STAT 474 Paper III - Classification & Regression Trees

Author: Juan Manubens

Professor: Richard Berk

Prompt - ***For many reasons, the death penalty in the United States has long been controversial. One of the complaints is that "illegitimate" factors have a substantial impact on the decision made by prosecutors to charge a defendant with a capital crime. Gender, race, ethnicity, and nationality are examples of illegitimate factors. Imagine you will be testifying before the Senate Judiciary Committee to advise them on whether illegitimate factors are associated with charging decisions. You are to use CART to examine the possible role of illegitimate factors in death penalty charging.***

1 Introduction

I will attempt to use CART to examine the possible role of illegitimate factors in death penalty charging. As such, the policy intent makes the analysis Level II. I will explore and examine the univariate and multivariate statistics and transition to a Level II analysis. For a level II regression analysis, statistical inference is the defining activity, and estimation will be undertaken using the results from a level I regression after cleaning and recoding the data when necessary. The validity of any statistical inference made will fundamentally depend on how the data was generated, and thus much care will be needed in the preliminary steps before applying the CART algorithm.

2 The Data

The R dataset used in this paper was provided to us by out professor, called "*DeathPenalty*". The data come from the federal system during the Clinton Administration, and includes all homicide cases for which it was legally permitted to seek the death penalty. The data were collected to consider federal death penalty charging practices with the goal of possible subsequent reform.

2.1 Format and Variables

The data frame contains a total of $N = 669$ observations. The following variables are used:

2.1.1 Defendant Variables

- **1. death:** capital charge (the response variable)
- **2. gender:** defendant male = 1, not =0
- **3. white:** defendant white = 1, not =0
- **4. black:** defendant black = 2, not=0
- **5. hisp:** defendant hispanic = 3, not = 0
- **6. education:** high school, professional school, college degree =1, no h.s. degree = 2, unknown = 3
- **7. birthplace:** U.S. born = 1, foreign born = 0
- **8. working:** working = 1, unemployed = 0
- **9. alcoholhistory:** drinking problem history = 1, none = 0
- **10. drughistory:** drug problem history = 1, none = 0
- **11. retarded:** retarded = 1, not = 0
- **12. mentalhistory:** mental illness history = 1, not =0

2.1.2 Other Variables

- **13. vmale:** victim male = 1, female = 0
- **14. vwhite:** victim white = 1, not = 0
- **15. vblack:** victim black = 1, not = 0
- **16. vhisp:** victim hispanic = 1, not = 0
- **17. bktorture:** victim tortured = 1, not = 0 ("bk" indicates before the killing)
- **18. bkhostage:** victim held hostage = 1, not = 0
- **19. bkbeaten:** victim beaten = 1, not = 0
- **20. bkplead:** victim pled for mercy =1, not = 0
- **21. bksexassault:** victim sexually assaulted = 1, not = 0

- **22. numbervictim:** number of homicide victims
- **23. autogun:** automatic firearm used = 1, not =0
- **24. handgun:** handgun used = 1, not = 0
- **25. residence:** homicide at victim's residence = 1, not = 0
- **26. business:** homicide at victim's business = 1, not = 0
- **27. store:** homicide at victim's store = 1, not = 0
- **28. stranger:** defendant did not know victim = 1, not = 0
- **29. rival:** defendant and victim rivals = 1, not = 0

2.2 Data Cleaning & Transformation

The first step I took was to remove the 23 observations with an unknown response variable. Next, after confirming that the defendant race variables were mutually exclusive (i.e. no defendant has multiple races), I joined these into one column, *defendant race*. I then examined missing values across columns as well as rows. I removed the duplicate *blkplead1* column, and vehicle given the substantial proportion of missing values and more importantly, is not present in the documentation provided. I also removed 15 observations with missing values in a quarter or more of the features. One observation had 25.8% missing features, while 14 others ad over 45% missing features. In most of these rows, there was little or no data about the victims, and two of them had no data regarding the defendants gender. This makes the quality of those cases doubtful, and in order to best determine the relevance of illegitimate factors, I decided to drop these. Finally I removed the redundant defendant race columns, since it is all encapsulated in the new variable. Ultimately I decided to proceed with listwise deletion, as the missing values for different columns were very rare occurrences, and all imputation options (such as joining the race columns) have been exhausted. Furthermore, the data loss is not major, and the proportions across factors before and after the cleaning remained roughly identical. We end up with $N = 669$ observations.

2.3 Univariate and Multivariate Statistics

On the univariate side, we see that roughly a quarter of these cases proceeded with a capital charge. A large majority of the defendants were male and not foreign, and most were either black ($\tilde{4}$8.32%) or hispanic ($\tilde{2}$7.01%). Education levels were more varied, but most defendants did not have a known education level. Homicide victims were overwhelmingly male, though it is not clear for cases with multiple murder victims (we cannot assume that all the victims had the same gender). Finally, most victims knew the defendant, and most cases had one ($\tilde{6}$5.6%) or two ($\tilde{1}$4.6%) victims.

While there is a large number of possible bivariate relationships with the response, I think a few, like those mentioned above, are likely more influential. I've displayed two-way frequency tables below for two of these variables, both of which are considered illegitimate factors.

	Gender			
	0		1	
Sentencing	Percent	All	Percent	All
0	3.761	17	96.24	435
1	2.778	4	97.22	140

Table 1: Frequency - Gender of Victims and Sentencing

Interestingly, the sentencing proportions are not too different across genders. Across different races, the proportions seem to differ more.

	Race							
	Black		Hispanic		Other		White	
Sentencing	Percent	All	Percent	All	Percent	All	Percent	All
0	48.67	220	30.31	137	5.973	27	15.04	68
1	47.22	68	16.67	24	9.028	13	27.08	39

Table 2: Frequency - Race and Sentencing

3 Classification & Regression Trees (CART) Analysis

To begin our estimation procedure we must first make the case that each observation in the dataset was independently realized from a relevant joint probability distribution. This requires a degree of subject-matter expertise and knowledge about how the data were collected. We have a credible case here, albeit with a few caveats, because we know our data comes from the federal system during the Clinton Administration, and includes all homicide cases for which it was legally permitted to seek the death penalty throughout this time period. Since the capital charge is within federal jurisdiction, it makes sense to only look at data from a single administration. Here, this is what we interpret as the single joint probability distribution. The second step is to define the target of estimation - in this case, our estimation target is a classification or regression tree, having the same structure as the tree derived from the data, but as a feature of the joint probability distribution which produced our data. The third step is to select an estimator - we will be using Least Sum of Squares with CART, where the "best" split for each predictor is defined as the split that reduces the sum of squares the most.

Because the partitions are determined empirically from the data, the partitioning process introduces a form of model selection. Here, data snooping is unavoidable and it creates some complications for our level II analysis.

Finally we apply the estimator to our training data and then apply it to the predicted data. The split used here is 60 to 40. Using the *rpart* R package, we fit the CART model using all predictors [1] Three models are used: a base model with no tuning, a tuned complexity model with an adjusted complexity parameter selected from the base model results, and a final higher-complexity model with tuned minimum splits and explicit priors. The confusion trees and confusion tables, including model, use and overall errors (bottom right cell) can be seen below.

[1] The code can be found in the Appendix.

3.1 Model 1: Default Parameters

The default model interestingly shows a heavy presence of illegitimate factors. If the victim is white, the main split, the model outputs a higher probability of capital sentencing. The presence of race in one of the final splits is also a relevant factor.

	No Charge Pred.	Capital Charge Pred.	Model Error
No Charge	160.00	17.00	0.25
Capital Charge	52.00	9.00	0.65
Use Error	0.10	0.85	0.29

Table 3: Default Parameters CART Model - Overall Error: approx. 28.99%

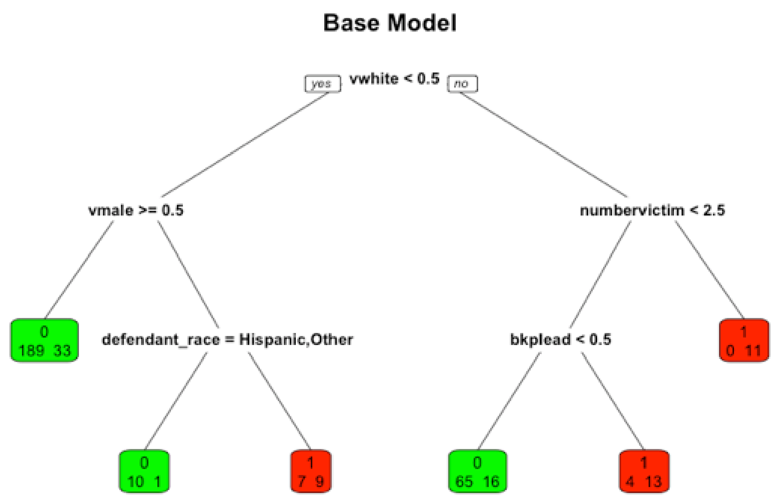

Figure 1: Default Parameters CART Model

3.2 Model 2: Tuned Complexity CART Model

The second model, while simpler, again shows a heavy presence of illegitimate factors. The victim being white or not is once again the main split.

	No Charge Pred.	Capital Charge Pred.	Model Error
No Charge	170.00	54.00	0.04
Capital Charge	7.00	7.00	0.89
Use Error	0.24	0.50	0.26

Table 4: Tuned Complexity CART Model - Overall Error: approx. 26%

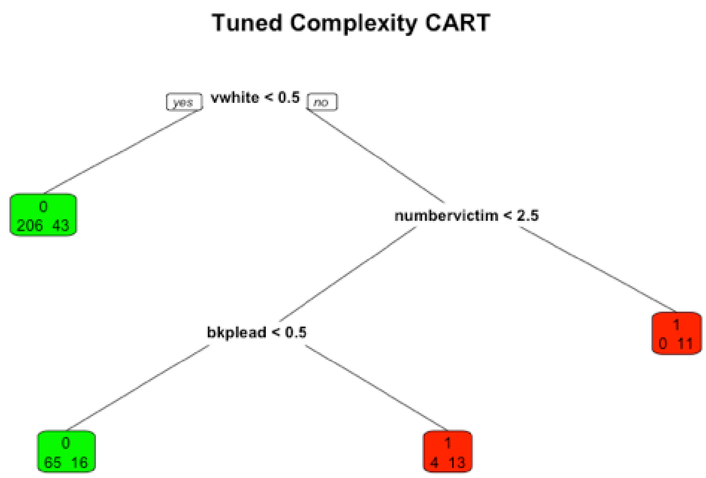

Figure 2: Tuned Complexity CART Model

3.3 Model 3: Final CART Model with Higher Complexity

	No Charge Pred.	Capital Charge Pred.	Model Error
No Charge	155.00	22.00	0.23
Capital Charge	45.00	16.00	0.58
Use Error	0.12	0.74	0.28

Table 5: Final CART Model with Higher Complexity - Overall Error: approx. 28.1%

The final model, shares many of the same splits and also uses illegitimate factors, such as the victim being male or white and the defendant being hispanic or other.

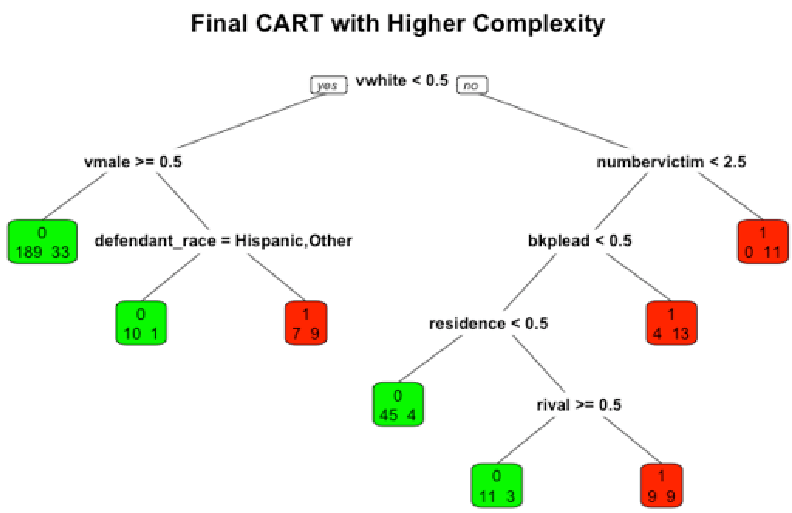

Figure 3: Final CART Model with Higher Complexity

4 Conclusion

For the sake of the hearing, conclusions can only be stated for the data at hand in a very cautious fashion - we would require more information about the data itself and substantial subject matter expertise to reach any broader conclusions. Nevertheless, the main conclusion to be reached here is that illegitimate factors play a non-minor role in sentencing, particularly when deciding if the defendant is not going to face the capital charge. While the data sample is not as large as one would hope, it is still representative, as indicated earlier, to an appropriate time frame covering a single administration. Repeating this exercise with data from other administrations could further corroborate this, but so far, illegitimate variables are present and particularly effective at predicting non-charges, as evidenced by the low use error in all three models.

5 Appendix

	VictimMale			
	0		1	
Sentencing	Percent	All	Percent	All
0	12.83	58	87.17	394
1	22.22	32	77.78	112

Table 6: Frequency - Victim Male and Sentencing

	Stranger			
	0		1	
Sentencing	Percent	All	Percent	All
0	70.58	319	29.42	133
1	59.72	86	40.28	58

Table 7: Frequency - Stranger and Sentencing

	NumVictims											
	1		2		3		4		5		6	
Sentencing	Percent	All	Percent	All	Percent	All	Percent	All	Percent	All	Percent	All
0	69.03	312	14.60	66	9.071	41	0.4425	2	1.991	9	0.000	0
1	54.86	79	14.58	21	15.278	22	2.0833	3	3.472	5	4.167	6

	NumVictims							
	7		9		10		14	
All	Percent	All	Percent	All	Percent	All	Percent	All
0	0.885	4	1.327	6	1.5487	7	1.1062	5
1	2.083	3	2.083	3	0.6944	1	0.6944	1

Table 8: Frequency - Number of Victims and Sentencing

	Employed			
	0		1	
Sentencing	Percent	All	Percent	All
0	18.58	84	81.42	368
1	28.47	41	71.53	103

Table 9: Frequency - Stranger and Sentencing

	DrugsHistory			
	0		1	
Sentencing	Percent	All	Percent	All
0	81.19	367	18.81	85
1	75.00	108	25.00	36

Table 10: Frequency - Drug History and Sentencing

6 R Code - CART

````
```{r}
out1 <- rpart(as.factor(death) ~., data = df_train, method = 'class')
get_ctbl(out1)
get_ratio(out1) # 3.058824
get_results(out1, "Base Model") %>% xtable
```


```{r}

out3 <- rpart(as.factor(death) ~., data = df_train, method = 'class',
              cp = 0.04)

summary(out3)
get_ctbl(out3)
get_ratio(out3) #7.714286
get_results(out3, "Tuned Complexity CART") %>% xtable

```


```{r}
out<-rpart(as.factor(death) ~., data = df_train, method="class",
           parms = list(prior = c(.75,.25)),cp=.004, control = (minsplit = 3))

get_ratio(out) #7.714286
get_ctbl(out)
get_results(out, "Final CART with Higher Complexity") %>% xtable

out$cptable

```
````

7 R Code - Preprocessing & EDA

```r
# Custom DF Operations

get_stats <- function(df_x){
  cols_df <- colnames(df_x)
  df_return <- as.data.frame(rep(0, cols_df %>% len)) %>% t %>% as.data.frame

  rownames(df_return) <- c(1)
  df_return[2,] <- rep(0, cols_df %>% len)
  df_return[3,] <- rep(0, cols_df %>% len)
  n_df <- nrow(df_return)
  for (i in 1:len(cols_df)){
    count <- df_x[cols_df[i]] %>% wna %>% len
    num_distinct <-  df_x[cols_df[i]] %>% unlist %>% as.vector %>% unique %>% len
    df_return[1,i] <- count
    df_return[2,i] <- 100*count / n
    df_return[3,i] <- num_distinct
  }
  colnames(df_return) <- cols_df
  rownames(df_stats) <- c('num_NAs', 'perc_NAs', 'num_distinct')
  return(df_return)
}

get_stats_rows <- function(df_x){
  n_df <- nrow(df_x)
  df_rows <- as.data.frame(rep(0, n_df))
  df_rows$c2 <- rep(0, n_df)
  df_rows$c3 <- rep(0, n_df)
  colnames(df_rows) <- c("ix", "num_NAs", "perc_NAs")

  for (i in 1:n) {
    count <- df_x[i, ] %>% is.na %>% unlist %>% as.numeric %>% sum
    df_rows[i, 1] <- i
    df_rows[i, 2] <- count
    df_rows[i, 3] <- 100*count / 31
  }
  return(df_rows)
}

ptable <- function(ix){
  cols <- colnames(df)
  ncols <- len(cols)
  n <- nrow(df)
  return(100*(table(df[cols[ix]]))/n)
}
```

```r
ctable <- function(ix){
  cols <- colnames(df)
  ncols <- len(cols)
  n <- nrow(df)
  return(1*(table(df[cols[ix]]))/1)
}

plot_tree <- function(mdl, title){
  rpart.plot::prp(mdl, extra = 1, faclen = 0, varlen = 0, cex = 0.8,
                  round = 1, main = as.character(title),
                  box.palette = c('green', 'red'))[mdl$frame$yval]
}


get_ctbl <- function(mdl){
  yactual <- df_test$death
  ypred <- predict(mdl, df_test, type="class")
  tdf <- table(yactual, ypred)
  ctbl <- tdf[1,] %>% as.data.frame() %>% t %>% as.data.frame
  ctbl[2,] <- tdf[2,]
  ctbl$c3 <- c(0,0)
  ctbl[3,] <- c(0,0,0)
  a <- ctbl[1,1]
  b <- ctbl[1,2]
  c <- ctbl[2,1]
  d <- ctbl[2,2]
  c13 <- c / (a + c)
  c23 <- b / (b + d)
  c31 <- b / (a + b)
  c32 <-  c / (c + d)
  c33 <- (b+c) / (a + b + c + d)
  row3 <- c(c31,c32,c33)
  ctbl[3,] <- row3
  ctbl[1,3] <- c13
  ctbl[2,3] <- c23
  rownames(ctbl) <- c('No Charge', 'Capital Charge', 'Use Error')
  colnames(ctbl) <- c('No Charge Pred.', 'Capital Charge Pred.', 'Model Error')
  return(ctbl)
}


get_results <- function(mdl, title){
  plot_tree(mdl, as.character(title))
  return(get_ctbl(mdl) %>% xtable)
}

get_ratio <- function(mdl){
  ctb <- get_ctbl(mdl)
  fn <- ctb[2,1]
  fp <- ctb[1,2]
  return (fn/fp)
}
```

````
```{r}
df <- DeathPenalty
na_Y <-  which(is.na(df$death))
na_Y %>% len #23
N <- nrow(df) # 3.437967 %

# (1) Eliminate observations with NAs
df <- df[-c(na_Y),]

str(df$white)
str(df$black)
str(df$hisp)

# (2) Check if victim race columns are mutually exclusive

(df$white + df$black) > 2 # Checks through
(df$white + df$hisp)  > 3 # Checks through
(df$black + df$hisp)  > 3 # Checks through

# (3) Recode into one column, as factor
ix_w <- which(df$white == 1)
ix_b <- which(df$black == 2)
ix_h <- which(df$hisp  == 3)

race_vals <- rep('Other', nrow(df))
race_vals[ix_w] <- rep('White', len(ix_w))
race_vals[ix_b] <- rep('Black', len(ix_b))
race_vals[ix_h] <- rep('Hispanic', len(ix_h))

colnames(df)[3] <- 'defendant_race'

df$defendant_race <- race_vals
df$defendant_race <- df$defendant_race %>% as.factor

# (4)  Remove blkplead1 since it's a duplicate, and vehicle given lack of
#      documentation. Also remove redundant columns

rm_ix <- c(4,5,22,27)
df <- df[,-rm_ix]

# Remove redundant columns
cols <- df %>% colnames



#ix_other <- which(df$victim_race == "Other")
#check <- DeathPenalty[ix_other,c(1:5)] %>% as.data.frame

````
````

```r
```{r}
# (5) Examine NAs in columns
df_stats <- get_stats(df)
#df_stats %>% View

numcols_na <- df_stats[1,]  %>% unlist %>% as.numeric > 0  %>% unlist
numcols_na <- numcols_na %>% as.numeric() %>% sum
numcols_na

# (6) Examine NAs in rows
df_rowNAs <- get_stats_rows(df)
#df_rowNAs %>% View
df_rowNAs <- df_rowNAs[order(-df_rowNAs$perc_NAs),]

bad_data_ix <- which(df_rowNAs$perc_NAs > 20)
bad_rows_ix <- df_rowNAs$ix[bad_data_ix]

# Save DF
#df_preclean <- df

df <- df[-c(bad_rows_ix),]

df_stats_clean <- get_stats(df)

numcols_na_clean <- df_stats_clean[1,]  %>% unlist %>% as.numeric > 0  %>% unlist
numcols_na_clean <- 29 - numcols_na_clean %>% as.numeric() %>% sum

t0 <- 29 - numcols_na
t1 <- 29 - numcols_na_clean

dif <- t1 - t0 # 12 - 4 = 8 columns cleaned

numd <- df_stats[3,] %>% unlist %>% as.vector %>% prod
# 4.597364e+12
numd/1000

669 - 631 # 38 columns lost from cleaning
631/669 # 5.68% loss , 94.32% left



# Save DF
df_postclean <- df

# Examine DF
df_stats <- get_stats(df)



```
```

```{r}
# (7) Check for odd rows
df <- df_save

df <- df[-c(df$gender %>% wna),]
df <- df[-c(df$birthplace %>% wna),]

df_stats <- get_stats(df)

df_stats[,which(df_stats[1,] > 0)]
df_stats[,which(df_stats[1,] > 0)] %>% colnames

one_victim <- which(df$numbervictim == 1)

df[one_victim,]$vblack + df[one_victim,]$vwhite
df[one_victim,]$vblack + df[one_victim,]$vhisp
df[one_victim,]$vwhite + df[one_victim,]$vhisp



# (8) Drop rows with no data on the victim's race, if all 3 are missing for cases with one victim

# Two rows dropped
df[df$vblack %>% wna,]

df <- df[-c(df$vblack %>% wna),]
df_stats <- get_stats(df)

# (9) Drop rows with no data on the weapons used, if all 2 are missing

df <- df[-c(df$autogun %>% wna),]
df_stats <- get_stats(df)


# Two rows dropped
df_stats[,which(df_stats[1,] > 0)] %>% colnames

# (9) Drop rows with no data on vmale, residence

df <- df[-c(df$vmale %>% wna),]
df <- df[-c(df$residence %>% wna),]
df_stats <- get_stats(df)


# Two rows dropped

df_stats[,which(df_stats[1,] > 0)]

613 / 669


```
```
```

```r
```{r}


# (10) Proceed to univariate and multivariate statistics


 df$defendant_race
races <- df$defendant_race %>% unlist %>% as.factor()
death_vals <- df$death %>% unlist %>% as.vector

all_ptables <- lapply(c(1:len(colnames(df))), ptable)
names(all_ptables) <- cols

all_tables <- lapply(c(1:len(colnames(df))), ctable)
names(all_tables) <- cols
all_ptables

304 + 170 + 45 + 112

if(!require('tables')) {
  install.packages('tables')
}
library('tables')


summary(df)

220 / 68 # 3.23
137 / 24 # 5.70833
68  / 39 # 1.74359
27 / 13 # 2.076923



tabular((Sentencing=as.factor(death)) ~ (Race=defendant_race)*(Percent("row") + 1), data = df) %>% latex

tabular((Sentencing=as.factor(death)) ~ (Gender=as.factor(gender))*(Percent("row") + 1), data = df) %>% latex

tabular((Sentencing=as.factor(death)) ~ (VictimMale=as.factor(vmale))*(Percent("row") + 1), data = df) %>% latex


tabular((Sentencing=as.factor(death)) ~ (Stranger=as.factor(stranger))*(Percent("row") + 1), data = df) %>% latex

tabular((Sentencing=as.factor(death)) ~ (NumVictims=as.factor(numbervictim))*(Percent("row") + 1), data = df) %>% latex

tabular((Sentencing=as.factor(death)) ~ (NumVictims=as.factor(numbervictim))*(Percent("row") + 1), data = df)

tabular((Sentencing=as.factor(death)) ~ (NumVictims=as.factor(numbervictim))*(Percent("row") + 1), data = df)

```
```