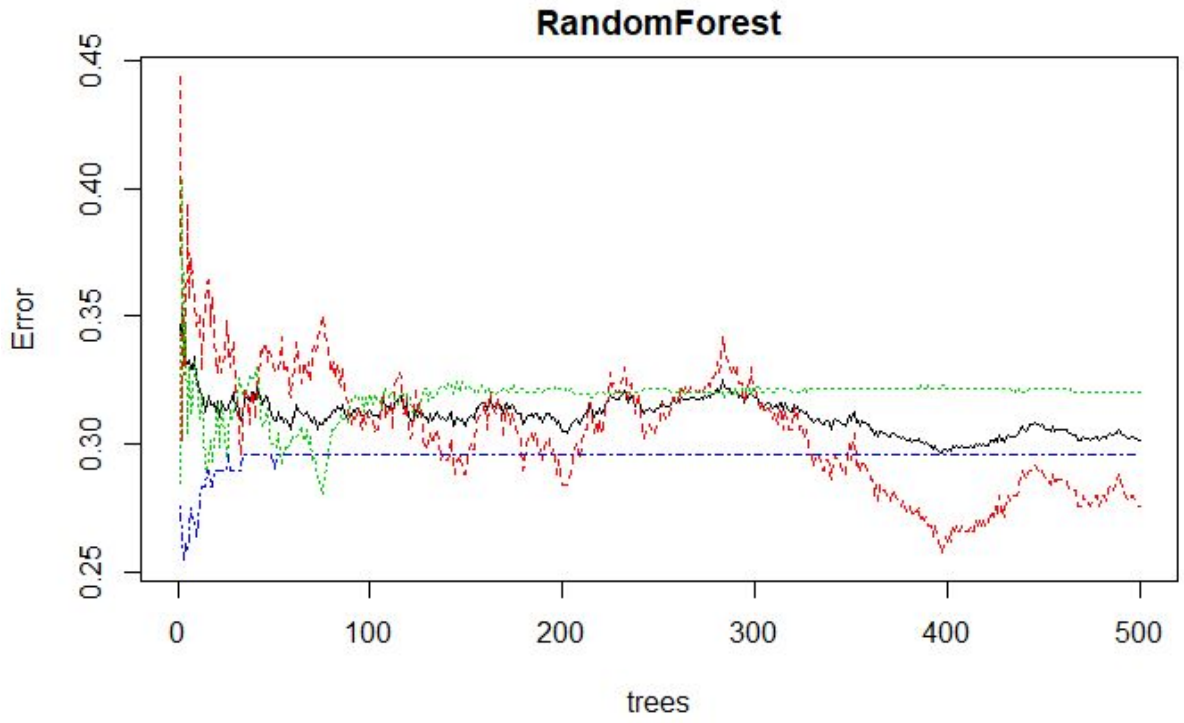
Many outside political factors (e.g. the September 11 attacks, the Global War on Terror, Presidential Administration) can affect how the government classifies crimes as terroristic acts or not. These classifications do change over time and the involving methods play a significant role in determining whether they are state speech acts of terrorism, not affiliated with state speech, or a combination of the two. In predicting what a government will classify a case as, the method by which a crime is committed and the President in office at the time of its being committed, in order of importance, have the most impact. The others do not provides as much information, but they are, in order of relevance, lethality of the crime and the Othered Status of the terrorist committing the crime. The splitting power of President in the characteristic trees drives home the finding of how the Reason for Inclusion changes in time. The random forest solidifies the importance of the variables presented in the pruned tree through the Mean Decrease Accuracy and the Mean Decrease Gini. Not surprisingly, the pruned tree ended with five terminal nodes,

two of which were No State Speech, two of which were Combination, and only one was State Speech. These results are consistent with the raw counts of each of the individual reasons for inclusion. With a 65% accuracy rate in the pruned tree and a 70% accuracy rate in the random forest, there is reason to believe that the variables in these trees make for important determining factors in whether a terroristic act will be classified as a state speech act, a non-state speech act, or a combination of the two.

Appendix

Prevalence of Tactic

Tactic	Number of Cases (%)
Providing Support	457 (29.85%)
Explosives	246 (16.07%)
Various Methods	189 (12.34%)
Non-Political	150 (9.80%)
Arson	117 (7.64%)
Firearms: Civilian	112 (7.32%)
Hostage Taking	52 (3.40%)
Perjury/ Obstruction	47 (3.07%)
Armed Intimidation	44 (2.87%)
Unknown/Unspecified	42 (2.74%)
Firearms: Military	14 (0.91%)
Blade/Blunt Weapon	12 (0.78%)
Chemical/Biological	12 (0.78%)
Vandalism/Sabotage	8 (0.52%)
Other	7 (0.46%)
Vehicle Ramming	7 (0.46%)
Unarmed Assault	6 (0.39%)
Animal Release	5 (0.33%)
Blockading	2 (0.13%)
Bomb Threat	2 (0.13%)



Code

```
library(tidyr)
library(tidyverse)
library(stringr)
library(rpart)
library(lubridate)
library(RColorBrewer)
library(tidyverse)
library(caret)
library(rpart)
library(rpart.plot)
library(randomForest)
base <- read.csv("file:///C:/Users/Elena/Documents/Statistics/STA475/tPP.csv")%>%
  select(Date, Date.descriptor,Reason.for.inclusion, X..killed, X..injured, Tactic,
Other..status)%>%
  rename(Reason = Reason.for.inclusion, Killed=X..killed, Injured = X..injured, Other =
Other..status) %>%
  mutate( Date= mdy(Date),Killed =as.numeric(Killed),
```

```
        Injured= as.numeric(Injured), Lethality = Killed + Injured)%>%
  filter(!is.na(Reason))  #Removing those without Reason for Inclusion
## Separating Reasons for Inclusion into Groups and separating Date
test <- base  %>%
  mutate(test = as.factor(str_detect(Reason, "State speech act")),
         test2 = as.factor(str_detect(Reason, "AND"))) %>%
  mutate( Group = ifelse(test=="TRUE" & test2=="TRUE","Combination" ,
                  ifelse(test=="TRUE" & test2=="FALSE", "State Speech",
                  ifelse(test=="FALSE" & test2=="FALSE", "No State Speech","No State
Speech"))))%>%
  mutate(Day = day(Date), Month = month(Date), Year= year(Date))%>%
  select(Date, Date.descriptor, Reason, Group, Tactic, Other, Lethality, Killed, Injured, Day,
Month, Year)
## Creating President, Party, and Method

tPP <- test %>%
  mutate(President = ifelse( Date >="1993-01-20" & Date < "2001-01-20","Clinton",
                             ifelse(Date >= "2001-01-20" & Date < "2009-01-20", "Bush",
                             ifelse(Date >= "2009-01-20"& Date < "2016-01- 20",
"Obama","Trump")))) %>%
  mutate(Party = ifelse(President == "Clinton" | President =="Obama",
                  "Democratic", "Republican"))%>%
  mutate(Method = ifelse(Tactic == "Explosives" | Tactic== "Bomb threat/hoax" ,
"Explosives",
                  ifelse(Tactic =="Providing material/financial support to terrorist organization",
"Provide Support",
                  ifelse( Tactic=="Various methods", "Various methods",
                  ifelse(Tactic=="Criminal violation not linked or motivated politically",
"Non-Politcal",
                  ifelse(Tactic=="Arson","Arson",
                  ifelse(Tactic=="Firearms: civilian" |Tactic== "Firearms: military", "Firearms",
                  ifelse(Tactic=="Hostage-taking" |Tactic== "Armed intimidation/standoff",
"Hostage/Standoff",
                  ifelse(Tactic=="Perjury/obstruction of justice ", "Perjury/obstruction of justice
", "Other")))))))))
##Lethality Summary
tPP %>%
  group_by(Group) %>%
```

```
  summarise(Lethality_AVG = mean(Lethality),
  Lethality_Med = median(Lethality),
  Lethality_sd = sd(Lethality),
  count= n())
tPP %>%
  group_by(President) %>%
   summarise(Lethality_AVG = mean(Lethality),
  Lethality_Med = median(Lethality),
  Lethality_sd = sd(Lethality),
  count=n())
## Lethality by Tactic/Method Summary

(sort(table(tPP$Tactic), decreasing = T)/length(tPP$Tactic))*100
##                                      tPP %>% group_by(Method)%>%
  summarise(round(mean(Lethality),digits = 2),
            round(sd(Lethality), digits=3),
            n())
tPP %>% group_by(President, Method)%>%
  summarise(mean(Lethality),
              sd(Lethality),
            median(Lethality))
## Other Summary
tPP %>% group_by(Other)%>%
  summarise(mean(Lethality),
            sd(Lethality),
            median(Lethality))

##By President
tPP %>% group_by(President,Other)%>%
  summarise(mean(Lethality),
            sd(Lethality),
            median(Lethality))

##Creating Bar Graph for Method Changing in Time
tt <-tPP %>%
  filter(Method!="Other")%>%
  group_by(President, Method)%>%
  mutate(n=n())
```

```
ttt<- tt[!duplicated(tt[,c('Method', 'President')]),]

barM <- ttt %>%
  group_by(President)%>%
  mutate(sum=sum(n),
         methProp = n/sum)

ggplot()+
  geom_bar(aes(x=President, y=methProp, fill=Method, group=Method), stat = "identity",
data=barM) +
  labs(title = "Methods in Time", y="Proportion of Method") +
  theme_minimal() +
  scale_fill_brewer(palette = "Dark2")
## Other Status Changing in Time
ttO <-tPP %>%
  filter(Method!="Other")%>%
  group_by(President, Other)%>%
  mutate(n=n())

tttO<- ttO[!duplicated(ttO[,c('Other', 'President')]),]

barO <- tttO %>%
  group_by(President)%>%
  mutate(sum=sum(n),
         OProp = n/sum)

ggplot()+
  geom_bar(aes(x=President, y=OProp,  fill=Other, group=Other), stat = "identity", data=barO)
+
  labs(title = "Other Status in Time", y="Proportion of Other Status") +
  theme_minimal() +
  scale_fill_brewer(palette = "Dark2")
## Lethality Changing in Time
ttL <-tPP %>%
  filter(Method!="Other")%>%
  group_by(President)%>%
  mutate(Total=sum(Lethality))
```

```
tttL<- ttL[!duplicated(ttL[,c('Total', 'President')]),]

cols <- c("Republican" = "red", "Democratic" = "blue")

ggplot()+
  geom_bar(aes(x=President, y=Total, fill=Party, group=Party), stat = "identity", data=tttL) +
  labs(title = "Lethality in Time", y="Lethality") +
  theme_minimal()  +
  scale_colour_manual(values = cols,aesthetics = c("colour", "fill"))
## Reason for Inlcusion chainging in Time
ttG <-tPP %>%
  filter(Method!="Other")%>%
  group_by(President, Group)%>%
  mutate(n=n())

tttG<- ttG[!duplicated(ttG[,c('Group', 'President')]),]

barG <- tttG %>%
  group_by(President)%>%
  mutate(sum=sum(n),
         GProp = n/sum)

ggplot()+
  geom_bar(aes(x=President, y=GProp,  fill=Group, group=Group), stat = "identity",
data=barG) +
  labs(title = "Reason for Inclusion in Time", y="Proportion of Inclusion") +
  theme_minimal() +
  scale_fill_brewer(palette = "Dark2")
## Characteristic Trees
TtPP <- tPP%>%
  filter(Method != "Other") %>%
  mutate(Group=as.factor(Group),
         President=as.factor(President),
         Method=as.factor(Method))

set.seed(345433)
training.samples <- TtPP$Group %>%
```

```
  createDataPartition(p = 0.75, list = FALSE)
train.data <- TtPP[training.samples, ]
test.data <- TtPP[-training.samples, ]


# Build the model
model1 <- rpart(Group ~ Method + Other + Lethality + President, data = train.data, method =
"class")
# Plot the tree
rpart.plot(model1, extra=104)
# Make predictions on the test data
predicted.classes <- model1 %>%
  predict(test.data, type = "class")
# Compute model accuracy rate on test data
mean(predicted.classes == test.data$Group)
# Confusion Matrix
table(test.data$Group, predicted.classes)
# Fit the model on the training set, forcing a dense tree
model2 <- rpart(Group ~ Method + Other + Lethality + President, data = train.data, cp=1e-04)

# Plot the tree (just FYI) and display cp values at various depths
rpart.plot(model2)
printcp(model2)
##
## Classification tree:
## rpart(formula = Group ~ Method + Other + Lethality + President,
##       data = train.data, cp = 1e-04)
##
# Prune the dense tree back
model3 <- prune(model2, cp=0.005)
# Plot the tree
rpart.plot(model3, extra=104)
# Make predictions on the test data
predicted.classes <- model3 %>%
  predict(test.data, type = "class")
# Compute model accuracy rate on test data
mean(predicted.classes == test.data$Group)
# Confusion Matrix
```

```
table(test.data$Group, predicted.classes)
        ## Random Forest
TtPP <- TtPP %>%
  mutate(Group=as.factor(Group),
         President=as.factor(President),
         Method=as.factor(Method))
RandomForest <- randomForest(Group ~ Method + Other + Lethality + President,
                  data=TtPP,
                  importance=TRUE)

# display variable importance metrics
varImpPlot(RandomForest)
plot(RandomForest)
# Make predictions on the test data
predicted.classes <- RandomForest %>%
  predict(test.data, type = "class")
# Compute model accuracy rate on test data
mean(predicted.classes == test.data$Group)
# Confusion Matrix
table(test.data$Group, predicted.classes)
```

References

Buntine, Wray. “Learning Classification Trees.” *Statistics and Computing*, vol. 2, no. 2, 1992, pp. 63–73., doi:10.1007/bf01889584.

Zhong, Yurong. “The Analysis of Cases Based on Decision Tree.” *2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, 2016, doi:10.1109/icsess.2016.7883035.

Software

caret (R):

Max Kuhn. Contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan and Tyler Hunt. (2019). caret: Classification and

Regression Training. R package version 6.0-82.
https://CRAN.R-project.org/package=caret

lubridate (R):

Garrett Grolemund, Hadley Wickham (2011). Dates and Times Made Easy with lubridate. Journal of Statistical Software, 40(3), 1-25. URL http://www.jstatsoft.org/v40/i03/.

randomForest (R):

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

rpart (R):

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

rpart.plot (R):

Stephen Milborrow (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